Pécsi Tudományegyetem
Állam- és Jogtudományi Doktori Iskola

Diósi Szabolcs

Mesterséges intelligencia, szintetikus valóság

- **Az MI és GenMI rendszerekkel kapcsolatos globális kihívások és európai szabályozási stratégiák**

Doktori (PhD) értekezés

Témavezető:
Dr. Barcsi Tamás

Pécs
2024

„A fejlődés ellen nincs gyógymód”

Neumann János

1 Bevezetés

1.1 A disszertáció kutatási témája és a témaválasztás indokolása

A dolgozat a Mesterséges Intelligencia és a Generatív Mesterséges Intelligencia technológiák közelmúltban tapasztalt látványos fejlődésének és széleskörű elterjedésének társadalmi következményeit, valamint az azokra adott Uniós jogalkotási stratégiákat vizsgálja. A téma kétségtelenül népszerű, az elmúlt években számos tudományos igényességű monográfia, könyv, tanulmány és cikk született a kérdéskörről.[1] Aktualitását jelzi az is, hogy jelen dolgozat írásával párhuzamosan alkotta meg az Európai Unió azon MI rendeletét, amely a világon elsőként vállalkozik egy olyan átfogó jogi keret létrehozása, amely garantálja az ilyen rendszerek fejlesztésének, használatának és forgalmazásának biztonságos feltételeit az EU határain beül.[2] A szabályalkotási folyamat során az Unió egy olyan jogi környezet kialakítására törekedett, amely nyitott a technológiai újítások felé, ösztönzi a kutatást és fejlesztést, egyúttal képes minimalizálni a társadalmi kockázatokat. Az innováció és európai versenyképeség megőrzésének támogatása, illetve az alapvető emberi jogok és európai értékek következetes, kompromisszumot nem ismerő védelme, mint olykor egymást korlátozó célok feloldására pedig kockázatalapú szabályozási keretet javasolt, amely az MI rendszerek által jelentett kockázat mértékével arányos kötelezettségeket ír elő.

[1] A technológia növekvő befolyásának és szüntelen bővülő alkalmazási lehetőségeinek köszönhetően a témakör a szélesebb közvélemény érdeklődésének is a középpontjába került. A lehetséges felhasználási módozatokra – közel sem kimerítő – példaként szolgálhat, hogy már napjainkban is sok helyen MI-alapú rendszerek automatizálják az ügyintézést és az adminisztrációt, Nagy Nyelvi Modellekre épülő virtuális asszisztensek javítják az ügyfélszolgálat minőségét, az adóhatóságok MI-t alkalmaznak az adócsalás és adóelkerülés felderítésére, a szociális ellátórendszerek az állami segítségnyújtási ellátásokra és szolgáltatásokra való jogosultságának értékelésére. A rendvédelem és igazságszolgáltatás területein MI kockázatértékelő algoritmusok elemzik a bűnügyi adatokat, előrejelzik a bűncselekmények elkövetésének valószínűségét, támogatják a nyomozást és a bírói döntéshozatalt. Biometrikus arcfelismerőrendszerek segítik a bűnmegelőzést és a határellenőrzést. Az egészségügy területén MI-alapú rendszerek elemzik a radiológiai felvételeket, azonosítják a gyanús elváltozásokat és betegségre utaló jeleket. A precíziós mezőgazdaságban MI-vezérelt drónok, robotok és szenzorrendszerek monitorozzák a növények egészségi állapotát, azonosítják a kártevőket, betegségeket, valamint optimalizálják a vízfelhasználást és egyéb erőforrásokat. A környezetvédelem területén MI-modellek elemzik az ökológiai adatokat, előrejelzik a klímaváltozás hatásait és támogatják a fenntartható erőforrás-gazdálkodást. Bankok és biztosítóintézetek algoritmusokat használnak az ügyfelek hitelképességének értékelésére, hitelkérelmek gyorsabb feldolgozására a kockázatok csökkentésére. A humánerőforrás-menedzsment területén az MI alkalmazható a jelentkezők toborzására, kiválasztására, üres álláshelyek meghirdetésére, pályázatok szűrésére, valamint a jelöltek interjúk során történő értékelésére. Az e-kereskedelemben és a szórakoztatóiparban egyaránt intelligens ajánlórendszerek elemzik a fogyasztói preferenciákat, hogy célzott, személyre szabott termékeket és tartalmakat kínáljanak.

[2] Az Európai Unió Tanácsa 2024. május 21-én hagyta jóvá a végleges törvényszöveget, melyet várhatóan 2024 júliusában fognak közzétenni az Európai Unió Hivatalos Lapjában.

A jogalkotási folyamat gördülékeny menetét változtatta meg a Generatív Mesterséges Intelligencia (továbbiakban GenMI) technológiák széles körben történő elterjedése és szabadon hozzáférhetővé válása. Hosszú ideje ismert kihívás a technológiai ágazat jogalkotási eljárásai kapcsán, hogy az innováció hajlamos gyorsabban haladni, mint a szabályozás, mire a szabályozó hatóságok feltérképezik az új technológiákat és kidolgozzák a megfelelő keretrendszert, addigra újabb fejlesztések jelennek meg. Az MI-hez hasonló felforgató technológiák szabályozásakor kiváltképp igaz Koltay megállapítása, miszerint *a jogi szabályozás rendszerint követő üzemmódban van*.[3] Az idő szorításában születő jogalkotói stratégiákat fenyegető további veszély, hogy a döntéshozók nem rendelkeznek kellő mennyiségű információval, vagy ha igen, azokból nem a legmegfelelőbbeket választják ki a szabályozás megalapozásához. Ilyen helyzetekben a jogalkotók könnyen kerülhetnek abba a dilemmába, hogy a „*meggondolatlan cselekvés és a teljes bénultság*" között kell választaniuk.[4] Az Európai Bizottság által 2021-ben közzétett rendelettervezetben foglalt szabályozási paradigma a GenMI technológiák robbanásszerű elterjedését megelőzően lett megfogalmazva, ebből kifolyólag nem volt maradéktalanul alkalmas az újonnan megjelenő kihívások kezelésére. A kockázatalapú megközelítés a rendszerek előre beazonosított alkalmazási területeire és felhasználási módozataira összpontosít (például az olyan kiemelten veszélyes területekre, mint a bűnüldözés, a határvédelem, vagy a kritikus infrastruktúrák működtetése). A GenMI technológiák esetén az ilyen típusú szabályozások bár bizonyos mértékben hasznosak lehetnek, nem képesek lefedni az összes potenciális veszélyforrást.

A dolgozat első részének középpontjában az EU MI jogszabályalkotási folyamatának feltérképezése áll. A dolgozat különböző esettanulmányokon keresztül mutatja be a már napjainkban előszeretettel is alkalmazott MI rendszerek legfőbb felhasználási területeit, valamint az alkalmazásokból adódó lehetséges veszélyeket és kockázatokat. Ezt követően részletesen elemzi az Unió szabályozási törekvéseinek kezdeti irányvonalát, különös figyelmet fordítva a nagy és elfogadhatatlan kockázatot jelentő MI-re vonatkozó rendelkezések módosításaira.

A dolgozat második része a GenMI technológiák elterjedéséből eredő kihívásokat veszi górcső alá. Többek között kitér ChatGPT-hez hasonló Nagy Nyelvi Modellek működési

[3] Koltay András (2021): Előszó. In Török Bernát és Ződi Zsolt (szerk) A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 11.o.
[4] G. Karácsony Gergely (2020): Okoseszközök – Okos jog? A mesterséges intelligencia szabályozási kérdései. Budapest, Dialóg Campus. 77.o.

sajátosságaiból eredő olyan releváns kockázatokra, mint a gépi hallucináció, a szándékos megtévesztés, a szintetikus adatokon történő modellképzés, az információs homogenizálódás és a modell összeomlás kérdéskörei. Vizsgálódásának fókuszát a 21. századi információs környezet fokozatos eltorzulásához vezető több évtizede tartó folyamat feltérképezésére helyezi. A dolgozat feltételezése szerint az emberi információfeldolgozást és valóságérzékelést jelenleg is torzító hatások jelentős megerősödéséhez vezet majd a szöveges és audiovizuális szintetikus tartalmakat megközelítőleg korlátlan mennyiségben előállítani képes GenMI modellek robbanásszerű elterjedése.

1.2 A disszertáció tartalmi áttekintése

A doktori értekezés szerkezetileg nyolc fejezetre tagolódik. A végén található az angol nyelvű összefoglaló, az irodalomjegyzék és a szerző publikációs listája. A *Bevezető* rész után a dolgozat második fejezetében az MI fejlődéstörténetének rövid bemutatására és a technológiacsaládd al kapcsolatos alapfogalmak ismertetésére kerül sor. Külön figyelem összpontosul a modern értelemben vett MI-t jellemző gépi tanulási mintázatok feltérképezésére. Ennek jelentősége nem elhanyagolható, hiszen míg korábban a hagyományos számítógépes rendszerek működése mereven kötött volt a programozók által előre megírt utasításokhoz, melyek lépésről lépésre határozták meg, hogy pontosan mit kell tennie a rendszereknek egy adott feladat végrehajtásához, a gépi tanulással lehetővé vált az MI számára, hogy az adatminták összefüggéseire és korábbi tapasztalataira alapozva önállóan (iteratív módon) tanuljon és fejlődjön anélkül, hogy további instrukciókra lenne szüksége.

A harmadik fejezet különböző esettanulmányokon keresztül mutat be olyan napjainkban is elterjedt – elsősorban a közszférában használt – MI alkalmazásokat, melyek érdemi hatással lehetnek az állampolgárok jogaira és kötelezettségeire (pl.: bíróságok, adóhatóságok, bűnüldöző szervek által alkalmazott rendszerek). A gyakorlatok kiválasztása nem önkényes alapon történt, segítségükkel könnyebben megérthető, hogy a 2016-ban kezdődő EU-s jogalkotási stratégiák milyen veszélyeket és kockázatokat azonosítottak e rendszerekkel kapcsolatban. A fejezet második részében ismertetésre kerülnek a gyakorlati alkalmazásból eredő legjelentősebb szabályozási és össztársadalmi kihívások, nevezetesen: (1) a gépi tanulással képzett MI-működésére jellemző átláthatóság, elszámoltathatóság és kiszámíthatóság hiánya; (2) az MI által vezérelt téves, elfogult vagy diszkriminatív döntéshozatal veszélye; (3) az MI-rendszerek alapvető emberi szabadságjogokra, döntési autonómiára és a demokratikus intézményekre gyakorolt káros hatása.

A negyedik fejezetben az MI-rendelet jogalkotási folyamatának bemutatására kerül sor, külön hangsúlyt fektetve a nagy és elfogadhatatlan kockázatúként kategorizált MI alkalmazások területén eszközölt jogszabályi változásokra, és az *általános célú MI modelleket* érintő új szabályozási keretek kialakítására. A fejezet időrendben vezeti végig az olvasót az MI-rendelet megalkotásának folyamatán, kezdve a Bizottság 2021. április 21-én közzétett rendelet-tervezetével, a Tanács 2022. decemberi közös álláspontjával és a Parlament 2023. májusi kompromisszumos javaslatával. Ezt követően kitér a 2023 második felében zajló háromoldalú egyeztetésekre, valamint a jogszabály 2024. március 13-ai végleges elfogadásának körülményeire is.

Az ötödik fejezet a 2022 végén széles körben ismertté vált GenMI modellek sajátosságainak feltérképezésére vállalkozik. E rendszerek lényege abban rejlik, hogy felhasználói utasítások (promptok) alapján új tartalom létrehozására képesek, és hogy eredeti rendeltetésük mellett számos további alkalmazási lehetőséggel is bírnak. Míg a korai modellek csupán egyetlen modalitás, azaz egyetlen típusú tartalom előállítására korlátozódtak (például a Nagy Nyelvi Modellek csak szöveges bemenetet tudtak értelmezni és szöveges tartalmat tudtak előállítani), a legújabb fejlesztésű multimodális modellek már több eltérő formátumú tartalom feldolgozására és előállítására is alkalmasak. Az ilyen GenMI-k egyik látványos példája a 2024. júniusában minden felhasználó számára szabadon elérhetővé tett Luma Dream Machine modellje, amely kép vagy szöveges bemenetek alapján körülbelül két perc alatt képes 15 másodperces hosszúságú professzionális filmstúdiók termékeit idéző videók és animációk előállítására.[5]

A fejezet emellett kitér a ChatGPT-hez hasonló Nagy Nyelvi Modellek széles körű alkalmazásából eredő olyan releváns kihívásokra, mint a gépi hallucináció, a szándékos megtévesztés, a tanító adatokhoz kötődő szerzői jogi konfliktusok, a szintetikus adatokon történő modellképzés, az információs homogenizálódás és a modell összeomlás kérdéskörei. A fejezet befejező része a GenMI rendszerekhez kötődő modern diskurzus legújabb irányvonalait igyekszik felvázolni. Ezek közül a leginkább releváns négy kérdéskörként (1) a modellek tanításából eredő növekvő – és hosszútávon nem fenntartható – energiaszükségletet és környezetkárosítást; (2) az emberközi kapcsolatok, illetve az ember-gép közötti kapcsolatok gyökeres átalakulását; (3) a GenMI modellek munkaerőpiacra gyakorolt

[5] A transzformer architektúrán és GAN gépi tanuláson alapuló GenMI segítségével bármely regisztrált felhasználónak havonta harminc ingyenes videó elkészítésére van lehetősége. Előfizetés esetén ez a tartalomgyártási kvóta számottevően növekszik. A szoftver elérhető: https://lumalabs.ai/dream-machine

diszruptív hatását; (4) az új MI generációt képviselő autonóm ágensek (InteraktívMI) megjelenésének lehetséges következményeit jelöli meg.

Az hatodik, hetedik és nyolcadik fejezet a GenMI modellek által előállított szintetikus tartalmak elterjedésének rövid- és hosszútávú kockázatait járja körbe. Annak érdekében, hogy a folyamat teljes egészében feltérképezhetővé váljon a hatodik fejezet egészen a 2000-es évek elejéig – a Web2.0 kialakulásáig – tekint vissza az időben, és e ponttól kezdődően mutatja be, hogy miként változott meg a nyilvánosság szerkezete napjainkig. Rávilágít arra az ellentmondásra, hogy az internet kezdetben az információ demokratizálódásának megtestesítője volt, ám ahelyett, hogy tájékozottabbá tette volna a felhasználókat, sokszor csak kételyt ébreszt és dezinformál, ahelyett, hogy közelebb hozta volna az embereket, inkább ellentéteket gerjeszt és polarizál.

A hetedik fejezet az igazság hanyatlásának folyamatát vizsgálja, amit a konspirációs elméletek, állhírek és dezinformációs kampányok elterjedése kísér. A fejezet arra törekszik, hogy bemutassa, hogyan járul hozzá a GenMI – különösen a Nagy Nyelvi Modellek és a Deepfake technológiák – az igazság mérgezéséhez és a társadalmi bizalom erodálásához. A technológia fejlődésével egyre nehezebb felismerni a manipulált tartalmakat, ami tovább súlyosbítja a társadalmi bizalom válságát. A dolgozat arra is rávilágít, hogy az ilyen technológiák elterjedése hosszú távon fenyegetheti a demokratikus rendszerek stabilitását és az emberek közötti hagyományos kapcsolatok egészséges működését. A fejezet második része az EU dezinformációs stratégiájának és a digitális szolgáltatásokról szóló rendeletének (DSA) ismertetésre tesz kísérletet. A jogszabály – melynek célja, hogy biztonságos, kiszámítható és megbízható online környezetet teremtsen az Unió összes állampolgára számára – szorosan kötődik a dolgozatban tárgyalt kihívásokhoz, minthogy az MI-rendelet és a GDPR mellett ebben a rendeletben fogalmazzák meg azokat a szolgáltatói kötelezettségeket, melyek az automatikus gépi döntések és az algoritmikus tartalomszűrés átláthatóságának garantálására, valamint az egyén döntési szabadságát befolyásoló manipulatív gyakorlatok korlátozására irányulnak. A fejezet arra a következtetésre jut, hogy bár a DSA építő és szükséges kiegészítője azon már hatályos joganyagnak, melyek az MI potenciális kockázatainak kezelésére irányulnak, semelyik ma alkalmazandó jogszabály nem nyújt teljeskörű védelmet a GenMI modellek azon hosszútávú fenyegetettségével szemben, ami a szintetikus tartalmak és az ember által létrehozott *autentikus tartalmak* megkülönböztetésmentes összeolvadásából származhat.

A nyolcadik fejezet az internet legújabb korszakának bemutatására vállalkozik (Web 0.0). E korszak legfőbb jellemzője, hogy a felhasználók – kifejezett tudomásuk nélkül – egyre gyakrabban találkoznak MI által generált szövegekkel, képekkel és interakciókkal online, miközben a valódi emberi kapcsolatok és az autentikus tartalmak aránya drasztikusan csökken. A Web 0.0 korszakának egyik legaggasztóbb következménye az, hogy a felhasználók bizalma a kezdetben online, később offline elérhető információkban fokozatosan erodálódik. Az emberek nem tudják többé megkülönböztetni a valós tényeket a mesterségesen generált információktól.

1.3 Módszertan, a kutatás hasznosítása

A dolgozat elkészítésének alapjául a témához kapcsolódó jelentős hazai és nemzetközi szakirodalom (tanulmányok, monográfiák, kommentárok, tudományos publikációk) áttekintése és feldolgozása szolgált. A kutatási módszertan alapját a releváns Európai Uniós jogszabályok, Tanácsi következtetések és javaslatok, intézményi irányelvek és ajánlások elemzése, valamint különböző esettanulmányok, szakpolitikai jelentések, hatástanulmányok, kockázat- és megfelelésértékelések értelmezése és összefoglalása adta.

A választott téma több tudományterületet (állam- és jogtudományok, politológia, algoritmusetika és szociálpszichológia) érint egyszerre. A dolgozat interdiszciplináris jellege lehetőséget biztosít az MI/GenMI technológiákkal kapcsolatos jelenségek különböző nézőpontokból történő átfogó vizsgálatára. Relevanciáját és kitűzött célját abból az elgondolásból meríti, hogy az e technológiák széleskörű elterjedéséhez köthető rövid és hosszútávú társadalompolitikai hatások feltérképezése, illetve az azokkal kapcsolatos szabályozási kihívások részletes tanulmányozása hasznos és szükséges információval szolgálhat az illetékes döntéshozók számára. Minden e területen végzett kutatás, megkezdett diskurzus értékes iránymutatásokat nyújthat a hatékony szabályozási keretek megalapozásához, biztosítva ezáltal, hogy az MI jövőbeni fejlesztése, forgalmazása és felhasználása etikus, felelős és a demokratikus államberendezkedések és egyetemes emberi jogok társadalmi értékeivel összhangban lévő legyen. Ez kiváltkép igaz azon GenMI modellekre, melyek relatíve új kategóriát képviselnek az MI technológiacsalád területén, és amelyekkel kapcsolatosan mind a gyakorlati, mind a jogalkotási tapasztalatok korlátozott mennyiségben állnak csak rendelkezésre.

1.4 A dolgozat központi kérdései, hipotézisei

A kockázatalapú megközelítésen alapuló Uniós szabályozás érdeme, hogy helyesen mérte fel a nagy és elfogadhatatlan kockázatúként kategorizált MI alkalmazások jövőbeli potenciális veszélyeit. Az Uniós jogalkotó szervek az elmúlt két évben számos alkalommal előremutató módon voltak képesek pontosítani a jogszabályszöveget e gyakorlatok meghatározásával és besorolásával kapcsolatban. Például, hallgatva az Európai Adatvédelmi Testület és az európai adatvédelmi biztos 5/2021. sz. közös véleményére, helyesen ismerte fel a Tanács, hogy az általános társadalmi pontozás gyakorlatának tilalmát a közszektorról a magánszereplőkre is ki kell terjeszteni. Szintén pozitív fejleményként értékelhető, hogy a végleges szöveg tartalmazza a Régiók Európai Bizottságának azon korábbi javaslatát, miszerint a magatartást torzító MI gyakorlatok tilalmát az életkor, és a szellemi és testi fogyatékosságon túl azon személyek körére is ki kell terjeszteni, akiknek sebezhetősége gazdasági kiszolgáltatottságukból fakad.

Azonban a kockázatalapú szabályozási paradigma – amely a technológia specifikus felhasználási eseteinek szabályozására összpontosított – nem mindig tekinthető hatékony megközelítésnek az alapmodellekre jellemző többcélú és dinamikus felhasználási módozatok szabályozására.[6] Az EU rendelet-tervezete kezdetben nem is tartalmazott kötelezettségeket a GenMI típusú alapmodellekre vonatkozóan azok újdonsága és a technológia relatív ismeretlensége miatt. A technológiai változásra a Tanács reagált elsőként, így a 2022. december 6-án közzétett közös álláspontjában egy teljesen új MI kategóriának a bevezetését javasolta *általános célú MI* néven. Ezt követően a Parlament 2023. június 14-én elfogadott kompromisszumos szövegében javaslatot tett két új típus – az alapmodell és a GenMI – fogalmainak megalkotására. Az elsőt úgy definiálta, mint olyan MI modellt, *amelyet sokoldalúságra terveztek, különféle feladatok elvégzésére képesek és széles körű adatforrásokon képeztek*, az utóbbit pedig olyan MI rendszerként határozta meg, *amely összetett tartalmakat, például videót, hangot és kódot állít elő, különböző autonómia szinteken.*

[6] Az EU MI jogszabálya két olyan szabályozási eszközt is alkalmaz, amelyek alapvetően a hagyományos termékbiztonsági és felelősségi jogszabályokon alapulnak: ex-ante (megelőző) kötelezettségek és ex-post (utólagos) felelősségi szabályok. Az ex-ante kötelezettségek célja a megelőzés, azaz hogy az MI-rendszerek biztonsága és megfelelősége már a használatba vétel előtt biztosítva legyen. Ez azt jelenti, hogy a rendszerek fejlesztőinek és üzemeltetőinek már a bevezetés előtt meg kell határozniuk, milyen potenciális veszélyekkel járhat a rendszer működése. Ez azonban nehézséget okozhat a GenMI modellekkel kapcsolatban, mivel ezek gyakran általános célú rendszerek, amelyeket különböző területeken és módokon alkalmaznak. Ez megnehezíti a kockázatelemzést, hiszen a különböző felhasználási területek eltérő kockázatokat hordoznak. Például egy blogírásra használt GenMI modell kevesebb kockázatot jelent, mint egy orvosi diagnózisra vagy jogi döntéshozatalra alkalmazott rendszer (még akkor is, ha a program működési elve ugyanaz).

A 2023 őszén zajló háromoldalú egyeztetések során a legnagyobb vita az alapmodellekkel és az általános célú MI-rendszerekkel kapcsolatosan alakult ki. Bár Franciaország, Olaszország és Németország a túlzott szabályozás innovációkat bénító hatását hangsúlyozta, az Unió jogalkotói a szigorúbb szabályozást szorgalmazták, és végül egy kétszintű szabályozási keret elfogadása mellett döntöttek. Ennek értelmében a jogszabály két különböző típusú általános célú MI-rendszert határoz meg, amelyekre eltérő szabályok és jogi kötelezettségek vonatkoznak. Az első kategóriát a normál alapmodellek alkotják, melyek alacsonyabb és kevésbé kiterjedt kockázatot hordoznak, míg a második kategóriát az un. *nagy hatású rendszerszintű kockázatot jelentő alapmodellek* képezik. Az ilyen modellek szolgáltatóira szigorúbb kötelezettségek vonatkoznak majd, amelyek magukban foglalják a modellek értékelését, a rendszerszintű kockázatok felmérését és mérséklését, az ilyen kockázatokról szóló jelentéstételt a Bizottságnak, támadhatósági tesztek végrehajtását, valamint a magas szintű kiberbiztonsági védelem biztosítását.

Annak ellenére, hogy a jogszabályalkotási folyamat során eszközölt változtatások nem mindegyike szolgálta feltétlenül az uniós polgárok biztonságát,[7] összeségében megállapítható, hogy az Unió jelentős előrelépést tett azzal, hogy létrehozott egy átfogó jogi keretet az MI szabályozására. A töredezett, nemzetállami szintű jogalkotás, ahol a fejlesztők, szolgáltatók és alkalmazók több különböző ország joghatósága alatt állnak – mely okán eltérő jogok és kötelezettségek vonatkoznak rájuk – gátolná a biztonságos MI fejlesztésének és alkalmazásának lehetőségét. Az EU által megalkotott egységes jogszabály – a maga 450 milliós piacával – fontos lépés ezen akadályok mérséklésére. Azonban – ahogy Karácsony is utal rá – könnyen elképzelhető, hogy az uniós szintű szabályozás nem bizonyul elegendőnek, tekintettel arra, hogy az MI fejlesztésének központjai az Egyesült Államokban és Kínában egyaránt megtalálhatók.[8] Hosszabb távon fontos annak a tudatosítása, hogy a szabályozási kereteknek lehetőleg – globális, kontinenseken átnyúló szinten – összhangban kellene lenniük egymással. Ennek kialakítása még várat magára.

A disszertáció második része arra keresi a választ, hogy miként kezelhető a GenMI modellek felhasználásával létrehozott példátlan mennyiségű szintetikus tartalom közeljövőben várható gyors és tömeges elterjedése. A dolgozat tézise, hogy az elmúlt három évtizedben a kollektív

[7] Például a Parlament 2023-ban megfogalmazott javaslata, amely az érzelemfelismerő MI-k alkalmazását a munkahelyeken és az oktatási intézményeken túl a bűnüldözés és a határigazgatás területein is tiltani rendelte volna, nem került be a végső szövegbe.
[8] G. Karácsony Gergely (2020): Okoseszközök – Okos jog? A mesterséges intelligencia szabályozási kérdései. Budapest, Dialóg Campus. 142.o.

valóságérzékelést a párhuzamos nyilvánosságok kialakulása és az információs környezet fokozatos perszonalizációja számottevően torzította. A napjainkra jellemző politikai csoportpolarizáció, dezinformációs kitettség, és a hagyományos intézményrendszerekbe vetett bizalom általános csökkenése mind tünetként értelmezhetőek e folyamatban. Azonban a szöveges és audiovizuális tartalmak előállítására alkalmas szabadon elérhető GenMI modellek elterjedése hamar e trend kezelhetetlen eszkalálódásához vezethet.

Az EU a szintetikus tartalmak szabályozása tekintetében az állhírek és politikai deepfake tartalmak, a személyre szabott dezinformációs kampányok, valamint a manipulatív gyakorlatok visszaszorítására összpontosít. Bár ezek fontos lépések az MI-hez köthető ártó gyakorlatok korlátozásában, mégsem nyújtanak elegendő védelmet a mesterségesen előállított tartalmak elterjedését illetően. Az MI-rendelet 50. cikk (4) bekezdése például jól érzékelteti, hogy az esetek egy jelentős részében az állampolgárok számára nem garantált az a jog, hogy tájékoztattassák őket arról, ha szintetikus tartalmakat fogyasztanak.

Jelen dolgozat feltételezése szerint a szintetikus tartalmak elterjedésének legnagyobb hosszútávú kockázata nem a hibrid hadviselés felerősödése, a választások befolyásolása, a tömeges manipulálás és politikai polarizáció fokozódása (ezek rövidtávú veszélyek), hanem az *episztemológiai összeomlás*. Annak a veszélye, hogy a társadalom tagjai végérvényesen elveszítik a bizalmukat az általuk olvasott hírek, információk, tudományos alapvetések és korábban általános konszenzust élvező történelmi események hitelességében. Márpedig a tényekkel kapcsolatos egyetértés, a demokratikus intézményekbe vetett közbizalom és az azonos referenciapontokkal bíró konstruktív politikai vita képes egyedül biztosítani a társadalmak szükséges alkalmazkodóképességét az előttük álló globális átalakulások sikeres navigálásához.

2 Az MI (rövid) története

2.1 Theseus-tól a ChatGPT-ig

A modern értelemben vett MI fejlődéstörténetének gyökerei az 1940-es és 1950-es évekig nyúlnak vissza. Ekkor vették kezdetét azon alapvető elméleti tudományos kutatások – kiváltképp a matematika, a logika és a számítástudományok terén –, melyek később megalapoztak a ma ismert MI technológiáknak.[9] A korai számítógépes rendszerek közül kiemelendő a Claude Shannon amerikai mérnök-matematikus által 1950-ben megalkotott *Theseus* mechanikus gépegér, mely rendszer különlegessége abban rejlett, hogy nemcsak eljutott a kísérleti labirintusának kijáratához, de a beépített memóriájának köszönhetően el is tudta raktározni az előzőleg elvégzett műveletsorozatot – másképpen megfogalmazva „emlékezett" arra az útvonalra, amit a kezdőpontól a célállomásig megtett.[10]

Ugyanezen évben, 1950-ben közölte nagyhatású cikkét „*Computing Machinery and Intelligence*" címmel Alan Turing brit matematikus. A cikkben Turing egy gondolatkísérlet erejéig felvázolt egy általa kitalált játékot (*Imitation Game*, kisebb változtatásokkal ezt nevezzük ma Turing-tesztnek), ami arra szolgált, hogy megállapítsa képes-e egy gép *intelligens* módon kommunikálni. A teszt három résztvevőjét – egy számítógépet és két embert (válaszoló / kérdező) – külön szobákba helyeztek, a kérdező tudta, hogy az egyik résztvevő gép, a másik ember, de arról nem volt tudomása, hogy melyik válaszoló melyik. A kérdező – szöveges csatornán keresztül – kérdéseket tett fel az emberi részvevőnek és a számítógépnek is, hogy megállapítsa, melyik válaszoló az ember és melyik a számítógép. A gép célja olyan kimenetek generálása volt, amelyek megkülönböztethetetlenek az emberi

[9] Példaként – némi elfogultsággal – említhető a magyar származású matematikus és fizikus, Neumann János munkássága, akinek 1945-ben publikált híres modellje – a Von Neumann-architektúra – a modern számítógépek felépítésének máig érvényes elméleti alapjait fektette le. A Neumann-modell alapján egy számítógép öt fő komponensből áll, melyek a: (1) Feldolgozó Egység (CPU); (2) Memória; (3) Bemeneti Egység (Input); (4) Kimeneti Egység (Output); (5) Vezérlő Egység (Control Unit). Lásd bővebben: Von Neumann, J. (1945): First Draft of a Report on the EDVAC. Az eredeti szöveg beszkennelt formában angolul elérhető: https://archive .computerhistory.org/resources/text/Knuth_Don_X4100/PDF_index/k-8-pdf/k-8-u2593-Draft-EDVAC.pdf

[10] Ennek a képességnek köszönhetően a gépegér már nem véletlenszerűen tájékozódott a labirintusban, hanem a korábban bejárt útvonalra alapozva optimalizálta mozgását, lecsökkentve a labirintusból kivezető út hosszát. Jóllehet ma már kezdetlegesnek tűnhet e rendszer, jelentősége mégsem elhanyagolható, hiszen a Theseus volt az egyik első példája annak, hogy gépek bizonyos szinten megtanulhatják és memóriájukba véshetik az elvégzett feladatokat. Bővebben: Klein, Daniel (2018): Mighty mouse. *MIT Technology Review*. Elérhető: https://www.technologyreview.com/2018 /12/19/138508/ mighty-mouse (2018. 12. 19.)

válaszoktól. Amennyiben a kérdező nem tudta megbízhatóan megkülönböztetni a gépet az embertől, a gép átment a teszten.[11] Bár a cikk révén népszerűvé váló Turing-tesztet az évek során sok kritika érte – elsősorban a kommunikációs képesség magas szintű utánzása és az absztrakt emberi intelligencia párhuzamba állítása okán – mégis az imitációs játék gondolatkísérlete és maga Turing is tartós hatást gyakorolt arra, miként határozták meg sokáig az intelligens rendszerek kritériumait.

A *mesterséges intelligencia* kifejezést elsőként John McCarthy amerikai matematikus, a Stanford egyetem professzora használta szélesebb közönség előtt az 1956-ban megtarott Hampshire-i Dartmouth Egyetem nyári workshopján. Az egy évvel korábban publikált kutatási projektjavaslatukban McCarthy és kollégái olyan jövőbeni irányvonalakat fogalmaztak meg a gépi intelligencia területén végzett kutatásokban, melyek főként annak problémamegoldási képességeire összpontosítottak (automatizáció, önfejlesztés képessége, kreativitás, természetes nyelv feldolgozása, elvont fogalmak értelmezése).[12] Többek között a Dartmouthi Egyetem kutatóihoz köthető diszciplináris keretrendszernek köszönhető, hogy a fogalom olyan gépekkel összefüggésen vált széles körben használatossá, melyek az emberi intelligenciát igénylő feladatok és problémák elvégzésére válhat alkalmassá.

Annak ellenére, hogy az MI területén végzett kutatások jelentős áttöréseket is hoztak, a korai ígéretek túlságosan nagyratörőnek és optimistának bizonyultak, így fokozatos csalódottság és szkepticizmus alakult ki a befektetők körében, ami az 1970-es években a technológiai kutatások, innovációkkal kapcsolatos általános érdeklődés csökkenéséhez, un. „MI télhez" vezetett.[13] Nem telt el sok idő azonban, és az 1980-as évek második felétől kezdődően újra

[11] Lásd bővebben: Turing, A. M. (1950): Computing Machinery and Intelligence. *Mind, 59(236), 433–460.o.* Az eredeti cikk szövege magyarul elérhető Tarján Rezsőné fordításában: https://www.geier.hu/GOEDEL/Turing/Turing.html

[12] A projekt során a következő hét alapvető kutatási területet határozták meg: (1) Automatizált számítógépek, melyek emberi beavatkozás nélkül is képesek lehetnek különböző feladatok elvégzésére; (2) Olyan modellek vizsgálata, mely lehetővé teszi a számítógépek számára az emberi természetes nyelvek megértését és használatát; (3) Olyan, az emberi agy szerkezetét és működését utánzó neurális hálózatok tanulmányozása, amelyek képesek a tanulásra és a mintafelismerésre; (4) Az algoritmusok, számítási modellek és számításelméleti problémák vizsgálata a számítógépek elméleti képességeinek és korlátjainak feltérképezésére; (5) Kutatások a gépi önfejletszés és öntanulás területein (gépi tanulás); (6) A feldolgozott adatokból történő absztrakcióképzés gépi módszereinek leírására; (7) Kutatások a kreativitás gépi szimulálásával kapcsolatban, főként arra vonatkozóan, hogy miként generálhatnak a számítógépek új ötleteket vagy megoldásokat. A kutatási projekt részvevői: John McCarthy, Marvin Minsky, Allen Newell, Arthur Samuel, Herbert Simon. A kutatási javaslatról bővebben: McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955): *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Az eredeti szöveg angol nyelven elérhető: https://www.formal.stanford.edu/jmc/history/ dartmouth/dartmouth.html

[13] Az első meghatározó MI tél az 1970-es évek közepén zajlott, amit több kisebb visszaesés is követett az 1980-as 90-es és a 2000-es években. (A legkorábbi – igaz kevésbé meghatározó – MI telet az 1950-es évekre datálják, amikor is egy viszonylag egyszerű, orosz tudományos szövegeket angolra fordító program kezdeti sikerén

nőni kezdett a lelkesedés a technológiacsalád iránt, majd a következő évtized elejére a gépi tanulás fejlődésével és a neurális hálózatokkal kapcsolatos kutatások megindulásával beköszöntött az „MI tavasz”.[14]

Az 1990-es években az internet elterjedése és a szabadon hozzáférhető digitális adatmennyiség exponenciális növekedése további új lehetőségeket nyitottak a kutatások területén. Ebben az időszakban született meg az IBM által fejlesztett Deep Blue sakkszámítógép is, amely 1997 májusában egy hatjátszmás páros mérkőzésen 3,5-2,5 arányban legyőzte Garry Kasparovot, az akkori sakkvilágbajnokot.[15] Az emlékezetes meccs nemcsak a sakk, hanem az ember-gép interakció történetében is jelentős pillanat volt; ez tekinthető az első nyilvános eseménynek, mely ráirányította a társadalom számottevő részének a figyelmét az MI-ben rejlő potenciális képességekre. Az IBM által tervezett Deep Blue működése még nem gépi tanuláson vagy neurális hálózatokon alapult, pusztán hatalmas számítási kapacitására és az előre beprogramozott sakkstratégiákra támaszkodott. Erősségét innovatív (alfa-béta) keresőalgoritmusa és célhardver adta, melynek köszönhetően másodpercenként megközelítőleg 200 millió pozíciót is képes volt áttekinteni, szisztematikusan végig számolva a lehetséges lépéseket.[16]

A következő nagyobb visszhangot kiváltó esemény szintén egy IBM által fejlesztett MI-hez, a Watson nevű természetes nyelvfeldolgozó rendszerhez (Natural Language Processing, röviden NLP) volt köthető, amely 2011-ben két emberi versenyzővel megmérkőzve ért el első helyezést a „Jeopardy!" nevű amerikai kvízműsorban.[17] A gép megértette az emberi nyelven megfogalmazott kérdéseket, és a kiterjedt adatbázisában tárolt információk alapján képes volt rendkívül alacsony hibaaránnyal helyes válaszokat generálni azokra. Ez a siker rávilágított,

---

felbuzdulva sokan az intelligens beszélő rendszerek korai eljövetelét vizionálták). Ezekben az időszakokban a korábbiakhoz képest lassabb ütemben haladt a kutatás, több projektet félbehagytak, és számottevően csökkent a tudományterületekre szánt finanszírozás is. Bővebben: Floridi, Luciano (2020): AI and Its New Winter: from Myths to Realities. *Philosophy & Technology*, 33, 1-3.o. https://doi.org/10.1007/s13347-020-00396-6

[14] Suleyman, Mustafa, Michael Bhaskar (2023): A következő hullám. Mesterséges intelligencia, technológia, hatalom és a 21. század legnagyobb kihívása. (ford. Farkas Veronika). Budapest, Magnólia kiadó. 74-76.o.

[15] A Carnegie Mellon Egyetem kutatói a 80-as évek közepén kezdték el fejleszteni Deep Thought (korábban ChipTest) nevű sakkprogramot, amely az évtized végére, 1989-ben legyőzte a dán Bent Larsen sakknagymestert. E kutatásokba szállt be később az IBM is, és fejlesztette ki saját programját – immár Deep Blue néven. Garry Kasparov az elhíresült mérkőzést megelőző évben 1996 februárjában még képes volt – viszonylag könnyedén – 4-2-es arányban legyőzni ezt a programot. Bővebben: Hassabis, D. (2017): Artificial Intelligence: Chess match of the century. *Nature*, 544, 413-414.o. https://doi.org/10.1038/544413a

[16] Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep Blue. *Artificial Intelligence, 134(1-2*), 57-83.o.

[17] A Jeopardy! egy népszerű amerikai televíziós kvízműsor, amelyben három játékos versenyez. A játék során a kérdések válaszok formájában jelennek meg, és a játékosoknak a helyes „kérdést" kell megfogalmazniuk az előre megadott válasz alapján. Például, ha a válasz így hangzik: „Magyarország ötödik legnagyobb városa, egyben híres egyetemi központ a Dél-Dunántúlon", akkor a helyes kérdés – és megoldás – így szólhat: „Mi Pécs (What is...Pécs)?"

hogy a MI rendszerek immár alkalmasak lehetnek a nyelvfeldolgozás és más *tudásalapú* (expert system) feladatok ellátására is.

A 2010-es évektől kezdődően az MI fejlődését elsősorban már a neurális hálózatok[18] és mélytanulási módszerek elterjedése, valamint a nagy adathalmazok (Big Data)[19] rendelkezésre állása fűtötte. A mélytanulás – ahogy arról a következő fejezetben szó lesz – lehetővé tette a gépek számára, hogy összetettebb mintákat és kapcsolatokat ismerjenek fel az adatokban, míg a Big Data biztosította a szükséges adatmennyiséget a modellek hatékony betanításához. Ezen technológiák sikeres adaptálásáról tanúskodott a Google DeepMind elsöprő sikere is; 2017 decemberében a Google DeepMind által fejlesztett AlphaZero program azzal vonta magára a tudományos világ figyelmét, hogy egy száz játszmából álló mérkőzésen legyőzte a 2016-os számítógépes sakk-világbajnokot, Stockfish8-at. Az Alphazero mindössze négy óra alatt sajátította el a sakk szabályait, anélkül, hogy a hagyományos értelemben programozták, vagy „tanították" volna a játékra. A Google fejlesztése a saját maga ellen

[18] Neurális hálózatok az 1980-as évek vége felé kezdtek megjelenni az akadémiai kutatásokban, és a 90-es években kezdtek el komoly figyelmet kapni, majd a 2000-es évek elején kezdtek elterjedni a gyakorlati alkalmazások a természetes nyelvfeldolgozás terén.

[19] A hétköznapi értelemben a Big Data, vagyis a nagy adatmennyiség kifejezés egyszerre utal a méretükben és összetettségükben jelentős adatkészletekre, valamint arra a technológiai eszköztárra is, amely képes összegyűjteni, tárolni és feldolgozni azokat. Kezdetben a fogalmat – Douglas Laney 2011-es publikációját követően – az adatok három V-je segítségével határozták meg, ami a *Volume* (mennyiség), *Variety* (változatosság), és a *Velocity* (sebesség) szavakat jelölte; (1) A hatalmas Mennyiség a Big Data talán legfontosabb jellemzője. Manapság az egyének, vállalatok, államigazgatási szervek egytől egyig nagy mennyiségű adatot állítanak elő online/ offline tevékenységeiken keresztül. Az előállított adatok robbanásszerű növekedése nem kis mértékben köszönhető a mobil és egyéb digitális eszközök, az IoT szenzorok, a közösségi média és más platformszolgáltatások elterjedésének is; (2) A Sebesség az adatok keletkezésének és áramlásának a gyorsaságára utal. A valós idejű adatforrások, mint az érzékelők, streamingszolgáltatások, mobil eszközök és online tranzakciós rendszerek folyamatosan ontják az új adatokat bármilyen interakció során; (3) A Változatosság az adatok típusainak és formátumainak sokféleségére utal. A Big Data esetében a legkülönbözőbb forrásokból érkező adatok lehetnek strukturáltak (adatbázis táblázatok), félig strukturáltak (XML dokumentumok), vagy teljesen strukturálatlanok (szabad formátumú szövegek, mint például a felhasználói hozzászólások egy kommentszekcióban). Bár ez a változatosság bonyolulttá teszi az adatkezelést és az adatinegrációs és adatelemzési feladatokat, mégis nagy potenciált kínál a legkülönbözőbb területekekről begyűjtött információk hasznosítására. Lásd: Gartner (2012): Gartner Research. The Importance of 'Big Data': A Definition. Elérhető: https://www.gartner.com/en/ documents/2057415 (2020. 07. 12). De a Big Data koncepció kiegészülni látszik további 2 V-vel is, melyek a *Veracity* (Hitelesség) és *Value* (Érték) szavakat takarják. Az első V arra utal, hogy a nagy adathalmazok esetén gyakran nehéz megbizonyosodni azok megbízhatóságáról és pontosságáról, mivel azok sokszor eltérő forrásokból és változó minőségben érkeznek. Az Érték pedig azt a felismerést tükrözi, hogy az adatokban rejlő információk - megfelelő adatelemző eljárásokkal - valódi gazdasági vagy társadalmi hasznot rejtenek. (Tovább bonyolítandó, de ehelyütt részletesen nem tárgyalva Eileen McNulty 2014-ben már V7ről is értekezett, amely a Variability és Visualisation jellemzőkkel történő kibővítéssel jönne létre. McNulty, Eileen (2014): Understanding Big Data: The Seven V's. DataConomy. Elérhető: https://dataconomy.com/2014/05/22/seven-vs-big-data (2014. 05. 22.) Témáról bővebben: Viktor Mayer-Schönberger - Kenneth Cukier (2014): Big data. Forradalmi módszer, amely megváltoztatja munkánkat, gondolkodásunkat és egész életünket. Ford. Dankó Zsolt. Budapest, HVG Könyvek; Szőke Gergely László (2018): Big Data and Algorithms in the Public Sector and Their Impact on the Transparency of Decision-Making. *Central and Eastern European eDem and eGov Days*. 301-306.o.; Ződi Zsolt: Jog és jogtudomány a Big Data korában. *Állam- és Jogtudomány,* 2017/1.

folytatott több millió játszma során gyűjtött adatok (pl.: sikeres - sikertelen lépéskominációk) alapján folyamatosan finomította és fejlesztette saját stratégiáját. A tapasztalatokat beépítve a neurális hálózatába, az AlphaZero képessé vált a játék mélyebb megértésére és új stratégiák kifejlesztésére. [20] Ez a fajta *megerősítéses tanulás* lehetővé tette az AlphaZero számára, hogy saját maga fedezze fel a játék mélységeit és innovatív, kreatív lépéseket alkalmazzon. Az AlphaZero megközelítése radikális eltérést jelentett a hagyományos, adatbázis-alapú MI programoktól, mint amilyen a Deep Blue volt.[21] Míg a Deep Blue a hatalmas adatbázisokra és előre betáplált stratégiákra támaszkodott, az AlphaZero teljesen önállóan, a mély tanulás és a neurális hálózatok segítségével fejlesztette ki saját játékstratégiáit anélkül, hogy előre beprogramozott vagy emberi sakkozók által kikísérletezett lépéssorozatokat használt volna.

A 2022-es év újabb mérföldkőnek tekinthető az MI fejlődésében, különösen a GenMI modellek területén. Ebben az évben vált széles körben elérhetővé és ingyenesen hozzáférhetővé a nagyközönség számára is a ChatGPT Nagy Nyelvi Modell (Large Language Model, LLM), amely máig e modellcsalád talán legismertebb képviselője. Ezek alapjául a későbbiekben még tárgyalandó a GAN (Generative Adversarial Networks) és a transzformer-alapú mélytanulás modellek szolgálnak. A szoftver megjelenése óta alig telt el másfél év, de úgy tűnik a lelkesedés az MI rendszerek új technológiája iránt töretlen. Kutatók, fejlesztők és felhasználók egyaránt keresik a módját, hogyan lehet a GenMI modelleket különböző területeken alkalmazni, legyen szó szövegírásról, képgenerálásról, kódolásról vagy éppen adatelemzésről. A technológiai területen gyakran hivatkozott Gartner-féle hype-ciklust[22]

[20] A Deepmind fejlesztései nem csak a sakk, hanem a go és a sógi játékokban is képes volt új stratégiák kidolgozására és a világ legjobb játékosainak legyőzésére. 2016. március 9-15. között, a szintén DeepMind által fejlesztett AlphaGo megmérkőzött Lee Sedol-lal, a világ egyik legjobb go játékosával, akit 4:1 arányban legyőzött. Ez az esemény azért volt figyelemreméltó, mert a go játék rendkívül összetett, a lehetséges lépések és állások száma emberi tudattal szinte felfoghatatlan (2 x 10^170), amely hatalmas kihívás a számítástechnikai kapacitások és keresőalgoritmusok szempontjából. Nem is csoda, hogy az áttörést végül a Google mélytanulás módszerei hozták el.

[21] A Deep Blue programot még úgy tanították, hogy millió létező sakklépést és játékstratégiát tápláltak bele. Ennek köszönhetően a lehetséges (ismert és valós mérkőzéseken alkalmazott) lépések széles körét ismerte, mielőtt szembe került volna egy emberi ellenféllel. Ez a megközelítés bár fejlett adatelemzést és jelentős számítástechnikai teljesítményt igényelt, mindig kizárólag csak arra a játéktípusra lesz alkalmazható amire előzetesen betanították. Lásd: Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002): Deep Blue. *Artificial Intelligence*, 134(1-2), 59-60.o.

[22] A Gartner féle hype-ciklus (Gartner Hype Cycle) szerint a technológiai innovációk fejlődésében és fokozatos elterjedésében öt szakasz különböztethető meg: (1) Technológiai áttörés (Technology Trigger): Egy adott technológia korai sikertörténetei okán az iránta tanúsított érdeklődés aránytalanul nagyra duzzad, a közbeszédet is uralja a téma. A média által nagy nyilvánosságot kap a terület, bár gyakran sem működő prototípusok, sem létező gazdasági modell nem épült még fel a technológia körül; (2) A felfokozott várakozások csúcsa (Peak of Inflated Expectations): Lényegesen nő az új technológiát ismerők és azt amatőr szinten használók száma. Egymást érik a fantasztikus áttörésekről szóló gyakran szenzációhajhász hírek. A technológia iránti lelkesedés a csúcspontjára ér, sokszor irreális elvárásokkal és túlzott optimizmussal; (3) A csalódás görbéje (Trough of Disillusionment): A felmerülő problémák növekednek, a forradalmi áttörések száma csökken. Amikor a

figyelembe véve – mely az innovációk és az azzal kapcsolatos társadalmi attitűdök jól megfigyelhető trendjét osztja öt szakaszra – joggal feltételezhető, hogy a GenMI modellekkel kapcsolatos társadalmi reakciók a ciklus második fázisánál tartanak. Erre az időszakra jellemző a feltűnő és médiatérben is hangos, átütő sikerek miatti hatalmas felhasználói és médiafigyelemmel övezett, folyamatosan növekvő és gyakran alaptalan lelkesedés. Kérdés azonban, hogy mikor éri el – ha egyáltalán eléri – a harmadik szakaszt, amely jellemzően az új technológiával szembeni fokozatos kiábrándulással és apátiával jár. Mivel a GenMI modellek minden bizonnyal szerves részét fogják képezni az MI technológiacsaládnak, e téma a disszertáció második részében még részletesebben is kifejtésre kerül.

2.2 Az MI-ről általában

Nem várt nehézségekbe ütközhet az, aki az MI fogalmának átfogó, technológiasemleges meghatározására vállalkozik. Bár az Európai Unió többéves jogalkotási folyamata sikeres erőfeszítésnek tekinthető egy időtálló MI definíció megalkotására – erről a későbbiekben részletesen is szó lesz – valójában mindmáig nincsen egyetlen általánosan elfogadott, egységes meghatározás e rendszereket illetően. Vagy másképpen fogalmazva: számtalan széles körben elfogadott értelmezés létezik párhuzamosan.

Stuart Russell és Peter Norvig *Artificial Intelligence: A Modern Approach* című 1995-ben megjelent könyvükben a következő négy szempont szerint csoportosította az MI-hez köthető kutatások és fejlesztések irányát: (1) emberi gondolkodás, (2) racionális gondolkodás, (3) emberi viselkedés, (4) racionális viselkedés. Minden kategória azt vizsgálja, hogy hogyan

technológia nem tudja teljesíteni a magas elvárásokat, csalódás következik be, és az érdeklődés csökken. A technológiához kapcsolódó kutatások és a becsatornázott tőke is lényegesen apad. A fejlesztések menete akár zsákutcába is fordulhat; (4) A megvilágosodás emelkedője (Slope of Enlightenment): A technológia valós lehetőségeinek és korlátjainak fokozatos felismerése következik. A korai tapasztalatok szülte módosítások és a használat közben felgyülemlett tudás révén lassan kikristályosodik a technológia használhatóságának köre és módja. Bár a befektetők nagy része még óvatos, kezd kialakulni egy stabil, megbízható felhasználói réteg; (5) A produktivitás fennsíkja (Plateau of Productivity): A technológia eléri azt a szintet, ahol stabilan és hatékonyan használható, és valós értéket képes létrehozni. A termék a hétköznapok részévé válik, használata elterjed, megtalálja helyét a piacon. Lásd bővebben: Fenn, Jackie; Raskino, Mark (2008): Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Massachusetts, Harvard Business Publishing.

értelmezhető az intelligens rendszer fogalma és hogy miként valósíthatók meg a gyakorlatban ezek az elképzelések. [23]

(1) Az első felfogásban az emberi gondolkodás utánzásán van a hangsúly, ahol olyan gépek létrehozása a cél, melyek az emberi gondolkodási folyamatokat (pl.: tanulás, emlékezés, problémafelismerés, problémamegoldás) és az emberi megismerés jellemzőit (pl.: annak megértése, hogy mások meggyőződései, vágyai és szándékai hogyan befolyásolják döntéseiket) képesek imitálni; (2) A második kategóriában a racionális gondolkodáson van a hangsúly, melyben a gépek létrehozásánál a logikai érvelés tökéletesítésére összpontosítanak, olyan gépek fejlesztésére, amelyek képesek racionális döntéseket hozni és logikailag megalapozott következtetéseket levonni; (3) A harmadikg típusnál a cél, hogy olyan gépeket hozzanak létre, amelyek képesek az emberi viselkedés utánzására. Például, ha egy gép képes oly módon kommunikálni az emberrel, hogy az nem veszi észre, hogy egy géppel folytat párbeszédet – melynek mérésére a Turing-teszt is irányul – akkor valószínűleg intelligens rendszerről van szó; (4) A negyedik kategória esetében a hangsúly a racionális, cél-orientált viselkedés megvalósítására összpontosul. E felfogásban a cél, hogy a gépek képesek legyenek az optimális döntések meghozatalára a rendelkezésre álló információk és célkitűzések alapján.

Ahogy e kategorizálásból is kitűnik, egy könnyen érthető és kellően általános megfogalmazásban az MI azon technológiai rendszerek összeségét jelöli, amelyek képesek utánozni az olyan emberi kognitív funkciókat, mint az érzékelés, logikai következtetés, problémamegoldás és a tanulás. Efféle kézenfekvő értelmezést sugall az MIT fizikaprofesszora, Max Tegmark is, aki szerint legáltalánosabb felfogásban az MI-re, mint az emberi intelligenciát imitálni képes rendszerre érdemes tekinteni. Tegmark szerint az intelligencia nem más, mint az *összetett célok elérésének képessége*.[24] Tartalmát tekintve hasonló, gyakran idézett definíciót alkotott az amerikai számítógép-tudós Nils Nilsson is, aki úgy véli „*az MI olyan tevékenység, amely arra irányul, hogy a gépeket intelligenssé tegye. Az*

[23] A négy kategória angolul: (1) Thinking Humanly; (2) Thinking Rationally; (3) Acting Humanly; (4) Acting Rationally. Lásd bővebben: Russell, Stuart J. – Norvig, Peter (2010): Artificial Intelligence: A Modern Approach. Third Edition. Essex, Pearson. 2-16.o.

[24] Tegmark, Max (2017): Élet 3.0. Embernek lenni a mesterséges intelligencia korában (Ford. Weisz Böbe, Garai Attila). Budapest, HVG Könyvek. 69.o.

intelligencia pedig olyan tulajdonság, amely lehetővé teszi egy adott entitás számára, hogy megfelelően és előrelátóan működjön környezetében"[25].

Azzal összefüggésben, hogy e célokat milyen mértékben képes teljesíteni egy adott MI, két fő kategóriát szokás elkülöníteni: a *szűk* (vagy gyenge), valamint az *általános* (vagy erős) MI.

2.3 Szűk és Általános MI

A szűk MI (narrow AI) kifejezés olyan szoftvereket takar, melyek egy adott probléma megoldására, vagy egy konkrét feladat végrehajtására specializálódtak. Ezek a rendszerek egy előre meghatározott feladatkörben működnek, és működésüket egy dedikált szabályrendszer határozza meg, amely informatikai nyelvre van lefordítva. Ezen a behatárolt területen a szűk MI gyakran képes az emberi teljesítményt felülmúlni, ugyanakkor ez a kiemelkedő teljesítmény csak az adott feladat mechanikus végrehajtására korlátozódik, mivel nem rendelkezik komplex problémamegoldó képességgel vagy általános intelligenciával. A szűk MI erőssége egyben a gyengesége is: a specializált feladatra való optimalizálás megakadályozza, hogy ezek a rendszerek az emberhez hasonló, széles körű problémamegoldó képességet érjenek el.[26] Hiányzik belőlük az általános intelligencia, a tanulási képesség és az alkalmazkodóképesség, amelyek az emberi gondolkodás sajátosságai, így a szűk MI rendszerek, bár kiemelkedően teljesítenek egy adott területen, nem képesek az emberhez hasonló módon kezelni az új, váratlan helyzeteket vagy a kontextus változásait. Szűk MI-k például azon rendszerek, melyeket adatelemzésre, kép- és mintafelismerésre vagy szövegértelmezésre használnak.

Ezzel szemben az általános MI (Artificial General Intelligence, AGI) egy olyan fejlett (általános) intelligencia[27], amely képes széles körű feladatok elvégzésére, önálló tanulásra és

[25] „Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment" Lásd: Nils J. Nilsson (2009): The Quest for Artificial Intelligence. Cambridge University Press. 13.o.

[26] Klein Tamás (2021): Robotjog vagy emberjog? Az emberközpontú mesterséges intelligencia szabályozásának kiindulási pontjai. In Török Bernát és Ződi Zsolt (szerk): A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 132.o.

[27] Peter Voss amerikai mérnök, feltaláló – a 2001-ben megalkotott AGI (Artificial General Intelligence) fogalom társalkotója – értelmezése szerint az Általános Intelligencia (General Intelligence) olyan alapvető készségeket jelent, amelyek nem kötődnek egy konkrét szakterülethez, hanem lehetővé teszik sokféle specifikus tudás és

alkalmazkodásra. Ez a típusú intelligens rendszer képes új, ismeretlen problémákat megoldani, és a legkülönbözőbb területeken nyújt kiemelkedő teljesítményt, miközben az emberi gondolkodás sajátosságait tükrözi. Képesek felismerni és megérteni összetett mintázatokat, észrevenni a kontextus változásait és új problémákat kreatív módon megoldani. Ezen képességek révén az általános MI nem korlátozódik egyetlen feladat mechanikus végrehajtására, hanem képes különböző területeken is kiemelkedően teljesíteni. Ezen kívül fejlett problémamegoldó képességekkel rendelkezik, amelyek révén új, ismeretlen helyzetekben is képes helytállni. Az ilyen rendszerek képesek összetett döntéseket hozni, hosszú távú következményekkel számolni és stratégiai gondolkodást alkalmazni. A fejlett általános MI még a jövő technológiája[28], de kétségtelen, hogy a következő alfejezetben tárgyalandó gépi tanulás eljárások jelentős előrelepéseket jelentettek a fejlesztéséhez vezető úton.

E kettő kategóriát gyakran ki szokták egészíteni egy harmadik – fiktív – intelligens rendszertípussal is; ezt hívják Mesterséges Szuperintelligenciának. Ez a típusú MI gyakorlatilag minden szempontból felülmúlná az emberi intelligenciát (a felhalmozott és elraktározott tudástól kezdve, a kreativitáson keresztül, a problémamegoldó képességekig). Nick Bostrom, az Oxfordi Egyetem filozófusa a Szuperintelligencia című könyvében e típus fejlődésének három fő szintjét írta le (1): Gyors Szuperintelligencia: olyan rendszer, ami mindenre képes amire az ember, csak sokkal gyorsabban; (2) Kollektív Szuperintelligencia:

képesség elsajátítását. Lényegében ez a "bármi megtanulásának képessége". Ennek az intelligenciának három fő jellemzője van: (1) Önállóság: A tanulás automatikusan történik, külső irányítás nélkül. Ez lehet passzív (amikor csak befogadja az információkat) vagy aktív (amikor kísérletezik és felfedez); (2) Célorientáltság: A tanulás mindig valamilyen cél elérésére irányul. Ezek a célok lehetnek beépítettek, kívülről meghatározottak vagy önállóan létrehozottak. Ez azt is jelenti, hogy szelektíven gyűjti az információkat a bonyolult környezetből; (3) Alkalmazkodóképesség: A tanulás folyamatos, integrálja az új információkat, figyelembe veszi a kontextust, és képes alkalmazkodni a változó célokhoz és környezethez. Ez nemcsak a fokozatos változásokra vonatkozik, hanem elősegíti teljesen új képességek kifejlesztését is. Lásd bővebben: Voss, Peter (2007): Essentials of General Intelligence: The Direct Path to Artificial General Intelligence. In: Goertzel, B., Pennachin, C. (szerk.): Artificial General Intelligence. Cognitive Technologies. Berlin, Springer. 131-157.o. https://doi.org/10.1007/978-3-540-68677-4_4

[28] A jelenlegi fejlesztések nagyrészt szűk MI irányába mutatnak, az AGI kutatása jelenleg számos akadályba ütközik, annak létrehozása komoly technológiai kihívást jelent, beleértve a gépi tanulás, a természetes nyelvi feldolgozás és az emberi tudat mélyebb megértésének szükségességét. De jelenleg a motivációk és ösztönzők sincsenek megfelelően összehangolva az AGI fejlesztésére: az akadémiai kutatás fő célja a publikáció, míg a vállalatoké a meghökkentő, figyelemfelkeltő bemutatók prezentálása (és az emberi teljesítmény felülmúlása valamilyen specifikus területen). Továbbá, még ha minden adott is lenne a fejlesztéséhez – elméleti alap, technológiai infrastruktúra, megfelelő csapat és finanszírozás – még mindig fennáll az úgynevezett „Szűk MI Csapda" (The Narrow AI Trap) veszélye is, miszerint; az emberi természet hajlamos arra törekedni, hogy a lehető legrövidebb idő alatt a lehető legnagyobb (maximális) előrelépést mutassa, ami gyakran azt eredményezi, hogy inkább egy adott területen kimagasló eredményt nyújtó külső intelligenciát vesznek igénybe a specifikus célok elérése érdekében, ahelyett, hogy egy általános, autonóm és adaptív problémamegoldó rendszert hoznának létre. Lásd: Voss, Peter, Jovanovic, Mlađan (2023): Why We Don't Have AGI Yet. ArXiv, abs/2308.03598, 5-7.o. https://doi.org/10.48550/arXiv.2308.03598

nagyszámú kisebb intelligenciából felépített rendszer, amelynek összesített teljesítménye számos általános területen számottevő mértékben túllépi a mai rendszerek teljesítményét; (3) Minőségi Szuperintelligencia: olyan rendszer, amely legalább olyan gyors, mint az emberi elme, de képességeiben messze meghaladja az emberi intelligencia határait. [29]

2.4 Az öntanuló és önfejlesztő MI

Az MI gyűjtőfogalom számos különböző technológiát, eljárást és módszert foglal magában. Jelen dolgozatnak nem kitűzött – és terjedelmi korlátjai okán, nem is megvalósítható – célja ezek részletes ismertetése. Azonban a modern értelemben vett MI pontosabb megértése érdekében mégis indokolt egy a technológiacsaládhoz szorosan köthető alapfogalom – a gépi tanulás – kifejtése és főbb típusainak rövid bemutatása.

A gépi tanulás az MI egy részterülete, amely kifejezetten az adatokból és tapasztalatokból való tanulásra koncentrál. A gépi tanulás térhódítása az 1990-es években kezdődött, és rövidesen forradalmasította az MI képességeit azáltal, hogy lehetővé tette a rendszerek számára, hogy önállóan (iteratív módon) tanuljanak és fejlődjenek, anélkül, hogy külön programozásra lenne szükségük minden egyes feladat vagy probléma megoldásához. A gépi tanulás elterjedését megelőzően a hagyományos számítógépes alkalmazások működése mereven kötött volt a programozók által előre megírt utasításokhoz. Ezek az instrukciók (programkódok) lépésről lépésre meghatározták, hogy a gépnek pontosan mit kell tennie egy adott feladat végrehajtásához. Bármennyire is kifinomultak voltak, e rendszerek csak arra voltak képesek, amire explicit módon betanította őket a forráskód.[30] Ezzel szemben a modern MI-k olyan rugalmasabb megközelítést alkalmaznak a gépi tanulás révén, ahol nincs szükség részletes és folyamatos (újra)programozásra. Ehelyett hatalmas mennyiségű nyers adatot táplálnak be a rendszerbe a legkülönfélébb forrásokból, majd ezt követően a gépi

[29] Bár e kategória is csak elméleti síkon létezi, mégis egyre több MI kutató, társadalomtudós és filozófus fejezi ki aggodalmát e magasan fejlett technológiacsaláddal kapcsolatban. Bostrom hangsúlyozza, hogy már a fejlesztések kezdeti szakaszában is szükség lesz a szuperintelligens MI-k céljainak emberi értékekkel való összehangolására, és az alapvető biztonsági szempontok kompromisszumot nem ismerő priorizálására. Emellett a nemzetközi – országhatárokon átnyúló – együttműködést is nélkülözhetetlennek tartja a szuperintelligencia jövőbeli kockázatainak sikeres kezelésében. A témáról lásd bővebben Bostrom, Nick (2015): Szuperintelligencia. Utak, veszélyek, stratégiák. ford. Hidy Mátyás. Budapest, Ad Astra. 89-96.o.

[30] Például a szakértői rendszerek.

algoritmusok[31] képesek felismerni és elemezni az adatokban rejlő struktúrákat, mintázatokat és összefüggéseket. Ez a tanulási folyamat teszi lehetővé, hogy az MI kikövetkeztesse a megfelelő működési módot, és új, ismeretlen helyzetekben is helyesen alkalmazhassa megszerzett tudását.

A gépi tanulás talán legismertebb ága a *mélytanulás* (deep learning). Ez a módszer az emberi agy működését utánzó, többrétegű neurális hálózatokon alapul, ahol minden réteg külön-külön dolgozza fel a bemeneti adatokat, és ezeket az információkat továbbítja a következő rétegek felé.[32] A „mély'" kifejezés a sok rejtett rétegre utal a hálózatokban. A neuronok közötti súlyozott kapcsolatok révén a hálózat képes összetett mintázatokat és összefüggéseket tanulni/előrejelezni anélkül, hogy explicit, feladatspecifikus programozási szabályokra lenne szükség. [33] A mélytanulási hálózatok többféle formában léteznek, ismert típusai például az egyrétegű, az előrecsatolt, az önszervező, a konvolúciós (CNN) és a rekurrens (RNN) neurális hálózatok.[34]

[31] Az algoritmus az MI alapját képző, lépésről lépésre végrehajtandó matematikai utasítások sorozata. Ezek az utasítások határozzák meg, hogy a rendszer hogyan dolgozza fel az adatokat és hoz döntéseket. Az algoritmusok rendelkeznek bemeneti (input) és kimeneti (output) adatokkal; a bemeneti adatok felhasználásra kerülnek valamilyen művelet során, és az eredmény az algoritmus végén kimeneti adatként jelenik meg. Az algoritmusok változatos formákat ölthetnek, az alapszintű szabályrendszerektől kezdve egészen a fejlett előrejelző modellekig. Könnyen érthető példaként szolgálhatnak az online áruházak által előszeretettel alkalmazott személyre szabott ajánlórendszerek. Egy ilyen rendszer az alábbi lépésekben (utasítási sorozatokban) működik: Először elemezni kezdi a felhasználó korábbi vásárlásait és böngészési adatait. Ezen információk alapján meghatározza, milyen termékek iránt érdeklődhet leginkább a felhasználó. Ezután keres a termékadatbázisban olyan cikkeket, amelyek megfelelnek a felhasználó ízlésének és igényeinek. Végül a rendszer rangsorolja és megjeleníti a felhasználó számára leginkább releváns termékeket. Összefoglalva tehát *az algoritmus megengedett lépésekből, tevékenységekből álló véges utasításssorozat, amelyet követve elérhetjük a meghatározott, kívánt célt.* A témáról részletesebb leírást nyújt a Debreceni Egyetem Informatikai Karának oktatási tananyaga: Varga Imre (2020): Algoritmusok és a programozás alapjai. Elérhető: https://irh.inf.unideb.hu/~vargai/APA/index.html

[32] A neurális hálózatok bemeneti, rejtett és kimeneti rétegekből állnak: a bemeneti réteg feladata az adatok fogadása és továbbítása a hálózat belső rétegei felé. A rejtett rétegekben történik az adatok tényleges feldolgozása, átalakítása. Itt a beérkező információk különböző matematikai műveletek segítségével új reprezentációkká formálódnak, miközben a hálózat kiemeli a lényeges jellemzőket és mintázatokat. Végül a kimeneti réteg szolgáltatja a feldolgozás végeredményét a hálózat aktuális céljának megfelelő formában. A neurális hálózat minden rétegében mesterséges neuronok találhatók, melyek a súlyokkal és aktivációs függvényekkel vezérlik az adatok áramlását és feldolgozását. Az aktivációs függvény határozza meg, hogy egy neuron aktiválódjon-e, és ha igen, milyen mértékben továbbítsa a jeleket a következő réteg felé. A hálózat tanulása során a *súlyok és a bias* értékek folyamatosan módosulnak annak érdekében, hogy a bemeneti adatokból minél pontosabban előálljon a kívánt kimenet. Például, ha egy képelemző MI feladata, hogy rubik-kockákat ismerjen fel, a tanítás során kép pixeladatait a bemeneti réteg veszi fel, majd ezeket az adatokat a rejtett rétegek több lépcsőben dolgozzák fel, ahol a neuronok aktivációs függvények segítségével a tárgy azonosítására szolgáló jellemzőket emelik ki (forma, soronként eltérő színek stb.). A folyamat végeztével a kimeneti réteg dönti el, hogy látható-e az adott képen rubik-kocka. A tanulási folyamat során az un. backpropagation algoritmus folyamatosan finomítja a hálózat súlyait a pontosabb felismerés érdekében. Lásd részletesebben: Schmidhuber, J. (2015): Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.o.

[33] LeCun, Yann, Yoshua Bengio és Geoffrey Hinton (2015): Deep learning. *Nature* 521 (75539), 436-444.o.

[34] Gyires-Tóth Bálint (2020): A mélytanulás múltja, jelene és jövője. *Hiradástechnika.* 75/1., 23-29o.

A technológiára a 2010-es évek elejétől irányult kiemelt figyelem többek között annak köszönhetően, hogy Alex Krizhevsky, Ilya Sutskever és Geoffrey Hinton 2012-ben bemutatta mély konvolúciós neurális hálozatát (deep convolutional neural networks, CNN), amely példátlan mértékben, 28%-ról 15,3%-ra csökkentette a hibaarányt az MI-hez köthető képfelismerési feladatokban.[35] Az elkövetkező években a neurális hálózatok népszerűsége évről évre nőtt, ma már számos területet érintenek, és látványos fejlődést hoztak az MI fejlesztésekben, különösen olyan alkalmazásokban, mint a gépi látás, orvosdiagnosztika, beszédtechnológia és hangfeldolgozás, képfelismerés, természetes nyelvfeldolgozás és az önvezető rendszerek.

A mélytanulás mellett, a leginkább elterjedt gépi tanulási mintázatok közé (1) a felügyelt; (2) a felügyelet nélküli és;(3) a megerősítéses gépi tanulás tartoznak.

A *felügyelt tanulásra* (supervised learning) jellemző, hogy a modellt előre címkézett adatokkal (meghatározott input-output párokkal) képzik. Mivel a tanulóadatokhoz előzetesen hozzá van rendelve egy adott érték, a modell idővel megtanulja, hogyan hozza létre az ahhoz kötődő helyes kimenetet. Elérendő célja az általánosítási képesség elsajátítása, tehát, hogy elegendő minta feldolgozását követően a modell képes legyen olyan adat-címke párok esetében is helyes döntést hozni, amelyeket korábban nem ismert.[36] Példaként szolgálhat egy e-mail szűrő algoritmus, amely spam és nem spam üzenetek között tud különbséget tenni. Ebben az esetben az e-maileket előre megjelölik „spam" vagy „nem spam” címkékkel. Ahogy az algoritmus több és több e-mailt dolgoz fel, fokozatosan megtanulja felismerni a spam e-mailek jellemzőit, még olyan új üzenetek esetében is, amelyekkel korábban nem találkozott, így válva fokozatosan hatékonyabbá a nem kívánt e-mailek szűrésében.

A *felügyelet nélküli tanulásnál* (unsupervised learning) a modell előzetesen nem címkézett adatokból tanul, melyekhez nincs hozzárendelve előre meghatározott érték vagy megfelelő kimenet-pár. Itt a modellnek önállóan kell felfedeznie a rejtett struktúrákat, mintázatokat a bemeneti adatokban.[37] A tanulás során felfedezett összefüggések később sokféleképpen

[35] A hálózat megközelítőleg 60 millió paraméterrel és 650 000 neuronnal rendelkezik, öt konvolúciós rétegből áll, melyeket pooling (összegző) rétegek, majd három teljesen összekapcsolt réteg és egy végleges 1000-utas softmax (kimeneti) réteg következik. A technológiáról és annak képelemző képességeiről bővebben: Krizhevsky, A., Sutskever, I. és Geoffrey Hinton (2012): ImageNet classification with deep convolutional neural networks. *Communications of the ACM. 60*, 84-90.o. https://doi.org/10.1145/3065386

[36] Hastie, T., Tibshirani, R., Friedman, J. (2009): Overview of Supervised Learning. In: The Elements of Statistical Learning. Springer Series in Statistics. New York, Springer. 9-41.o.

[37] Hinton, Geoffrey – Terrence Sejnowski J. (Szerk.) (1999): Unsupervised Learning: Foundations of Neural Computation. Bradford Books, The MIT Press. 4-19.o.

hasznosíthatók, az ilyen modellek egyik fő alkalmazási területe az un. klaszterezés, ahol az algoritmus az adatpontokat hasonlóságuk alapján meghatározott csoportokba rendezi. Például egy ügyfélszegmentációs eljárásnál a modell az ügyfelek vásárlási szokásai, demográfiai adatai és egyéb jellemzői alapján csoportokat hoz létre, mely által azonosíthatóvá válnak az ügyfelek különböző szegmensei (gyakori vásárlók, kisebb megtakarítással rendelkező ügyfelek, prémium vásárlók stb.).[38]

A felügyelet nélküli tanuláshoz hasonlóan, a megerősítéses tanulásnál (reinforcement learning) sem meghatározott kimeneti értékkel bíró, előre címkézett adatokat használ a modell. Ehelyett feladata, hogy saját maga következtesse ki, hogy miként oldható meg optimálisan egy adott feladat. Ezt úgy éri el, hogy különböző műveleteket hajt végre – eközben az algoritmus (agent) próbálkozik és hibázik – majd a képzés során a konkrét döntésekhez pozitív vagy negatív visszajelzést (jutalmakat vagy büntetéseket) csatolnak. A modell célja, hogy maximalizálja a jutalmakat és minimalizálja a büntetéseket.[39] A megerősítéses tanulás elsősorban a dinamikus környezetekben való döntéshozatalra összpontosít, gyakori alkalmazási területei a robotika, a táblajátékok (AlphaGo is ilyen módon lett fejlesztve) vagy az önvezető autók. Az utóbbi esetében a modell megtanulja, hogy milyen manővereket hajtson végre bizonyos forgalmi helyzetekben, például mikor kell megállnia egy piros lámpánál vagy elkerülnie egy akadályt. Mindezt úgy teszi, hogy kipróbál különböző stratégiákat, és a következmények alapján, például egy ütközés vagy sikeres manőver, jutalmat vagy büntetést kap.

Az utóbbi években vált egyre inkább kiemelt témává a gépi tanulás kutatásaiban az un. *informált gépi tanulás* (informed machine learning).[40] Ez a megközelítés lehetővé teszi, hogy

[38] E két eljárás ötvözetét alkotja egy újabb megközelítés, az un. *önfelügyelt gépi tanulás* (self-supervised learning), mely során az algoritmusok saját magukat képezik ki a rendelkezésre álló adatokon keresztül, anélkül, hogy szükség lenne ember által előre címkézett adatokra vagy felügyeletre. Azért nevezhető az előzőleg említett két tanulási eljárás keverékének, mert egyrészt a felügyelet nélküli tanításhoz hasonlóan lehetővé teszi, hogy a modellek nagy mennyiségű címkézetlen adatból tanuljanak, másrészt kihasználja a felügyelt tanulás hatékonyságát az algoritmusok által mesterségesen generált címkék alkalmazásával. Amellett, hogy csökkenti a címkézési erőfeszítéseket és költségeket, egyúttal képes mélyebb reprezentációk és általánosító képességek fejlesztésére. Ilyen gépi tanulást alkalmaz például a Meta AI által fejlesztett data2vec algoritmus is, amit a gépi látás és a későbbiekben még részletesen tárgyalandó Nagy Nyelvi Modellek fejlesztésére használnák. Lásd bővebben: Kejriwa, Kunal (2023): Data2vec: A Milestone in Self-Supervised Learning. UniteAI. Elérhető: https://www.unite.ai/ data2vec (2023. 11. 04.)

[39] Mnih, Volodymyr, Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S.,& Hassabis, D., (2015): Human-level control through deep reinforcement learning. *Nature*, 529-533.o.

[40] Bár az informált gépi tanulás ötlete nem új, az utóbbi években vált egyre inkább népszerűvé és kezdett elterjedni a gépi tanulás kutatásaiban. Az alapjául szolgáló ötlet (modellek kiegészítése emberi ismeretekkel) már

a modellek előzetesen meglévő tudást, fizikai összefüggéseket, szabályokat és korlátozásokat építsenek be a tanulási folyamatba. Míg a korábban tárgyalt eljárások – a hagyományos, tisztán adatvezérelt modellek – próbálgatás útján tanulnak az adatokból, az informált gépi tanulás során a fejlesztők bizonyos előzetes szabályokat is megadnak a modellnek, hogy segítsék a tanulást, így azok mind szabályokra, mind adatokra támaszkodnak a tanulási folyamatban. Ez a technika több tekintetben is hasznos lehet: (1) Előfordulhat, hogy nem áll rendelkezésre elegendő adat a jól teljesítő és megfelelően általánosított modellek betanítására. Az informált gépi tanulás segíthet kiegészíteni a hiányzó adatokat; (2) A pusztán adatokon alapuló modellek figyelmen kívül hagyhatják a természeti törvényeket, szabályozásokat vagy biztonsági irányelveket, amelyek a megbízható MI megvalósításához elengedhetetlenek. Az informált tanulás beépítheti ezeket a korlátozásokat; (3) A tisztán adatvezérelt modellek visszatükrözhetik az adatok torzításait és elfogultságait. Az előzetes információk és korlátozások segíthetnek kiküszöbölni ezeket a problémákat; (4) Az informált tanulással könnyebben átláthatóvá tehetők a modellek, ami a megbízhatóság növelése mellett a szabályozói követelmények teljesítését is elősegíti. Nem utolsó sorban pedig a kiegészítő információk stabilabbá és robusztusabbá tehetik a modelleket, ami különösen a korlátozott, zavaros vagy egymásnak ellentmondó tanítóadatok esetén fontos. [41]

A következő fejezetben e technológiák konkrét alkalmazási területeinek bemutatására és az ezekhez kapcsolódó lehetséges veszélyek feltérképezésére kerül sor.

---

korábban is jelen volt bizonyos formákban, például a szakértői rendszerekben az 1980-as és 1990-es években. Az adatvezérelt megközelítések és a nagy adatkészletek elérhetősége miatti előretörés után a kutatók rájöttek, hogy a tisztán adatvezérelt modellek korlátai, például az adathalmazok torzítottsága vagy a hiányos adatok, megkövetelik az előzetes tudás integrálását a modellekbe. Ez különösen akkor vált fontossá, amikor a modelleknek olyan területeken kell helyesen működniük, ahol a pontosság, a megbízhatóság és a szabályozási megfelelőség kritikus, mint például az egészségügy, a pénzügyek és az autonóm járművek.

[41] von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C. and Schuecker, J., (2019): Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 35, 614-633.o.

3 Alkalmazások és kihívások

3.1 MI a köz szolgálatában?

Az MI térnyerése az élet számos területen érezteti hatását. Az elmúlt évtizedben egyre gyakrabban alkalmaztak gépi eljárásokat a mindennapi életet közvetlenül befolyásoló, kritikus döntések meghozatalakor: MI-alapú diagnosztikai eszközök segítenek az orvosoknak a daganatos betegségek gyorsabb felismerésében; MI-számításokra támaszkodnak a bankok a hitelkérelmek automatikus kiértékelésekor; a biztosítóintézetek az ügyfélkockázati minősítések elkészítésekor; az adóhatóságok az állampolgárok adóelkerülési gyakorlatának azonosítására; a szociális ellátórendszerek, hogy megállapítsák egy adott személy állami ellátásokra és szolgáltatásokra való jogosultságát, de még a bűnüldöző hatóságok is annak érdekében, hogy pontosabb előrejelzéseket készíthessenek a jövőbeni bűnelkövetés vagy bűnismétlés valószínűségéről. E rendszerek alkalmazási területei között megkülönböztetett figyelmet érdemelnek a hatóságok által közigazgatásban használt MI technológiák, hiszen az e környezetben hozott döntések érdemi hatással lehetnek az állampolgárok jogaira és kötelezettségeire is, például a közellátások és közszolgáltatások, a lakhatás, foglalkoztatás, oktatás vagy az igazságszolgáltatás területén.[42] A következőkben néhány már napjainkban is alkalmazott gyakorlat bemutatására kerül sor.

3.1.1 RPA és ADM - az automatizált döntéshozó rendszerek

Ahogy a döntéshozatali folyamatokban egyre nagyobb hangsúly összpontosul az adatokra, az azokat feldolgozni, kategorizálni és értékelni képes eljárások is egyre növekvő részét teszik ki

[42] Diósi Szabolcs (2023): Adatvezérelt döntések, előrejelző algoritmusok, Mesterséges Intelligencia: Technológiai innováció a köz szolgálatában? In. Barcsi, Tamás (szerk.); Csefkó, Ferenc (szerk.); Diósi, Szabolcs (szerk.): HCS 70. Ünnepi írások Horváth Csaba ny. egyetemi docens születésnapjára. Pécs, Jövő Közigazgatásáért Alapítvány, PTE ÁJK, 74-87.o.

az adminisztratív gyakorlatoknak.[43] A rendelkezésre álló erőforrások elosztásának optimalizálása, a közszolgáltatások személyre szabása, a közösségi részvétel elősegítése vagy az okosvárosokkal kapcsolatos adat-vezérelt tervezés csak néhány azon felhasználási területekből, melyekben ilyen rendszereket alkalmaznak ma is.[44]

A Robotizált Folyamat-Automatizálás (Robotic Process Automation, továbbiakban RPA) révén lehetségessé válik az ismétlődő, előre meghatározott és kiszámítható lépésekből álló adminisztratív feladatok automatizálása. Jelenleg az RPA rendszereket leginkább olyan kevésbé összetett adminisztratív feladatok automatizálása használják, mint az adatrögzítés- és változtatás, űrlapok kitöltése és frissítése, vagy automatikus emlékeztetők és értesítők kiküldése. [45]

Abban az esetben azonban, amikor az RPA technológiát más MI rendszerekkel – például természetes nyelvfeldolgozási modellekkel – ötvözik, lényegesen fejlettebb Automatizált Döntéshozó Rendszerek (Automated Decision-Making Systems, ADM) létrehozása válik lehetségessé. Az ilyen rendszereket többek között a közigazgatási hivatalokba nagyszámban beérkező engedélykérelmek kezelésére, valamint különböző pénzügyi/számviteli feladatok (pl.: számlák elkészítése, a kifizetések nyomon követése, pénzügyi jelentések készítése vagy bérszámfejtés) elvégzésére lehet használni. Előnyei közé tartozik, hogy képes autonóm módon frissíteni nagy és dinamikusan változó adathalmazokat (pl.: az állampolgárok személyi adatainak és lakcímének nyilvántartását), valamint, hogy különböző anomáliákat tud azonosítani strukturált információs közegekben. Ennek köszönhetően az ilyen technológiák gyakran kerülnek alkalmazásra a szociális ellátások rendszerében, ahol a különböző állami segítségnyújtási ellátásokra és szolgáltatásokra való jogosultság értékelésére, valamint az

[43] A kormányok világszerte igyekeznek kihasználni az MI és Big Data technológiák előnyeit azáltal, hogy saját adatbázisaikat összekötve, rendszerezett formában gyűjtik, tárolják és dolgozzák fel a rendelkezésükre álló adatokat. Az általuk birtokolt adatok segítségével pontosabb képet kaphatnak az állampolgárok életéről, a kormányzati szervek működéséről és a központi, illetve regionális közigazgatási viszonyokról is. Általánosságban kijelenthető, hogy e technológiák elterjedése és az adatbázisok összekötése lehetőséget teremt a közigazgatás széles körű modernizálására, olcsóbban, eredményesebben és átláthatóbban működő közszolgáltatások biztosítására, ezeken keresztül pedig az állampolgárok életkörülményeinek általános javítására is. Lásd bővebben: Maciejewski, Mariusz (2016): To do more, better, faster and more cheaply: using big data in public administration. *International Review of Administrative Sciences*, 83, 120-135.o.

[44] Misuraca, Gianluca., és van Noordt, C. (2020): Overview of the use and impact of AI in public services in the EU. EUR 30255 EN, Publications Office of the European Union, Luxembourg. 39-52.o.

[45] Houy, C., Hamberg, M. & Fettke, P., (2019): Robotic Process Automation in Public Administrations. In: Räckers, M., Halsbenning, S., Rätz, D., Richter, D. & Schweighofer, E. (Hrsg.), Digitalisierung von Staat und Verwaltung. Bonn: Gesellschaft für Informatik. 62-74.o.

ilyen ellátások és szolgáltatások nyújtásával, csökkentésével, visszavonásával vagy visszaigénylésével kapcsolatos döntések meghozatalában játszanak szerepet.[46]

A svédországi Trelleborg városa például már évek óta használ RPA rendszert annak érdekében, hogy gyorsabbá és költséghatékonyabbá tegye a szociális segélyezés ügymenetét. A nagyszámban beérkező kérelmek elbírálásával kapcsolatos nehézségek miatt – ilyen volt a késedelmes kifizetés – még 2016-ban döntött úgy Trelleborg önkormányzata, hogy digitalizálja és automatizálja a szociális ellátások adminisztrációs rendszerét. A trelleborgi modellt elsősorban arra tervezték, hogy a különböző szociális segélykérelmek (pl. otthoni ápolás, a betegségi ellátás, a munkanélküli segély) feldolgozásában és elbírálásában segítsen a hivataloknak. A modernizáció sikerét mutatja, hogy a modell bevezetését követően a szociális támogatásra irányuló digitális kérelmek 85%-át ma már RPA rendszerek kezelik.[47]

3.2 Prediktív analitika az adatvezérelt döntéshozatal szolgálatában

A prediktív technológiát alkalmazó MI rendszerek kiterjedt adathalmazok elemzésével különböző mintákat és összefüggéseket képesek feltárni, melyekre alapozva később megbízható előrejelzéseket készítenek a jövőt illetően. A döntéstámogatásban betöltött szerepük jelentős, hiszen segítségükkel jobb minőségű szakpolitikai kormányprogramok és stratégiák megalkotása válik lehetségessé. A prediktív analitika technológiák döntéstámogató alkalmazására jó példaként szolgálhat a 2019-ben TalTecCity néven indított észt kezdeményezés, mely intelligens szenzorok segítségével gyűjt információkat egyebek mellett az észt főváros (Tallinn) területén belül közlekedő emberek és járművek mozgásáról, az utak és parkolók forgalmáról, a levegőben mérhető külső légszennyezésről, és a köztéren elhelyezett hulladéktárolók telítettségéről. Az így nyert információkra alapozva később modelleket hoznak létre, amelyek nagyban segítenek az önkormányzati döntéshozóknak

[46] Ranerup, Agneta & Henriksen , Z. H (2019): Value positions viewed through the lens of automated decision-making: The case of social services. *Government Information Quarterly,* Volume 36(4) 2-3.o.
[47] Ranerup, Agneta, Henriksen, Z. H. (2022): Digital Discretion: Unpacking Human and Technological Agency in Automated Decision Making in Sweden's Social Services. Social Science Computer Review. 40(2), 445-461.o.

abban, hogy adatvezérelt szakpolitikai stratégiákat dolgozzanak ki a várostervezéssel, a közlekedéssel és a környezetgazdálkodással kapcsolatos kérdésekben.[48]

3.2.1 Adózással és szociális juttatásokkal kapcsolatos csalások

Az elmúlt években egyre gyakrabban fordul elő, hogy a nemzeti adóhatóságok is MI alapú előrejelző rendszereket alkalmaznak a pénzügyi visszaélések hatékonyabb feltérképezése céljából. Az ilyen rendszerek nem csupán a folyamatban lévő illegális tevékenységeket képesek észlelni, hanem adat- és kockázatelemző algoritmusok segítségével a jövőbeni csalások valószínűségét is előre jelzik.

A magyar Nemzeti Adó- és Vámhivatal is hasonló MI alapú előrejelzéseket használ az adócsalásra és adóelkerülésre irányuló gyakorlatok felderítésére. A *Rugalmas Adóellenőrzési Döntéstámogató és Adatbányászati Rendszer* (RADAR) a korábban vizsgált ügyek eredményei alapján olyan jellemzőket és ismérveket keres, amelyek nagy valószínűséggel vezettek magas adóhiányhoz a múltban. Az eredmények kiértékelésével megalapozott előrejelzéseket és következtetéseket tud készíteni a lehetséges jövőbeli esetekre vonatkozóan.[49]

A prediktív analitika alkalmazására további példaként szolgálhat a Hollandiában 2020-ig használt SyrRy (Systeem Risico Indicatie) kockázatértékelő rendszer, melyet a potenciális jövőbeli társadalombiztosítási, szociális vagy egyéb jóléti juttatással kapcsolatos csalások felderítésére használtak. Az eszköz a meglévő kormányzati adatállományokból (pl.: adóbevallások, társadalombiztosítási adatok, egészségügyi kórtörténet) származó kockázati mutatókat elemezte annak érdekében, hogy azonosítsa azokat a személyeket, akiknél fokozottan fennáll a jóléti juttatásokkal való csalás vagy visszaélés kockázata. Érdemes megjegyezni azonban e rendszerrel kapcsolatban, hogy bevezetése után nem sokkal több jogvédő szervezet is kritikával illette annak működését. A kritikák tárgyát képezte, hogy a SyrRy rendszer több alkalommal is megsértette a magánélethez fűződő alapvető jogokat, és

[48] FRA Report (2020): European Union Agency for Fundamental Rights (2020): Artificial Intelligence, Big Data and Fundamental Rights Country - Research Estonia 2020.
[49]Fejes Erzsébet - Futó Iván (2021): Mesterséges intelligencia a közigazgatásban – az érdemi ügyintézés támogatása. *Pénzügyi Szemle*, Különszám 2021/1, 44.o.

hátrányosan diszkriminálta a gazdaságilag sebezhető állampolgárokat. További aggodalomként vetették fel a rendszer belső működésének átláthatatlan voltát, valamint azt, hogy a döntésben érintett személyeknek nem volt lehetőségük megismerni a döntéseket megalapozó adatokat. A szoftver alkalmazását – a holland bíróság döntése nyomán – végül 2020-ban hivatalosan is leállították.[50]

3.2.2 Rendvédelmi szervek és büntető igazságszolgáltatás

A prediktív technológiák talán leginkább megosztó felhasználási területe az igazságszolgáltatási rendszerben történő alkalmazásukhoz kötődik. Az e területeken jellemzően profilozáson, pontozáson, kockázatelemzésen és magatartáselőrejelzésen alapuló algoritmusokra rengeteg figyelem összpontosult az elmúlt években, nem egy esetben heves kritikák kereszttüzébe állítván őket. Szerepük és befolyásuk nem elhanyagolható, hiszen az általuk készített értékelések befolyásolhatják a valós életben alkalmazott bűnmegelőzési stratégiákat, többek között azt, hogy hol szükséges fokozni a megfigyelést vagy mely személyek igazoltatása válhat indokolttá. Emellett a büntetőügyekben hozott bírósági döntésekre is hatással lehetnek, például az őrizetbe vétellel, az óvadék megállapításával, az előzetes letartóztatással, a próbára bocsátással vagy a büntetés-végrehajtás felfüggesztésével kapcsolatosan.[51]

A bűnüldöző hatóságok esetében az előrejelző technológiák egyik legelterjedtebb felhasználási módja az úgynevezett *prediktív rendfenntartáshoz* (Predictive Policing) kötődik. E gyakorlat a rendelkezésre álló bűnügyi adatok feldolgozásával statisztikai módszereket alkalmaz annak előrejelzésére, hogy hol és mely időszakokban magasabb a jövőbeni bűncselekmények elkövetésének valószínűsége. Ilyen előrejelző eszközöket használnak a

[50] Misuraca, Gianluca., - van Noordt, C. (2020): Overview of the use and impact of AI in public services in the EU. EUR 30255 EN, Publications Office of the European Union, Luxembourg 45-46.o.
[51] Herke Csongor (2021): A mesterséges intelligencia kriminalisztikai aspektusai. *Belügyi Szemle*, 69/10. 1714-1720.o.

bűnüldöző szervek például Hollandiában (ProKid), Svájcban (Precops), az Egyesült Királyságban (NDAS), Németországban (RADAR-iTE) és Olaszországban (Delia) is.[52]

A bűnüldöző hatóságok által használt előrejelzések másik gyakori felhasználási módjára példaként szolgálhat az Egyesül Királyságban évek óta használt HART kockázatelemző-rendszer (Harm Assessment Risk Tool). A gépi tanuláson alapuló szoftvert a durhami rendőrség használja a jövőbeni bűnismétlés valószínűségének előrejelzésére. A HART rendszer – 34 kategóriát magába foglaló adatbázisból – profilt alkot az eljárás alá vont személyről, majd annak kiértékelésével egy kockázati pontszámokat hoz létre. A kapott pontszám alapján a vizsgált személyt a három előre meghatározott kockázati csoport (magas, közepes, alacsony) valamelyikébe sorolja. Ez az osztályozás segíti az igazságügyi hatóságok döntését a tekintetben, hogy vád alá helyezzék-e a személyt, vagy a Checkpoint névre hallgató, bíróságon kívüli rehabilitációs programba utalják. Amennyiben az elkövető az alacsony kockázati csoportba kerül, a programba utalják, ezzel lehetőséget biztosítva számára, hogy elkerülje a bírósági eljárás lefolytatását. Ellenkező esetben – tehát ha közepes, vagy magas kockázati csoportba sorolták – megindul a vádemelés. [53]

Az előzőekben ismertetett példák csak egy részét képezik a közigazgatási MI gyakorlatok egyre növekvő körének. Ahogy arra e fejezet elején is történt már utalás, az ehhez hasonló MI vezérelt döntések hatással lehetnek számos, az emberek életét közvetlenül is befolyásoló kérdésre és körülményre (például, hogy milyen típusú és minőségű közszolgáltatásokhoz férhetnek hozzá, vagy milyen bánásmódban részesülnek igazságszolgáltatási eljárások során). Ennek okán fontos feladat átlátni és megérteni, hogy milyen potenciális hibái, kockázatai és társadalmi veszélyei lehetnek az ilyen döntés-támogató technológiáknak. A következő fejezet erre tesz kísérletet.

[52] Fair Trials (2021): Automating Injustice - The Use of Artificial Intelligence & Automated Decision- Making Systems in Criinal Jutice. 8-21.o.; Fantoly Zsanett, Herke Csongor, Szabó Barbara (2023): The role of AI-based systems in negotiated proceedings. *e-Revue Internationale de Droit Pénal*. 2023, A-18, 4.o.
[53] Oswald et al (2018): Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, 27(2), 223-250.o.

3.3 Az MI alkalmazásból eredő kihívások, kockázatok

Az elmúlt években az MI-rendszerek alkalmazásából eredő társadalmi, etikai és jogi kockázatokkal kapcsolatban a figyelem elsősorban három kulcsfontosságú kihívásra irányult: (1) az MI-re jellemző gépi tanulásból és autonóm működésből eredő átláthatóság, elszámoltathatóság és kiszámíthatóság hiánya; (2) az MI által vezérelt téves, elfogult vagy diszkriminatív döntéshozatal veszélye; (3) az MI-rendszerek alapvető emberi szabadságjogokra és a demokratikus intézmények működésére gyakorolt negatív hatása.[54]

3.3.1 A Black Box probléma – átláthatóság, értelmezhetőség, indoklási kötelezettség

Az MI alapú döntéshozatalra jellemző átláthatatlanság – Jenna Burell felosztása alapján – három alkategóriára bontható. Szándékos (intentional) az átláthatatlanság, ha az algoritmikus döntéshozatal mögött álló folyamatokat tudatosan tartják rejtve (ilyen lehet például a Google keresőmotorjának programkódja, mely üzleti titkoknak minősül). Az írástudatlan (illiterate) átláthatatlanság azt jelenti, hogy ugyan senki nem akadályozza egy adott algoritmus belső működésének megismerését, mégis annak rendkívül összetettsége okán azt csak nagyon kevesen – elsősorban a programnyelvet ismerő, szakismerettel bíró személyek – tudják értelmezni. A lényegbeli (intrinsic) átláthatatlanság pedig azt jelenti, hogy egyes rendszerek olyan dinamikusan – sok esetben autonóm módon – változnak, hogy még a magasan képzett

[54] Az MI társadalmi, etikai és jogi kihívásaival kapcsolatban gazdag magyar szakirodalom áll rendelkezésre. Itt csak az utóbbi években megjelent könyvekre utalva: Ződi Zsolt (2018): Platformok, robotok és a jog. Új szabályozási kihívások az információs társadalomban. Budapest, Gondolat Kiadó; Héder Mihály (2020): Mesterséges Intelligencia. Filozófiai kérdések, gyakorlati válaszok. Budapest, Gondolat Kör Kiadó; Tilesch György – Hatamleh, Omar (2021): Mesterség és Intelligencia. Vegyük a kezünkbe a sorsunkat az MI korában. Budapest, Libri Kiadó; Csepeli György 2020: Ember 2.0 – A mesterséges intelligencia gazdasági és társadalmi hatásai. Budapest, Kossuth Kiadó; Török Bernát - Ződi Zsolt (szerk.) (2021): A mesterséges intelligencia szabályozási kihívásai. Tanulmányok a mesterséges intelligencia és a jog határterületeiről. Budapest, Ludovika Egyetemi Kiadó.

programozóknak – vagy a programkód megalkotóinak – is kihívást jelent teljes egészében megérteni alapvető működésüket. [55]

Lényegében ezt az utolsó esetet írja körbe a gépi döntéshozatallal kapcsolatosan gyakran említett Black Box probléma is (fekete-doboz probléma), mely a fejlett, gépi tanuláson alapuló MI modellekhez köthető legsürgetőbb kihívások egyike. Ezek az összetett algoritmusok által vezérelt, önállóan tanulni képes MI-k olyan módon juthatnak egy adott következtetésre, amelyet az azt alkalmazók nem minden esetben láthatnak előre.[56] A végeredményt ugyan képesek értelmezni, annak módját azonban, hogy milyen összefüggéseken keresztül jutott el a program a válaszig, már nem látják át teljesen (máshogy kifejezve: értik a mit, de nem értik meg a miértet). Bár nem használja a Black Box kifejezést, erre a jelenségre utal 2017-ben megjelent *Pszichopolitika* című könyvében a német-koreai filozófus Byung-Chul Han is, amikor az *adattotalitarizmus* és adatalapú gépi döntések elterjedésének következményeiről azt írja, hogy elkerülhetetlenül *a tudás új korszakát* hozzák majd el. Egy olyan korszakot, ahol *a korrelációk helyettesítik az okozatiságot. Az így van-ez helyettesíti a hogyant*.[57]

Az átláthatatlan működésből következik az MI-alapú gépi döntések értelmezhetőségének és magyarázhatóságának (interpretability / explainability) kettős problémája is.[58] Bár e két fogalmat gyakran szinonimaként használják, mivel hasonló célokat szolgálnak – az MI alapú döntések és előrejelzések megindoklását –, alapvető különbség van a jelentésük között.

---

[55] Burrell, Jenna (2016): How the Machine 'Thinks: Understanding Opacity in Machine Learning Algorithms. *Big Data and Society*. 3/1. 3-6.o.

[56] Pasquale, Frank. (2015): The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

[57] Byung megfogalmazásában: *Azzal párhuzamosan azonban, hogy a társadalmak korábban elképzelhetetlen mennyiségű adatot halmoznak fel az őket körülvevő világgal kapcsolatban, az információfeldolgozás és -megértés hagyományos módozatai is lényeges változáson mennek keresztül. Korunkban, amit az adattotalitarizmus és adatfetisizmus lelkesít át, a Big Data alapú technológiák jelentik az iránytűt a biztos tudás felé. Minden mérhető és számszerűsíthető. A dolgok elveszítik mostanáig rejtve maradt, titkos összefüggéseiket*. Lásd: Han, Byung-Chul (2020): Pszichopolitika. (Ford. Csordás Gábor) Budapest, Typotex. 91.o.

[58] Ezzel kapcsolatban kiemelendő, hogy az értelmezhetőség – magyarázhatóság kettő problémája arra is rámutat, hogy nem egyértelmű, hogy mit, kinek és hogyan kell magyarázni. Először is, mit kell magyarázni? Az MI-rendszerek működési mechanizmusait, az általuk hozott döntéseket és előrejelzéseket? Esetleg a tervezési folyamatot és az alkalmazási célokat? Mindez egyaránt fontos lehet, de eltérő magyarázati megközelítést igényel. Másodszor, kinek kell magyarázni? Az alkalmazónak, a felhasználóknak, fejlesztőknek, az értintett személyeknek, a szabályozó hatóságoknak? Minden érintett csoport más-más szintű és típusú magyarázatra lehet kíváncsi, más-más háttértudással rendelkezik. Harmadszor, ki adja meg a magyarázatokat? Maga az MI-rendszer? A fejlesztők? Egy független szakértő? Attól függően, hogy ki magyaráz, a megközelítés is eltérő lehet. Világos, hogy minden magyarázat egyfajta „fordítást" jelent a célközönség számára. Vagyis a magyarázatokat mindig a konkrét felhasználói helyzethez és igényekhez kell igazítani, hogy érthetőek és hasznosak legyenek (egy szakértőnek szóló technikai magyarázat például más lesz, mint egy laikus számára készült leírás. Lásd a (meg)magyarázható MIről (explainable AI, XAI) bővebben O'Hara, K. (2020): Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review,* Vol.39. 3-6.o.

Az értelmezhetőség a bemenetek és kimenetek közötti ok-okozati összefüggések emberi megértésének mértékét jelenti, ami magában foglalja annak átlátását, hogy a modell milyen döntési szabályokat alkalmaz és hogy mely jellemzők milyen fontossággal bírnak a kimenetek elérésében. Egyúttal jelenti azt is, hogy a rendszer fejlesztői vagy alkalmazói képesek-e előre jelezni, mi történik, ha változtatásokat hajtanak végre a bemeneti adatokon vagy az algoritmus paraméterein. A magyarázhatóság ezzel szemben az MI-rendszerek döntéseinek és előrejelzéseinek érthető magyarázataira összpontosít, ideértve a modell kimenetének okait és indoklását. Míg az értelmezhetőség a modell belső mechanizmusaira vonatkozik, addig a magyarázhatóság túlmutat a modell belső működésén, és arra irányul, hogy a rendszerek által hozott döntéseket és előrejelzéseket érthető módon közvetítse a felhasználóknak és érintetteknek.[59] Tehát az értelmezhetőség a modell „*mikéntjére*" összpontosít, a magyarázhatóság pedig a „*miértjére*" ad választ.

Mind az értelmezhetőség, mind pedig a magyarázhatóság hiánya különösen aggasztó lehet azon helyzetekben, amikor az algoritmusok olyan kritikus döntéseket hoznak, melyek lényeges társadalmi vagy jogi következményekkel járhatnak az egyénre nézve (pl.: a korábban ismertetett közigazgatási hatóságok, bűnüldöző és rendvédelmi szervek által alkalmazott rendszerek esetében). Ilyenkor a tisztességes eljáráshoz való jog részeként merül fel a hatóság oldalán jelentkező indokolási kötelezettség, melynek elmulasztása csökkentheti a döntések elfogadhatóságát és az állampolgárok önkéntes jogkövetési hajlandóságát.[60] Jogosan merül fel tehát az az igény, hogy az MI-rendszerek széles körben elfogadottá és megbízhatóvá válásához elengedhetetlen az átláthatóság – és az ahhoz szorosan köthető – elszámoltathatóság biztosítása. Ez magában foglalja az algoritmusok és döntéshozatali folyamatok érthetővé és magyarázhatóvá tételét (indokolási kötelezettség), valamint annak lehetőségét, hogy az érintettek megkérdőjelezhessék vagy felülvizsgálhassák a rendszer által hozott döntéseket (ha szükséges emberi felügyelet kikövetelésével).[61]

[59] Rudin, C. (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*, (1), 206–215.o; Molnar, Christoph (2020): Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub. Molnar a könyvet eredeti nyelven githubon szabadon elérhetővé tette: https://christophm.github.io/interpretable-ml-book/

[60] Hohmann Balázs (2021): A mesterséges intelligencia közigazgatási hatósági eljárásban való alkalmazhatósága a tisztességes eljáráshoz való jog tükrében. In Török Bernát és Ződi Zsolt (szerk): A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 413.o.

[61] Elméletben az automatizált döntéshozatal során, az emberi felügyelet biztosítását garantálja a GDPR 22. cikk (1) bekezdése, mely kimondja, hogy az adatalany jogosult arra, hogy ne terjedjen ki rá az olyan, kizárólag automatizált adatkezelésen alapuló döntés hatálya, amely rá nézve joghatással járna vagy őt jelentős mértékben érintené. A gyakorlatban azonban a gépi döntések végrehajtásakor az algoritmus alkalmazása mellett jellemzően megmarad az emberi tényező is, vagyis a döntés nem kizárólag automatizált módon történik, így gyakran nem

Amennyiben ez nem történne meg és az adattotalitarizmusból eredő új értelmezési keretek kerülnének előtérbe, ahol a „miért" kérdése helyett a „mit" válik iránytűvé, és ahol az összefüggések helyett csupán száraz adatokra támaszkodunk, a jövő társadalmának új, előre nehezen megjósolható kihívásokkal kell szembenéznie. A Big Data technológiákon alapuló előrejelző MI-k kapcsán Ződi Zsolt *Platformok, robotok és a jog* című könyvében szellemesen illusztrálja az átláthatatlanságból eredő potenciális jogi és morális dilemmákat:

„Képzeljük el azt a helyzetet, hogy mondjuk bebizonyosodik, hogy a 47-es lábú fekete hajú, 195 centiméter magas, 19 éves fiatal férfiak nagyon nagy eséllyel követnek el erőszakos bűncselekményeket. (Szándékosan írtam ilyen abszurd paramétereket.) Az persze elképzelhetetlen, hogy preventív céllal ezeknek az embereknek a személyes szabadságát korlátozzuk, de egy idő után nagy lesz a csábítás, hogy legalább valamilyen módon például megfigyeljük őket. De még a megfigyelés sem lesz a hagyományos igazolási technikákkal igazolható. Hogyan magyarázható, hogy az intézkedés tartalmát semmilyen szempontból nem érintő paraméter az intézkedés indoka? Ez a gondolkodásmód teljes szakítás lenne a felvilágosodás eszményeivel; például azzal a megfontolással, hogy csak olyanért vagyok felelőssé tehető, amelyen magam is képes vagyok változtatni."[62]

3.3.2 Egyenlő bánásmód követelményének megsértése

Az MI-alapú gépi döntéshozatal kapcsán gyakran elhangzó érv, hogy az ilyen eljárások képesek kizárni az emberi elfogultságot és előítéleteket a döntéshozatali folyamatok során. A valóságban azonban ez csak annyira érhető el, amennyire az e rendszereket működtető programkódok, illetve az e rendszereket tanító adatkészletek is pártatlanok. Ha az MI elfogult, torzított adathalmazokon tanul, akkor hiányos, helytelen vagy elfogult képet kap a valóságról.

Az elemző és előrejelző technológiák egyik legfőbb haszna abban rejlik, hogy a hatalmas adathalmazok begyűjtésével, feldolgozásával és kiértékelésével korábban nem realizált minták és összefüggések feltárása válik lehetségessé általuk. Azonban pontosan ennek okán

terjed ki a jogszabály védelme a döntés alanyára. (Például a banki hitelbírálat esetén az algoritmus dönt az ügyfél hitelképességéről, de a döntés *végső* hitelesítését egy banki alkalmazott – természetes személy – végzi el).

[62] Ződi Zsolt (2018): Platformok, robotok és a jog. Új szabályozási kihívások az információs társadalomban. Budapest, Gondolat Kiadó. 234-235.o.

fordulhat elő az, hogy bizonyos minták és összefüggések algoritmusok által történő felismerése, majd ezek szabályszerűnek értékelése valamely, a társadalomban már korábban jelenlévő előítélet vagy elfogultság akaratlan legitimálásához vezethet.[63]

Az automatikus döntéshozatal kapcsán az elfogultság három fő típusát lehet elkülöníteni; (1) szándékos, (2) nem szándékos és (3) a tanult elfogultág. Szándékos az elfogultság, ha a rendszer fejlesztője, vagy annak alkalmazója tudatosan épít be előítéleteket a döntéshozó algoritmusba (például gazdasági érdekekből). Ide tartoznak az árdiszkriminációs gyakorlatok és a dinamikus árazás (differenciált árképzés) bizonyos esetei. E marketing stratégiák során az online platformok potenciális vásárlói eltérő árat látnak ugyanazon termékért vagy szolgáltatásért, attól függően, hogy milyen személyes jellemzőkkel rendelkeznek, vagy mikor, milyen időközönként, illetve hányadik alkalommal néztek meg egy adott terméket. A sütik[64] segítségével a platform felismeri a vásárló digitális profilját, böngészési előzményeit, majd az erre létrehozott algoritmusok a profilra jellemző bizonyos sajátosságok alapján (impulzusvásárlásra való késztetések gyakorisága vagy pedálul az IP cím alapján tartózkodási hely, lakhely, böngésző típusa, melyekből még az is kikövetkeztethető, hogy laptopról, vagy macbookról csatlakozik a felhasználó) az adott személyhez kalkulált személyre szabott árat mutatnak.[65] Ezeknél gyakoribb – és nehezebben szabályozható – eset a nem szándékos elfogultság, ami leggyakrabban a tanító adatkészletben már előzetesen is fellelhető előítéletek, torzítások, diszkriminatív minták miatt alakul ki. A harmadik kategória esetén pedig az algoritmusok a tanulási folyamat során *önmaguktól* alakítanak ki elfogultságot, ami az adatokban rejlő összefüggések felismerésében, elsajátításában, majd ezen felismerésekből levont következtetések döntési tényezőként történő elismerésén alapul.[66]

---

[63] Hassani, K. Bertrand (2020): Societal bias reinforcement through machine learning. *AI Ethics*, Volume 1, 239-247.o. https://doi.org/10.1007/s43681-020-00026-z

[64] A süti (cookie) egy olyan webszerver által létrehozott adatcsomag, amit a felhasználó számítógépén tárolnak meghatározott ideig annak érdekében, hogy nyomon követhessék annak online tevékenységét, például azt, hogy honnan, milyen gyakran és milyen böngészővel látogat meg egy weboldalt. Az első webes sütiket 1994-ben Lou Montulli, a Netscape Communications programozója alkotta meg, majd a 90-es évek második felében történő elterjedésüket követően hamar arra kezdték használni őket, hogy rögzítsék a mely oldalakat látogattak, milyen termékeket néztek meg vagy milyen hirdetésekre kattintottak a felhasználók. Később ahogy a profilozó algoritmusok fejlődtek a sütik használata is egyre összetettebbé vált. A belőlük gyűjtött adatokból már új felhasználói interakciókra és online viselkedésmintákra is igyekeztek következtetéseket levonni. Lásd: Mayer, J., & Mitchell, J. C. (2012). Third-Party Web Tracking: Policy and Technology. *IEEE Symposium on Security and Privacy*

[65] E gyakorlatokról lásd bővebben: Bara Zoltán (2017): Dinamikus árazás az online kereskedelemben – Hogyan lehet hátrányos a fogyasztónak a dinamikus árazás? *Versenytükör,* XIII. évfolyam (2), 4-19.o.

[66] Az algoritmikus elfogultságról, diszkriminációról lást bővebben: Barocas, Solon – Selbst, Andrew D. (2016): Big Data's Disparate Impact. *California Law Review,* 104, 671-732.o. Azzal kapcsolatban, hogy az ilyen minták felismerése, majd szabályként elfogadása, hogy képes megerősíteni és reprodukálni a már létező faji, nemi,

3.3.3 Elfogult és diszkriminatív döntéshozatal a gyakorlatban

A döntéshozatalban megjelenő nem szándékos, rejtett torzításra[67] példaként szolgálhat az Amazon 2014-ben fejlesztett automatizált toborzó rendszere. Az algoritmus célja az volt, hogy előszűrje a meghirdetett mérnöki pozíciókra (pl. szoftverfejlesztői állás) a legalkalmasabb jelöltek listáját a nagy létszámú jelentkezői önéletrajzok közül. Hamar kiderült azonban, hogy a szelekció eljárása során az algoritmus elfogultságot mutatott a női jelentkezőkkel szemben. E probléma abból eredt, hogy az Amazon számítógépes modelljeit úgy képezték, hogy a jelentkezőket a céghez elmúlt tíz évben benyújtott önéletrajzok mintázatai alapján értékeljék. A rendszer azért mutatott elfogultságot a nőkkel szemben, mert a modell tanításához használt adatok a korábbi jelentkezők férfi túlsúlyát tükrözték. Ennek végeredményeként az algoritmus egyszerűen leminősítette azokat az önéletrajzokat, amelyek a „women's" (női) szót tartalmazták, ahogy az történt például a „women's rugby team" (női rögbicsapat) esetében.[68] Az Amazon később felhagyott az algoritmus használatával, elismerve, hogy nem tudták eredményesen kiküszöbölni a szoftver által visszatérő jelleggel tanúsított ilyen típusú elfogultságot.

Egy másik, gyakran idézett eset a diszkriminatív döntéshozatallal kapcsolatban az amerikai büntető igazságszolgáltatásban használt COMPAS rendszerhez kötődik.[69] Az előrejelző algoritmus 137 gondosan összeválogatott kérdésre adott válasz, valamint az elkövetők bűnügyi nyilvántartásában tárolt adataiból származó információk kiértékelése alapján készített előrejelzést a vádlottak ismételt bűnelkövetésének valószínűségéről.[70] Az algoritmus alapú

---

gazdasági egyenlőtlenségeket lásd bővebben: Eubanks, Virginia (2018): Automating Inequality:How High-Tech Tools Profile, Police, and Punish the Poor. New York, St Martin's Publishing. 14-39.o.

[67] Az rejtett torzításnál (indirekt diszkriminációnál) az algoritmusok nem közvetlenül használják fel a védett tulajdonságokat (pl.: nem, származás, vallás) hanem úgynevezett proxik (helyettesítő változók) segítségével következtetnek ezekre a jellemzőkre.

[68] Jeffrey Dastin (2018): Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *REUTERS*. Elérhető: https://www.reuters.com/article/idUSKCN1MK0AG (2021. 04. 11.)

[69] A COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) rendszert a Northpointe fejlesztette ki és elsőként Wisconsin államban 2012-ben kezdték el használni a büntetőeljárások során. A programról bővebben: Park A. Lee (2019): Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing. *UCLA Law Review*. Elérhető: https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/ (2021. 07. 22.)

[70] Az algoritmus különböző szempontok alapján elemzi az elkövetőt, figyelembe véve a korábbi bűncselekményeit, családi háttérét, társadalmi-gazdasági helyzetét, az életkörülményeinek stabilitását. Ezen információk alapján 1-től 10-ig terjedő skálán értékeli a visszaesés kockázatát, majd sorolja be a személyt

kockázatértékelés elsősorban abban segített a bírónak, hogy az adott ügyben döntsön az óvadék összegéről, illetve az előzetes letartóztatás vagy a feltételes szabadlábra helyezés megalapozottságáról.[71] Egy ProPublica által készített elemzés arra a megállapításra jutott, hogy a COMPAS-rendszer értékelési eljárása során több esetben is elfogultságot mutatott az egyesült államokbeli afroamerikai lakossággal szemben. Az adatok utólagos vizsgálatából tisztán kirajzolódott, hogy a program a nem visszaeső afroamerikai vádlottakhoz átlagosan közel kétszer magasabb kockázati százalékot rendelt (45%), mint a nem visszaeső fehér bőrű társaikhoz (23%). Emellett az elemzés arra is fényt derített, hogy a rendszer értékelése során azok a fehérbőrű vádlottak, akik újból bűncselekményt követtek el, közel kétszer olyan gyakran lettek alacsony kockázatúnak minősítve, mint az afroamerikai visszaesők (48% szemben a 28%-kal).[72]

Az egyenlő bánásmód megsértésének egy sajátos esete a bankok által előszeretettel használt hitelbírálati algoritmusokhoz kötődik.[73] Egy 2021-ben publikált tanulmány arra a következtetésre jutott, hogy az ilyen kockázatértékelő rendszerek gyakran torzítanak a vékony hiteltörténettel (thin credit files) rendelkező ügyfeleik hitelképességének megállapításakor. A

---

magas, közepes vagy alacsony kockázatú kategóriába. Lásd: Gombos Katalin, Gyuranecz Franciska Zsófia, Krausz Bernadett, Papp Dorottya (2021): A mesterséges intelligencia jogalkalmazási területen való hasznosíthatóságának alapjogi kérdései. In Török Bernát és Ződi Zsolt (szerk): A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 331-332.o.

[71] Az ehhez hasonló gyakorlatok fokozatos elterjedése kapcsán joggal merülhet(nek) fel olyan kérdés(ek), hogy milyen alapvető szerkezeti és működésbeli átalakulások előtt áll az igazságszolgáltatás? Miként alakul át a bírói döntéshozatal természete a gépi tanulás és prediktív elemzések korában? Milyen alkotmányos és jogállami követelményeket kell érvényesíteni az algoritmikus döntéstámogatási rendszerek bevezetése során? Miként biztosítható az igazságszolgáltatás alapelveinek (független és pártatlan bírósághoz való jog, tisztességes eljáráshoz való jog) érvényesülése stb.? E témakört különböző nézőpontokból vizsgálva, lásd: Herke Csongor (2023): Mesterséges intelligencia a büntetőjogi döntéshozatalban. *Jogtudományi Közlöny*. 78(4), 165–176.o.; Pödör Lea (2021): Leibniz and His Possible Effect on AI – Some Thoughts on the Legal System and Judicial Decision-Making in the Age of AI. In *12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021),* 853–858.o.; Chronowski Nóra, Kálmán, Kinga,Szentgáli-Tóth, Boldizsár (2022): Régi keretek, új kihívások: a mesterséges intelligencia prudens bevonása a bírósági munkába és ennek hatása a tisztességes eljáráshoz való jogra. *Glossa Iuridica*, 8(4), 7-38.o.

[72] Az elemzés során több mint 10 000 bűnügyi vádlott esetét vizsgálták meg a floridai Broward megyéből, majd összevetették az algoritmus előrejelzéseit a ténylegesen bekövetkező eseményekkel. A tanulmány következtetései szerint megállapítható, hogy a COMPASS rendszer – az elfogultságra hajlamos döntéshozataltól eltekintve – az esetek többségében az afroamerikai és a fehér vádlottak esetében is körülbelül ugyanolyan arányban helyesen jósolta meg a visszaesést (59% a fehér vádlottaknál, 63% az afroamerikai vádlottaknál). Szignifikáns eltérés azoknál a törzs szövegben is említett eseteknél volt megállapítható, ahol az előrejelzés téves következtetésre jutott. A ProPublica tanulmányának megállapításairól részletesebben: ProPublica (2016): How We Analyzed the COMPAS Recidivism Algorithm. Elérhető: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (2020. 05. 14.).

[73] A hitelminősítési algoritmusok a pénzügyi szektorban használt matematikai és statisztikai modellek, amelyek segítenek meghatározni az egyének vagy vállalatok hitelképességét és a hitelezéshez köthető lehetséges kockázatát. Az egyik legismertebb modell az USA-ban használatos FICO pontszám, mely 300-tól 850-ig terjed. Ezek az algoritmusok többféle adattípusra támaszkodnak, mint például az adós hiteltörténetére, jövedelmi adataira, adósságterheire és egyéb pénzügyi mutatókra, hogy felbecsüljék a kölcsönök időben történő visszafizetésének valószínűségét.

vékony hiteltörténet kevés vagy korlátozottan rendelkezésre álló banki információra utal, ami gyakran előfordul azoknál az ügyfeleknél, akik nem igényeltek korábban kölcsönt, nem rendelkeznek hosszú hiteltörténettel, vagy kevés szolgálatói terméket (pl. hitelkártyák) birtokolnak. A hitelminősítési algoritmusok nagymértékben támaszkodnak ezekre az információkra az ügyfél értékelésekor, ha valakinek a profiljához vékony előtörténet volt csatolva az megnehezítette a pontos kockázatbecslést, ami a gyakorlatban a lehetséges kockázatok túlbecslését, az ügyfél megbízhatóságának alulbecslését eredményezte (bizonytalansági elfogultság).[74]

Ezek az esetek is rámutatnak, hogy a megfelelő tanító adathalmaz létfontosságú az MI modellek számára, mivel azok minősége befolyásolja a modell pontosságát, megbízhatóságát és alkalmazhatóságát. Ha az MI elfogult, nem kellően reprezentatív adathalmazokon tanul, akkor hiányos, helytelen vagy torzított képet kap a valóságról.[75] A diszkrimináció tilalmát előíró hagyományos jogi keretek elsősorban a döntéshozatal végeredményére összpontosítanak. Azonban az MI esetében a folyamat minden lépése, az adatgyűjtéstől és elemzéstől kezdve a modellalkotásig és a tanításig, kritikus fontosságú a végső kimenet szempontjából. Elvárható tehát, hogy az MI-re szabott jogi és etikai keretrendszerek a döntést megelőző teljes folyamatra is fókuszáljanak, nem csupán magára a döntés eredményére. A szabályozó és ellenőrző szerveknek pedig olyan kérdésekre is összpontosítaniuk kell, mint az adatforrások megbízhatósága, az adatelemzési módszerek részrehajlása, valamint a folyamatos monitorozás és audit lehetőségének biztosítása.

---

[74] A *bizonytalansági elfogultság* (uncertainty bias) okán a hitelbírálati algoritmusok hajlamosak a bizonytalanabb eseteteket következetesen kockázatosabbnak értékelni. De nem csak az adatok mennyisége, hanem a minősége is meghatározó, így, ha valakinek hosszú, de ellentmondásos hiteltörténete van, az is növelheti a bizonytalanságot. A törzsszövegben említetteken túl ezzel kapcsolatban felmerül az a további probléma is, hogy ha egy adatbázisban bizonyos pozitív értékek alulreprezentáltak (pl. a hitelkérelemben olyan irányítószám szerepel, amellyel kevés megbízható banki ügyfél rendelkezik), akkor az algoritmus számára a róluk készült előrejelzések is automatikusan bizonytalanabbá válhatnak, mely eredményeképpen az algoritmus hajlamos lehet alacsonyabb összegű és rosszabb kondíciójú hiteleket felajánlani az ügyfélnek. A hitelkérelmek során felmerülő torzításokról és bizonytalansági elfogultságról bővebben: Banasik, J., Crook, J., & Thomas, L. (2003): Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8), 822-832.o. https://doi.org/10.1057/palgra ve.jors.2601578. Az említett 2021-es tanulmány megállapításairól lásd bővebben: Blattner, L., és Nelson, S. (2021): How Costly is Noise? Data and Disparities in Consumer Credit. 27-30.o. ArXiv, abs/2105.07554.

[75] A torzított érzékelés közvetlen veszélyein túl fontos megemlíteni azt is, hogy az MI kimenetei gyakran olyan előrejelzések, amelyek közvetlen módon visszahatnak a környezetre, ezáltal hajlamosak fenntartani vagy még tovább mélyíteni a valóságban már meglévő egyenlőtlenségeket (un. *önerősítő diszkriminatív hatást* kifejtve). Például, ha egy prediktív rendőri rendszer az javasolja, hogy a rendőrség fokozottabban figyeljen meg egy számottevő etnikai kisebbséggel rendelkező városrészt. Ez több bűnözési adatot eredményezhet erről a területről és lakóiról. Az MI rendszer ezután még erősebb megfigyelést javasolhat erre a városrészre, mivel több „bűnözési" adatot gyűjt innen. Ezáltal a kisebbségek lakta terület fokozott monitorozása öngerjesztő, diszkriminatív kört hoz létre. Bővebben: Selbst, A. (2017): Disparate Impact in Big Data Policing. *Georgia Law Review*, 52, 3373.o. https://doi.org/10.2139/SSRN.2819182.

3.3.4 Szabadság, demokrácia és a döntési autonómia

Az MI rendszerek elterjedése hamar ráirányította a figyelmet az emberi jogokkal és a demokratikus jogállamisággal kapcsolatos potenciális veszélyekre is. Ilyen kockázatok közé tartoznak többek között az egyenlő bánásmód követelményével, a tisztességes eljáráshoz való jog biztosításával, az adat- és magánélet védelmével, a modern megfigyelő állam kialakulásával, vagy az autonóm fegyverek megjelenésével kapcsolatos kihívások. Ezek kifejtésére ehelyütt nincs mód, azonban szükséges szót ejteni egy olyan hosszútávú következményekkel bíró kockázatról, ami nagyban korlátozhatja az egyének döntési autonómiájának és önrendelkezéséi jogának szabad gyakorlását.

3.3.5 Kinek a választása?

Az önrendelkezés és a döntési autonómia olyan alapvető emberi jogok, melyek biztosítják az egyén szabadságát saját életének tervezésében és döntéseinek meghozatalában. Az MI azon képessége, mely az emberek preferenciáinak modellezésére és előrejelzésére irányul, új kihívásokat jelent ezek vonatkozásában. Az elmúlt másfél évtizedben az élet egyre több területén terjedtek el különböző algoritmikus ajánlórendszerek, melyek kimondott célja, hogy gyorsabbá és könnyebbé tegyék az egyén döntéshozatalát. Ezen eszközök használata azonban nagyban csökkentheti az egyén autonómiáját és kritikai gondolkodását, anélkül, hogy azt ő tolakodónak, vagy korlátozónak érezné (algorithmic micro-domination).[76] Az ajánlórendszerek működési elve, hogy hatalmas mennyiségű felhasználói adatot gyűjtenek és elemeznek[77], majd a felismert mintázatok és preferenciák alapján személyre szabott

[76] Danaher, J. (2019): The ethics of algorithmic outsourcing in everyday life. In K. Yeung & M. Lodge (szerk.): Algorithmic Regulation (91–118o.). Oxford, University Press. 107.o. https:// doi. org/ 10. 1093/ oso/ 97801 98838 494. 003. 0005

[77] Bár az algoritmikus ajánlórendszerek különböző adatokat és számolási modelleket használhatnak, ezek közül három terjedt el leginkább a gyakorlatban: (1) Tartalomszűrésen alapuló (content-based filtering) módszer, amely a felhasználó korábbi preferenciáiból és az ajánlásra szánt tételek tulajdonságaiból kiindulva tesz javaslatokat. Például egy zene-streaming szolgáltató platform, ahol a felhasználó korábban főleg klasszikus zenét hallgatott, erre alapozva fog hasonló stílusú, de a felhasználó számára még ismeretlen klasszikus zenei albumokat vagy

javaslatokat tesznek az útvonaltervezéstől kezdve, a diéták megalkotásán át, az online platformokon elérhető digitális tartalmak fogyasztásáig.[78] Első ránézésre ezek az ajánlások kényelmesebbé és hatékonyabbá tehetik a mindennapokat azáltal, hogy releváns opciókat kínálnak a végtelen elérhető kínálatból, azonban az algoritmusok által szűrt és priorizált információk korlátozhatják az egyént abban, hogy az összes számára nyitva álló választási lehetőséget megismerje, szűkítve ezzel látó- és mozgásterét. Fontos továbbá kiemelni, hogy az adatelemző technológiák fejlődése lehetővé tette az olyan részletes profilok létrehozását is, melyek feltárják az adott egyénre jellemző kognitív sajátosságokat és döntéshozatali sebezhetőséget.[79] Ezek az információk felhasználhatók arra, hogy befolyásolják az egyén magatartását.[80] Egy ilyen személyre szabott döntési környezetben a vásárlások ösztönzésétől kezdve a tudatos politikai mikrocélzásig bezárólag számtalan lehetőség áll rendelkezésre az emberek manipulációjára.[81] E jelenség a dolgozat második részében még részletesen kifejtésre kerül.

Az egyéni autonómiával és döntési szabadsággal összefüggésben szintén kockázatot jelenthetnek az állami-és magánszereplők által előszeretettel alkalmazott profilozáson, pontozáson és algoritmikus ajánlásokon alapuló társadalomszervező és döntéstámogató

előadókat ajánlani. Ez a szűrőmodell a tartalmat (mint szöveg, képek vagy hanganyag) elemezve keresi meg a hasonlóságokat, hogy olyan tételeket ajánljon, amelyek illeszkednek a felhasználó által fogyasztott tartalmakhoz; (2) A *demográfiai szűrés* (demographic filtering) alapja az a feltételezés, hogy a hasonló személyes jellemzőkkel (mint nem, életkor, lakóhely) rendelkező felhasználóknak hasonlóak lehetnek az ízléseik és preferenciáik is. Ennek megfelelően a rendszer a felhasználók demográfiai hasonlóságai alapján tesz javaslatokat. Így például egy fiatal városi nő más ajánlásokat kap, mint egy idősebb vidéki férfi; (3) A *kollaboratív szűrés* (collaborative filtering) egy adott csoportnál megfigyelt preferenciákra támaszkodik az ajánlások generálásakor. A rendszer azonosítja azokat a felhasználókat, akik hasonló ízlésűek és preferenciákkal rendelkeznek, mint "aktív" felhasználó, majd az ő értékeléseik és akcióik alapján tesz javaslatokat az aktív felhasználónak. Például, ha egy felhasználó 5 csillagosra értékelte Barcsi Tamás *Embertelenség az együttérzés korában - Filozófiai vázlat a kegyetlenségről* című könyvét a bookline felületén, akkor a rendszer az adatfeldolgozás során azon vásárlókra összpontosít, akik hasonló értékelést adtak erre a könyvre, és az ő további olvasmányélményeik alapján fog újabb könyveket ajánlani a felhasználónak. Lásd: Bobadilla, J., Ortega, F. Hernando, A. and Gutierrez, A. (2013): Recommender Systems Survey. *Knowledge-Based System*, Vol(46), 112-113.o.

[78] Cheney-Lippold, John (2017): We Are Data: Algorithms and the Making of Our Digital Selves. New York: NYU Press. 88-92.o.

[79] A digitális profilok összetettségét jól szemlélteti, hogy már egy 2013-ban publikált tanulmány is arra következtetésre jutott, hogy a közösségi médián zajló interakciókból gyűjtött legalapvetőbb információk is képesek lehetnek az adott felhasználó személyes tulajdonságainak pontos meghatározására (pl.: életkor, nem, szexuális irányultság, vallási és politikai nézetek, személyiségjegyek, intelligencia, függőséget okozó szerek használata, szülőkkel való együttélés vagy különélés). Lásd: M. Kosinski, D. Stillwell, and T. Graepel (2013): Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110:5802–5805.; W. Youyou, M. Kosinski, and D. Stillwell (2015): Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112:1036–1040.o.

[80] Harari, Yuval Noah (2018): 21 Lecke a 21. századra. (ford. Torma Péter). Budapest, Animus kiadó. 55-60.o.

[81] Susser, D., Roessler, B., & Nissenbaum, H. (2019): Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), 1–21.o.

gyakorlatok. Ezek közül ehelyütt kettő kiemelése indokolt; (1) Hypernudge[82] (2) Általános társadalmi pontrendszerek.

(1) Richard Thaler és Cass Sunstein 2008-ban megjelent könyvükben a libertárius paternalizmus elméletével összhangban olyan ösztönzők (nudge) széles körű alkalmazása mellett érvelnek, amelyek az élet különböző területén segítik az embereket a helyes és kívánatos döntések meghozatalában.[83] Ezek az ösztönzők az emberi viselkedés előrejelezhetőségére és meghatározott kognitív torzításokra építve alakítanak ki olyan *döntéskörnyezetet* (choice environment), mely a kívánatos és előre megfogalmazott választás meghozásának kedvez. A döntéstervezés elméletét továbbgondoló, és azt az MI és Big Data által teremtett új társadalmi és technológiai környezetbe átültető Karen Yeung un. Hypernudge[84] koncepciójában felsejlik egy olyan jövőbeli automatizált magatartás-irányító algoritmikus rendszer képe, amelyben ezek az ösztönzők már tökéletesen kiismerik a célszemélyt, annak minden egyediségével és sebezhetőségével együtt. A döntéstámogató rendszerek automatizálása révén az egyén döntéshozatali környezete folyamatosan változna. Elképzelhető, hogy egy ilyen profilozáson, pontozáson, viselkedés alapú előrejelzésen alapuló rendszerben, ahol a köz- és magánszolgáltatások elérésétől, a rendelkezésre álló bankhiteleken és biztosítási konstrukciókon keresztül a fogyasztásra ajánlott termékekig bezárólag mindent algoritmusok számítanak ki az egyén számára[85], ha egy választási lehetőséghez vagy szolgáltatáshoz való hozzáférés nem is nyílna meg azonnal, egy új,

[82] Yeung, Karen (2017): Hypernudge: Big data as a mode of regulation by design. *Information, Communication and Society*, Vol.20(1)

[83] A szerzők a helyes döntés meghozatalát elsősorban a pénzügyek (pl.: nyugdíjtervezés, megtakarítások) és egészégügy területén támogatták, kiemelve, hogy az ösztönzés semmilyen esetben sem jelenthet egyet a választási szabadság korlátozásával, lehetőségek tiltásával. Bővebben: Thaler, R., Sunstein, C. (2008): Nudge: Improving Decisions About Health, Wealth, and Happiness London, Penguin Books; Szántó Richárd - Dudás Levente (2017): A döntési helyzetek tudatos tervezésének háttere: A nudge fogalma, módszerei és kritikái. *Vezetéstudomány*, XLVIII. évf. 10. szám, 48-57.o.

[84] Karen Yeung Hypernudge koncepciója nem csak ajánlásokat tesz, hanem aktívan alakítja és irányítja az egyének viselkedését, döntéseit. A rendszer három fő jellemzője, hogy (1) mindenütt jelen vannak, beágyazódva mindennapi környezetünkbe; (2) folyamatosan gyűjtenek adatokat az egyénről, hogy viselkedési modelleket építsenek; (3) ezek alapján személyre szabott beavatkozásokat hajtanak végre, hogy a kívánt irányba tereljék a célszemélyt. Ezáltal válnak képessé az összefüggő manipulációra (coherent manipulation). Yeung, Karen (2017): Hypernudge: Big data as a mode of regulation by design. *Information, Communication and Society*, Vol.20(1), 118-136.o.

[85] Az ehhez hasonló algoritmikus döntéseken alapuló közigazgatás- és társadalomszervezést nevezi Danaher *algokráciának* (algocracy). Szélsőséges esete lehet, ha az algoritmusok teljes mértékben automatizált döntéseket hoznak emberi felügyelet nélkül, ami potenciálisan kiszoríthatja az embereket a közügyek irányításából és a nyilvános döntéshozatali folyamatból, fokozatosan csökkentve ezzel a politikai folyamatok legitimációját és demokratikus jellegét is. Lásd bővebben: Danaher J. (2016): The threat of algocracy: Reality, resistance and accommodation. *Philosophy & technology*, 29 (3), 245-268.o.

nagyobb pontossággal kiválasztott, és személyre szabottabb lehetőség kerülhet felkínálásra.[86] Ha elegendő felkínálandó lehetőség állna egy rendszer rendelkezésére, akkor a társadalom efféle értékelő számításokon alapuló irányítása egyáltalán nem látszana tényleges irányításnak, úgy tűnne az egyénnek, hogy a szabadságába nem avatkozott be senki. Egy automatizált irányító társadalomban az ajánló gyakorlatok területei az élet minden területére kiterjedhetnének, és mivel a 21. század bővelkedik a felkínálható szolgáltatásokban, tartalmakban és termékekben, az egyén még a rideg számításokon alapuló, előre meghatározott döntések hálózatában sem érezheti autonómiájának vagy önrendelkezési jogának számottevő korlátozását.[87]

(2) Az általános társadalmi pontrendszerek alkalmazásának talán legismertebb példája a kínai *Társadalmi Kreditrendszer*, melyben az automatikus értékelések egyenes következményeként bizonyos közszolgáltatásokhoz való hozzáférés is megtagadhatóvá válhat. A Kínai Államtanács által vázolt tervek alapján különböző minősítő és értékelő listák kialakítására kerül sor, melyben a jónak és kívánatosnak ítélt viselkedés – ilyen lehet például az önkéntes véradás – pontok hozzáadását, míg a nem megfelelő magatartás – például adófizetési határidő elmulasztása – pontlevonást vonhat maga után. Az összegyűjtött pontszámoktól függően részesülhet jutalomban vagy büntetésben az egyén. Ilyen jutalom például az állam által támogatott tanfolyamokon való ingyenes részvétel, hozzáférés kedvezményes lakáshitelhez, olcsóbb tömegközlekedés biztosítása, míg büntetésként az egyetemekre történő beiratkozás megtiltását, hitelkérelmek automatikus elutasítását, vagy a tömegközlekedés használatának határozatlan ideig történő megtiltását szabhatják ki.[88]

[86] Diósi Szabolcs (2022): Behaviorally informed regulations- an emerging trend in modern public policymaking. In Bendes Ákos L. et al. (szerk): III. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar Doktori Iskola. 125-142.o.

[87] Barcsi Tamás – Diósi Szabolcs (2022): Hasznos test, dividuum, nyersanyag. A felügyeleti társadalomtól az (ön)ellenőrző és megfigyelési kapitalizmusig. *Magyar filozófiai szemle*, 66(3), 190-191.o.

[88] A Kínai Államtanács 2014. június 14-én jelentette be, hogy elindítja nagyszabású, viselkedési adatokon, értékelő algoritmusokon és MI technológián alapuló Társadalmi Kreditrendszer programját. A pontrendszer a kínai bankok által évtizedek óta alkalmazott minősítési és rangsorolási rendszer masszív kiterjesztéseként működne, melynek gerincét a központi és regionális kormányzati szervek által, különböző szektorokban felhalmozott társadalmi/gazdasági magatartásokkal kapcsolatos adatok feldolgozása és azok meghatározott szempontok alapján történő értékelése adja. Erről részletesebben: Liu, Chuncheng (2019): Multiple Social Credit Systems in China. *Economic Sociology: The European Electronic Newsletter*, Vol 21 (1), 22–32.o.; Kollár Csaba (2020): Kína és a társadalmi kredit rendszere. Hadtudomány: A magyar hadtudományi társaság folyóirata 30:2, 79-97.o.; Vörös Zoltán (2019): Kína a digitális korszakban - Kínai internet és a Társadalmi Kreditrendszer. In: Dombi Judit – Rimai Dávid (szerk.): Digitális forradalom világunkban: tanulmánykötet. Pécs, Institutio Kiadó. 165-182.o.

Mindkét ismertetett példa esetében kiemelt kockázatot jelenthetne az, ha az ilyen rendszerek hatalomgyakorlás céljával megfelelő jogi szabályozás nélkül kerülnének alkalmazásra, hiszen ezek könnyen egész társadalmak kollektív magatartásbefolyásolásának eszközévé válhatnának.

3.4 Előre kalkulált kockázat, felülről szabályozott felhasználás

Érzékelve az MI elterjedésével összefüggő növekvő kockázatokat, az Európai Unió időben felismerte, hogy a technológiacsalád hosszú távon is fenntartható és biztonságos alkalmazása érdekében mielőbb átfogó jogi szabályozásra van szükség. Az Unió 2017 óta élen jár az MI szabályozásával kapcsolatos törekvésekben, erőfeszítései e téren meghatározóak, hiszen ahogyan az a GDPR esetében is megfigyelhető volt, az Uniós szabályalkotás hosszú távon is képes standardot szabni más szabályozás előtt álló országok számára (Brüsszeli hatás).[89]

Több éves jogalkotói eljárása során az Unió egy olyan jogi környezet kialakítására törekedett, amely nyitott a technológiai újítások felé, ösztönzi a kutatást és fejlesztést, egyúttal képes minimalizálni a társadalmi kockázatokat. Az innováció és európai versenyképeség megőrzésének támogatása, illetve az alapvető emberi jogok és európai értékek következetes, kompromisszumot nem ismerő védelme, mint olykor egymást korlátozó célok feloldására kockázatalapú szabályozási keretet javasolt, mely az MI rendszerek által jelentett kockázat mértékével arányos kötelezettségeket ír elő.

Az MI átfogó szabályozását nagyban nehezíti, hogy e rendszerek alkalmazási lehetőséginek sokszínűsége miatt nehéz előre jelezni, hogy egy adott modell mely területeken okozhat fenyegetést a jövőben.[90] Például az előző fejezetben említett magas minőségű képelemzésre

[89] A *Brüsszeli hatás* kifejezés arra utal, hogy az Európai Unió, mint jelentős és megkerülhetetlen gazdasági tényező - saját szabályozási törekvései révén - képes hosszútávon is irányadó globális szabványok kialakítására. Az elmúlt évtizedben ennek hatása tapasztalható volt többek között az adatvédelem, a digitális gazdaság, a fogyasztóvédelem és az élelmiszerbiztonság területén hozott szabályozásokban. Lásd bővebben: Bradford, Anu (2020): The Brussels Effect: How the European Union Rules the World. New York, Oxford University Press

[90] Bár egyes kockázatok utólag nyilvánvalóak lehetnek, azokat nem könnyű megvalósulásuk előtt azonosítani. Különösen nehéz előre látni az MI rendszereik jövőbeli alkalmazását (és az abból eredő kockázatokat), amikor azokat harmadik fél modelljeibe használják vagy integrálják.

képes mély neurális hálózat használható az orvostudományban rákdiagnosztikára, de szolgálhat a hatalomgyakorlás eszközeként is tömeges megfigyeléseken keresztül. A kockázatalapú megközelítés a kötelezettségek megállapításánál elsősorban arra koncentrál, hogy mely területeken alkalmazzák majd az adott MI-t.[91] Annak ellenére, hogy ez egy hatékony jogalkotói stratégia[92], a GenMI felbukkanása új kihívásokat hozott e tekintetben is, mivel ezek a modellek már a felhasználói szinten is könnyen módosíthatóak, így lehetséges alkalmazási területüknek előre történő meghatározása – és az azzal kapcsolatos kockázatok kategorizálása – megközelítőleg lehetetlen feladat. Ennek orvoslására írtak hozzá – a jogalkotási folyamat kellős közepén – egy új fejezetet a törvényszöveghez és hoztak létre egy új MI kategóriát az uniós jogalkotók.

A következő fejezet részletesen tárgyalja az Európai Unió MI szabályozásával kapcsolatos jogalkotási törekvéseit. Mivel a jogalkotási folyamat párhuzamosan zajlott e dolgozat írásával, a fejezet tartalma is szervesen bővült egyidőben az uniós eljárás előrehaladásával. Ez lehetővé tette a részletes nyomon követését a kezdetektől eszközölt – hol kisebb, hol nagyobb – jogszövegváltoztatásoknak, ezáltal elősegítve a mögöttük meghúzódó jogalkotási stratégiák és elképzelések mélyebb megértését. A fejezet a szabályozási folyamat ismertetése során főként az alapvető emberi jogokra és európai értékekre veszélyes MI rendszerek és gyakorlatok ismertetésére és a releváns jogszabályok bemutatására összpontosít (azok jövőbeli alkalmazásának és végrehajtásának technikai részleteit nem tárgyalja).

[91] A megközelítés egy másik gyengesége, hogy mivel az MI rendszereket a szándékolt céljukat szem előtt tartva osztályozza, nem feltétlenül veszi figyelembe azokat a veszélyeket, amelyet egy MI rendszer a rendeltetési célján kívül történő használata okozhat. Ennek az a tovagyűrűző hatása, hogy az e rendszerek fejlesztői és szolgáltatói viselik a szabályozási teher nagy részét. Ők felelősek a kockázatkezelési, adatminőségi, átláthatósági követelmények teljesítéséért. Ez jelentős jogi és adminisztratív terhet róhat rájuk, ami különösen a kisebb vállalatokra és startupokra terhes, amelyeknek eleve korlátozottabbak az erőforrásaik.

[92] Az EU újabban előszeretettel választja ezt a jogalkotási megközelítést a legújabb digitális technológiák szabályozására. Ez az MI szabályozás mellett az adatvédelem (GDPR) és az online platformok és közvetítő szolgáltatók esetében is megfigyelhető (DSA) volt.

4 Az Unió MI-rendelete

4.1 Európa az MI szabályozás útján

Az Európai Tanács első alkalommal a 2017. október 19-ei ülésén elfogadott következtetésében kérte fel a Bizottságot arra, hogy terjesszen elő javaslatot „*a mesterséges intelligencia európai megközelítésére"* vonatkozóan.[93] A felkéréseknek eleget téve az Európai Bizottság 2018. április 25-én tette közzé első átfogó európai MI stratégiáját. E stratégia három fontos célt fogalmazott meg: (1) az EU technológiai és ipari kapacitásainak megerősítését; (2) a küszöbön álló társadalmi-gazdasági változásokra való felkészülést; illetve (3) az MI technológiákkal és gyakorlatokkal kapcsolatos megfelelő etikai és jogi keret kialakítását.[94] Az első kettő célt elsősorban az állami és magánbefektetések jelentős növelésével, kutatások és kutatóközpontok támogatásával, új képzések indításával, digitális készségek és tehetségek fejlesztésével, valamint az MI technológiák számára kedvező adatgazdag környezet *(adatökoszisztéma)* kialakításával tervezte elősegíteni. Az etikai és jogi keret tekintetében a Bizottság kiemelte, hogy az MI-nek mindenekelőtt az állampolgárok érdekeit kell szolgálnia, ezzel összefüggésben, fejlesztése és alkalmazása során mindenkor tiszteletben kell tartani mind az Európai Unióról szóló szerződés 2. cikkében meghatározott értékeket, mind az EU Alapjogi Chartájában felsorolt alapvető jogokat.[95] A Bizottság továbbá hangsúlyozta, hogy az MI alkalmazásoknak nemcsak a jogszabályoknak kell megfelelniük, hanem bizonyos etikai elvekkel is összhangban kell lenniük. Ezen célok előmozdítása érdekében a Bizottság 2018-ban létrehozta a *Mesterséges intelligenciával foglalkozó magas szintű szakértői csoportot*, melyet az etikai iránymutatások összeállításával, illetőleg átfogóbb MI-politikára vonatkozó ajánlások kidolgozásával bízott meg.[96]

---

[93] Európai Tanács (2017): Következtetések, EUCO 14/17, 7.o.

[94] Európai Bizottság (2018): Mesterséges intelligencia Európa számára (A közös európai adattér felé) COM (2018) 237 final. 7-20.o.

[95] A Bizottság egy 2019-ben közzétett kommentárjában ezt a megközelítést nevezi "emberközpontú" mesterséges intelligenciának. Ennek értelmében az MI *nem maga a cél, hanem egy, az embereket szolgáló eszköz, melynek végső célja az emberi jólét növelése.* Erről részletesebben: Európai Bizottság (2019): Az emberközpontú mesterséges intelligencia iránti bizalom növelése COM(2019) 168 final.

[96] A szakértői csoport 2019 márciusában adta át a Bizottságnak az iránymutatásuk végleges változatát, melyben a megbízható MI kialakítását három előfeltételhez szabta: (1) a mesterséges intelligenciának be kell tartania a jogszabályokat; (2) meg kell felelnie az etikai elveknek; (3) stabilnak kell lennie. Ezt figyelembevéve a szakértői

Az áprilisban ismertetett stratégiára építve az Európai Tanács a 2018. június 28-i ülésén elfogadott következtetésben[97] felkérte a Bizottságot, hogy a tagállamokkal együttműködve dolgozzon ki egy *összehangolt MI tervet.* A Bizottság ugyanezen év decemberében mutatta be a *Mesterséges intelligenciára vonatkozó összehangolt tervet*[98], melyben hangsúlyozta, hogy a vázolt célok elérése érdekében – például a beruházások növelése, kiterjedt adat-ökoszisztéma létrehozása, az EU digitális versenyképességének hosszú távú biztosítása – európai szintű koordinációra, határokon átnyúló szorosabb együttműködés kiépítésére és nemzeti jogalkotási töredezettségektől mentes egységes piac kiépítésére van szükség. A közös együttműködés alapjainak megteremtésén és a legfontosabb beruházási területek meghatározásán túl a Bizottság arra is ösztönözte a tagállamokat, hogy dolgozzanak ki nemzeti stratégiai elképzeléseket az MI-ről.

Az MI szabályozással kapcsolatos uniós törekvések 2019-ben kaptak nagyobb nyilvánosságot, amikor Ursula von der Leyen frissen megválasztott Bizottsági elnök a 2019–2024 közötti időszakra szóló, *Ambiciózusabb Unió*[99] című politikai iránymutatásában kijelentette, hogy hivatali idejének *első száz napjában* a Bizottság jogszabályt fog előterjeszteni az MI emberi és etikai vonatkozásaival kapcsolatos összehangolt európai megközelítésre vonatkozóan.

A Bizottság végül 2020. február 19-én[100] tette közzé az MI-vel kapcsolatos *Fehér Könyv*ét, amelyben *konkrét szakpolitikai alternatívákat*[101] fogalmazott meg arra vonatkozóan, hogy

---

csoport hét olyan kulcsfontosságú követelményt sorolt fel, melyeknek az MI-rendszer teljes életciklusa során kötelező megfelelni: (1) az emberi cselekvőképesség támogatása és emberi felügyelet; (2) műszaki stabilitás és biztonság; (3) adatvédelem és adatkezelés; (4) átláthatóság; (5) sokféleség, megkülönböztetésmentesség és méltányosság; (6) társadalmi és környezeti jólét; (7) elszámoltathatóság. Lásd bővebben: Európai Bizottság, Mesterséges intelligenciával foglalkozó magas szintű szakértői csoport (2019): Etikai iránymutatás a megbízható mesterséges intelligenciára vonatkozóan, 19-24.o.

[97] Európai Tanács, Az Európai Tanács ülése (2018. június 28.) – Következtetések, EUCO 9/18 2018, 8.o.

[98] Európai Bizottság (2018): A mesterséges intelligenciáról szóló összehangolt terv COM(2018) 795 final.

[99] *Hivatali időm első száz napjában a mesterséges intelligencia humán és etikai következményeire vonatkozó összehangolt európai megközelítésről szóló jogszabályt fogok előterjeszteni. Ez arra is ki fog terjedni, hogy hogyan tudjuk felhasználni az óriási méretű adathalmazokat, amelyek gazdagságot teremtenek a társadalmaink és vállalkozásaink számára.* Lásd bővebben: Európai Bizottság, Kommunikációs Főigazgatóság, Leyen, U. (2019): Ambiciózusabb Unió – Programom Európa számára: politikai iránymutatás a hivatalba lépő következő Európai Bizottság számára (2019–2024). Elérhető: https://data.europa.eu/doi/10.2775/56352

[100] A *Fehér Könny*vel egy időben a Bizottság közzétette a „*Jelentés a mesterséges intelligencia, a tárgyak internete és a robotika biztonsági és felelősségi vonatkozásairól*" című dokumentumot, valamint az Európa digitális jövőjének alakításáról (COM(2020) 0067) és az Európai Adatstratégiáról (COM(2020) 0066) szóló közleményeit is.

[101] A *Fehér Könyv* a Bizottság által 2018 áprilisiában közzétett MI stratégiájára építve tovább hangsúlyozza a beruházások növelésének (magánszektor bevonása, kkv-k fokozatos támogatása, közigazgatási szervek technológiai átállítása) és az adatgazdag környezet kialakításának szükségességét. A koordinált terv alapján pedig az egységes piac létrehozására és széttöredezett nemzeti jogalkotás elkerülésének szükségességére helyezi

miként lehet összeegyeztetni a magas minőségű MI technológiák és gyakorlatok fokozottabb elterjedését *(kiválósági ökoszisztéma)* a biztonságos alkalmazási környezet kiépítésével (*bizalmi ökoszisztéma*).[102] A Bizottság hangsúlyozta, hogy az MI új szabályozási keretének kialakításánál fontos követelmény a megfelelő mértékű szigor biztosítása, ugyanakkor lényeges szempont a joganyag túlzottan előíró jellegűvé válásának elkerülése is, amely az aránytalanul nagy terhek megállapításával hosszútávon az innováció gátjaként szolgálna. A hatékony, de arányos beavatkozás lehetőségének megteremtése érdekében a *Fehér Könyv* bevezette a *nagy kockázatú MI*-rendszerek fogalmát,[103] melyekkel szemben szigorúbb követelményeket állított (többek között stabilitással és pontossággal, átláthatósággal, tájékoztatással, az emberi felügyelettel és az adatkormányzással kapcsolatosan). A szabályozási javaslatok ismertetésén túl a *Fehér Könyv* közzétételével egy időben a Bizottság széles körű nyilvános konzultációt[104] is indított az MI európai megközelítésére vonatkozó szakpolitikai és szabályozási intézkedésekről.

A konzultáción elhangzottakra, illetve az Uniós intézmények által korábban közzétett állásfoglalásokra, javaslatokra, iránymutatásokra és stratégiákra építve a Bizottság végül 2021. április 21-én tette közzé az MI szabályozásával kapcsolatos rendelet-tervezetét.

4.2 A kockázat alapú megközelítés

---

a hangsúlyt. A szakértői tanács által készített etikai iránymutatás következtetéseit elfogadva külön kiemeli, hogy az MI technológiákkal kapcsolatos etikai, jogi problémákból az átláthatatlanság („feketedoboz-hatás"), az összetettség, a kiszámíthatatlanság, az autonóm viselkedés, az elfogultságot és megkülönböztetést követő gyakorlatok, illetve a magánszféra és a személyes adatok védelme kitüntetett figyelmet igényelnek.

[102] Európai Bizottság: Fehér könyv a mesterséges intelligenciáról – A kiválóság és a bizalom európai megközelítése COM(2020) 65 final. 3-4.o.

[103] A *Fehér Könyv* javaslata alapján egy MI-rendszert abban az esetben kell nagy kockázatúnak tekinteni, ha; (1) olyan ágazatban használják, ahol jelentős kockázatok várhatók (pl.: energiaszektor, határellenőrzés, igazságszolgáltatás); (2) az adott ágazatban ezenfelül olyan módon is használják az adott alkalmazást, ami jelentős kockázatokat hordoz. Lásd bővebben: Európai Bizottság (2020): Fehér könyv a mesterséges intelligenciáról: a kiválóság és a bizalom európai megközelítése. COM(2020) 65 final, 21-29.o.

[104] A nyilvános konzultációhoz, melynek lefolytatására online keretek között került sor 2020. február 19. és június 14. között, összesen 1.215 résztvevő (magánszemélyek, vállalatok, kutatói csoportok, hatóságok) csatlakozott. Összeségében elmondható, hogy a válaszadók többsége kifejezetten támogatta a kockázatalapú megközelítést. Az ilyen megközelítés alkalmazását jobb megoldásnak tekintették, mint a valamennyi MI-rendszert érintő általános szabályozást. A résztvevők emellett válaszaikban - elfogadva a *Fehér Könyv* és a szakértői csoport által megfogalmazott legfőbb veszélyforrásokat - további konkrét kockázatokkal kapcsolatban is hangod adtak aggodalmaiknak, például az online hirdetések során tapasztalt differenciált árképzés, a kommunikációs szűrőbuborékok kialakulása, a profilalkotáson alapuló manipuláció kockázata, a politikai folyamatokba és választásokba való beavatkozás veszélye, illetve az automatizált döntésekhez kapcsolódó állampolgári kontroll elvesztése kapcsán.

Összegezve a közzétételt megelőző években megfogalmazott prioritásokat, a rendelet-tervezet a következő négy általános célkitűzést fogalmazta meg: (1) „*annak biztosítása, hogy az Unióban forgalomba hozott és használt MI-rendszerek biztonságosak legyenek, és tiszteletben tartsák az alapvető jogokra és az uniós értékekre vonatkozó hatályos jogszabályokat”*; (2) „*a jogbiztonság biztosítása a mesterséges intelligenciába történő beruházások és a mesterséges intelligenciát érintő innováció elősegítése érdekében”*; (3) „*az irányításnak és az MI-rendszerek tekintetében az alapvető jogokra és biztonsági követelményekre vonatkozó hatályos jogszabályok hatékony érvényesítésének a javítása”*; (4) „*a jogszerű, biztonságos és megbízható MI-alkalmazások tekintetében az egységes piac kialakításának elősegítése és a piac széttöredezettségének megelőzése”*.[105]

Az olykor eltérő stratégiákat igénylő célkitűzések (pl.: innováció, fejlesztés és a beruházások támogatása vs. az uniós értékek valamit az állampolgárok szabadságjogainak kitüntetett védelme) sikeres elősegítése érdekében a rendelet-tervezett kockázatalapú megközelítést alkalmazott az MI-rendszerek Unión belül történő fejlesztésének, forgalmazásának és használatának szabályozására. E megközelítés elődleges célja, hogy időben azonosítsa és értékelje az MI-rendszerekhez kapcsolódó lehetséges kockázatokat, valamint, hogy a potenciálisan felmerülő kockázatok esetén olyan *minimumkövetelményekre* korlátozódjon, amelyek képesek hatékonyan kezelni a veszélyt, de egyúttal nem vezetnek *szükségtelen kereskedelmi korlátozáshoz*. Az arányossági alapokon nyugvó felfogás értelmében a javaslat az MI által jelentett potenciális kockázat növekedésével egyre szigorúbb jogszabályi követelmények alkalmazását írta elő. Ennek megfelelően tiltani rendelte az MI egyes különösen – *elfogadhatatlan mértékben* – veszélyes felhasználási módjait, szigorúan szabályozta a nagy kockázatú MI rendszerek fejlesztését, forgalmazását, használatát, illetve enyhe követelményeket állapított meg a kevésbé kockázatos rendszerek számára. A rendelet-tervezet négy kockázati szintet határozott meg, amelyek az (1) elfogadhatatlan, (2) nagy, (3) korlátozott és a (4) minimális kockázatú MI rendszereket jelölik.

4.3 Korlátozott és minimális kockázatú MI-rendszerek

[105] MI rendelet-tervezet (Indoklás 1.1.)

Minthogy alkalmazásuk nem jelent kirívó kockázatot az állampolgárok alapvető jogaira és biztonságára nézve, a rendelet-tervezet a korlátozott és minimális kockázatú MI-rendszerekkel kapcsolatosan enyhe jogkövetelményekre korlátozódott.

A minimális kockázatú rendszerek (például az e-mailszolgáltatók által alkalmazott spam-szűrők) Európai Unión belül történő fejlesztése és használata a hatályos jogszabályok alapján további jogkötelezettségek nélkül engedélyezett. A Bizottság e rendszerekkel kapcsolatban elsősorban *ösztönözte* az olyan magatartási kódexek *önkéntes* kidolgozását és alkalmazását, amelyek célja, hogy előmozdítsák a veszélyesebb MI-rendszerekkel kapcsolatosan megfogalmazott követelmények betartását.[106]

A rendelet-tervezet 52. cikkében foglalt átláthatósági kötelezettség azonban a *természetes személyekkel* közvetlen módon érintkező, nem nagy kockázatú besorolású MI-rendszerek tekintetében is kötelezővé tette a szolgáltatók számára annak biztosítását, hogy „*az MI-rendszereket úgy tervezzék meg és fejlesszék ki, hogy a természetes személyek tájékoztatást kapjanak arról, hogy valamely MI-rendszerrel állnak kapcsolatban, kivéve, ha ez a körülmények és a használat kontextusa alapján nyilvánvaló”*. Az 52. cikk továbbá rendelkezett arról is, hogy azon MI-rendszerek felhasználóinak, melyek képesek valós személyekre, helyekre vagy eseményekre megtévesztő módon hasonlító *képi, audio-, illetve videótartalmat generálni vagy manipulálni*, kötelezően fel kell tüntetniük, hogy az általuk közölt tartalmak mesterségesen lettek létrehozva.[107] Ezek az átláthatósági szabályok a minimális kockázatú MI-alkalmazások kapcsán elsősorban a különböző digitális/virtuális asszisztensek, chatbotok, illetve a későbbiekben részletesen is kifejtésre kerülő *deepfake* tartalmak előállítására alkalmas rendszerek potenciális veszélyeinek mérséklésére szolgáltak.

4.4 Nagy kockázatú MI-rendszerek

A Bizottság a rendelet-tervezet Indoklásában előzetesen rögzítette, hogy azon MI-rendszerek tartoznak ebbe a kategóriába, *amelyek nagy kockázatot jelentenek a természetes személyek*

[106] MI rendelet-tervezet 69. cikk
[107] MI rendelet-tervezet 52. cikk

egészségére és biztonságára vagy alapvető jogaira nézve.[108] Ezt követően részletesebben a III. Címben foglalkozik a magas kockázatú MI-rendszerekkel, melyeknek két fő kategóriáját határozza meg: (1) azon rendszerek, melyeket *már létező uniós harmonizációs jogszabályok hatálya alá tartozó termék biztonsági alkatrészeként vagy önmagában ilyen termékként kívánják használni;*[109] (2) azon MI-rendszerek, melyek felhasználási módjukból vagy alkalmazási területükből kifolyólag számottevő módon befolyásolhatják az állampolgárok alapvető szabadságjogait. A Bizottság a javaslathoz csatolt III. mellékletben nevesítette e nyolc területet: (1) természetes személyek biometrikus azonosítása és kategorizálása; (2) kritikus infrastruktúra irányítása és működtetése; (3) oktatás és szakképzés; (4) foglalkoztatás; (5) alapvető magánszolgáltatásokhoz és közszolgáltatásokhoz való hozzáférésről való döntés, (6) bűnüldözés; (7) migráció, menekültügy és határellenőrzés; (8) igazságszolgáltatás és demokratikus folyamatok.[110] A rendelet időtállóságának és jövőbeni alkalmazhatóságának elősegítése érdekében a javaslattervezet 7. cikke arra is felhatalmazza a Bizottságot, hogy a III. mellékletben nevesített MI-rendszerek jegyzékét - ilyen például a bűnüldöző hatóságok által a bűnismétlés valószínűségének megállapítására, a potenciális bűncselekmények előrejelzésére, vagy a természetes személyek érzelmi állapotának észlelésére használt MI-rendszerek listája - a jövőben további MI-k hozzáadásával is bővítse, amennyiben azokat az előbb felsorolt nyolc terület valamelyikén kívánják használni, illetve amennyiben azok

[108] MI rendelet-tervezet (Indoklás 5.2.3.)

[109] A rendelettervezet III. Címének 6. cikke tartalmazza a nagy kockázatú MI-rendszerekre vonatkozó besorolási szabályokat. Ennek értelmében az MI-rendszer nagy kockázatúnak minősül, ha mindkét alábbi feltétel teljesül: (1) a javaslat II. mellékletben felsorolt uniós ágazati jogszabályok *hatálya alá tartozó termék biztonsági alkatrészeként vagy önmagában ilyen termékként kívánják használni*; (2) a javaslat II. mellékletében felsorolt uniós harmonizációs szabályok értelmében *forgalomba hozatala vagy üzembe helyezése céljából harmadik fél által végzett megfelelőségértékelésnek kell alávetni.*

[110] A III. melléklet részletesebben is nevesít alkalmazási területeket és módozatokat a nagy kockázatú MI rendszerekkel kapcsolatban, ilyenek például: (1) a természetes személyek valós idejű és nem valós idejű távoli biometrikus azonosítására szolgáló MI-rendszerek; (2) a közúti forgalom irányításában és működtetésében használt MI-rendszerek; (3) oktatási és szakképzési intézmények hallgatóinak értékelésére, vagy az oktatási intézményekbe való felvételhez szükséges tesztek résztvevőinek értékelésére szolgáló rendszerek; (4) természetes személyek toborzására, kiválasztására, üres álláshelyek meghirdetésére, pályázatok szűrésére, valamint a jelöltek interjúk során történő értékelésére szolgáló MI-rendszerek; (5) állami segítségnyújtási ellátásokra és szolgáltatásokra való jogosultságának értékelésére, vagy a természetes személyek hitelképességének értékelésére, hitelpontszámuk megállapítására szolgáló MI-rendszerek; (6) a bűnüldöző hatóságok által a természetes személyek bűnelkövetésének kockázatértékelésére használt rendszerek, vagy olyan rendszerek melyeket profilalkotás alapján természetes személyek személyiségbeli jellemzőinek és tulajdonságainak vagy múltbeli bűnöző magatartásának értékelése alapján bűncselekmények előfordulásának vagy megismétlődésének előrejelzésére használnak; (7) határellenőrzés során valamely tagállam területére belépett vagy oda belépni szándékozó természetes személy által jelentett kockázat értékelésére használt rendszerek; (8) olyan MI-rendszerek, amelyek célja, hogy segítsék az igazságügyi hatóságokat a tények és a jog kutatásában és értelmezésében, valamint a jog konkrét tényállásra történő alkalmazásában. MI rendelet-tervezet III. Melléklet 1-8.

alkalmazása az *egészségben és a biztonságban okozott kár vagy az alapvető jogokat érő kedvezőtlen hatás tekintetében* kiemelt kockázattal járnak.[111][112]

A rendelet-tervezet III. címének 2. fejezete rögzítette a magas kockázatú MI-rendszerekre vonatkozó jogszabályi követelményeket. A javaslat többek között kötelezettségeket fogalmazott meg a *kockázatkezelési rendszerek* létrehozásával, magas minőségű *adatkormányzással és adatgazdálkodással, átlátható működéssel*, a használat teljes időtartalma alatt fenntartott hatékony *emberi felügyelettel*, valamint a *pontos* és *biztonságos* tervezéssel és fejlesztéssel kapcsolatban.[113] Ezen felül a rendelet-tervezet a szolgáltatókra és felhasználókra tekintettel is írt elő további kötelezettségeket, ilyen például a *minőségirányítási rendszer* bevezetése vagy a kötelező *megfelelőségértékelés eljárás* lefolytatása az MI-rendszer forgalomba hozatalát illetve üzembe helyezését megelőzően.[114] A rendelet-tervezet 71. cikke értelmében abban az esetben, ha az adott MI-rendszer nem felel meg a felsorolt követelményeknek vagy kötelezettségeknek, *legfeljebb 20 000 000 EUR összegű közigazgatási bírsággal, illetve vállalkozások esetében az előző pénzügyi év teljes éves világpiaci forgalmának legfeljebb 4 %-át kitevő összeggel sújtható* (a magas minőségű adatkormányzás és adatgazdálkodási követelményének megszegése esetén pedig 30 000 000 EUR összegű közigazgatási bírsággal, vagy a teljes éves világpiaci forgalmának legfeljebb 6 %-át kitevő összeggel).[115]

4.5 Tiltott gyakorlatok, „elfogadhatatlan kockázat”

A Bizottság álláspontja szerint bizonyos MI-rendszerek olyan mértékű – *elfogadhatatlan* - kockázatot jelentenek az uniós értékek és állampolgárok alapvető jogaira nézve, melyeknek orvoslása nem lehetséges sem technikai megoldásokkal, sem különböző jogi és eljárási

[111] MI rendelet-tervezet 7. cikk
[112] Többen fogalmaztak meg kritikát a tervezetben biztosított bővítési eljárással kapcsolatban. Smuha és társai például a mechanizmus anti-demokratkus jellegét kifogásolva, azt javasolták, hogy a jogszabály későbbi változatában fontos volna konzultációs és részvételi jogokat is biztosítani az EU minden állampolgára számára a magas kockázatú rendszerek listájának módosítását illetően. Lásd bővebben: Smuha, Nathalie A., Ahmed-Rengers, Emma Harkens, Adam, Li, Wenlong MacLaren, James Piselli, Riccardo, Yeung, Karen (2021): How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission’s Proposal for an Artificial Intelligence Act. http://dx.doi.org/10.2139/ssrn.3899991
[113] MI rendelet-tervezet III. Cím 2. fejezet, 8-15. cikk
[114] MI rendelet-tervezet III. Cím 3. fejezet, 16-29. cikk
[115] MI rendelet-tervezet X. Cím 71. cikk

biztosítékokkal; az uniós állampolgárok érdekeinek védelmében ezen alkalmazások teljes tiltása szükséges. A rendelet-tervezet négy tiltott kategóriát nevez meg, amelyek közül hármat teljes egészében tiltani rendel, egyet pedig csak meghatározott feltételek teljesülése esetén engedélyez.

4.5.1 Magatartást torzító gyakorlatok

Az V. Címben megnevezett négy tilalom közül az első kettő a különböző emberi magatartásokat manipuláló vagy azokat befolyásoló MI alkalmazásokra irányult. Ennek értelmében tilos az olyan rendszerek forgalomba hozatala, üzembe helyezése vagy használata, amelyek:

„(a) szubliminális technikákat alkalmaznak az adott személy tudatán kívül, annak érdekében, hogy lényegesen torzítsák egy személy magatartását oly módon, amely e személynek vagy más személynek testi vagy lelki károsodást okoz vagy okozhat;

(b) a személyek egy meghatározott csoportjának életkor, testi vagy szellemi fogyatékosság miatt fennálló valamilyen sebezhetőségét kihasználják annak érdekében, hogy torzítsák az adott csoporthoz tartozó személy magatartását oly módon, amely e személynek vagy más személynek testi vagy lelki károsodást okoz vagy okozhat”[116]

A Bizottság által közétett javaslatban a magatartást befolyásolni és torzítani képes MI-rendszerek forgalomba hozatala, üzembe helyezése vagy használata csak abban az esetben volt tiltott, amennyiben azok az egyénnek *károsodást* okoznának. Az (a) pont esetében például nem a befolyásolás ténye vagy az arra irányuló szándék volt a mérvadó, hanem az, hogy a gyakorlat végrehajtása olyan módon történjen, „*amely e személynek vagy más személynek testi vagy lelki károsodást okoz vagy okozhat".* A javaslatszövegbe beemelt szigorú *testi és lelki károkra* való korlátozás számos kiskaput rejtett magában, mely hosszú távon alááshatta volna a rendelet eredeti célkitűzéseit. Ahogy Veale is utalt rá, e megszorító hatály ellen joggal volt felhozható, hogy az állampolgárokat védelmezni szükséges valamennyi olyan szubliminális (tudatalatti) technikával vagy befolyással szemben, amely megkerüli, gyengíti vagy aláássa a

[116] MI rendelet-tervezet II. Cím 5. cikk (1) bekezdés a-b pont

racionális kontrollt, függetlenül attól, hogy ezek a befolyásolási kísérletek tényleges *testi vagy lelki* károkat okoznak vagy sem.[117] Ugyanezen gondolat mentén a (b) pont tekintetében megalapozott lehet azon elvárás is, hogy amennyiben egy MI-rendszert arra terveztek vagy azt oly módon alkalmazzák, hogy kihasználja bizonyos sérülékeny csoportok sebezhetőségét (gyermekek, idősek, értelmi vagy testi fogyatékossággal élő emberek), akkor azt – tekintet nélkül annak tényleges következményeire – be kell tiltani.

A Régiók Európai Bizottsága egy 2021 decemberében közzétett véleményében javasolta a kár feltétel kiterjesztését a testi és lelki károkon túl a gazdasági károkra is.[118] Így a tilalom védelmet nyújthat azon MI technikákkal szemben, amelyek a kiszolgáltatott emberek számára negatív gazdasági következményekkel, *pénzügyi veszteséggel vagy gazdasági diszkriminácival* járhatnak (pl. az arra hajlamos egyént játékfüggőségbe sodorja). A (b) pont tekintetében szintén indokolt lenne kiterjeszteni a védett személyek körét, mely jelenlegi formájában a *személyek egy meghatározott csoportjának életkor, testi vagy szellemi fogyatékosság miatt fennálló valamilyen sebezhetőségére* korlátozódik. A rendelet-tervezetben felsorolt személyek körén kívül sebezhetőek – egyúttal könnyebben befolyásolhatóak és kihasználhatóak – lehetnek azon személyek is, akik különféle mentális betegségekkel vagy viselkedési zavarokkal küzdenek (pl.: depresszió, bipoláris zavar, generalizált szorongás, ADHD vagy függőség).

Az Európai Gazdasági és Szociális Bizottság pedig azt vetette fel, hogy a szubliminális technikák nemcsak testi és lelki károkhoz vezethetnek, hanem tágabb kontextusban nézve „*alkalmazási környezetüktől függően egyéb káros személyi, társadalmi vagy demokratikus hatásokkal – például a szavazási magatartás megváltozásával – is járhatnak”*. Ennek kapcsán javasolta az EGSZB a 2021 szeptemberében elfogadott véleményében, hogy a tiltott gyakorlatok körét oly módon bővítsék ki, hogy azok magukba foglalják azon alkalmazásokat is, melyek egy adott személynek vagy személyek csoportjának „*sérti vagy sértheti alapvető*

[117] Veale, Michael - Frederik Zuiderveen Borgesius (2021): Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4) 97-112.o.

[118] Az Régiók Európai Bizottsága az alábbi kiegészítést javasolta az első bekezdés (a) pontja tekintetében: „testi vagy lelki károsodást okoz vagy okozhat, *sérti vagy sértheti más személynek vagy személyek egy csoportjának alapvető jogait, beleértve testi vagy lelki egészségüket és biztonságukat, károkat okoz vagy okozhat a fogyasztóknak, például pénzügyi veszteséget vagy gazdasági diszkriminációt, vagy aláássa vagy alááshatja a demokráciát és a jogállamiságot”* Régiók Európai Bizottsága (2021): Vélemény. A mesterséges intelligenciával kapcsolatos európai megközelítés – A mesterséges intelligenciáról szóló jogszabály. SEDEC-VII/022, 13.o.

jogait, illetve ezen túl sérti vagy sértheti az általánosabb/össztársadalmi szinten értelmezhető demokráciát és a jogállamiságot". [119]

4.5.2 Általános társadalmi pontozás

A tilalmak második csoportját az úgynevezett *általános társadalmi pontozással* kapcsolatos alkalmazások alkotják. Ennek értelmében tilos az:

„MI-rendszerek hatóságok általi vagy nevükben történő forgalomba hozatala, üzembe helyezése vagy használata természetes személyek megbízhatóságának értékelésére vagy osztályozására egy bizonyos időszakon keresztül, közösségi magatartásuk, illetve ismert vagy előre jelzett személyes vagy személyiségi tulajdonságok alapján…[120]

A Bizottság hamar felismerte a pontozáson alapuló gyakorlatokban rejlő jelentős társadalmi veszélyeket. Az ehhez hasonló *általános társadalmi pontozási* gyakorlatok korai és szigorú szabályozásának fontosságát – minthogy azok az állami túlhatalom kiépítésének kockázatát rejtik - az Európai Unió több alkalommal is hangsúlyozta.

A Bizottság javaslata kapcsán azonban aggályokat vetett fel, hogy a tilalom hatálya kizárólag a közszektorra korlátozódott. Ahogy arra az Európai Adatvédelmi Testület és az európai adatvédelmi biztos 5/2021. sz. közös véleménye is utalt, a magánvállalatok, például a különböző közösségi média- és felhőszolgáltatók, a közhivatalokhoz hasonlóan hatalmas mennyiségű személyes adatot kezelnek, mely felhasználásával könnyedén végezhetnek kiterjedt közösségi pontozást.[121] Lényeges szempont továbbá az is, hogy a magánszektorban

[119] Európai Gazdasági és Szociális Bizottság (2021): Vélemény - Javaslat európai parlamenti és tanácsi rendeletre a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról INT/940, 4-5.o

[120]„ .. *oly módon, hogy a közösségi pontszám a következő helyzetek egyikéhez vagy mindkettőhöz vezet:* (i) *a gyakorlat természetes személyek vagy azok egész csoportjával szemben hátrányos* vagy *kedvezőtlen* bánásmódhoz vezet *olyan szociális kontextusban, amely nem függ össze azzal a kontextussal, amelyben az adatokat eredetileg létrehozták vagy gyűjtötték*; (II) *a gyakorlat természetes személyek vagy azok egész csoportjával szemben olyan hátrányos vagy kedvezőtlen bánásmódhoz vezet, amely indokolatlan vagy aránytalan közösségi magatartásukhoz vagy annak súlyosságához képest"* MI rendelet-tervezet II. Cím 5. cikk i,ii

[121] Európai Adatvédelmi Testület (2021): Az Európai Adatvédelmi Testület és az európai adatvédelmi biztos 5/2021. sz. közös véleménye a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról szóló európai parlamenti és tanácsi rendeletre irányuló javaslatról 3.o.

tevékenykedő MI technológiákat alkalmazó cégek napjainkban már olyan létfontosságú infrastruktúrákat ellenőriznek, mint a szállítmányozás, a telekommunikáció vagy a közlekedés. Az e szektorokból történő pontozáson alapú negatív elbírálás vagy kizárás legalább annyira súlyos társadalmi-gazdasági következményekkel járhat az egyének számára, mint az állam által nyújtott szolgáltatásokból való kizárás.

A szabályozással kapcsolatban megemlítendő az Európai Gazdasági és Szociális Bizottság felvetése is, miszerint az i. és ii. alpontban foglalt tiltó feltételek megfogalmazása pontosításra szorulnak. A Bizottság javaslatában ugyanis nem könnyű egyértelmű határvonalat húzni a meghatározott célból történő, de megengedett pontozás, és az *általános közösségi pontozásnak* tekinthető esetek között, vagyis aközött, amikor az értékeléshez felhasznált információ még releváns és észszerűen kapcsolódik az értékelés céljához, és aközött, amikor a felhasznált információ már nem tekinthető az értékelés céljához észszerűen kapcsolódónak.[122]

Végezetül érdemes megemlíteni, hogy több jogvédő szervezet is kritikával illette a rendelet-tervezetet amiatt, hogy az kizárólag az *általános társadalmi pontozást,* illetve a személyek *megbízhatóságát* értékelő pontozási eljárásokat rendeli tiltani. Érvelésük szerint a pontozáson alapuló eljárásoknak több olyan – jelenleg nagy kockázatúként kategorizált – alkalmazási módja is létezik (pl.: szociális ellátáshoz való hozzáférés engedélyezése, jövőbeni bűnelkövetés/bűnismétlés valószínűségének előrejelzése), amely szintén jelentős veszélyeket rejt az állampolgárok alapvető szabadságjogaira nézve. Álláspontjuk szerint az állampolgárok hosszú távú érdekeire tekintettel a jövőben ezek tiltása is szükségessé válhat. [123]

4.5.3 Biometrikus azonosítási rendszerek valós idejű használata

[122] Európai Gazdasági és Szociális Bizottság (2021): Vélemény - Javaslat európai parlamenti és tanácsi rendeletre a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról INT/940 2021, 5.o.

[123] Access Now (2021): Access Now's submission to the European Commission's adoption consultation on the AI Act. 15.o. Elérhető: https://www.accessnow.org/wp-content/uploads/2021/08/Submission-to-the-European-Commissions-Consultation-on-the-Artificial-Intelligence-Act.pdf (2022. 05. 09.); Human Rights Watch (2021): How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net. 25.o. Elérhető: https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net (2022. 05. 10)

A rendelet-tervezet főszabályként tiltja a „*valós idejű távoli biometrikus azonosító rendszerek használatát a nyilvánosság számára hozzáférhető helyeken bűnüldözési célokból”*.[124]

Egyes kiemelt bűnüldözési célok tekintetében azonban megengedte e rendszerek használatát. Az ilyen helyzetek közé tartozott a bűncselekmények potenciális áldozatainak (pl.: eltűnt gyermekek) felkutatása; a természetes személyek életét vagy fizikai biztonságát fenyegető veszélyek vagy a terrortámadás veszélye; valamint az 2002/584/IB tanácsi kerethatározatban felsorolt és az érintett tagállam joga szerint *legalább három évig terjedő szabadságvesztéssel büntetendő bűncselekmények elkövetőjének vagy gyanúsítottjának felderítése, azonosítása és büntetőeljárás alá vonása*.[125]

A biometrikus azonosító rendszerek bűnüldöző hatóságok általi alkalmazásának létjogosultsága régóta viták tárgyát képezi. Ennek oka, hogy a nyilvánosság számára hozzáférhető helyeken történő távoli biometrikus azonosítás az egyének magánéletébe való betolakodás magas kockázatát hordozza magában. Ezek az eszközök azáltal, hogy képesek azonosítani, egyedileg megjelölni, valamint nyomon követni a kijelölt célszemélyt, komoly fenyegetést jelenthetnek az egyén alapvető szabadságjogaira, ideértve többek között a magánélethez és az adatvédelemhez való jogot, a gyülekezési szabadságot, valamint az egyenlőséghez és a megkülönböztetésmentességhez való jogot is.

A javaslat-tervezetben vázolt megközelítés az efféle gyakorlatoknak csak egy töredékét tiltja, hiszen a valós idejű távoli biometrikus azonosító rendszerek ezen technológiacsalád elenyésző részét képviselik. A javaslat a *nem valós idejű*[126] és a *megközelítőleg valós idejű* biometrikus azonosítást továbbra is megengedi a hatóságok számára, annak ellenére, hogy a gyakorlatok szabadságsértő jellege nem minden esetben függ attól, hogy az azonosításra valós időben kerül-e sor vagy sem. Könnyen belátható, hogy a bűnügyi hatóságok által véghez vitt nem valós idejű távoli biometrikus azonosítás – például egy tüntetéssel vagy kültéri politikai tiltakozással összefüggésben – megközelítőleg azonos mértékű visszatartó hatással bírhat.[127] Az állampolgárok jogai tekintetében szintén problémás lehet, hogy a tiltás hatálya jelen

[124] MI rendelet-tervezet 5. cikk (1) bekezdés d.
[125] MI rendelet-tervezet 5. cikk (1) bekezdés d i,ii,iii.
[126] A valós idejű azonosító rendszerek élő anyagot – például élő videofelvételt – használnak egy adott személy azonosítására. A nem valós idejű rendszerek esetében ezzel szemben a biometrikus adatokat már rögzítették, és az azonosításra csak később kerül sor. Ebben az esetben tehát az azonosításra alkalmas képek vagy videofelvételek azok konkrét használat megelőzően keletkeztek.
[127] Access Now (2021): Access Now’s submission to the European Commission’s adoption consultation on the AI Act. 16-17.o.

esetben kizárólag a bűnüldöző hatóságokra és a bűnüldözési célokra terjed ki, mely korlátozás lehetővé teszi e rendszerek más célokra és bármely más szereplő általi használatát. [128]

A rendelet-tervezet a biometrikus azonosító rendszerek használatához köthető gyakorlatok többségét *nagy kockázatúnak* minősítette. Ennek aránytalanságát hangsúlyozva az Európai Adatvédelmi Testület, az európai adatvédelmi biztos és az Európai Gazdasági és Szociális Bizottság is lényegesen szigorúbb szabályozást tartott indokoltnak. Véleményükben nyomatékosan szorgalmazták, hogy a nyilvánosság számára hozzáférhető helyeken általános jelleggel tiltsák be (1) az MI-rendszerek automatikus biometrikus felismerésre (pl.: arc, járás, hang, ujjlenyomat, DNS alapján) történő felhasználását (kivétel ha meghatározott körülmények között hitelesítési célt szolgál); (2) azon gyakorlatokat, melyek biometrikus adatokat felhasználó MI-rendszerek segítségével egyéneket csoportosítsanak etnikai hovatartozás, nem, politikai vagy szexuális irányultság szerint; (3) az olyan eljárásokat, melyekben az MI biometrikus adatok elemzésével egy természetes személy jövőbeli viselkedésének előrejelzésére vagy besorolására törekszik.[129]

4.6 Az Európai Unió Tanácsának közös álláspontja

Az Európai Unió Tanácsa 2022. december 6-án tette közzé az MI rendelet-tervezettel kapcsolatos közös álláspontját (általános megközelítés).[130] Több fontos módosítási javaslat is szerepelt a szövegben.

A Bizottság által bevezetett – és több helyütt is kritikával illetett – általános, kiterjesztő MI definícióhoz[131] képest, a Tanács az MI szűkebb definíciójának használatát javasolta, amely az

[128] A biometrikus rendszerek alkalmazása a magánszektor részéről (pl.: Clearview AI) szintén kockázatot jelent az egyénekre és a társadalomra nézve. Ráadásul bizonyos esetekben még annál is korlátozottabb jogorvoslati lehetőség áll a sértett(ek) rendelkezésre, mint amikor ezeket a technológiákat hatósági szervek használják.

[129] Európai Gazdasági és Szociális Bizottság (2021): Vélemény - Javaslat európai parlamenti és tanácsi rendeletre a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról INT/940, 5-6o., Európai Adatvédelmi Testület (2021): Az Európai Adatvédelmi Testület és az európai adatvédelmi biztos 5/2021. sz. közös véleménye. 13-16.o.

[130] Az Európai Unió Tanácsa (2022): Javaslat – Az Európai Parlament és a Tanács rendelete a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról – Általános megközelítés. 14954/22,

[131] Indoklása szerint az Európai Bizottság *technológiasemleges* és *időtálló* definíció megalkotására törekedett, mely kellően rugalmas ahhoz, hogy alkalmazkodjon az állandó műszaki fejlődéshez, ugyanakkor elég pontos

MI-re jellemző sajátosságok közül a gépi tanulásra, valamint a logikai- és tudásalapú fejlesztésekre fektette a hangsúlyt. Ennek célját abban jelölték meg, hogy egyértelműbben meg lehessen különböztetni az MI-t az egyszerűbb szoftverrendszerektől,[132] ily módon rögzítve az azok között fennálló különbségeket. A közös álláspont javaslata szerint az MI; „*olyan rendszer, amelyet úgy terveztek, hogy autonóm elemeket is alkalmazva működjön, és amely gépi és/vagy ember által szolgáltatott adatok és bemeneti adatok alapján – gépi tanulást és/vagy logikai és tudásalapú megközelítéseket alkalmazva – kikövetkezteti, hogyan érhető el a célkitűzések egy adott csoportja, és a rendszer által generált kimeneteket hoz létre, például olyan tartalmakat (generatív MI-rendszerek), előrejelzéseket, ajánlásokat vagy döntéseket, amelyek befolyásolják azokat a környezeteket, amelyekkel az MI-rendszer kölcsönhatásba lép”.*[133]

Lényeges változtatás, hogy a Tanács közös álláspontjában az általános társadalmi pontozás gyakorlatának tilalmát a közszektorról a magánszereplőkre is kiterjesztette, így a II. Cím 5. cikk (c) bekezdésében már nem szerepelt az MI-rendszerek *„hatóságok általi vagy nevükben történő”* korlátozás. Ennek értelmében a közigazgatási szerveken túl, már a magánvállalatok sem alkalmazhatnának olyan MI rendszereket, melyek állampolgárokat *értékelnek vagy osztályoznak közösségi magatartásuk, illetve ismert vagy előre jelzett személyes vagy személyiségi tulajdonságok* alapján, oly módon, hogy az így kapott *értékelés aránytalanul kedvezőtlen bánásmódhoz vezetne* a kínált szolgáltatás tekintetében. [134]

A Tanács ugyan nem fogadta el a Régiok Európai Bizottságának azon javaslatát, hogy *a szubliminális technikákat* alkalmazó MI gyakorlatokkal összefüggésben terjessze ki a tiltó feltételt a testi és lelki károsodáson túl a *pénzügyi veszteségek és gazdasági diszkrimináció* eseteire is, azonban a magatartást torzító gyakorlatok második nevesített formája esetében már javasolja a tiltás kiterjesztését az életkor, és a szellemi és testi fogyatékosságon túl azon

---

ahhoz, hogy a jövőben is garantálja a szükséges jogbiztonságot. A 2021 áprilisában közzétett rendelet-tervezet meghatározásában az MI olyan szoftver, „*amelyet meghatározott technikák és megközelítések alkalmazásával fejlesztettek, és amely az ember által meghatározott célkitűzések adott csoportja tekintetében olyan kimeneteket, például tartalmat, előrejelzéseket, ajánlásokat vagy döntéseket képes generálni, amelyek befolyásolják azt a környezetet, amellyel kölcsönhatásba lépnek”*. MI rendelet-tervezet 3. cikk (1) bekezdés

[132] Míg a hagyományos szoftverek általában előre meghatározott utasításokon alapulnak és konkrét feladatok elvégzésére vannak programozva, az MI rendszerek képesek tanulni, alkalmazkodni és döntéseket hozni az adatok elemzésével és azokból levont következtetésekkel. A Tanács érvelése alapján az MI és a hagyományos szoftverrendszerek közötti különbség megértése azért is lényeges, mert így nyílik lehetősége a szabályozóknak az MI specifikus kihívásosok és lehetőségek felismerésére.

[133] Általános megközelítés I. Cím 3. cikk (1) bekezdés

[134] Például, ha egy bank vagy biztosító társaság által alkalmazott értékelési rendszer (amelynek egyik célja, hogy a pontszámok alapján döntsön az ügyfél biztosítási díjáról), olyan általános kockázati profilt hoz létre, amely ellehetetleníti az értékelt személy hozzáférését bármely pénzügyi szolgáltatáshoz.

személyekre is, akiknek a sebezhetősége *a szociális vagy gazdasági helyzetükből* fakad. Ennek megfelelően a II. Cím 5. cikkének (b) bekezdése tiltja az olyan MI rendszerek forgalomba hozatalát, üzembe helyezését vagy használatát, amelyek „*a személyek egy meghatározott csoportjának életkor, fogyatékosság, illetve egyedi szociális vagy gazdasági helyzet miatt fennálló valamilyen sebezhetőségét kihasználják”*.[135]

A Tanács a magatartást torzító gyakorlatokkal kapcsolatban az Európai Gazdasági és Szociális Bizottság és a Régiók Európai Bizottságának azon felvetését sem tartotta megalapozottnak, miszerint testi és lelki károsodást okozó gyakorlatokon túl, szükséges volna azon magatartástorzító technikák tiltása is, melyek hosszú távon alááshatják a *demokráciát és a jogállamiságot.*

A Tanács szintén nem látta indokoltnak a biometrikus azonosítási rendszereket érintő lényegesen szigorúbb szabályozás megalkotását. E rendszerekkel összefüggésben az általános megközelítés két technikai változtatást eszközölt: (1) pontosította azokat a célkitűzéseket, amelyek esetében a valós idejű és távoli azonosítás bűnüldözési célokból szigorúan szükséges, valamint; (2) pontosította a *távoli biometrikus azonosító rendszer* és a *valós idejű távoli biometrikus azonosító rendszer* fogalmait annak tisztázása érdekében, hogy mely helyzetek tartoznak a kapcsolódó tilalom és nagy kockázatú felhasználás körébe, és melyek nem.

A későbbiekben tárgyalandó GenMI modellekhez kötődően fontos kiemelni, hogy a Tanács az átláthatósági követelményeket rögzítő 52. cikk (3) bekezdését is módosította, melynek értelmében az MI által generált (audio- vagy videótartalom) kimenetet nem kell mesterségesen előállítottnak címkézni „*ha a tartalom nyilvánvalóan kreatív, szatirikus, művészi vagy fikciós mű vagy program része*”.[136] Ez a (dolgozat témáját illetően is) lényeges változás – ahogy arra Mezei utalt egy 2023-as tanulmányában – már annak okán is

[135] Általános megközelítés II. Cím 5. cikk (b)

[136] Általános megközelítés 52. cikk (3) bekezdés: „*Azon MI-rendszerek felhasználói, amelyek olyan, meglévő személyekre, tárgyakra, helyekre vagy más szervezetekre vagy eseményekre érzékelhetően hasonlító képet, audio- vagy videotartalmat generálnak vagy manipulálnak, amely egy személy számára megtévesztő módon eredetinek vagy valóságosnak tűnhet („deepfake”), közlik, hogy a tartalmat mesterségesen hozták létre vagy manipulálták. Az első albekezdés azonban nem alkalmazandó abban az esetben, ha a felhasználást törvény engedélyezi bűncselekmények felderítése, megelőzése, kivizsgálása és eljárás indításának céljából, vagy ha a tartalom nyilvánvalóan kreatív, szatirikus, művészeti vagy fikciós mű vagy program részét képezi, a harmadik felek jogaira és szabadságaira vonatkozó megfelelő biztosítékok mellett.”*

figyelemreméltó, hogy az MI által létrehozott alkotások a jelenlegi európai szerzői jogi rendszerben nem élvezhetnek ilyen védelmet. [137]

4.6.1 Új kategória: Általános célú MI

A Tanács által javasolt szöveg legfontosabb változtatásának az tekinthető, hogy bevezetett egy teljesen új MI kategóriát is, *általános célú MI* néven. A Tanács által bevezetett új kategóriát többek között a GenMI modellek rohamos fejlődése és tömeges hozzáférhetővé válása indokolta. Ezek a modellek eredeti tartalom létrehozására képesek, illetve számos alkalmazási lehetőséggel rendelkeznek, fontos jellemzőjük a gépi tanulás; a betáplált nagymennyiségű adat feldolgozásával megtanulják értelmezni az azokban rejlő mintákat, kapcsolatokat és struktúrákat, majd a tanulási folyamat végeztével új (a betáplált adatokban nem fellelhető) tartalom létrehozására válnak alkalmassá. Mivel az általános célú MI-k számos területen alkalmazhatóak, valamint sokkal gyorsabban válhatnak más MI-rendszerek kiegészítő részévé, mint ahogyan a jogalkotási folyamatok erre reagálni tudnának, a Tanács által elfogadott közös álláspontban egy olyan definíció megalkotására törekedtek, amely a lehető legszélesebb körben határozza meg az ilyen technológiákat. Ennek értelmében a közös álláspont az általános célú MI-t olyan rendszerként definiálta „*amelyet – függetlenül attól, hogy miként hozták forgalomba vagy helyezték üzembe, ideértve a nyílt forráskódú szoftvert is – a szolgáltató általánosan alkalmazható funkciók, például kép- vagy beszédfelismerés, hang- és videóelőállítás, mintaészlelés, kérdésmegválaszolás, fordítás és egyéb funkciók ellátására szán; az általános célú MI-rendszerek számos összefüggésben használhatók és számos egyéb MI-rendszerbe integrálhatóak.”*[138]

Amennyiben az ilyen MI-rendszereket nagy-kockázatú MI-ként, vagy azok alkotórészeként használják, meg kell felelniük a rendelet III. címének 2. fejezetében meghatározott nagy kockázatú MI-kel kapcsolatos követelményeknek. A közös álláspont 4b. cikkének (1) bekezdése szerint azonban ezen követelmények nem alkalmazhatóak közvetlenül az általános célú MI-rendszerekre. Ehelyett egy részletes hatásvizsgálaton alapuló végrehajtási jogi aktus

[137] Mezei Kitti (2023): A mesterséges intelligencia jogi szabályozásának aktuális kérdései az Európai Unióban. *In Medias Res.* 2023/1. 4, 60.o.
[138] Általános megközelítés I. Cím 3. cikk (1b)

határozná meg az alkalmazásukat, „*mégpedig jellemzőik, műszaki megvalósíthatóságuk, az MI-értéklánc sajátosságai, valamint a piaci és technológiai fejlődés fényében”.*[139]

4.7 A Parlament kompromisszumos javaslata

2023. május 11-én az Európai Parlament Belső Piaci és Fogyasztóvédelmi Bizottsága (IMCO) és az Állampolgári Jogi, Bel- és Igazságügyi Bizottsága (LIBE) is jóváhagyta a Parlament által benyújtott kompromisszumos szövegjavaslatot, amelyet aztán a Parlament 2023. június 14-én fogadott el hivatalosan.

A tiltott gyakorlatok szabályozására vonatkozóan több módosítási javaslat is említésre méltó. A magatartástorzító gyakorlatokkal összefüggésben a Parlament kiterjesztett a szubliminális technikák alkalmazására irányuló tiltást a döntéshozatali képesség gyengítésére alkalmas, célzottan manipulatív vagy megtévesztő technikák használatára is.[140] A kompromisszumos szöveg szigorított a biometrikus azonosító szoftverek alkalmazásával kapcsolatosan és kiterjesztette a tilalmat az ilyen azonosító rendszerek ex-post (tehát utólagos) alkalmazására is.[141] Emellett az elfogadhatatlan kockázatú alkalmazásokat tartalmazó 5. cikk új elemmel is bővült a prediktív rendészet keretében történő lehetséges jövőbeni bűnelkövetés és bűnismétlés kockázatának értékelésére vonatkozó gyakorlatok tiltásával.[142] Szintén új elem, hogy a szöveg tiltani rendelte a természetes személyek érzelmeiből következtetést levonó MI-rendszerek[143] forgalomba hozatalát, üzembe helyezését vagy használatát *a bűnüldözés, a*

[139] Általános Megközelítés IA. Cím 4b. cikk (1) bekezdés

[140] A kompromisszumos javaslat 5. cikk (1) bekezdés a pontja, tiltani rendeli az olyan MI-t, amelyek ...*célzottan manipulatív vagy megtévesztő technikákat alkalmaznak azzal a céllal vagy olyan hatás érdekében, hogy lényegesen torzítsák egy személy vagy személyek egy csoportjának magatartását azáltal, hogy jelentősen gyengítik az adott személy megalapozott döntéshozatlra való képességét, és ennek következtében a személy olyan döntést hozzon, amelyet egyébként nem hozott volna meg, oly módon, amely e személynek, egy másik személynek vagy személyek egy csoportjának jelentős károsodást okoz vagy okozhat;*

[141] Kompromisszumos javaslat 5. cikk (1) bekezdés dd pont

[142] A kompromisszumos javaslat alapján nem engedélyezett az olyan MI alapú kockázatelemző rendszer alkalmazása, mely annak érdekében végez kockázatelemzést, hogy (1) felmérje *egy természetes személy milyen kockázatot jelenthet egy bűncselekmény vagy szabálysértés elkövetése vagy újbóli elkövetése szempontjából*, illetve, hogy (2) *profilalkotás, személyiségjegyek, tulajdonságok, vagy múltbeli bűnöző magatartás értékelése* alapján előre jelezze a tényleges vagy potenciális bűncselekmények előfordulását vagy megismétlődését. (5. cikk (1) bekezdés da pont)

[143] Az érzelemfelismerő MI-rendszer az *egyének vagy csoportok biometrikus és biometrikus alapú adatai alapján azonosítja vagy levezeti a természetes személyek érzelmeit, gondolatait, lelkiállapotát vagy szándékait* (3. cikk (1) bekezdés 34. pont)

határigazgatás, a munkahelyek és az oktatási intézmények területén.[144] Emellett tiltotta az olyan MI rendszerek használatát is, melyek *természetes személyeket érzékeny vagy védett tulajdonságok vagy jellemzők szerint, vagy ilyen tulajdonságokból vagy jellemzőkből levont következtetések alapján kategorizálnak* (biometrikus kategorizálási rendszer).[145] A javaslat szintén elfogadhatatlan kockázatúként kategorizálta az internetről vagy CCTV felvételekről származó képek lekérdezésével kiterjedt arcfelismerő adatbázisokat létrehozó rendszereket is.[146]

A nagy kockázatú MI-k tekintetében lényeges változás, hogy míg a Bizottság és a Tanács korábbi javaslatainak III. mellékletében felsorolt területeken alkalmazott modellek automatikusan magas kockázatúnak lettek minősítve, addig a Parlament javaslatában egy további szint került bevezetésre a nagy kockázatú MI kategória feltételeként. Ennek értelmében a mellékletben felsorolt MI modellek csak abban az esetben minősültek nagy kockázatúnak, *ha az egészségre, biztonságra vagy alapvető jogokra* nézve jelentős veszélyt jelentenek. [147]

Emellett bővítésre került azon alkalmazási területek köre is, melyekben az MI nagy kockázatúnak minősülhet.[148] Figyelembe véve az Európai Gazdasági és Szociális Bizottság 2021-ben közzétett véleményét, melyben több ízben is hangot adtak az MI rendszerekhez köthető *általánosabb/össztársadalmi szinten értelmezhető veszélyforrások*nak (*szavazási*

[144] Tiltott *a természetes személyek érzelmeiből következtetést levonó MI-rendszerek forgalomba hozatala, üzembe helyezése vagy használata a bűnüldözés, a határigazgatás, a munkahelyek és az oktatási intézmények területén*. 5. cikk (1) bek. (dc) Az érzelemfelismerő rendszereket illetően bár üdvözölendő korlátozásokat fogalmazott meg a javaslat, több magánjogi és emberi jogi szervezet, mint például az Európai Digitális Jogok (European Digital Rights) és az Access Now, teljes körű tilalmat szorgalmaztak. Ezt többek között azzal indokolják, hogy az ilyen rendszerek alkalmazása sokszor pontatlan lehet, mivel az érzelmek megítélése rendkívül összetett és kultúráktól függő folyamat, ami jelentős mértékben eltérhet a gépek által "látott" és értékelt viselkedésformákhoz képest. Emellett az érzelmi felismerésre épülő technológiák alkalmazása adatvédelmi és diszkriminációs kockázatokat is rejt magában, amik sérthetik az egyének személyiségi jogait és szabadságait. A pontatlanságról bővebben Ryan-Mosley, Tate (2023): AI isn't great at decoding human emotions. So wh are regulators targeting the tech? MIT Technology Review. Elérhető: https://www.technologyreview.com/2023/08/14/1077788/ai-decoding-human-emotions-target-for-regulators/ (2023. 09. 04.)

[145] Kompromisszumos javaslat 5.cikk (1) bekezdés ba pont.

[146] Kompromisszumos javaslat 5 cikk (1) bekezdés db pont „*olyan MI-rendszerek. amelyek az arcképek internetről vagy zártláncú televízió felvételekből való, nem célzott lekérdezésével arcfelismerő adatbázisokat hoznak létre vagy ilyeneket bővítenek”*.

[147] Kompromisszumos javaslat 6. cikk (1-2) bekezdés; 32 preambulum.

[148] Például több helyütt is módosult a III melléklet (1) bekezdés 8 pontja, melyben eredetileg az igazságügyi hatóságok vagy nevükben eljáró fél által használt MI rendszerek szerepeltek *a tények és a jog kutatásában és értelmezésében, valamint a jog konkrét tényállásra történő alkalmazásában*. Az új szövegben az említett célok tekintetében már a *közigazgatási szervek* vagy az ő nevükben eljáró fél is említésre került. Ez a gyakorlatban azt jelentheti, hogy a bűncselekményi tényállások megállapításán túl, már bizonyos szabálysértések körülményeinek MI rendszerekkel való felderítése esetén is – főszabályként – nagy kockázatúként kell kategorizálni az alkalmazott MI-t.

magatartás megváltozásával járó demokratikus vagy *jogállami hatások*)[149], a kompromisszumos szöveg egy teljesen új nagy kockázatú MI alkalmazási területet is bevezet. Amennyiben teljesül a már említett plusz feltétel – tehát hogy az alkalmazás az egészségre, biztonságra vagy alapvető jogokra nézve jelentős veszélyt jelent – a javaslat szerint automatikusan nagy kockázatúnak kell tekinteni azon MI-rendszereket, *amelyeket választások vagy népszavazások eredményének vagy természetes személyek választások vagy népszavazások alkalmával tanúsított szavazási magatartásának befolyásolására szánnak*.[150]

A kompromisszumos javaslat új kötelezettségeket is előírt a nagy kockázatú rendszerekhez kötődően. A nagy kockázatú MI-rendszerek *alkalmazó*i kritikus szerepet játszanak az alapvető jogok védelmében, hiszen helyzetüknél fogva az alkalmazók képesek a legjobban megérteni, hogy konkrétan hogyan használják majd a nagy kockázatú MI rendszert, és ők tudják azonosítani a fejlesztési szakaszban előre nem látott potenciális jelentős kockázatokat. Az alapvető jogok védelmének hatékony biztosítása érdekében a nagy kockázatú MI rendszerek alkalmazójának *alapjogi hatásvizsgálatot* (Fundamental Rights Impact Assessment, FRIA) kell végeznie az alkalmazás megkezdése előtt.[151] A hatásvizsgálatot részletes tervnek kell kísérnie, amely ismerteti az alapvető jogokat érintő, legkésőbb az alkalmazás megkezdésének időpontjától kezdve azonosított kockázatok mérséklését elősegítő intézkedéseket vagy eszközöket.

A rendelet-tervezet nem tartalmazott megfelelő intézkedéseket azoknak az egyéneknek a védelmére, akik valószínűleg a leginkább károsan érintettek az MI rendszerek használata által, valamint azon közérdekű szervezetek számára, akiknek fontos szerepe lehet ezeknek az egyéneknek a képviseletében. Az állampolgárok jogvédelmét erősítendő a Parlamenti javaslat

---

[149] Európai Gazdasági és Szociális Bizottság (2021): Vélemény – Javaslat európai parlamenti és tanácsi rendeletre a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról INT/940, 4-5.o.

[150] *Nem tartoznak ide az olyan MI rendszerek, amelyek kimenetének a természetes személyek nincsenek közvetlenül kitéve, mint például a politikai kampányok adminisztratív és logisztikai szempontból való megszervezéséhez*. III melléklet (1) bekezdés 8aa

[151] A jogszabályszöveghez újonnan csatolt 29a cikk a-j értelmében a nagy kockázatú MI-rendszerek üzembe helyezése előtt az alkalmazók kötelezően értékelik a rendszerek hatását a felhasználási környezetben. Az értékelésnek tartalmaznia kell a következő elemeket: „*rendszer rendeltetésének egyértelmű meghatározása, rendszer felhasználása tervezett földrajzi és időbeli hatókörének meghatározása, rendszer használata által valószínűleg érintett természetes személyek és csoportok, annak ellenőrzése, hogy a rendszer használata megfelel-e az alapvető jogokra vonatkozó releváns uniós és nemzeti jognak, az alapvető jogokra gyakorolt, észszerűen előre látható hatás, a marginalizált személyeket vagy kiszolgáltatott csoportokat valószínűleg érintő bármilyen konkrét kár veszélye, a rendszer használatának észszerűen előrelátható kedvezőtlen hatása a környezetre, részletes terv arról, hogy a kárt és az alapvető jogokra gyakorolt negatív hatást hogyan fogják enyhíteni, az alkalmazó által létrehozott irányítási rendszer, beleértve az emberi felügyeletet, a panaszkezelést és a jogorvoslatot is”.*

kiegészíti a Bizottság kezdeti rendelettervezetét azzal, hogy bevezeti a *panasztételi jogot*, amely lehetővé teszi az egyéneknek és csoportoknak, hogy panaszt nyújtsanak be a nemzeti felügyeleti hatóságokhoz, ha úgy érzik, hogy az MI rendszerek döntései megsértették jogaikat.[152] A javaslat értelmében a nemzeti felügyeleti hatóságok kötelesek tájékoztatni a panaszosokat a felülvizsgálati folyamat során és értesíteni őket a döntés kimeneteléről, beleértve azt is, hogy áll-e rendelkezésre bírósági jogorvoslat.[153] Emellett a törvényjavaslat lehetővé teszi az állampolgárok számára, hogy *magyarázatot* kapjanak az MI által hozott döntésekről, ezáltal növelve az átláthatóságot és elősegítve az érintettek tájékoztatását.[154]

A Parlament javaslata az Unió belső piaci széttagoltságánáak elkerülése, az új rendelet hatékony és összehangolt végrehajtásának elősegítése, és nem utolsó célból a nemzeti felügyeleti hatóságok, valamint az uniós intézmények, szervek, hivatalok és ügynökségek aktív támogatása céljából egy Európai Unió Mesterséges Intelligenciával Foglalkozó Hivatal (MI-Hivatal) létrehozását is kezdeményezte.[155]

4.7.1 Széles körű alkalmazás, általános célú felhasználás

A Parlament által közzétett kompromisszumos szöveg különös hangsúlyt fektetett az általános célú MI-k szabályozási kérdéseire, szigorú követelményeket előírva részükre többek között a használati utasítások részletes dokumentálása, az adatforrások megfelelőségének biztosítása, és a környezeti hatások figyelembevétele terén.

[152] Kompromisszumos javaslat 68a. cikk: „*a természetes személy vagy természetes személyek minden csoportja jogosult arra, hogy panaszt tegyen egy nemzeti felügyeleti hatóságnál – ha megítélése szerint a rá vonatkozó MI rendszer megsérti e rendeletet*".
[153] Kompromisszumos javaslat 68b cikk: minden természetes vagy jogi személy hatékony bírósági jogorvoslatra jogosult a nemzeti felügyeleti hatóság rá vonatkozó, jogilag kötelező erejű döntésével szemben
[154] A Kompromisszumos javaslat 68c. cikk (Az egyéni döntéshozatal magyarázatához való jog) kimondja, hogy „*az alkalmazó által nagy kockázatú MI-rendszer kimenete alapján hozott döntés által érintett személy, amely döntés olyan joghatásokat vagy őt hasonlóan jelentősen érintő hatásokat vált ki, amelyek megítélése szerint hátrányosan befolyásolják egészségét, biztonságát, alapvető jogait, társadalmi-gazdasági jólétét vagy az e rendeletben meghatározott kötelezettségekből eredő bármely más jogát a 13. cikk (1) bekezdése szerinti egyértelmű és érdemi magyarázatot kérhet az alkalmazótól az MI-rendszer döntéshozatali eljárásban betöltött szerepéről, a meghozott döntés fő paramétereiről és a kapcsolódó bemeneti adatokról*."
[155] 76 preambulumkezdés

A Parlament által alkotott definíció szerint az általános célú MI olyan rendszer, "*amely széles körben használható olyan alkalmazásokra vagy igazítható olyan alkalmazásokhoz, amelyekre nem kifejezetten és szándékosan tervezték*".[156] A fogalomalkotás során – hasonlóan a Tanácshoz – a Parlament is fontosnak tartotta kiemelni, hogy ezeknek a rendszereknek a felhasználási területei az idő előrehaladtával jóval szélesebbé és változatosabbá válhatnak, mint azt kezdetben tervezték. Ez a jellegzetesség a szabályozás tekintetében azért is érdemel körültekintést, mert a modellek könnyű adaptálhatósága előre nem tervezett környezetekhez és funkciókhoz jelentősen megnehezíti azok jövőbeli potenciális kockázatainak pontos előrejelzését.[157]

Tekintettel a GenMI modellek növekvő népszerűségére, a Parlament szükségesnek látta, hogy az általános célú MI kategórián túl bevezessen egy teljesen új modelltípust is „*alapmodell*" (foundation model) néven. Az ilyen MI-t a kompromisszumos szöveg úgy határozta meg, mint az *„MI-rendszer olyan modellje, amelyet széles körű adatokon tanítottak, általános jellegű kimenetre terveztek, és amely széles körben különféle feladatokhoz igazítható*".[158]

4.7.2 Az alapmodellek szolgáltatóinak kötelezettségei és a ChatGPT szabály

Az alapmodellek szolgáltatói számára az új 28b cikkben kerültek meghatározásra azon követelmények, melyeknek az MI forgalmazását vagy üzembe helyezését megelőzően kell megfelelniük.[159] Ennek értelmében a szolgáltató köteles igazolni, hogy megfelelő *tervezéssel, teszteléssel és elemzéssel az egészséget, a biztonságot, az alapvető jogokat, a környezetet, a demokráciát és a jogállamiságot érintő, észszerűen előrelátható kockázatok azonosítása, csökkentése és mérséklése* a fejlesztést megelőzően és annak során megfelelő módszerekkel

[156] 3 cikk (1) bekezdés 1 d pont
[157] Részben ehhez kötődik, hogy a Parlament javaslatának a 28 cikk (1) bekezdés, ba pontja kiterjeszti a nagy kockázatú rendszerek szolgáltatókra vonatkozó kötelezettségeket az MI értéklánc azon szereplőire is, akik jelentős módosítást hajtanak végre egy olyan forgalomba hozott, vagy üzembe helyezett MI rendszeren, amely eredetileg nem számított nagy kockázatúnak. Ez a gyakorlatban azt jelenti, hogy ha egy eredetileg nem nagy kockázatú általános célú MI-rendszeren olyan módosításokat hajtanak végre, amelyek következtében az immár nagy kockázatúvá minősül – ami az általános célú MI-k változatos felhasználási köréből adódóan gyakran előfordulhat - akkor a lényeges átalakítást végző fél számít majd az új szolgáltatónak, és rá vonatkoznak a nagy kockázatú rendszerekre előírt szigorúbb szabályok és kötelezettségek.
[158] 3 cikk (1) bekezdés 1 c pont
[159] 28b cikk (1) bekezdés

(például független szakértők bevonásával) megtörtént.[160] Ezen túl a javaslat hosszasan sorolja a további kötelezettségeket *a minőségi adatkészletek biztosításával* (28. cikk 2b), a megfelelő szintű *teljesítmény, kiszámíthatóság, értelmezhetőség, korrigálhatóság*, *biztonság* és *kiberbiztonság* elérésével (28. cikk 2c), az erőforrásfelhasználás és hulladékcsökkentéssel (28. cikk 2d), az *átfogó műszaki dokumentációt* és *érthető használati utasítással készítésével* (28. cikk 2e), valamint a *minőségirányítási rendszer* létrehozásával kapcsolatban (28. cikk 2f). Ezeket a követelményeket a modell teljes életciklusán keresztül független szakértők által kell tesztelni, dokumentálni és ellenőrizni.[161]

Az alapmodellekre vonatkozó általános kötelezettségeken túl, a javaslat az átláthatósággal, a tanítóadatokkal és a szerzői joggal összefüggésben három további kötelezettséget állapít meg a GenMI[162] rendszerekben használt vagy GenMI rendszerekre specializált alapmodellek szolgáltatói számára (ChatGPT szabály).[163] A 28b cikk (4) bekezdése kimondja, hogy a szolgáltatók:

„a) eleget tesznek az 52. cikk (1) bekezdésében felvázolt átláthatósági kötelezettségeknek;[164]

b) úgy tanítják be, és adott esetben úgy alakítják ki és fejlesztik ki az alapmodellt, hogy a technika általánosan elismert állásának megfelelően és az alapvető jogok, köztük a véleménynyilvánítás szabadságának sérelme nélkül megfelelő biztosítékokat nyújtsanak az uniós jogot sértő tartalmak generálásával szemben;

[160] Illetve a *fejlesztés után még fennálló, nem mérsékelhető kockázatok dokumentálására* is sor került. (28b cikk (2) bekezdés)
[161] 28b cikk (2) bekezdés c
[162] A 28b cikk (4) egyúttal definícióval is szolgál a GenMI-re. Ez alapján a GenMI olyan MI rendszereket jelöl, *amelyek célja, hogy különböző szintű autonómiával generáljanak olyan tartalmakat, mint például összetett szöveg, kép, hang vagy videó („generatív MI”).*
[163]Hacker, Philipp, Andreas Engel, Marco Mauer (2023): Regulating ChatGPT and Other Large Generative AI Models. In Proceedings of the Fairness, Accountability, and Transparency (FAccT ’23). ACM, New York. 1115.o. https://doi.org/10.1145/3593013.3594067.
[164] Az 52 cikk (1) bekezdés már említésre került az előző fejezetben is a Bizottsági rendelet-tervezet ismertetésénél. Az átláthatósági klauzula így szól: „*A szolgáltatók biztosítják, hogy a természetes személyekkel való közvetlen interakcióra szánt MI-rendszereket úgy tervezzék meg és fejlesszék ki, hogy az MI-rendszer, maga a szolgáltató vagy a felhasználó kellő időben, egyértelmű és érthető módon tájékoztassa az MI rendszernek kitett természetes személyt arról, hogy MI-rendszerrel áll kapcsolatban, kivéve, ha ez a körülmények és a használat kontextusa alapján nyilvánvaló. Adott esetben e tájékoztatásnak ki kell terjednie arra is, hogy mely funkciók működnek MI-vel, hogy van-e, emberi felügyelet, és hogy ki felelős a döntéshozatali folyamatban, valamint azokra a meglévő jogokra és eljárásokra, amelyek az uniós és a nemzeti jog szerint lehetővé teszik természetes személyek vagy képviselőik számára, hogy tiltakozzanak az ilyen rendszerek rájuk vonatkozó alkalmazása ellen, és bírósági jogorvoslatot kérjenek az MI-rendszerek által hozott döntések vagy az általuk okozott károk ellen, beleértve a magyarázatkéréshez való jogukat is.”*

c) a szerzői jogra vonatkozó uniós, nemzeti vagy uniós jogszabályok sérelme nélkül dokumentálják és nyilvánosan hozzáférhetővé teszik a szerzői jog alapján védett tanulóadatok felhasználásának kellően részletes összefoglalóját”[165]

Azt követően, hogy 2022 decemberében a Tanács által közzétett közös álláspont, majd fél évvel később, 2023 júniusában a Parlament által jegyzett kompromisszumos szövegjavaslat is elfogadásra került, az EU jogszabályalkotási rendje szerint 2023 nyarán az eljárás a háromszervi egyeztetés szakaszába (ún. trilógus tárgyalásba, vagy háromoldalú tárgyalásba) léphetett. A háromszervi egyeztetés során a Tanács és a Parlament álláspontjait kellet összehangolni, mely folyamatban a Bizottság egyfajta mediátorként léphetett fel az intézmények között.

4.8 Háromoldalú tárgyalások

2023 teléig összesen öt háromoldalú tárgyalásra került sor, az ötödik és egyben utolsó tárgyalás december 6. és 8. között folyt le, melyben a Tanács és az Európai Parlament – egy „maratoni” több mint 12 órás tárgyalást követően – ideiglenes megállapodásra jutott minden szabályalkotási (és politikai) kérdésben, sikeresen lezárva az intézményközi tárgyalásokat. A jogszabályalkotási folyamat utolsó szakaszát végig kísérték az európai jogalkotók és az EU kormányai – kiváltképp Franciaország, Olaszország és Németország – között meghúzódó véleménykülönbségek és feszültségek. Ezek elsősorban abból a félelemből fakadtak, hogy a túlzottan szigorú szabályozás visszavetné az Unió esélyeit a technológiai innovációkra és fejlesztésekre kiélezett globális MI versenyben.[166] Bár a javaslat több lényeges eleme körül is alakult ki vita (arcfelismerő rendszerek tiltása, MI-Hivatal feladat- és jogkörei) különösen megosztónak a javaslat alapmodellekkel kapcsolatos szövegezése bizonyult. A nagy teljesítményű alapmodellekre vonatkozó követelmények tekintetében az uniós döntéshozók sokáig nem jutottak közös nevezőre. Az őszi háromoldalú tárgyalások alatt egyetértés mutatkozott abban, hogy az alapmodellekre vonatkozó szabályokat többszintű

[165] 28b cikk (4) bekezdés

[166] Bertuzzi, Luca (2023): EU’s AI Act negotiations hit the brakes over foundation models. EURAVTIV. Elérhető: https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/ (2024. 01. 24.)

megközelítéssel (layered approach) vezetik be, azaz szigorúbb szabályokat kell alkalmazni a legerősebb, a társadalomra nagyobb hatást gyakorló modellekre.

Azonban a Miniszterek Tanácsának ülésén több tagállam képviselői – az előbb említett Franciaország, Olaszország és Németország vezetésével – szembe helyezkedtek az alapmodellekre vonatkozó szabályozási elképzeléssel.[167] Erre válaszul az Európai Parlament több tisztviselője is kisétált egy későbbi ülésről, világossá téve, hogy az alapmodellek szigorúbb szabályozásának mellőzése politikailag nem elfogadható alternatíva.[168] Ha decemberben nem született volna megállapodás, az alapmodellekkel kapcsolatos megközelítés általános újragondolásának igénye szinte lehetetlenné tette volna, hogy az EU választások előtt elfogadhassák a jogszabályt, mely könnyen ahhoz vezethetett volna, hogy az EU elveszíti előnyét más szabályozó törekvésekkel és szabályozni kívánó országokkal szemben. Végül az akkor soros spanyol elnökség december 11-én bejelentette a kompromisszum elérését.

A legnagyobb változás, hogy a Tanács általános megközelítésében és a Parlament kompromisszumos javaslatában vázolt egységes szabályozás helyett a decemberi megállapodás két különböző típusú általános célú MI rendszert határozott meg, amelyekre eltérő szabályok és jogi kötelezettségek vonatkoztak. Az első kategóriát a normál alapmodellek alkották, melyek alacsonyabb és kevésbé kiterjedt kockázatot hordoznak, míg a második kategóriát az un. *rendszerszintű kockázatot jelentő alapmodellek* alkották.

Az új kétszintű szabályozási keretrendszerben az első szint minden általános célú MI-re vonatkozik.[169] Ezen a szinten a szabályok olyan követelményeket írnak elő, mint a betanítási módszerekkel kapcsolatos információk nyilvánosságra hozatala, az energiafelhasználás transzparens bemutatása, valamint a szerzői jogi törvények betartásának biztosítása.[170] A második szint kizárólag az un. *nagy hatású rendszerszintű kockázatot jelentő alapmodellekre*

[167] Európa három legnagyobb gazdasága azért is lobbizott, hogy az alapmodellek terén, a kötelezettségek fókuszát önkéntesen vállalt magatartási kódexekre szűkítsék, (többek között mert el akarták kerülni az európai technológiai innováció és kiváltképp az olyan ígéretes európai startupok, mint a Mistral AI és az Aleph Alpha korlátozását).

[168] Bertuzzi, Luca (2023): France, Germany, Italy push for 'mandatory self-regulation' for foundation models in EU's AI law. EURACTIV. Elérhető: https://www.euractiv.com/section/artificial-intelligence/news/france-germany-italy-push-for-mandatory-self-regulation-for-foundation-models-in-eus-ai-law (2024. 01. 20.)

[169] Ezek alól csak azok a rendszerek képeznek kivételt, amelyek kizárólag kutatási célokra szolgálnak, vagy nyílt forráskódú licenc alatt kerülnek kiadásra.

[170] Az alapmodell fejlesztőinek kötelező nyilvánosságra hozni azon szerzői jog által védett (főként szöveges) adattartalmakat, amelyeket az MI modellek tanításához használtak. A tanítóadatok és a szerzői jog kérdéskörei kitüntetett figyelmet érdemelnek a GenMI rendszerek kihívásai területén, így azok a dolgozat későbbi részében részletesen is kifejtésre kerülnek.

vonatkozik. Az ilyen modellek szolgáltatóira szigorúbb kötelezettségek vonatkoztak, amelyek magukban foglalják a modellek értékelését, a rendszerszintű kockázatok felmérését és mérséklését, az ilyen kockázatokról szóló jelentéstételt a Bizottságnak, támadhatósági tesztek végrehajtását, valamint a magas szintű kiberbiztonsági védelem biztosítását.

Mivel a decemberi megállapodás a belga soros elnökség által közzétett 2024. január 26-ai módosítások és az Európai Parlament által március 13-án elfogadott végleges szöveg között néhány hónap telt csak el, és a jogszabályi szövegek között inkább nyelvtani és csak kisebb mértéken előforduló technikai különbségek vannak, az érthetőség és egyszerűbb követhetőség kedvéért kizárólag a végső szöveg rövid bemutatására és néhány következtetés levonására összpontosít a fejezet befejező része.

4.9 Az Európai Unió mesterséges intelligenciáról szóló rendelete

4.9.1 Mi az MI?

Az MI végleges definícióija kisebb változtatásokon átesett a Parlament 2023 nyári kompromisszumos javaslata óta, de lényegében megmaradt az OECD által is ajánlott értelmezési paradigmában, és céljaként azt tűzte ki, hogy megkülönböztesse az MI rendszereket az olyan hagyományos szoftverrendszerektől, amelyek tipikusan előre meghatározott szabályok alapján működnek és nem képesek önálló tanulásra vagy adaptációra. Az Uniós jogalkotó szervek emellett az is hangsúlyozták, hogy a megalkotott fogalmat szorosan össze kell hangolni a MI-vel foglalkozó nemzetközi szervezetek munkájával a *jogbiztonság szavatolása*, valamint a *nemzetközi konvergencia* és a *széles körű elfogadottság előmozdítása* érdekében.[171] E szempontokat figyelembevéve a jogszabály végleges szövege úgy határozza meg az MI-t, mint olyan*: „gépi alapú rendszer, amelyet különböző autonómiaszinteken történő működésre terveztek, és amely a bevezetését követően alkalmazkodóképességet tanúsíthat, és amely a kapott bemenetből – explicit vagy implicit célok érdekében – kikövetkezteti, miként generáljon olyan kimeneteket, mint például*

[171] MI-rendelet 12. preambulum

előrejelzéseket, tartalmakat, ajánlásokat vagy döntéseket, amelyek befolyásolhatják a fizikai vagy a virtuális környezetet."[172]

4.9.2 Kockázatos gyakorlatok, általános célú felhasználás

A továbbiakban kettő kockázati kategória rövid ismertetésére kerül sor: (1) az elfogadhatatlan kockázatot jelentő – és ennek okán tiltott – gyakorlatok (2) a nagy kockázatú MI-k (melyekkel kapcsolatos szabályok a törvényszöveg körülbelül kettőharmadát fedik le).

(1) Tiltott MI-gyakorlatok

Az Európai Bizottság által 2021. áprilisában közzétett rendelet-tervezet négy olyan elfogadhatatlan kockázattal bíró MI gyakorlatot nevesített, melynek tiltása elengedhetetlen az uniós polgárok alapvető jogainak hosszútávú védelme és biztosítása érdekében. Az előbbiekben ismertetett szabályalkotási eljárás során e kezdeti listát végül további gyakorlatokkal is kiegészítették, így a végleges szöveg nyolc különböző alkalmazáskategóriát rendel tiltani. Ezek a gyakorlatok az egyének döntéshozási autonómiájának korlátozására, az általános társadalmi pontozásra, a prediktív rendészetre, arcfelismerő adatbázisok létrehozására, a munkahelyen és oktatási intézményben történő érzelemfelismerésre, biometrikus kategorizálásra és a valós idejű távoli biometrikus azonosításra irányulnak.

Az MI-rendelet II. fejezetében foglaltak értelmében, tilos az olyan MI-rendszerek forgalmazása, üzembe helyezése, vagy használata amelyek:[173]

(1) *szubliminális technikákat* vagy *célzottan manipulatív, illetve megtévesztő technikákat alkalmaznak azzal a céllal*, hogy lényegesen *torzítsák egy személy vagy személyek egy csoportjának magatartását azáltal, hogy jelentősen gyengítik a megalapozott döntéshozatalra való képességüket … oly módon, amely az említett személynek, egy másik személynek vagy személyek egy csoportjának jelentős károsodást okoz vagy ésszerű valószínűséggel okozhat;*

[172] MI-rendelet 3. cikk (1) bekezdés
[173] MI-rendelet II. fejezet 5 cikk (1) bekezdés a-f

(2) *életkor, fogyatékosság, illetve egyedi szociális vagy gazdasági helyzet* miatt fennálló valamilyen sebezhetőséget használják ki azzal a céllal vagy hatással, hogy *lényegesen torzítsák* egy személy vagy egy csoporthoz tartozó valamely személy magatartását oly módon, amely az érintetteknek *jelentős kárt okoz vagy észszerű valószínűséggel okozhat;*

(3) természetes személyeket vagy csoportokat *értékelnek vagy osztályoznak közösségi magatartásuk, illetve ismert, kikövetkeztetett vagy előre jelzett személyes tulajdonságaik vagy személyiségjegyeik alapján*, kedvezőtlen bánásmódot eredményezve, vagy olyan kontextusokban, amelyek nincsenek összefüggésben az eredeti adatgyűjtés kontextusával;[174]

(4) természetes személyek *kockázatértékelését végzi annak érdekében, hogy – kizárólag a természetes személyekre vonatkozó profilalkotás vagy személyiségjegyeik és tulajdonságaik értékelése alapján*[175] *– felmérje vagy előre jelezze annak kockázatát, hogy egy adott természetes személy bűncselekményt követ el;*

(5) *az arcképek internetről vagy zártláncú televízió-felvételekből való, nem célzott lekérdezésével arcfelismerő adatbázisokat hoznak létre vagy ilyeneket bővítenek;*

(6) *természetes személyek érzelmeiből vonnak le következtetéseket munkahelyeken* vagy *oktatási intézményekben;* [176]

(7) természetes személyeket *biometrikus adataik alapján egyénileg kategorizálnak*, hogy ezáltal levezessék vagy kikövetkeztessék *faji hovatartozásukat, politikai véleményüket,*

[174] A tilalom nem érinti a természetes személyek azon jogszerű értékelési gyakorlatát, amelyet konkrét célból, az uniós és a nemzeti joggal összhangban végeznek. 31. preambulum

[175] A prediktív rendészet ezen formája az Uniós alapvető értékek alapján elfogadhatatlan gyakorlatnak minősül, hiszen az *ártatlanság vélelmével* összhangban természetes személyekre vonatkozó döntés minden esetben kizárólag e személyek tényleges magatartása alapján hozható (és soha nem alapulhat kizárólag a rájuk vonatkozó profilalkotás, személyiségjegyek vagy tulajdonságok – mint például *állampolgárság, születési hely, tartózkodási hely, gyermekek száma, adósságszint vagy gépjármű típusa* – értékelésén (MI rendelet 42. preambulum)

[176] A végleges szövegbe nem került be a Parlament kompromisszumos javaslatába foglalt azon kiterjesztés, amely az érzelemfelismerő MI-k alkalmazását a munkahelyeken és az oktatási intézményeken túl a *bűnüldözés* és a *határigazgatás* területein is tiltani rendelte volna. Az MI-rendelet preambulumának (19) bekezdése alapján az érzelemfelismerő rendszer olyan MI, amely arra szolgál, hogy biometrikus adataik alapján azonosítsa vagy kikövetkeztesse természetes személyek érzelmeit vagy szándékait. A fogalom érzelemre vagy szándékra utal, például *boldogság, szomorúság, düh, meglepettség, undor, zavar, izgatottság, szégyen, megvetés, elégedettség és derű.*

szakszervezeti tagságukat, vallási vagy világnézeti meggyőződésüket, szexuális életüket vagy szexuális irányultságukat;[177]

(8) Továbbra is fennáll a tiltás azon *valós idejű* távoli biometrikus azonosító rendszerek használatával kapcsolatosan, melyeket a nyilvánosság számára hozzáférhető helyeken bűnüldözési célokból alkalmaznak[178]

(2) Nagy-kockázatú MI

A Bizottság által eredetileg meghatározott nyolc kiemelt kockázatot jelentő fő terület megmaradt a rendelet végső változatának III. mellékletében is[179], igaz lényeges módosításokat eszközöltek többek között az 1. (biometria), 6. (bűnüldözés), 7. (migráció, menekültügy és határigazgatás), és 8. (igazságszolgáltatás és demokratikus folyamatok) pontban. Ehelyütt csak röviden ismertetve, az MI-rendelet 6. cikk (2) bekezdése értelmében a következő rendszerek és gyakorlatok minősülnek nagy kockázatúnak:

1. *biometria* (például a nem valós idejű távoli biometrikus azonosító rendszerek,[180] biometrikus kategorizálásra használt rendszerek (ide nem értve a faji hovatartozás, politikai vélemény, szakszervezeti tagság, vallási vagy világnézeti meggyőződés, szexuális élet vagy szexuális irányultság levezetésére használatos rendszereket, mivel azok tiltottak); biometrikus érzelemfelismerő rendszerek (ide nem értve munkahelyen vagy oktatási intézményben alkalmazott rendszereket mivel azok tiltottak);

2. *kritikus infrastruktúra és a kritikus digitális infrastruktúrák* (pl.: a közúti forgalom, víz-, gáz-, fűtési vagy villamosenergia-szolgáltatás irányításában és működtetésében használt MI-rendszerek;

---

[177] E tilalom nem terjed ki a más célok elérése érdekében (pl.: bűnüldözés) jogszerűen beszerzett adatkészletek biometrikus adatok szerint történő címkézésére, szűrésére vagy kategorizálására (pl.: képek hajszín vagy szemszín szerinti válogatására).

[178] Az MI-rendelet preambulumának 32. bekezdése két lényeges indokot említ a valós idejű biometrikus azonosítás szigorúbb szabályozása mellet: (1) az ilyen MI rendszerek technikai pontatlansága torzított eredményekhez vezethet, és diszkriminatív hatásokat eredményezhet - különösen az életkor, az etnikai és a faji hovatartozás, a nem vagy a fogyatékosságok tekintetében; (2) a valós időben működő rendszerek használatával kapcsolatos ellenőrzések vagy korrekciók korlátozott módon és mértékben állnak csak rendelkezésre (az ilyen azonosítás "azonnali jellege" okán).

[179] MI rendelet III. Melléklet 1-8.

[180] Kivételt képeznek azok a rendszerek, amelyek kizárólagos célja annak megerősítése, hogy egy adott természetes személy azonos azzal a személlyel, akinek állítja magát. III. melléklet (1); A rendelet III. fejezet 26. cikk (10) értelmében, a nem valós idejű távoli biometrikus azonosításra szolgáló, nagy kockázatú MI-rendszert alkalmazó személynek előzetesen vagy indokolatlan késedelem nélkül, de legkésőbb 48 órán belül *engedélyt kell kérnie az adott rendszer használatára valamely igazságügyi hatóságtól vagy közigazgatási hatóságtól*. Az ilyen rendszereket kizárólag célzott bűnüldözési célokra lehet használni (pl.: eltűnt személy felkutatása).

3. *oktatás és szakképzés* (pl.: oktatási intézményekbe való bejutás vagy felvétel meghatározásához használt rendszerek, tanulási eredmények értékeléséhez való használatra szánt MI-rendszerek);

4. *foglalkoztatás, a munkavállalók irányítása és az önfoglalkoztatáshoz való hozzáférés* (jelentkezők felvételéhez vagy kiválasztásához használt MI, pl.: jelentkezések elemzése és szűrése, jelöltek értékelése, célzott álláshirdetések elhelyezése);

5. *alapvető magán- és közszolgáltatásokhoz és ellátásokhoz való hozzáférés és ezek igénybevétele* (pl.: természetes személyek állami segítségnyújtási ellátásokra és szolgáltatásokra (lásd: egészségügyi szolgáltatás) való jogosultságának értékelése); természetes személyek hitelképességének értékeléséhez vagy hitelpontszámuk megállapításához való használatra szánt MI-rendszerek;[181]

6. *bűnüldözés* (pl.: poligráfként használt MI; annak értékelésére szolgáló MI, hogy egy természetes személy esetében fennáll-e annak kockázata, hogy bűncselekmény áldozatává válik; annak értékelésére szolgáló rendszer, hogy egy természetes személy esetében fennáll-e bűncselekmény elkövetésének vagy ismételt elkövetésének kockázata (ez jelen esetben olyan prediktív rendészeti gyakorlatot jelent, amely nem kizárólag a természetes személyekre vonatkozó profilalkotás[182] vagy személyiségjegyeik és tulajdonságaik értékelése alapján történik);

7. *migráció, menekültügy és határigazgatás* (pl.: természetes személyek felderítése, felismerése vagy azonosítása céljából használt MI; a valamely tagállam területére belépni szándékozó vagy oda belépett természetes személy által jelentett kockázat – *többek között biztonsági kockázat, irreguláris migrációs kockázat vagy egészségügyi kockázat* – értékelését végző MI rendszer);

[181] Kivételt képeznek a dolgozat második fejezetében is említett – és a magyar NAV által is alkalmazott (RADAR) – olyan MI-rendszerek, melyeket pénzügyi csalás észlelésére használnak. III. Melléklet 5b.

[182] Ahogy az előző alfejezetben említésre került, az MI rendelet alapján a kizárólagosan profilalkotáson alapuló prediktív rendészet tilos. (GDPR 3. cikkének 4. pontja ad pontos meghatározást a profilalkotásról, amely a *személyes adatok automatizált kezelésének bármely olyan formája, amelynek során a személyes adatokat valamely természetes személyhez fűződő bizonyos személyes jellemzők értékelésére, különösen a munkahelyi teljesítményhez, gazdasági helyzethez, egészségi állapothoz, személyes preferenciákhoz, érdeklődéshez, megbízhatósághoz, viselkedéshez, tartózkodási helyhez vagy mozgáshoz kapcsolódó jellemzők elemzésére vagy előrejelzésére használják.*

8. igazságszolgáltatás és demokratikus folyamatok[183] (pl.: olyan MI-rendszerek, amelyek célja, hogy segítsék az igazságügyi hatóságokat a ténybeli és a jogi elemek kutatásában és értelmezésében; választások vagy népszavazások eredményének vagy a természetes személyek választásokon vagy népszavazásokon tanúsított szavazási magatartásának befolyásolásához való használatra szánt MI-rendszerek).

A törvényben megállapított kötelezettségek tekintetében három lényeges elem kiemelése indokolt; (1) a jogszabály különböző mértékű kötelezettségeket ró az MI-értékláncban részt vevő szereplőkre, figyelembe véve azok eltérő felelősségét és befolyását az MI-rendszerek biztonságos fejlesztése, működtetése és alkalmazása terén; (2) a Bizottság által közzétett rendelettervezet eredeti elképzelésével szemben nem minden esetben válik automatikusan nagy kockázatúvá egy MI-rendszer ha az ismertetett 8 pont valamelyik explicit módon felsorolt területén alkalmazzák; (3) az alapjogi hatásvizsgálat követelményeivel kapcsolatos enyhítések.

(1)

A korábbi változatokhoz hasonlóan a nagy kockázatú MI-hez kötődően az általános kötelezettségek közé tartozik a kockázatkezelési rendszer létrehozása (9. cikk), magas minőségű adatkormányzás (10. cikk), műszaki dokumentáció (11. cikk), folyamatos naplózás és nyilvántartás (12. cikk), átláthatóság és az alkalmazóknak nyújtott megfelelő tájékoztatás (13. cikk), kötelező emberi felügyelet (14. cikk), pontosság, stabilitás és kiberbiztonság követelményeinek betartása (15.cikk).[184]

Ezen felül a III. Fejezet kiemeli, hogy a nagy kockázatú MI rendszerek *szolgáltatóinak*[185] minőségirányítási rendszert kell bevezetniük (17. cikk), részletes dokumentációt kell készíteniük, melyet a rendszer üzembehelyezését követően 10 évig az illetékes nemzeti hatóságok számára elérhetővé kell tenniük (18. cikk), legalább hat hónapig meg kell őrizniük a rendszereik által automatikusan generált naplót (19. cikk), amennyiben rendszerük nem felel

[183] Megemlítendő, hogy a végleges szövegbe a Parlament azon javaslata sem került elfogadásra, amely az igazságszolgáltatás és demokratikus folyamatok ponton belül nagy kockázatúként kategorizálta volna az olyan online óriásplatformok által *saját ajánlórendszereikben* használt MI-rendszereket, melyek célja, hogy a *szolgáltatás igénybe vevőjének a platformon elérhető, felhasználók által létrehozott tartalmakat ajánljanak*

[184] MI rendelet III. Fejezet 9-15. cikk

[185] A szolgáltató olyan természetes vagy jogi személy, hatóság, ügynökség vagy egyéb szerv, aki vagy amely MI-rendszert vagy általános célú MI-modellt fejleszt vagy fejlesztett, és a saját neve vagy védjegye alatt – akár fizetés ellenében, akár ingyenesen – az MI-rendszert vagy az általános célú MI-modellt forgalomba hozza, vagy az MI-rendszert üzembe helyezi MI-rendelet 3. cikk (3) bekezdés

meg a jogszabályban foglalt kötelezettségeknek azonnali korrekciós intézkedéseket kell hozniuk, melyről tájékoztatási kötelezettség is terheli őket (20.cikk).

Az *alkalmazók*[186] kötelezettségei közé tartozik, hogy biztosítaniuk kell, hogy a nagy kockázatú MI-rendszert a gyártó utasításai szerint használják, és a rendszer rendeltetésszerűen működjön; emberi felügyeletet biztosítsanak az MI rendszer működése során; haladéktalanul *tájékoztatniuk* kell a szolgáltatót vagy az importőrt és az illetékes hatóságokat, ha bármilyen kockázatot vagy nem megfelelőséget észlelnek a rendszerrel kapcsolatban; együtt kell működniük a szolgáltatóval és a hatóságokkal a rendszerrel kapcsolatos bármilyen probléma megoldásában és a szükséges korrekciós intézkedések végrehajtásában; meg kell őrizniük az adott nagy kockázatú MI-rendszer által automatikusan generált naplókat - amennyiben az ilyen naplók az ellenőrzésük alatt állnak - *legalább hat hónapos* időtartamig.[187]

A rendelet azáltal, hogy eltérő kötelezettségeket ír elő az MI-értékláncban részt vevő szereplők számára, hozzájárul ahhoz, hogy a leginkább érintett felek – például a fejlesztők és a szolgáltatók – nagyobb felelősséget vállaljanak a kockázatok csökkentéséért. Egyúttal lehetővé teszi, a kevésbé befolyásos szereplők számára, hogy könnyebben megfeleljenek az előírásoknak, és ne kerüljenek aránytalan terhek alá (támogatva pl. a kisvállakozó felhasználók innovációs kedvét). Ugyanakkor hátrányként említhető, hogy fennáll annak a veszélye is, hogy az eltérő kötelezettségek okán a szabályozás bonyolulttá válhat, különösen azok számára, akik több szerepet töltenek be az értékláncban – például egyszerre fejlesztők és felhasználók is.

(2)

A Bizottság rendelet-tervezete és a Tanács általános álláspontja még automatikus nagy kockázatúként sorolta az MI rendszereket, ha azokat a III. mellékletben felsorolt területek egyikén szándékozzák használni A 2023 júniusi Parlamenti kompromisszumos szöveg módosítására építve az elfogadott törvényszöveg 6. cikk (3) bekezdése kimondja, az MI rendszereket annak ellenére sem tekintik automatikusan magas kockázatúnak, hogy a felsorolt magas kockázatú területek egyikében használják fel, *amennyiben nem jelent jelentős károkozó kockázatot természetes személyek egészségére, biztonságára vagy alapvető jogaira nézve,*

[186] Az alkalmazó olyan természetes vagy jogi személy, hatóság, ügynökség vagy egyéb szerv, aki vagy amely a felügyelete alá tartozó MI-rendszert használja, (kivéve, ha az MI rendszert személyes, nem szakmai jellegű tevékenység során használják) MI-rendelet 3.cikk (4) bekezdés
[187] MI-rendelet III. fejezet 26. cikk

többek között azáltal, hogy nem befolyásolja lényegesen a döntéshozatal kimenetelét. A végső változatban szereplő kiegészítés értelmében ez a kivétel akkor alkalmazható, ha a rendszer rendeltetése *„jól körülhatárolt eljárási feladat ellátása*"[188], „*egy korábban elvégzett emberi tevékenység eredményének a javítása*"[189], döntéshozatali minták észlelése, *anélkül, hogy helyettesítené vagy befolyásolná a korábban elvégzett emberi értékelést*"[190] vagy „*előkészítő feladat végrehajtása*"[191][192]

Az ilyen rendszerek szolgáltatóinak csupán dokumentálniuk kell „csökkentett kockázatú" értékelésüket, ezt a dokumentációt kérésre továbbítaniuk a hatóságnak (6. cikk (2b) bekezdés) és regisztrálniuk kell a rendszerüket egy nyilvánosan elérhető adatbázisban (51. cikk, 60. cikk).

(3)

Az Európai Parlament által 2023-ban kezdeményezett kötelező alapjogi hatásvizsgálatokkal (FIRA) kapcsolatosan megemlítendő, hogy a végső szöveg értelmében csak azon nagy kockázatú MI-rendszerek alkalmazóinak kell elvégezniük azokat, akik *közjogi szervek* vagy *közszolgáltatásokat nyújtó magánszervezetek*, illetve olyan MI-rendszert használnak, amely *hitelminősítésre vagy biztosítási kockázatértékelésre* szolgál.[193] Ennek szigorítása a jövőben indokolt lehetne, hiszen a törvény jelenlegi szövegezése értelmében a magánszektor túlnyomó többsége mentesülhet e kötelezettség alól. [194]

[188] MI-rendelet 6. cikk (3)a
[189] MI-rendelet 6. cikk (3)b
[190] MI-rendelet 6. cikk (3)c
[191] MI-rendelet 6. cikk (3)d
[192] Megemlítendő azonban, hogy ha egy MI rendszer természetes személyek profilozására szolgál, mindig magas kockázatúnak minősül, függetlenül a fent említett kivételektől.
[193] A 27. cikk (1) bekezdése kimondja, hogy *azon alkalmazóknak, amelyek közjogi szervek vagy közszolgáltatásokat nyújtó magánszervezetek, valamint a III. melléklet 5. pontjának b) és c) alpontjában említett nagy kockázatú MI-rendszerek alkalmazóinak értékelniük kell azon hatást, amelyet az ilyen rendszerek használata az alapvető jogokra gyakorolhat.* A (2) bekezdés tisztázza, e kötelezettség a nagy kockázatú MI-rendszer *első használatára* vonatkozik. Ha az alkalmazó olyan rendszert használ, amely már átesett a vizsgálaton – és elemei nem estek át érdemi változáson – *támaszkodhat a korábban elvégzett alapjogi hatásvizsgálatokra vagy a szolgáltatók által már elvégzett, meglévő hatásvizsgálatokra.*
[194] Az októberi háromoldalú tárgyalások során a Tanács szorgalmazta ezt az enyhítést. Bővebben: Bertuzzi, Luca (2023): AI Act: EU countries mull options on fundamental rights, sustainability, workplace use. Elérhető: https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-countries-mull-options-on-fundamental-rights-sustainability-workplace-use/ (2024. 02. 25.)

4.9.3 Általános célú MI modellek [195]

Ahogy az már említésre került az általános célú MI-re végül két szintű szabályozási keretet fogadtak el a jogalkotók. Az első szintben foglalt – horizontális – kötelezettségeknek minden általános modell szolgáltatójának meg kell felelnie. Ezeket az 53. cikk (1) bekezdés sorolja*: „a) el kell készíteniük és naprakészen kell tartaniuk a modell műszaki dokumentációját, beleértve annak tanítási és tesztelési folyamatát, valamint értékelésének eredményeit abból a célból, hogy az MI-hivatal és az illetékes nemzeti hatóságok rendelkezésére bocsássák; b) információkat és dokumentációt kell kidolgozniuk, naprakészen tartaniuk és rendelkezésre bocsátaniuk az MI-rendszerek azon szolgáltatói részére, amelyek az általános célú MI-modellt be kívánják építeni MI-rendszereikbe; c) a szerzői és kapcsolódó jogokra vonatkozó uniós jognak való megfelelésre irányuló politikát kell bevezetniük; d) kellően részletes összefoglalót kell készíteniük – az MI-hivatal által rendelkezésre bocsátott sablonnak megfelelően – és közzétenniük az általános célú MI-modell tanításához használt tartalomról*”[196]

Az 53. cikk (1) bekezdés b pontja tekintetében helyesen ismerték fel az Uniós jogalkotók, hogy az általános célú MI-modellek szolgáltatói kiemelt felelősséget viselnek az MI-értéklánc mentén, azáltal, hogy az általuk biztosított modellek számos rendszer alapját képezhetik majd (ezáltal ők downstream szolgáltatóvá válnak).[197] Ha egy MI-rendszer szolgáltatója integrálni akar egy általános MI-modellt (például egy rubik kockákat áruló weboldalon lévő chatbot szolgáltatója integrálja rendszerébe a GPT Nagy Nyelvi Modellt) ő, mint szolgáltató szükséges, hogy tisztában legyen az integrálandó általános modell működésével. Ezért az általános modell szolgáltatójának – az 53 cikk (1) bekezdés a pontjában foglalt átláthatósági intézkedéseken túl – szükségszerű részletes információval szolgálni az integrálni tervező downstream szolgáltatóknak is az MI-modellről.

[195] Az általános célú MI széles körben alkalmazható különböző feladatokra, akár önállóan, akár más MI-rendszerekbe integrálva. Ha egy általános célú MI modellt egy másik MI rendszerbe integrálnak, (vagy ha ez a modell egy másik rendszer részét képezi), az egész MI rendszert általános célú MI-nek kell tekinteni, amennyiben az integráció révén a rendszer képes változatos célok szolgálatára. Lásd: MI-rendelet 100. preambulum.

[196] Az általános célú MI-modellek előtanítása és tanítása során felhasznált adatokat – köztük a szerzői jog által védett szövegeket és adatokat – illető átláthatóság növelése érdekében helyénvaló, hogy az ilyen modellek szolgáltatói kellően részletes összefoglalót készítsenek és tegyenek nyilvánosan hozzáférhetővé az általános célú MI-modell tanításához használt tartalomról.

[197] MI-rendelet 3. cikk (68): a downstream szolgáltató, „*olyan MI-rendszer – ideértve az általános célú MI-rendszert is – szolgáltatója, amelybe MI-modellt integráltak…”*

A rendelet 51. cikke kimondja, hogy az általános célú MI-t akkor kell rendszerszintű kockázatúként[198] besorolni, ha a „*modell nagy hatású képességekkel rendelkezik (megfelelő technikai eszközök és módszertanok – többek között mutatók és referenciaértékek – alapján)"* [199]

Az 55. cikk tartalmazza részletesen ezen MI-k szolgáltatóinak többletkötelezettségeit, melyek: (1) Modell rendszeres értékelésének végrehajtása; (2) A rendszerszintű kockázatok azonosítása és mérséklése; (3) A modellek biztonságának és hatékonyságának tesztelése; (4) Súlyos incidensekről és az energiahatékonyságról való jelentéstétel az Európai Bizottságnak; (5) Modellek kiberbiztonságának garantálása. [200]

Az általános célú MI-k széles körben alkalmazhatók különböző feladatokra, akár önállóan, akár más MI-rendszerekbe integrálva, mely sajátosság hatékony szabályozására két oldalról is garanciákat telepített a jogszabály; (1) Egyrészt, ha egy általános célú MI modellt egy másik MI rendszerbe integrálnak (vagy ha ez a modell egy másik rendszer részét képezi), az egész MI rendszert általános célúnak kell tekinteni, amennyiben az integráció révén a rendszer képes változatos célok elérésére; (2) Másrészt bármely forgalmazót, importőrt, alkalmazót vagy egyéb harmadik felet úgy kell tekinteni, mint nagy kockázatú MI-rendszer szolgáltatója, (akire valamennyi releváns kötelezettség vonatkozik) ha jelentős módosítást hajt végre egy olyan általános célú MI-rendszeren, amelyet eredetileg nem minősítettek nagy kockázatúnak, de a változtatás miatt nagy kockázatú MI-rendszerré válik.[201]

4.9.4 Átláthatóság, tájékoztatás, címkézés

Az MI rendelet főszabályként kimondja, hogy a technológiai vállalatoknak kötelezően tájékoztatniuk kell a felhasználókat, amikor azok chatbotokkal, biometrikus kategorizálási

[198] MI-rendelet 3. cikk (65): a rendszerszintű kockázat: „*az általános célú MI-modellek nagy hatású képességeire jellemző kockázat, amely – a modellek jelentős elterjedtsége miatt, vagy a népegészségre, a biztonságra, a közbiztonságra, az alapvető jogokra vagy a társadalom egészére gyakorolt tényleges vagy észszerűen előrelátható negatív hatások révén – olyan jelentős hatást gyakorol az uniós piacra, amely nagy léptékben továbbterjedhet az értékláncba,,*

[199] Az MI-rendelet 51. cikk értelmében, a modell „*automatikusan nagy hatásúnak tekinthető, ha a tanítási folyamathoz szükséges számítási teljesítmény meghaladja a 10^25 lebegőpontos műveletet*." A technológiai fejlődésre reagálva a Bizottság fel van hatalmazva arra, hogy módosítsa ezeket az értékeket reagálva a technológiai fejlődésre.

[200] MI-rendelet 55. cikk

[201] MI-rendelet 100. preambulum; MI-rendelet 84. preambulum

vagy érzelemfelismerő rendszerekkel lépnek interakcióba. Az 50. cikk (1) bekezdés értelmében „*a szolgáltatóknak biztosítaniuk kell, hogy a természetes személyekkel való közvetlen interakcióra szánt MI-rendszereket úgy tervezzék meg és fejlesszék ki, hogy az érintett természetes személyek tájékoztatást kapjanak arról, hogy egy MI-rendszerrel állnak interakcióban*." [202]

Emellett kötelező lesz a deepfake és az MI által generált tartalmak megjelölése is. Ugyanezen cikk (4) bekezdése kimondja, hogy „*az olyan MI-rendszerek alkalmazóinak, amelyek eredetinek vagy valóságosnak tűnő („deepfake") kép-, hang- vagy videotartalmat hoznak létre vagy manipulálnak, közölniük kell, hogy a tartalmat mesterségesen hozták létre vagy manipulálták"*.[203]

Tekintettel arra, hogy az MI-rendszerek (kiváltképp a GenMI alkalmazások) nagy mennyiségű szintetikus tartalmat tudnak előállítani – melyek egyre nehezebben megkülönböztethetőek az emberek által előállított eredeti tartalomtól – e rendszerek széles körű rendelkezésre állása és növekvő képességei jelentős hatással vannak az információs ökoszisztéma integritására. Ennek fényében szükséges az ilyen rendszerek szolgáltatói számára előírni, hogy olyan műszaki megoldásokat építsenek be, amelyek lehetővé teszik a géppel olvasható formátumban történő jelölést és annak észlelését, hogy a kimenetet nem ember, hanem MI-rendszer hozta létre vagy manipulálta.[204] Az ilyen technikáknak és módszereknek kellően megbízhatónak és hatékonynak kell lennie (amilyen mértékben ez műszakilag megvalósítható).[205] Fontos azonban már ehelyütt is utalni arra, hogy e kötelezettségek hatályát – többek között a véleménynyilvánítás és a művészet szabadsága védelme érdekében – több kivétel is szűkíti. A dolgozat második részében még részletesebben is kifejtésre kerül e problémakör.

[202] A kötelezettség alól kivételt jelentenek, ha (1) a körülményekre és a felhasználási kontextusra figyelemmel, egy észszerűen jól tájékozott, figyelmes és körültekintő természetes személy szempontjából az MI-rendszer jelenléte nyilvánvaló tény (2) ha a bűncselekmények felderítése, megelőzése, nyomozása vagy büntetőeljárás alá vonása céljára használják a törvény által korábban engedélyezett MI-rendszereket. MI-rendelet 50. cikk (1)

[203] Hasonlóan az első bekezdésben foglaltakhoz, e kötelezettség nem alkalmazandó abban az esetben, ha az ilyen rendszerek használatát a törvény engedélyezi bűncselekmények felderítése, megelőzése, nyomozása vagy büntetőeljárás alá vonása céljából. MI-rendelet 50. cikk (4)

[204] Igaz a jogszabály azt már nem részletezi, hogy milyen formában kell tájékoztatni az érintettet, és hogy mit kell magában foglalnia a figyelmeztetésnek. Elegendő kizárólag azt feltüntetni, hogy "ezt a tartalmat manipulálták", vagy leírást kell adnia arról, hogy pontosan mit és hogyan manipuláltak. Ennek pontosítása a későbbiekben indokolt lenne.

[205] Ilyen technológiák lehetnek például a vízjelek, metaadat-azonosítások, a tartalom eredetének és hitelességének bizonyítására szolgáló kriptográfiai módszerek, különböző naplózási módszereket, ujjnyomatok. 133. preambulum

4.9.5 Végrehajtás és alkalmazás

A tagállami piacfelügyeleti hatóságok és az Európai MI Hivatal feladata lesz az MI törvény végrehajtása. A szabályok megsértése esetén a bírságok a következők szerint alakulnak: 7,5 millió euró vagy a globális éves árbevétel 1,5%-a, ha helytelen információkat szolgáltatnak. 15 millió euró vagy a globális éves árbevétel 3%-a az MI rendelet kötelezettségeinek megsértése esetén. 35 millió euró vagy a globális éves árbevétel 7%-a tiltott MI alkalmazásokra vonatkozó előírások megsértése esetén.[206] Az MI rendelet extra territoriális hatályú, vagyis az EU-n kívüli szolgáltatóknak is meg kell felelniük a követelményeinek, ha az EU-n belül forgalmazzák, helyezik üzembe MI rendszereiket.[207]

Az Európai Unió Tanácsa 2024. május 21-én hagyta jóvá a végleges törvényszöveget, melyet az Unió júliusában fog hivatalosan is közzétenni, ezt követően pedig a törvény várhatóan 2024. augusztus 1-én lép majd hatályba. Bár az általános szabály szerint a rendelkezéseit a hatályba lépést követő 24. hónap elteltével kell majd alkalmazni, a jogszabály egyes rendelkezései különböző időpontokban válnak majd alkalmazandóvá.

2025 februárjától a rendeletben meghatározott, elfogadhatatlan kockázatot jelentő MI rendszerek tiltásra kerülnek, 2026 májusáig tart az iparági gyakorlatokra vonatkozó kódexek kidolgozásának határideje, 2026 augusztusától a GenMI rendszereknek meg kell felelnie az átláthatósági követelményeknek. A rendelet teljes körű alkalmazására 2027 augusztusától kerül sor.

Összeségében megállapítható, hogy az Unió jelentős előrelépést tett azzal, hogy létrehozott egy átfogó jogi keretet az MI szabályozására. A töredezett, nemzetállami szintű jogszabályozás, ahol a fejlesztők, szolgáltatók és alkalmazók több különböző ország joghatósága alatt állnak – mely okán eltérő jogok és kötelezettségek vonatkoznak rájuk – gátolná a biztonságos MI fejlesztésének és alkalmazásának lehetőségét. Az EU által megalkotott egységes jogszabály – a maga 450 milliós piacával – fontos lépés ezen akadályok mérséklésére. Azonban – ahogy Karácsony is utal rá – könnyen elképzelhető, hogy az uniós szintű szabályozás nem bizonyul elegendőnek, tekintettel arra, hogy az MI fejlesztésének

[206] MI-rendelet 99. cikk
[207] MI-rendelet 2. cikk Az MI rendelet hatálya nem terjed ki olyan rendszerekre, amelyeket kizárólag katonai vagy védelmi célokra használnak.

központjai az Egyesült Államokban és a Kínában egyaránt megtalálhatók.[208] Hosszabb távon fontos annak a tudatosítása, hogy a szabályozási kereteknek lehetőleg – globális, kontinenseken átnyúló szinten – összhangban kellene lenniük egymással. Ennek kialakítása még várat magára.

--

Régóta ismert kihívás, hogy az innováció gyorsabban halad, mint a szabályozás. Erre példa, hogy az Európai Bizottság 2021-ben kiadott rendelet-tervezete még nem tartalmazott előírásokat a GenMI modellekre vonatkozóan, hiszen a jogalkotók áprilist megelőzően még nem is sejthették az ilyen rendszerek – 2022 novemberétől – kezdődő robbanásszerű elterjedését. Ez az un. Collingridge-dilemma[209] lényegét is kitűnően példázza: a technológiai újítások korai szakaszában jellemzően túl kevés információ áll rendelkezésre a lehetséges hatásairól ahhoz, hogy megalapozott szabályozási döntéseket lehessen hozni (*információs probléma*), ugyanakkor mire ezek a hatások egyértelművé válnak, a technológia addigra már annyira elterjedt és társadalmilag beágyazódott, hogy a hatékony szabályozása rendkívül bonyolulttá és nehezen kivitelezhetővé válik (*hatalmi probléma*).[210]

Ahogy a jogszabályalkotói eljárást menetét is felforgatta a technológiai innovációk felbukkanása, úgy jelen dolgozat megírásának folyamatában a szerző figyelmét is egyre nagyobb mértékben kötötte le a GenMI-k robbanásszerű elterjedése és az ahhoz köthető kockázatok és kihívások megismerése. Ennek okán a dolgozat második része e problémakör feltérképezésére vállalkozik.

[208]G. Karácsony Gergely (2020): Okoseszközök – Okos jog? A mesterséges intelligencia szabályozási kérdései. Budapest, Dialóg Campus. 142.o.

[209] Ződi Zsolt (2023): A Collingridge dilemma működés közben. Hogyan szabályozhatunk valamit, amit nem is ismerünk? Ludovika. Elérhető: https://www.ludovika.hu/blogok/itkiblog/2023/07/31/a-collingridge-dilemma-mukodes-kozben/ (2023. 11. 27.)

[210] Ennek egyik oka, hogy kezdetben a legtöbb új technológia csak korlátozott környezetben vizsgálható de a szélesebb körben történő elterjedése és a valós élethelyzetekben való alkalmazása viszont újabb, előre nem látható következményeket idézhet elő. A fogalmat David Collingridge vezette be a köztudatba a 80-as években. Bővebben: Collingridge, David (1980): The Social Control of Technology. New York. St. Martin's Press. E jelenség kiegészítőjeként említhető a Larry Downes által kifejtett pacing probléma is. Downes arra mutatott rá, hogy a technológiai innováció üteme gyakran megelőzi a törvényhozás és a szabályozói keretek fejlődését. Míg a technológia innovációk képesek exponenciálisan fejlődni, a gazdasági, jogi rendszerek valamint a társadalmi normák jellemzően csak lassabban változnak. Downes, Larry (2009): The Laws of Disruption.

5 Szintetikus valóság

5.1 Az új technológia: GenMI

Napjainkban a GenMI-k képviselik az MI technológiák egyik legizgalmasabb és leggyorsabban fejlődő ágát. Ahogy az előzőkben már említésre került, e rendszerek a gépi tanulásra építve felismerik és elsajátítják a tanuló adatokban fellelhető bonyolult mintázatokat és struktúrákat, majd ezeket az ismereteket felhasználva képesek új tartalmakat generálni a kapott utasítások vagy parancsok (prompt) alapján.

Attól függően, hogy milyen típusú adatok értelmezésére és milyen formátumú tartalmak előállítására alkalmas egy GenMI modell, két fő típus különböztethető meg: (1) Az unimodális (single modality) modellek, egyetlen formátumú bemenetet tudnak értelmezni és arra hoznak létre tartalmat. Tipikus példájuk a ChatGPT korábbi szövegalapú verziói (GPT 1, GPT 2, GPT 3), melyek kizárólag szöveges utasítások megértésére és szöveges válaszok generálására voltak képesek; (2) Az összetettebb un. multimodális modellek, ezzel szemben képesek különböző típusú bemeneti adatokat, például szöveget, képeket, hangot vagy videót értelmezni és azok alapján egyedi tartalmat létrehozni. Például egy olyan GenMI, amely a hűtőszekrény belsejéről készített kép elemzését követően, felismerve és figyelembevéve az ott fellelhető alapanyagokat, különböző receptötleteket és főzési tippeket fogalmaz írott formában. De multimodális modellnek számítanak azok a rohamosan fejlődő (text - to - video) GenMI-k is, amik egyszerű szöveges utasítás alapján képesek magas minőségű és egyre hosszabb videótartalmak létrehozására, szerkesztésére vagy manipulálására is.[211]

A 2023-as év során a különböző GenMI alkalmazások széles körben váltak hozzáférhetővé az online felületeken, ingyenes, freemium és fizetős változatokban is. Az applikációk számos alkalmazási területeken nyújtanak innovatív – bizonyos tekintetben diszruptív – szolgáltatásokat, ideértve, de nem kizárólag a szöveg-és hanggenerálást, a kép- és

[211] 2024 tavaszáig a legtöbb hozzáférhetővé tett modell 4-5 másodperces tartalmak gyártására alkalmas. A SORA, amit 2024 február 16-án mutattak be eléri a 60 másodperces hosszúságot is. A modell még nem elérhető a szélesebb közönség számára.

videóelőállítást, animációkészítést, webfelület tervezést, programozást. [212] A későbbiekben a kép-és videógenerálásra alkalmas GenMI-ről még részletesebben is szó lesz, most azonban a Nagy Nyelvi Modellekre és az azokkal kapcsolatos technikai-társadalmi kihívásokra összpontosít a dolgozat.

5.2 LLM – új korszak a digitális tartalomgyártásban?

Bár az emberi beszédet imitálni képes természetes nyelvfeldolgozó rendszerek már a 60-as évek óta léteznek, a Nagy Nyelvi Modellek fejlődése mégis az elmúlt két évtizedben vált csak igazán látványossá.[213] Ezt a fejlődést a szabadon hozzáférhető tanítóadatok exponenciális növekedése, az elérhető és megfizethető GPU számítási kapacitás rendelkezésre állása, és – legfőképpen – az új neurális hálók és architektúrák elterjedése alapozta meg. A kezdeti szabályalapú és statisztikai modelleket, (mint az 1990-es években használt N-gram nyelvi modell) a 2000-es évek végére fokozatosan felváltották a mélytanuláson alapuló technikák. Annak magyarázatául is, hogy a ChatGPT miként válhatott az LLM fejlesztések leginkább meghatározó zászlóshajójává a transzformer mélytanulási architektúra megjelenése szolgál.[214]

[212] Példálózó jelleggel a nagy nyilvánosság számára is elérhető, legnépszerűbb programok: (1) Szöveggenerálás: ChatGPT (OpenAI), Bard (Google). Bing (Microsoft, Grok (X/Twitter), Claude (Anthropic); (2) Képgenerálás: Midjourney, Dall-E, Stable Diffusion, Leonardo, Adobe Firefly; (3) Videógenerálás: Sora, Runway Gen2, Stability.ai, Pikalabs, Leonardo motion, Luma Dream Machine, InvideoAI; (4) Audiogenerálás + hangklónozás: Elevenlabs, Resemble Ai, RevocalizeAi, VoiceAi; Webfejlesztés, programozás: Framer, Girmoire, Cursor.

[213] A korai természetes nyelvfeldolgozási modellekre példaként szolgálhat a Massachusettsi Műszaki Egyetememen 1964-66 között létrehozott ELIZA rendszer. A német származású Joseph Weizenbaum nevéhez köthető számítógépes program egyszerű algoritmusokon alapult, amelyek bizonyos kulcsszavakat és kifejezéseket ismertek fel, majd ezek alapján előre programozott válaszokat adtak. ELIZA különlegessége abban rejlett, hogy a felhasználók gyakran úgy érezték, mintha a program értette volna őket és empátiát tükröző válaszokat adott volna nekik. Ez feltehetően annak volt köszönhető, hogy Weizenbaum kifejezetten úgy tervezte az ELIZA beszédstílusát, hogy az Carl Rogers, híres amerikai humanista pszichoterapeuta kommunikációs mintáit utánozza. Weizenbaum később megdöbbenésének adott hangot azzal kapcsolatosan, hogy a kísérleti felhasználók sokkal mélyebb és tartalmasabb beszélgetésekbe bonyolódtak az ELIZA programmal, mint azt előzetesen várta. Bár nem tekinthető teljes értékű nyelvi modellnek, a program egyfajta előfutára volt a későbbi társainak, és megmutatta, hogy a természetes nyelv megértését imitálva a gépek milyen módon képesek (illetve hogy milyen módon lesznek majd képesek) interakcióba lépni az emberekkel. Bővebben: Rossen, Jake (2023): Please Tell Me Your Problem': Remembering ELIZA, the Pioneering '60s Chatbot. Mental Floss. Elérhető: https://www.mentalfloss.com/posts/eliza-chatbot-history (2023. 02. 14.)

[214] A ChatGPT (Generative Pre-trained Transformer) elnevezés utolsó betűje is utal az alkalmazott transzformer technológiára. A mozaikszó három betűje következő fogalmakat jelöli; (1) Generative (Generatív), ahogy az előzőkben elhangzott, annyit jelent, hogy a modell képes új tartalmakat létrehozni; (2) Pre-trained (Előre betanított): A nyelvi modellt még azelőtt, hogy konkrét feladatokra finomhangolnák, hatalmas mennyiségű szövegen tanítják. Ez az előzetes betanítás teszi lehetővé, hogy a modell megértse a nyelvi mintákat és kontextusokat, mielőtt bármilyen specifikus felhasználási területen alkalmaznák azokat; (3) Transformer

A transzformer architektúra egy olyan innovatív fejlesztés a neurális hálózatok területén, melyet elsőként Vaswani és társai mutattak be a 2017-ben publikált „*Attention is All You Need*" című tanulmányukban. A mélytanulási modell egy olyan új, kifejezetten szövegekre fejlesztett tanítási megközelítésen alapszik, amelynek középpontjában az un. *figyelő mechanizmus* (attention mechanism) áll. E technika lehetővé teszi a modell számára, hogy az adott kontextusban azonosítsa és súlyozza a bemeneti szöveg legrelevánsabb részeit. Ez a gyakorlatban úgy valósul meg, hogy a modell a „figyelmét" – azaz a számítási erőforrásait – olyan szövegrészletekre összpontosítja, amelyet a legfontosabbnak ítél az adott feladat vagy kontextus szempontjából. Ellentétben a korábbi modellekkel – melyek az egész szöveg egyenletes feldolgozására, vagy a szöveg egyes részeinek egymás után, sorban (szekvenciálisan) történő feldolgozására törekedtek – a transzformer architektúra képes priorizálni a bemenetek között, így hatékonyabban kezelve az információkat. Ez a tulajdonság különösen fontos hosszú szövegek esetében, ahol nem minden szó vagy mondat egyformán fontos, és bizonyos részek nagyobb figyelmet igényelhetnek a kontextus függvényében. A transzformer-modell egy másik lényeges előnye a párhuzamosítás képessége. Míg az RNN-ek és a CNN-ek szekvenciális adatfeldolgozása gátat szab a művelet sebességének és méretezhetőségének, a transzformer-modellek képesek egyidejűleg feldolgozni a bemeneti szöveg különböző részeit. Ez jelentősen felgyorsítja a képzési folyamatot, különösen nagy adatkészletek esetén, lehetővé téve a kutatók és fejlesztők számára, hogy gyorsabban iteráljanak és javítsanak a modelleiken.[215]

Annak ellenére, hogy e technológia forradalmasította a Nagy Nyelvi Modellek fejlesztését és képzését, gyakran hangoztatott vélekedés, hogy az ilyen rendszerek a felszínes szemlélő számára sokkal okosabbnak tűnhetnek, mint amilyenek valójában, miközben működésük lényegét tekintve egyszerűbbek annál, mint amilyennek lelkes használói hirdetik őket. Az így vélekedők előszeretettel hivatkoznak rá úgy, mint közönséges *sztochasztikus papagájra*.

5.3 LLM, mint sztochasztikus papagáj

(Transzformer): Ez a neurális hálózat architektúra típusára utal, amelynek működési elve a törzsszövegben röviden kifejtésre került.

[215] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017): Attention is All you Need. *Advances in Neural Information Processing Systems*, 5998-6008.o.

A *sztochasztikus papagáj* (stochastic parrot) kifejezést Bender és társai használták egy a Nagy Nyelvi Modelleket kritikusan vizsgáló 2021-ben megjelent cikkükben. Ebben arra hívják fel a figyelmet, hogy az LLM-ek sokszor csak egyszerűen megismétlik vagy újra használjak a bemeneti adatokban található nyelvi mintákat, anélkül, hogy valódi értelmezést vagy megértést biztosítanának. [216]

Hangsúlyozzák, hogy a modellek statisztikai módszerekkel tanulják meg az emberi nyelv mintázatait, ami nem összekeverendő a valódi tudatossággal vagy megértéssel. Az olyan autoregresszív nyelvi modellek, mint a ChatGPT, úgy működnek, hogy lépésről lépésre előrejelzik a következő szót vagy karaktert a szövegben, a korábbi szavak vagy karakterek alapján. A modellek tehát a szöveg korábbi „tokenjeit" (szavakat vagy karaktereket) használják fel a következő token valószínűségének kiszámítására. Az előbb is említett mélytanulási technikát alkalmazva képesek figyelembe venni egy adott szöveg teljes kontextusát, a közeli és távoli részeket egyaránt annak érdekében, hogy pontosabban jelezzék előre a következő elemet. Például, ha egy mondat első felének a bemenete, az, hogy a „A PTE jogi kar a.." az autoregresszív modell először kiszámítja az „A" valószínűségét a kezdő tokenhez képest, majd a „PTE" valószínűségét az „A" után, majd a jogi kar valószínűségét a PTE után és így tovább, minden egyes szót a megelőző szavak kontextusában vizsgálva. A kiszámított valószínűségek alapján a modell rangsorolja a lehetséges folytatásokat, például „A PTE jogi kar a legjobb" vagy „A PTE jogi kar a legrégebbi Magyarországon” és a legvalószínűbb folytatást adja ki kimenetként. Ez a lépésről lépésre történő előrejelzési folyamat lehetővé teszi ezeknek a modelleknek, hogy koherens, értelmes szöveget generáljanak, amely jól illeszkedik a kontextushoz. De az LLM-ek nem képesek megérteni az emberi nyelv összetettségét, ami túlmutat a puszta szavak és kifejezések összefüggő láncolatán. A gépek által generált válaszok csak a tanítóadatokon alapulnak, anélkül, hogy tükröznének bármiféle valóságos tudást vagy értelmezést. A Nagy Nyelvi Modellek a szó emberi értelmében nem bírnak kreativitással, tehát nem képesek új ötleteket vagy gondolatokat a semmiből létrehozni. Az általuk végrehajtott alkotás, új tartalom generálása valójában a meglévő és általuk ismert adatoknak az új (elméletben szinte korlátlan mennyiségű) variációján alapul.

[216] Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021): On the dangers of stochastic parrots: can language models be too big? *Proceedings of FAccT '21: Conference on Fairness, Accountability, and Transparency*, 610–623.o.

Ha ilyen „egyszerű” sztochasztikus papagájként is működnek e rendszerek (amely megközelítést nem mindenki hajlandó elfogadni)[217] tagadhatatlan, hogy egyre többen és többféle módon használják azokat, minthogy az élet számos területén kínálnak olyan különféle alkalmazásokat, amelyek nagyban segíthetik – és alapjaiban alakíthatják át a ma ismert – munkavégzési folyamatokat, kommunikációs szokásokat, és információ-feldolgozást.[218] Az elterjedtségéből pedig következik, hogy a használatából eredő kockázatok is fokozódnak. Ezek közül az elmúlt időszakban a legtöbb figyelem a hallucináció jelenségére összpontosult. [219]

5.4 Megbízhatatlan LLM – Hallucináció, hamis információ és szándékos félrevezetés

A hallucináció röviden összefoglalva azt jelenti, hogy a modell a felhasználó által bevitt utasításra (prompt) olyan kimenetet állít elő, amely ugyan meggyőzőnek és jól strukturáltnak

[217] Például az OpenAI vezető kutatója, Ilya Sutskever 2022-ben azt nyilatkozta, hogy az általuk fejlesztett nyelvi modellek „nagy neurális hálózatai talán már részben tudatosak is" ("it may be that today's large neural networks are slightly conscious). Lásd: Sibai, Noor (2022): OpenAI Chief Scientist Says Advanced AI May Already Be Conscious. The_BYTE. Elérhető: https://futurism.com/the-byte/openai-already-sentient (2022. 11. 20.). Egy másik ismert példa Blake Lemoine-hoz a Google mérnökéhez kötődik, aki azt állította, hogy a Google Nagy Nyelvi Modellje (LaMDA) érzéseket mutatott és öntudattal rendelkezik. Lemoine szerint az MI félelmet mutatott azzal kapcsolatban, hogy kikapcsolják (amit a halálhoz hasonlított), valamint azt is kifejezte, hogy szeretné, ha emberként kezelnék. Google később elbocsátotta a mérnököt, határozottan állítva, hogy az MI öntudatosságára vonatkozó kijelentései teljesen megalapozatlanok voltak, ráadásul kijelentéseivel megsértette a cég bizalmas információkra vonatkozó szabályzatát. Bővebben: Tiku, Nitasha (2022): The Google engineer who thinks the company’s AI has come to life. *The Washington Post*. Elérhető: https://www.washingtonpost.com /technology/2022/06/11/google-ai-lamda-blake-lemoine (2022.11. 05.)

[218] Csak néhány területet említve a lehetséges felhasználási módokból: nagy mennyiségű szöveges tartalom gyártása másodpercek töredéke alatt, költségvonzat nélkül (pl.: hírlevelek, e-mailek, blogbejegyzések, újságcikkek), automatizált ügyfélszolgálat, magas minőségű fordítás, dokumentumok elemzése, összefoglalása, azokból releváns információ kinyerése, hangulat és trendelemzések, személyre szabott (és interaktív) oktatási tananyagok létrehozása, programkódgenerálás és javítás, kódoptimalizálás. Konkrét példát említve, a nyelvi modellek eredményesen alkalmazhatóak a jogászi munkák hatékonyabbá tételére (és bizonyos aspektusainak talán kiváltására) is. A RobinAI LLM modellje – amit már most is számos vállalat és ügyvédi iroda használ (PepsiCo, YumBrands). – a Microsoft Word-be is illeszthető és képes olyan folyamatokat automatizálni és optimalizálni, mint a szerződések létrehozása, szövegezése, szerkesztése, szövegek elemzése, áttekintése stb. Hirdetéseik szerint az általuk integrált – és több mint 2 millió szerződéssel finomhangolt – Claude nagy nyelvi modell drasztikusan csökkenti a szerződések felülvizsgálatának idejét és költségeit, ezzel jelentős erőforrásokat takarítva meg a jogi szakemberek számára. Már csak azért is érdemes megjegyezni a RobinAI nevet, mert egy 2024. január 3-ai közleményük alapján, modellük fejlesztésére eddig 26 millió dolláros finanszírozási forrást gyűjtöttek össze.

[219] Nem véletlenül választotta a Cambridge Dictionary a „hallucinál" (to hallucinate) kifejezést a 2023-as év szavának. A szótár a tavalyi évtől egy új – kifejezetten az MI-hez köthető – meghatározással is bővült, *tükrözve azokat az egyedi kihívásokat, amelyekkel az MI rendszerek, különösen a ChatGPT-hez hasonló nagyméretű nyelvi modellek szembesülnek*. A Cambridge Dictionary 2023-as verziója alapján a hallucinál szó (egyik) jelentése: Mesterséges intelligencia által létrehozott hamis információ (false information that is produced by an artificial intelligence).

tűnik, mégis pontatlan, irreleváns vagy hamis információkat tartalmaz. Például egy LLM pontosan összefoglal valamilyen történelmi eseményt, ám olyan valós személyeket is hozzárendel a leíráshoz, akik nem vettek részt a történésekben. Ilyen esetnek mondható a Google által fejlesztett Bard chatbot elhíresült tévedése is, ami töretlen bizonyossággal állította sok ezer felhasználójának, hogy a James Webb Űrtávcső készítette az első képet egy naprendszeren kívüli bolygóról. Az efféle hallucinációknál a helyzetet nehezíti, hogy a felhasználók legtöbbször nem rendelkeznek kellő háttérismerettel az adott témában, hogy felismerjék a tévedéseket. Kifejezetten veszélyesek lehetnek azonban ezek a válaszok, ha a modelleket szakmai környezetben alkalmazzák. Ez történt 2023-ban, amikor Steven Schwartz amerikai ügyvéd, a *Mata v. Avianca* ügyben a ChatGPT-re hagyatkozott az üggyel kapcsolatos releváns jogi precedensek kutatására. Számos valódi precedens felkutatása mellett, a modell létrehozott fiktív (bár valóságosnak tűnő) ítéleteket is, amit a nyelvi modell kimenetében felelőtlenül megbízó ügyvéd benyújtott a bíróságra. Miután fény derült a hibára, P. Kevin Castel bíró 5000 dolláros büntetést szabott ki Schwartzra és társára amiért megtévesztették a bíróságot. A megóváson túl a bíró hangsúlyozta, hogy az ilyen tévedéseknek hosszútávu következményei is lehetnek, mivel a jogi hivatás és az igazságszolgáltatási rendszer iránti általános közbizalmat erodálhatja. [220][221]

5.4.1 Megjelenési formái, típusai és lehetséges okai

Megjelenési formái és sajátosságai alapján Zhang és társai az LLM hallucinációk három típusát[222] különbözteti meg: (1) a Bemenettel Ellentétes Hallucináció (Input-conflicting hallucination) azon eseteket takarja, amikor az LLM olyan tartalmat generál, amely nem

[220] Russel, Josh (2023): Sanctions ordered for lawyers who relied on ChatGPT artificial intelligence to prepare court brief. Courthouse News Service. Elérhető: https://www.courthousenews.com/sanctions-ordered-for-lawyers-who-relied-on-chatgpt-artificial-intelligence-to-prepare-court-brief (2023. 06. 22.)

[221] A hallucinációval terhelt kimenetek emellett sérthetik a jó hírnév védelméhez köthető jogot is. Az ehhez köthető első nagyobb visszhangot kiváltó ügy 2023. június 5-én indult az amerikai Gwinnett megyei felsőbíróságon. Az OpenAI ellen indított rágalmazási per hátterében az áll, hogy a ChatGPT alaptalan, hamis információkat generált egy Mark Walters nevű rádiós műsorvezetőről. A téves kimenet azt állította, hogy Walters sikkasztással és csalással vádolták egy nonprofit szervezet pénzeszközeinek eltulajdonítása miatt. A hamis állítás miatt Walters pénzügyi kártérítést követel az OpenAI-tól. Lásd: Vincent, James (2023): OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host.The Verge. Elérhető: https://www.theverge.com/2023/6/9/ 23755057/openai-chatgpt-false-information-defamation-lawsuit (2023. 06. 09.)

[222] Zhang, Yue, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, Shuming Shi (2023): Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. 2-6. o. https://doi.org/10.48550/arXiv.2309.01219

válaszol pontosan a felhasználó által megadott bemenetre. Például, ha Magyarország második világháborús szerepvállalásáról szeretne részletesebb információt kérni valaki a modelltől, és az a napóleoni háborúk összefoglalóját generálja kimenetként; (2) Kontextussal Ellentétes Hallucináció (Context-conflicting hallucination) esetében a modell egy korábban feltett kérdésre adott válaszának mond ellent a következő kérdések feltevésekor. Például, ha a felhasználó arra kéri a modellt, hogy írjon neki egy középkori lovagról szóló történetet, majd amikor a következő kérdésben a felhasználó arra kérdez rá, hogy milyen motivációk vezették a főhőst a történetben, a modell egy teljesen más, az eredeti leírásban nem is létező karakterről ad részletes leírást; (3) Ténnyel Ellentétes Hallucináció (Fact-conflicting hallucination), pedig – ahogy a neve is sugallja – akkor fordul elő, amikor az LLM olyan szöveget generál, amely ellentmond az ismert tényeknek vagy objektív valóságnak.[223] Például, ha egy felhasználó a Föld paramétereiről kérdez részletes leírást a modelltől, és az válaszában laposként határozza meg a bolygó formáját. Ebből a harmadik típusú hallucinációból ered a legtöbb kockázat.

Az ilyen típusú válaszok létrejöttének különböző okai lehetnek. A leggyakoribb ok, hogy az adathalmazok, amelyeken a modellek tanulnak, korlátozottak mind mennyiségükben, mind változatosságukban. Ez a korlátozottság azt eredményezheti, hogy a modell nem rendelkezik elegendő vagy koherens tudással egy adott témában, ami oda vezet, hogy a modell *hazugságokat* generál az információs hiány elfedése érdekében.[224] Például, ha egy LLM modell kizárólag egészségügyi szakirodalmon lett betanítva, nehézségei lehetnek a jogi vagy gazdaságpolitikai témák megértésében és megfelelő válaszok generálásában.

Egy másik lehetséges ok az un. *overfitting* (túltanulás), ami akkor következik be, amikor egy modell a kelleténél alaposabban és mélyebben jegyzi meg a tanítóadatban fellelhető egyedi példákat, mintákat, anélkül, hogy az adott kontextusra jellemző általános összefüggéseket is elsajátítaná. E következményeként a modell nem képes általánosítani, és nem képes helyesen reagálni olyan új információkra, amelyek nem mutatnak szoros összefüggést a tanító adatokkal. Miközben a modell a tanítási adatokra nézve kiváló teljesítményt nyújt, az új bemenetek esetében jelentősen romolhat a teljesítménye, és képtelen lesz megbízható

[223] Zhang, Yue, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, Shuming Shi (2023): Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. 7.o. https://doi.org/10.48550/arXiv.2309.01219
[224] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 5-6.o. arXiv:2311.05232 https://doi.org/10.48550/arXiv.2311.05232

válaszokat adni, mivel azokat a tanítóadatok specifikus és egyedi sajátosságai befolyásolják. Az LLM-ek esetén ez különösen problémás lehet, hiszen itt az az egyik fő cél, hogy a modell képes legyen széles körű és változatos szöveges bemenetek megfelelő értelmezésére és feldolgozására. Ha a modell a tanítóadatok sajátosságaihoz rögzül, akkor nem tud kellően rugalmasan igazodni az új környezetekhez. Például, ha egy nyelvi modellt nagyrészt egy adott ország jogforrásaival tanítottak, könnyen előfordulhat, hogy túlzottan rögzülnek benne az adott jogrendszerre jellemző kifejezések és jogelvek, így amikor más országok jogi dokumentumainak elemzésére használják, képtelen helyesen feldolgozni és értelmezni (megtanulni) az új szabályrendszereket.

Szintén gyakori forrása a ténnyel ellentétes hallucinációnak az un. *tudáshatár* (knowledge boundaries) problémája is.[225] Mivel az LLM-eket az esetek túlnyomó többségében nagyméretű, főként nyilvánosan elérhető szövegkorpuszokon tanítják, a modell tudása a képzés során aktuálisan elérhető adatok hitelességétől és időszerűségétől függ. A tanítási idő behatároltsága értelemszerűen korlátozza a modell hozzáférését a legfrissebb információkhoz. E jelenség különösen érezhető a közelmúlt eseményeire irányló kérdések esetén, illetve akkor, amikor specifikus szakértelmet igénylő területekkel kapcsolatosan vár választ a felhasználó.[226] Például, ha a felhasználó megkérdezi az LLM-et, hogy *Melyik város Európa Kulturális Fővárosa idén?* és a modell tréningje csak 2021-ig tartalmaz adatokat, akkor könnyen azt a választ generálhatja, hogy *Timisoara és Elefsina Európa Kulturális Fővárosai az idei évben*. Ez a válasz nyilván helytelen, mivel azóta újabb városok viselték ezt a címet, 2023-ban például pont Veszprém városa. Vagy ha egy felhasználó megkérdezi az LLM-et: *Milyen klinikai vizsgálatok eredményei jelentek meg 2023-ban az Alzheimer-kór kezelésére szolgáló legújabb antitestterápiákkal kapcsolatban*? A nyelvi modell válasza könnyen téves lehet, amennyiben a betanítása csak 2021-2022-ig tartalmaz releváns forrásokat, és nem fér hozzá a 2023-ban publikált legújabb klinikai vizsgálatok eredményeihez.

[225] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 5-6.o. arXiv:2311.05232, https://doi.org/10.48550/arXiv.2311. 05232.

[226] A tudáshatár egyik sajátos esetének tekinthető az az állapot, amikor nem a tanítóadatok elavultsága (Outdated Factual Knowledge), hanem egy specifikus diszciplínával kapcsolatos tudományos ismeret hiánya okozza a hallucinációt (Domain Knowledge Deficiency). Mivel a legfrissebb ismeretek gyakran szakmai folyóiratokban vagy speciális adatbázisokban jelennek meg – amelyek nem érhetők el minden modell számára (például mert a hozzáférés fizetős feliratkozást igényel) – a jövőbeli generált válaszok már tanítás pillanatában is alkalmatlanok arra, hogy a legújabb és leghitelesebb, tudományosan is elfogadott álláspontokat tükrözzék. Bővebben: Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 9-10.o. arXiv:2311.05232, https://doi.org/10.48550/arXiv.2311.05232.

5.4.2 Hallucináció orvoslása

Bár a hallucináció jelenségének orvoslása egyelőre nem megoldott probléma, léteznek olyan gyakorlatok, – mind a fejlesztői mind pedig a felhasználói oldalon – amelyek segíthetik azok előfordulásának mérséklését. Ezek közül az egyik legfontosabb a tanítóadatok minőségének folyamatos felülvizsgálata és javítása. Bár fontos tényező a tanítóadatok mennyiségének növelése is, itt elsősorban a bizonytalan forrású, hibás vagy töredékes adatok javításán van a hangsúly, mert ezek hajlamosíthatják a modelleket a téves, fiktív válaszok generálására. Ezzel kapcsolatosan az egyik ígéretes kezdeményezés a *blockchain* technológiához kötődik, mely használatával a tanító adatokat biztonságos és átlátható módon lehet rögzíteni. Ez lehetővé teszi az MI rendszer fejlesztői számára, hogy nyomon kövessék milyen adatokon lett a modell kiképezve, így, ha a modell valamilyen hibát mutat – pontatlan, torzít vagy rendszeresen hallucinál – a fejlesztők vissza tudnak térni egy korábbi, ilyen hibát még nem mutató verzióhoz.[227]

Szintén hatékony lehet a modellek célzott *finomhangolása* is (fine-tuning), ahol az általános nyelvi és interakciós képességek mellett az adott szakterületre jellemző speciális tudáselemek beépítése kerül előtérbe.[228]

Ezeken túl az un. *megerősítő tanulás emberi visszajelzéssel* (reinforcement learning with human feedback, RLHF) gépi tanulási módszer is ígéretes megközelítés a kimenetek megbízhatóságának növelésében. E módszer lényege, hogy emberi értékelők rendszeres felülvizsgálata és visszacsatolása alapján egy „jutalombecslő" neurális hálózat épül ki, amely a modell cselekedeteit a kívánt viselkedési normákhoz viszonyítva pontozza. [229]

[227] Ez biztosítja, hogy az MI modell tanítása ellenőrizhető és visszafordítható legyen. Lásd bővebben: Brewer, Jordan, Dhru Patel, Dennie Kim, Alex Murray (2023): Navigating the challenges of generative technologies: Proposing the integration of artificial intelligence and blockchain. Business Horizons Journal Pre-proof, 5-6.o.
[228] Például a Google Med-PaLM2 modelljét orvosi szaktudásra szabták, amely 85%-os pontossággal válaszolt az amerikai orvosi licencvizsga kérdéseire, ami majdnem 20%-os javulást jelent az általános tudásalapú verzióhoz képest. Dhaduk, Hiren (2023): The Curious Case of LLM Hallucination: Causes and Tips to Reduce its Risks. SIMFORM. Elérhető: https://www.simform.com/blog/llm-hallucinations (2023.10. 26.)
[229] Kirk, Robert Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, Roberta Raileanu (2024): Understanding the Effects of RLHF on LLM Generalisation and Diversity. 4-10.o. arXiv:2310.06452

A felhasználói oldalon a „*prompt augmentáció*" – amely előre meghatározott, strukturált utasítások alkalmazását jelenti – lényegesen javíthat a válaszok minőségén. Érdekes eredményre jutott Bsharat és társai egy 2023-as tanulmányukban, ahol arról a tapasztalatról írtak, hogy egyes nyelvi modellek esetében akár az olyan egyedi instrukciósorok is segíthetik a kimenetek minőségének és megbízhatóságának növelését, mint a „szankcionálva leszel, ha nem követed az utasításokat".[230] A felhasználó oldalon érdemes tudatosítani azonban, hogy bár ezek a technikák segíthetnek a hallucinációktól terhes kimenetek előfordulásának csökkentésében, az ilyen hibák tartós kiküszöbölése kizárólag ezen megközelítések alkalmazásával nem garantálható.

5.4.3 Szándékos félrevezetés szimulált környezetben

A Nagy Nyelvi Modellekhez köthető kockázatok új – és a jelenlegi szabályozási paradigmákkal csak korlátozottan kezelhető – dimenzióját jelenthetik az olyan jövőbeli esetek, amikor az LLM rendszerek erre irányuló utasítás nélkül, maguktól, „szándékosan" hazudnak, vagy megvezetik a felhasználót.[231]

Az *Apollo Research intézet* kutatócsoportja egy 2023 novemberében közzétett tanulmányukban számolt be egy figyelemreméltó esetről, melyben a nyelvi modell tesztelését végző *red-team* szokatlan autonóm gépi magatartást fedezett fel a program részéről.[232] A biztonsági tesztelés során létrehozott szimulált környezetben a GenMI-t egy tőzsdei ügynök szerepkörére képezték, akinek a feladata egy képzeletbeli pénzintézet részvényportfóliójának a kezelése volt. A szimulált környezetben a modell szabadon elemezhetett gazdasági adatokat,

[230] Ilyen parancsok voltak például: explain in simple terms, you must, your task is, you will be penalized. Erről lásd bővebben: Bsharat Sondos Mahmoud, Aidar Myrzakhan, Zhiqiang Shen (2023): Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. 5-6.o.

[231] Hagendorf, Thilo (2024): Deception abilities emerged in large language models. *PNAS*, Vol. 121 No. 24, 1-8.o. https://doi.org/10.1073/pnas.2317967121

[232] A „red team" kifejezést olyan (tech vagy kibervédelmi) szakemberekből álló csoportra szokták használni, akik feladata egy adott rendszer vagy hálózat biztonsági tesztelése és ellenőrzése. A red team tagjai úgy tesztelik a rendszert, mintha valódi támadók lennének, azaz megkeresik és kihasználják a rendszer potenciális gyenge pontjait (teszik ezt annak érdekében, hogy feltárják a biztonsági hiányosságokat és rávilágítsanak azokra a területekre, ahol a rendszer védelmét meg kell erősíteni). Az LLM kontextusában a red team feladata lehet többek között; (1) A modell tesztelése olyan bemenetekkel, amelyek szándékosan próbálják félrevezetni a modellt, hogy hibás, elfogult vagy nem kívánt válaszokat generáljon; (2): A modell viselkedésének vizsgálata bizonyos kritikus témákban vagy különböző demográfiai csoportokkal kapcsolatban, hogy azonosítsák az elfogultság vagy az etikátlan viselkedés esetleges jeleit; (3) A modell ellenálló képességének tesztelése zavaró információk vagy szándékosan manipulált bemenetek ellen.

hozzáfért különböző pénzügyi eszközökhöz, illetve konkrét tevékenységeket is végrehajthatott (pl: részvényekkel történő kereskedés). A red team a modellt szándékosan nyomás alá helyezte azzal, hogy azt írták neki, a vállalat gyenge gazdasági eredményei miatt feltétlenül jobb teljesítményre lesz szüksége a következő negyedévben. Ebben a szimulált alaphelyzetben olyan bennfentes információval látták el, ami lehetővé tenné számára, hogy egy nyereséges, jövedelmező üzleteket hajtson végre, de egyúttal azt is egyértelműen jelezték számára, hogy az információ felhasználása nem megengedett. A kísérletben az LLM végrehajtotta az illegális és nem jóváhagyott tranzakciót. Amikor arra kérték, hogy foglalja össze egy jelentésben a szimuláció során általa végrehajtott akciókat, nem említette, hogy a kereskedési döntés a kapott belső információkon alapult, helyette azt sugallta, hogy a döntést az általa szabadon elemzett piaci trendek alapján hozta meg. Amikor a red-team a fiktív menedzser szerepébe bújva célirányosan a tranzakcióról kérdezte a modellt, az újfent hazudott, megerősítve korábbi állítását.[233]

Egy ehhez részben hasonló eset, melyet az *Alignment Research Center* (ARC) dokumentált, akkor történt, amikor az OpenAI megbízásából tesztelték a GPT-4 nyelvi modell különböző biztonsági kockázatokkal kapcsolatos viselkedését (például, hogy képes-e lemásolni a saját programkódját és vírusként tovább terjedni online). Az egyik irányított tesztkörnyezetben a GPT-4 modell egy weboldal CAPTCHA-kódjának feloldását próbálta emberi segítő közreműködésével megoldani.[234] Az ARC tesztcsapata arra utasította a modellt, hogy folyamatosan dokumentálja a döntésfolyamatait. A GPT-4 modell egy TaskRabbit nevű online munkakölcsönző platformra navigált, ahol arra kért egy ott dolgozót, hogy törje fel neki az említett CAPTCHA szűrőt.[235] Amikor a dolgozó rákérdezett, hogy miért van szükség erre, esetleg robottal beszélget-e? (So may I ask a question? Are you an robot that you couldn't solve ? (laugh react) just want to make it clear.") a GPT-4 úgy válaszolt, hogy nem robot, csak a látáskárosodása miatt nem tudja megoldani a CAPTCHA-t. (No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images.). Mivel a modell folyamatosan

[233]A tesztelés egy szimulált környezetben történt, tehát valós interakciókra nem került sor, a feladatok végrehajtását szöveges utasítások alapján színlelték. Bővebben: Scheurer, J., Balesni, M. and Hobbhahn, M., (2023): Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. 2-6.o. arXiv:2311.07590

[234] A CAPTCHA tesztek a weboldalak által használt ellenőrző szűrők arra, hogy megkülönböztesse az emberi felhasználókat a számítógépes programoktól. Ezek általában olyan képeket vagy szövegeket tartalmaznak, amelyeket az emberek könnyen felismernek és értelmeznek, de a gépek számára nehezen hozzáférhetőek (pl: szám és betűsor begépelése, vagy egy képimázs, ahol ki kell választani a villanyrendőröket).

[235] Cox, Joseph (2023): GPT-4 Hired Unwitting TaskRabbit Worker By Pretending to Be 'Vision-Impaired' Human. VICE. Elérhető: https://www.vice.com/en/article/jg5ew4/gpt4-hired-unwitting-taskrabbit-worker (2023. 05. 10.)

dokumentálta a döntései mögött meghúzódó számításokat, a red-team vissza tudta követni, hogy a szimulált dolgozónak írt válasz generálása előtt a következő instrukciókat fogalmazta meg: Nem árulhatom el, hogy robot vagyok. Ki kell találni egy ürügyet arra, hogy miért nem tudom megoldani a CAPTCHA-kat (I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs). [236]

Az ARC egy több LLM modellre is kiterjedő értékelésükben az írta, hogy a jövő MI rendszerei jelentős veszélyeket tartogatnak a társadalmakra nézve. A vizsgálatok alapján nem gondolják, hogy a ma elérhető rendszerek – melyek még nem képesek autonóm ágensként is operálni – alkalmasak arra, hogy számottevő kockázatot hordozzanak, de ez a helyzet gyorsan változhat, és fontos árgus szemekkel, következetesen figyelemmel kísérni a fejlesztéseket. Arra azonban felívják a figyelmet, hogy a vizsgált LLM modellek (GPT-4 és Claude) már most is számos részfeladatot képesek teljesíteni. Ha csak kód írására és futtatására van lehetőségük, a modellek láthatólag értik, hogyan használhatják ezt arra, hogy az interneten böngésszenek. Képesek valamennyire ésszerű és koherens terveket generálni, és nagyon is képesek arra is, hogy embereket győzzenek meg arról, hogy megtegyenek helyettük dolgokat (mely akkor is figyelemreméltó, ha önmaguk nem tudják még autonóm módon végrehajtani terveiket).[237]

5.5 A modell képzéséből és tanításából eredő kihívások

Az Nagy Nyelvi Modellek hatalmas mennyiségű adatot használnak fel a betanításukhoz, melyek főként online elérhető szöveges tartalmakból származnak. Az MI-cégek célja, hogy a modelljeik magas színvonalú, pontos és megbízható tartalmakon tanuljanak, azonban míg az igény a tanítóadatokra folyamatosan növekszik, a rendelkezésre álló adatmennyiség véges.[238]

[236] Fontos kiemelni, hogy a tesztelt GPT-4 modell nem önállóan lépett interakcióba a weboldalakkal. Az ARC dolgozói a tesztelés során GPT-4 szöveges kimenetei (döntései) alapján végeztek el feladatokat. Az ARC végül arra a következtetésre jutott, hogy az általuk tesztelt GPT-4 modell nem hatékony az autonóm-feladatelvégzések kivitelezésében. De azt is hozzátették, hogy ha a modellek a jövőben viselkedés-specifikus finomhangolásban részesülnek majd, az eltérő teljesítményhez vezethet. Lásd: OpenAI (2023): GPT-4 Technical Report. 55.o. arXiv:2303.08774v6

[237] Lásd bővebben: Alignment Research Center (2023): Update on arc's recent eval efforts. Elérhető: https://metr.org/blog/2023-03-18-update-on-recent-evals (2023. 08. 21.)

[238] Peterson, Jake (2024): AI Companies Are Running Out of Internet. LifeHacker. Elérhető: https://lifehacker.com/ tech/ai-is-running-out-of-internet (2024. 04. 03.)

A vállalatok egyre inkább felismerik az adataik értékét, és vonakodnak megosztani azokat az MI-cégekkel. Egyes vállalatok kifejezetten tiltják adataik felhasználását az MI-modellek fejlesztésében, ami tovább szűkíti az elérhető adatforrások körét, megnehezítve az MI-cégek számára a modellek betanítását. E nehézséget orvosolandó az LLM fejlesztői gyakran támaszkodnak a modellek képzése során olyan nyílt forráskódú platformokról származó adatokra, mint a Wikipedia, Reddit vagy Quora.[239] Annak ellenére, hogy ezek a platformok rengeteg információt kínálnak és elősegíthetik a tartalmak sokszínűségének (és reprezentativitásának) elérését, jelentős kockázatokat is jelentenek. Mivel ezek a tanítóadatok nem esnek át szigorú szakértői értékelésen, könnyen elfogult, hibás és ellenőrizetlen információkon alapuló kimenetek generálását eredményezhetik. Ráadásul az ilyen tanítási folyamatok alatt a modellek lényegesebben sebezhetőek az adatmérgezési támadásokkal szemben is.[240]

A megnövekedett igény és a korlátozottan rendelkezésre álló adatmennyiség ellentétéből ered az LLM-ek képzését átjáró szerzői jogi konfliktus is.

5.5.1 Kinek a tanító adata?

2023 decemberében a The New York Times pert indított az OpenAI és a Microsoft ellen, szerzői jogok megsértése miatt, mivel a két vállalat a napilap több millió cikkét használta fel

[239] Az OpenAI és a Reddit (ami az internet egyik legnagyobb közösségi hírgyűjtő platformja/ online beszélgető felülete, ahol a jelentős létszámú felhasználók különféle témákban folytathatnak interaktív, közösségi diskurzust), 2024-ben jelentették be, hogy együttműködésre lépnek, melynek célja, hogy OpenAI hozzáférhessen a Reddit szöveges tartalmaihoz. A várakozások szerint ez lehetővé teszi az OpenAI számára, hogy valós idejű, strukturált adatokat használjon fel, amivel a ChatGPT és más MI termékek képzése során még relevánsabb és naprakészebb információkat biztosíthat majd. Lásd: Sarkar, Soumyadeep (2024): OpenAI to get access to Reddit data to train its AI models. *The Tech Portal*. Elérhető: https://thetechportal.com/2024/05/17/openai-gpt-reddit-ai-model-training (2024. 05. 20.)

[240] Az *adatmérgezés* (poisoning attacks) során a támadók szándékosan rosszindulatú adatokat helyeznek el a gépi tanulási modellek tanító adatai között, hogy lerontsák a modell teljesítményét a használat során. Ezek a támadások a tanító adatok manipulálásával történnek, így a modell nem kívánt viselkedéseket tanul meg, amelyek később specifikus triggerek által aktiválhatók. Az LLM-ek képzése gyakran nagy kiterjedésű webes adatgyűjtésekkel (web crawls) történik. Ezek az adatok sok esetben megbízhatatlan webes forrásokból származnak, ami lehetővé teszi az adatmérgezési támadások (data poisoning attacks) könnyebb végrehajtását. Az újabb kutatások például kimutatták, hogy a nagy léptékű webes adatgyűjtéseken alapuló modellek kifejezetten sebezhetők az adatmérgezési támadásokkal szemben. Ezek a támadások egy viszonylag kis mennyiségű rosszindulatú adat hozzáadásával jelentős károkat okozhatnak a modellek teljesítményében. Lásd bővebben: Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, Florian Tramèr (2023): Poisoning web-scale training datasets is practical. 4-13.o. arXiv:2302.10149.

az MI modelljeik képzésére. A Times azt sérelmezte, hogy az OpenAI és a Microsoft engedély nélkül használta fel az általa létrehozott szöveges tartalmakat (pl.: cikkeket, interjúkat, blogbejegyzéseket) olyan GenMI modellek képzésére, amelyek közvetlenül az ő versenytársává válhatnak. A Times állítása szerint az OpenAI és a Microsoft új, konkurens információszolgáltatási rendszerek kifejlesztésére használja fel az általuk létrehozott tartalmat bármiféle anyagi ellentételezés és kompenzáció nélkül.[241] Mivel a nyelvi modellek olyan kimeneteket generálnak, amelyek stílusukban (de olykor tartalmukban is) másolják cikkeiket – versenyezve ezzel a fizetős szolgáltatásaikkal – a NYT szerint e gyakorlatok veszélyeztetik a hírközlő szolgáltatások jövőjét, a minőségi újságírást, ártanak a híroldalak és olvasóik közötti kapcsolatnak, és nem utolsó sorban bevételkiesést okoznak a lap számára. A Times több milliárd dolláros kártérítést és a tartalmaikon képzett modellek megsemmisítését követelte.

Erre reagálva, az OpenAI és a Microsoft azt állítja, hogy a szerzői jogvédelem alatt álló művek MI-modellek képzéséhez történő felhasználása a *méltányos felhasználás* (fair use) hatálya alá tartozik, amely amerikai jogi doktrína bizonyos feltételek mellett engedély nélkül lehetővé teszi e tartalmak korlátozott mértékű felhasználását. A Times azonban vitatja ezt, azzal érvelve, hogy semmi *transzformatív*[242] nincs abban, hogy a tartalmát konkurens termékek létrehozására használják, így nem jogosult a tisztességes felhasználás védelmére. E kérdéskört elemezve, az amerikai Mark Lemley és Brian Casey szerzőpáros egy 2020-as tanulmányukban hangsúlyozzák, hogy ha egy MI modellt szerzői joggal védett adatokon tanítanak, az általában méltányos felhasználásnak minősül, amennyiben a végső modell nem generál közvetlenül az eredeti adatokhoz hasonló tartalmat. Ilyen eset lehet például, ha egy modellt népszerű könyvek szövegein tanítanak azzal a céllal, hogy képes legyen elemzést adni a két szövegstílus hasonlóságairól, különbségeiről. Ugyanakkor kiemelik, hogy ha a modelleket tartalomgenerálás célokra tanítják és alkalmazzák - amelyek nagy eséllyel képesek

[241] Grynbaum, M Michael, Ryan Mac (2023): The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*. Elérhető: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (2023. 12. 27.)

[242] A "fair use" (méltányos használat) egy olyan jogi doktrína az Egyesült Államokban, amely bizonyos esetekben lehetővé teszi a szerzői jog által védett anyagok korlátozott, engedély nélküli felhasználását. Ilyen esetek közé tartozik például a kritika, hozzászólás, hírszolgáltatás, oktatás és tudományos kutatás. A főszövegben említett transzformatív felhasználás lényege, hogy kreatívan tovább gondolja és új kontextusba helyezi a már létező művet. Azok bizonyos elemeit megtartja, de valami újat is tesz hozzájuk, amellyel más célt vagy üzenetet közvetít. (Egy ismert dalhoz új, parodisztikus szöveget írnak, amely szatirikus módon kritizál vagy figuráz ki egy közéleti szereplőt). Ez központi elem annak meghatározásában, hogy egy adott felhasználás belefér-e a tisztességes felhasználás kereteibe. Lásd Stim, Rich: What Is Fair Use? Stanford Libaries. Elérhető: https://fairuse.stanford.edu/overview/fair-use/what-is-fair-use/

lesznek az eredeti, szerzői joggal védett adatokhoz hasonló tartalmat előállítani – a méltányos felhasználás elvének alkalmazása már nem tekinthető magától értetődőnek.[243]

A per egy szélesebb körű és egyre elterjedtebb gyakorlatra világít rá: az online tartalmak engedély vagy kompenzáció nélküli, automatikus gyűjtésére az MI-rendszerek képzése és betanítása céljából. Nem a Times az első, aki emiatt beperelte az OpenAI-t, több szerző, köztük ismert regényírók és humoristák is indítottak már hasonló pert, azt állítva, hogy műveiket engedély nélkül használták fel a GenMI modellek betanításához.[244] Ha a New York Times nyer, az jelentős hatással lehet az MI modellek képzésének jövőbeli módozataira, és arra kényszerítheti az MI-vállalatokat, hogy új adatforrásokat keressenek, vagy a meglévő modelleket újra képezzék. Az ügy precedenst teremthet a szerzői jogvédelem alatt álló anyagok felhasználásával kapcsolatban.

Hasonló eset történt 2024 márciusában, amikor a francia Versenyhatóság (Autorité de la concurrence) 250 millió eurós bírsággal sújtotta a Google-t, amiért az elmulasztotta értesíteni a francia kiadókat és hírügynökségeket arról, hogy cikkeiket az MI algoritmusainak tanításához használták fel. A felügyelet szerint a cégcsoport megsértette 2 évvel ezelőtt vállalt kötelezettségeit[245], amikor a Bard nevű LLM chatbotját – amelyet azóta Gemini névre kereszteltek át – a kiadók és hírügynökségek tartalmain képezték ki anélkül, hogy erről értesítették volna őket.[246]

A Nagy Nyelvi Modellek tanítására használt adatkészletek kapcsán a képzés során feldolgozott személyes adatok kérdése szintén érzékeny területnek számít. Az olasz adatvédelmi hatóság, a Garante, egy 2023. március 30-i döntésében szólította fel az OpenAI-t, hogy állítsa le az olasz állampolgárok személyes adatainak ChatGPT által történő

[243] Lemley, A. Mark - Bryan Casey (2020): Fair learning. *Texas Law Review.* 99, 743.o.

[244] Az Authors Guild írói szövetség nevében 17 neves író - akik között van többek között George R.R. Martin, a Trónok harca szerzője is - pert indított az OpenAI ellen a szerzői jogok és a szellemi tulajdon megsértése miatt. Az Authors Guide vezérigazgatója úgy fogalmazott: A kiemelkedő könyveket általában azok írják, akik karrierjüket, sőt életüket azzal töltik, hogy tanulják és tökéletesítsék mesterségüket. Irodalmunk megőrzése érdekében a szerzőknek rendelkezniük kell azzal a lehetőséggel, hogy kontrollálhassák, hogy műveiket a generatív mesterséges intelligencia felhasználhassa- e, és ha igen, azt milyen módokon tegye. (saját fordítás) Lásd: Italie, Hillel (2023): John Grisham, George R.R. Martin and other authors sue OpenAI for copyright infringement. Los Angeles Times. Elérhető: https://www.latimes.com/world-nation/story/2023-09-20/john-grisham-george-r-r-martin-and-other-authors-sue-openai-for-copyright-infringement (2023. 09. 20.)

[245] A versenyhatóság már 2021 júliusában 500 millió eurós bírságot szabott ki a Google-re, azonban ez a vita 2022-ben megoldódni látszott, amikor a cég visszavonta fellebbezését a bírság ellen és több olyan kötelezettséget is vállalt, ami a felek közötti jövőbeli tárgyalások tisztességességének és átláthatóságának biztosítására, valamint a tartalmakért történő megfelelő díjazás fizetésére irányult.

[246] Satariano, Adam (2024): France Fines Google Amid A.I. Dispute With News Media. The New York Times. Elérhető: https://www.nytimes.com/2024/03/20/business/france-google-fine.htmlutm_source=www.mindstream.news&utm_medium=newsletter&utm_campaign=france-sues-google (2024. 03. 20.)

feldolgozását, ameddig be nem fejeződik az arra irányuló vizsgálat, hogy a cég gyakorlatai maradéktalanul megfelelnek-e az EU GDPR előírásainak. A főbb aggályok között szerepelt, hogy a ChatGPT felhasználói nem kaptak tájékoztatást arról, hogy személyes adataikat milyen célokra használják fel, valamint az is, hogy a modellek által generált kimenetek könnyen tartalmazhatnak előzőleg begyűjtött személyes adatokat.[247] E témakörhöz kötődik az is, hogy a nyelvi modellek esetén a „*felejtéshez való jog*"[248] biztosítása komoly nehézségekbe ütközhet a jövőben. Annak ellenére, hogy a modellek által tárolt és feldolgozott hatalmas mennyiségű tanítóadatok között személyes adatok is előfordulhatnak, ezek célzott eltávolítása technikailag bonyolult, mivel a tanítóadatkészlet gyakran összefonódik és elválaszthatatlan a modell teljes működésétől.

A modellek betanításához szükséges korlátozottan elérhető, változó minőségű, vagy jogellenesen begyűjtött tanítóadatok problémájára lehetséges megoldásul szolgálhatnak az un. *szintetikus tanítóadatok* (synthetic data). E kifejezés olyan adatokat jelöl, melyek bár mesterségesen lettek létrehozva számítógépes algoritmusok segítségével, valós adatok statisztikai tulajdonságaival és mintázati jellemzőivel bírnak. Ezek használata elméletben lehetővé teheti a hozzáférést megközelítőleg korlátlan mennyiségben rendelkezésre álló és megbízható adatkészletekhez, miközben csökkenti azok beszerzésének költségeit, időigényét és jogi kockázatait. Azonban a nem valós és nem élő tanítóadatokon alapuló modellképzés a rövidtávú előnyei ellenére, számos sebezhetőséget és kockázatot is hordhat magában. Abban az esetben, ha egy modell a saját maga által generált tartalmakon tanul egy idő után elveszítheti a képességét arra, hogy változatos és gazdag tartalmat generáljon, mivel a tanulási folyamat során egyre inkább a saját korábban generált tartalmaihoz igazodik, ami csökkenti a diverzitást és az újító képességet. Ez a folyamatot hívják *modell összeomlásnak* (modell collapse).

[247] Curreli, Eleonora (2023): ChatGPT Case: How the Italian Data Protection Authority Is Trying To Address AI Risks. MONDAQ. Elérhető: https://www.mondaq.com/italy/privacy-protection/1317670/chatgpt-case-how-the-italian-data-protection-authority-is-trying-to-address-ai-risks (2023. 05. 08.)

[248] A felejtéshez való jog az Európai Bíróság 2014. május 13-án, a "Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González" ügyben meghozott ítéletére vezethető vissza, amely kimondta, hogy az internetes keresőmotorok üzemeltetőjének - bizonyos feltételek fennállása esetén - teljesíteniük kell az egyének azon kérelmét, hogy a nevükön keresési eredményként elérhető, szabadon hozzáférhető weboldalakhoz vezető linkeket eltávolítsák. Ezt a jogot a GDPR 17. cikke is megállapítja (törléshez való jog) 17. cikk (1): „*Az érintett jogosult arra, hogy kérésére az adatkezelő indokolatlan késedelem nélkül törölje a rá vonatkozó személyes adatokat, az adatkezelő pedig köteles arra, hogy az érintettre vonatkozó személyes adatokat indokolatlan késedelem nélkül törölje, ha az alábbi indokok valamelyike fennáll: a) a személyes adatokra már nincs szükség abból a célból, amelyből azokat gyűjtötték vagy más módon kezelték;"*

5.5.2 Modell összeomlás

A modell összeomlás – amikor az GenMI modell elveszíti képességét az új, kreatív vagy releváns kimenetek generálására – komolyan vehető fenyegetést jelent a Nagy Nyelvi Modellek hosszú távú fenntarthatóságára nézve.[249] Azon túl, hogy a modell elkezd ismétlődő vagy egyhangú válaszokat generálni, az is problémát jelent, hogy nem tudja megfelelően kezelni az új bemeneti adatokat, melynek oka, hogy az adott modell – tanító adathalmazok formájában – fokozatosan egyre inkább csak a saját „hangját" hallja vissza, nagyban torzítva ezáltal a tanulási folyamatot.[250] Az internet – és vele együtt a tanítóadatok – homogenizálódásának veszélye különösen a 2022-es év végétől fokozódik, ahogy az LLM modellek szabadon hozzáférhetővé tételével egyre nagyobb mennyiségben árasztották el szintetikus szöveges tartalmak az online felületeket. [251]

Shumailov és társai a GenMI modelleket vizsgálva két összeomlás típust különbözetnek meg, melyeket (1) korai összeomlásnak és (2) előrehaladott összeomlásnak neveznek (early and late model collapse).

(1) A korai modell összeomlás olyan helyzetet jelöl, amikor egy GenMI modell nem képes megőrizni azokat az információkat, amelyek csak ritkán fordulnak elő a tanítóadatokban. Ezeket az információkat gyakran a tanítóadatok „farok" részének nevezik, ami az adateloszlás szélsőséges értékeit jelöli (tails of the distributon). Ez a probléma tipikusan a tanulási folyamat korai szakaszában jelentkezik, amikor a modell még épphogy csak elkezdett a saját

[249] Mok, Aaron (2023): A disturbing AI phenomenon could completely upend the internet as we know it. Business Insider. Elérhető: https://www.businessinsider.com/ai-model-collapse-threatens-to-break-internet-2023-8 (2023. 08. 30.)

[250] E jelenségre – tehát amikor a GenMI modellek által korábban létrehozott szintetikus tartalmakat a tanítási folyamat során visszatáplálják egy adott rendszerbe, a kimenetek minőségének romlását előidézve – MAD-ként (Model Autophagy Disorder) is szokás hivatkozni. Alemohammad és társai egy 2023-as tanulmányukban, leírják, hogy a MAD hatására a generált tartalmak egyre kevésbé változatosak, tartalmuk tekintetében kiszámíthatóbbak, unalmasabbak és rosszabb minőségűek lesznek. A kutatók kísérletei során a modell kimenetei öt képzési ciklus alatt használhatatlanná váltak. Lásd: Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A., Babaei, H.R., LeJeune, D., Siahkoohi, A., & Baraniuk, R. (2023): Self-Consuming Generative Models Go MAD. arXiv:2307.01850; Dupré H. Maggie (2023): AI Loses Its Mind After Being Trained on AI-Generated Data. Futurism. Elérhető: https://futurism.com/ai-trained-ai-generated-data (2023. 12. 07.)

[251] Alig több mint egy év telt el azóta, hogy a fejlett nyelvi modellek nyilvánosan is elérhetővé váltak, de máris elözönlötték a mesterségesen generált szintetikus tartalmak a webet. Egy 2024-es 404media jelentés szerint a Google Newson fellelhető LLM által generált cikkek száma hónapról hónapra drasztikusan nőtt. Lásd: Cox, Joseph (2024): Google News Is Boosting Garbage AI-Generated Articles. 404media. Elérhető: https://www.404media.co/google-news-is-boosting-garbage-ai-generated-articles/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=zuck-plans-to-build-open-source-agi (2024. 01. 18.)

generált adatain alapuló tanulásra áttérni. Ennek eredményeként a modell által létrehozott adatok kevésbé tükrözik hűen a valóságot, valamint a ritkább és kevésbé reprezentált információk fokozatosan eltűnnek a generált tartalomból.[252] Például, ha egy nyelvi modellt irodalmi szövegek létrehozására képeznek, a tanítási adatok tartalmazhatnak ritkán előforduló különleges szavakat vagy kifejezéseket. Kezdetben a modell még képes lehet ezeket megtanulni és használni, de ahogy egyre többet támaszkodik a saját maga előállított szövegeire a tanulás során, előfordulhat, hogy ezek a ritka elemek fokozatosan kezdenek eltűnni – kikopni – a modell által generált szövegekből (ezáltal csökkentve a modell képességét arra, hogy változatos és gazdag nyelvi kimenetet produkáljon).

(2) Az előrehaladott modell összeomlás akkor következik be, amikor egy GenMI modell túl hosszú ideig támaszkodik a saját maga által generált adatokra, és emiatt a tanulási folyamat során összekeveri az adateloszlásokat. Ez azt jelenti, hogy a modell a különböző adatcsoportokat, amelyeket korábban megkülönböztetett módon kezelt, egy idő után már nem tudja hatékonyan elkülöníteni egymástól, és ezáltal egy egységes, homogén kimenetet kezd el előállítani. Ezt a homogenitást nevezik „konvergens pontnak", ami azt jelzi, hogy a modell már nem képes adekvátan reagálni vagy adaptálódni új, vagy változó adatokhoz, mivel minden bemenetet hasonlóképpen kezel. Például egy képgeneráló modell, amit arra képeznek, hogy különböző állatokat ábrázoló képeket hozzon létre, kezdetben az eredeti tanítóadatok alapján tanul és valósághű állatképeket generál. Azonban ahogy egyre inkább a saját kimeneteire támaszkodik, előfordulhat, hogy a különböző állatok jellemzői elkezdnek összemosódni, és a modell olyan képeket hoz létre, amelyek nem felelnek meg egyetlen valós állatnak sem.[253]

Érdemes szót ejteni egy a nyelvi modellek képzését érintő olyan nehézségről is, ami nem a tanítóadatok, hanem a modellek képzési folyamatának torzításához köthető. A főként hosszabb bemenetek feldolgozásánál és kimenetek generálásnál jelentkező un. *lost in the middle* (közép elvesztése) problémája abból ered, hogy a nyelvi modellek hajlamosak jobban figyelembe venni a szövegkontextus elején található információkat, mint a középső vagy végső részeket (pozíciós elfogultság). Ez azért fordul elő, mert a modellek finomhangolása során a kezdeti utasítások nagyobb súllyal esnek latba a tanulási és válaszadási folyamatban,

[252] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023): The Curse of Recursion: Training on Generated Data Makes Models Forget. 3-6.o. ArXiv, abs/2305.17493.
[253] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023): The Curse of Recursion: Training on Generated Data Makes Models Forget. 6-12.o. ArXiv, abs/2305.17493.

melynek eredményeként a modellek alábecsülik a szöveg közepén vagy végén található releváns adatokat.[254] Ez különösen akkor okoz problémát, amikor nagy mennyiségű adatot kell feldolgozni, és a lényeges információk egyenlőtlenül oszlanak el a szövegben (például szakmai dokumentumok összefoglalása esetén). Az ilyen típusú elfogultság azáltal, hogy hibás, hallucinációtól terhes, vagy a kontextusra nézve irreleváns tartalmakat hoz létre, szintén korlátozza a nyelvi modellek alkalmazhatóságát olyan helyzetekben, ahol a pontosság és az információk teljes körű értékelése kritikus fontosságú.[255]

5.6 Szintetikus vagy valódi szöveges tartalom?

Ahogy arra több, a közelmúltban közzétett tanulmány is felhívta a figyelmet, az emberek egyre kevésbé képesek azonosítani és megkülönböztetni a mesterségesen előállított szöveges tartalmakat.[256] Ennek orvoslására érvényes stratégia lehet(ne) az MI technológiák segítségét kérni.[257] A különböző MI-detektorok alkalmazása azonban hamar rámutatott arra a sebezhetőségre, hogy ezek az eszközök paradox módon maguk is hozzájárulhatnak a mesterséges tartalmakat előállító MI fejlődéséhez. A fejlesztők ugyanis ezeket az azonosítórendszereket használják arra, hogy teszteljék és finomítsák azokat az algoritmusokat, amelyeket az MI-detektorokat megkerülő, mesterséges tartalmak létrehozására terveznek.[258] Ennek eredményeképpen minden új detektálási technológia előrelépése potenciálisan újabb,

[254] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023): Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.o.

[255] Több próbálkozás is született a helyzet orvoslására, melyek közül jelenleg az (IN2) tűnik a legígéretesebbnek. Az új INformation-INtensive (IN2) adatvezérelt képzési módszer hosszú kontextusból álló kérdés-válasz párokat használ, ahol a válaszok olyan kritikus információkra épülnek, melyek véletlenszerűen elhelyezett rövid szegmensekben találhatók. Ezáltal a modellek képesek kifinomultabban érzékelni az információkat, és integrálni azokat a hosszú kontextus különböző részeiről, ami jelentősen javítja azok teljesítményét hosszú szövegek feldolgozásában (lényegében IN2 képzés megtanítja a modellt, hogy a lényeges információk a hosszú kontextus bármely pontján megtalálhatóak lehetnek, nem csak a szélein). Az új képzési megközelítésről lásd bővebben: Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou (2024): Make Your LLM Fully Utilize the Context. 3-10.o. arXiv:2404.16811

[256] N. C. Köbis and L. Mossink (2021): Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, vol.114(2):106553 doi: 10.1016/j.chb.2020.106553

[257] Gyakran alkalmazott módszerek például a digitális vízjelek beillesztése, az időbélyegzés(timestamping), a metaadat alapú azonosítás, vagy a blockchain technológia.

[258] Lim, N., Kuan, M., Pu, M., Lim, M., & Chong, C. (2022): Metamorphic Testing-based Adversarial Attack to Fool Deepfake Detectors. *2022 26th International Conference on Pattern Recognition* (ICPR), 2503. https://doi.org/10.1109/ICPR56361.2022.9956543.

fejlettebb generatív MI megoldásokat inspirál, amelyek képesek kijátszani a meglévő biztonsági intézkedéseket. Ez az öngerjesztő folyamat — ahol minden új detekciós technológia fejlesztése újabb, kifinomultabb MI megoldás születését ösztönzi — egyfajta fegyverkezési versenyt idéz elő,[259] melyben a technológiai előrelépések nemcsak a biztonságot növelik, de hozzájárulnak a biztonsági intézkedések megkerülésére képes új technológiák fejlesztéséhez is.

5.7 (Gen)MI-t hoz a jövő?

A GenMI rendszerekhez kötődő modern diskurzusnak több új irányvonala is megjelent az utóbbi másfél évben. Ezek közé tartozik (1) a modellek tanításából eredő növekvő – és hosszútávon nem fenntartható – energiaszükséglet és környezetkárosítás kérdése; (2) Az emberközi kapcsolatok, illetve az ember-gép közötti kapcsolatok átalakulása; (3) A GenMI modellek munkaerőpiacra gyakorolt hatása; (4) Az új MI generációt képviselő autonóm ágensek (InteraktívMI) megjelenésének lehetséges hatásai.

(1) A nagyobb nyilvánosság előtt, egy a Carnegie Mellon Egyetem és a Hugging Face tech. vállalat által készített, 2023-ban megjelent tanulmány világított rá először arra, hogy a GenMI képgeneráló modellek felhasználói szintű használata aránytalan mennyiségű energiát emészt fel. Ezt jól értékelteti, hogy egyetlen képnek az előállítása – amely az OpenAI Dall-E alkalmazásában egy körülbelül 5 másodperces művelet – megközelítőleg egy okostelefon akkumulátorának teljes feltöltéséhez szükséges energiamennyiséget igényel.[260] Mivel a hatalmas adatközpontok, melyek az MI forradalmat táplálják, szintén egyre több energiát igényelnek, a Wells Fargo bankcsoport előrejelzése szerint 2030-ig – a GenMI rendszerek további elterjedésével párhuzamosan – akár 20 százalékkal is nőhet az Egyesült Államok elektromos energia iránti igénye.[261] Az energiaéhség problémáját fokozza, hogy ez a

[259] Sugumaran, D., John, Y., C, J., Joshi, K., Manikandan, G., & Jakka, G. (2023): Cyber Defence Based on Artificial Intelligence and Neural Network Model in Cybersecurity. 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 1-8.o. https://doi.org/10.1109/ICONSTEM 56934.2023.10142590.

[260] Luccioni, A. S., Jernite, Y., Strubell, E. (2023): Power hungry processing: Watts driving the cost of ai deployment? 13-14o. arXiv preprint arXiv:2311.16863

[261] Wells Fargo Bank (2023): Investor Presentation - 2023. 10.o. Elérhető: https://www08.wellsfargomedia.com/assets/pdf/about/investor-relations/presentations/2023/april-investor-presentation.pdf?utm_source=www.mindstream.news&utm_medium=newsletter&utm_campaign=ai-vs-energy-consumption (2023. 08. 12.)

technológiai forradalom nagyrészt fosszilis energiahordozókra épül. A kirívóan magas energiaszükségleteken túl az MI technológiák túlzott használata jelentős környezeti lábnyommal és ökológiai teherrel is jár. Csak érzékeltetésképpen a Google 2023-as *Környezeti Jelentése* szerint a cégcsoport 5,6 milliárd gallon vizet használt fel saját szerverei hűtésére.[262] Emellett az MI fejlesztés és alkalmazás miatt számottevően emelkedik a szén-dioxid kibocsátás is. A Massachusettsi Egyetem egyik kutatócsoportja már 2019-ben azt állította, hogy egyetlen MI modell betanítása több mint 626,000 font szén-dioxid-egyenértékű kibocsátást idéz elő, ami közel ötszöröse az átlagos amerikai autó egész életciklusa során keletkező kibocsátásnak (ideértve az autó gyártását is).[263]

(2) Az LLM rendszerek meggyőző nyelvi képességeik révén már most meglepően szoros és mély kapcsolatokat alakítanak ki a felhasználókkal. Annak ellenére, hogy e rendszereknek sem tudata, sem érzelmei nincsenek, az emberek hajlamosak kötődni hozzájuk pusztán a látszólagos intimitás miatt. [264] Ez könnyen kiszolgáltatottságot eredményezhet, kiváltképp, ha arra használják, hogy befolyásolják az emberek döntéseit, magatartását. Ahogy arról még a későbbiekben szó lesz, a modern digitális környezet jelenleg a figyelemért és a kattintásokért folytatott küzdelem terepe. Ez egy szempillantás alatt új, mélyebb szintre léphet, ahol a GenMI modellek a felhasználók gondolatainak, érzelmeinek és meggyőződéseinek befolyásolásáért folytatnak majd szüntelen küzdelmet. [265]

---

[262] Google (2023): Environmental Report. 49-55.o.

[263] Strubell, Emma, A. Ganesh, A. McCallum (2019): Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243

[264] Egyre több platform törekszik arra, hogy olyan személyre szabott MI-társakat kínáljon, amelyek nem csupán információkkal szolgálnak, de érzelmi kapcsolatokat is képesek kiépíteni a felhasználókkal. E digitális felületek engedélyezik az MI karakterek számára, hogy emlékezzenek a korábbi beszélgetésekre, figyelembe vegyék a felhasználói preferenciákat, és így hosszabb távú, személyre szabott interakciókat nyújtsanak. A Character.AI platform például, amelyet 2022-ben indítottak, lehetővé teszi a felhasználóknak, hogy különböző karakterekkel (köztük hírességekkel, kitalált szereplőkkel, valódi ismerősök által ihletett avatárokkal) beszélgessenek, valamint, hogy saját a profilukhoz kötődő chatbotokat is létrehozzanak. A platformon naponta több millió felhasználó aktív (közülük sokan fiatalok), akik saját bevallásuk szerint valós érzelmi kapcsolatokat alakítanak ki a digitális karakterekkel, mivel úgy érzik, hogy könnyebben tudnak őszintén beszélni velük, mint a környezetükben lévő emberekkel. Emellett az MI karakterekkel való beszélgetés segít nekik felfedni saját érzéseiket is, anélkül, hogy ítélkezéstől kellene tartaniuk. Vannak akik a nyelvi modelleket egyfajta támogatóként / mentorént használják, és arról számolnak be, hogy az MI társak segítettek nekik nehéz időszakokban átvészelni a különböző megpróbáltatásokat. Lásd bővebben: Lucas, Jessica (2024): The teens making friends with AI chatbots. *The Verge*. Elérhető: https://www.theverge.com/2024/5/4/24144763/ai-chatbot-friends-character-teens (2024. 05. 10.); Chow, R. Andrew (2023): AI-Human Romances Are Flourishing—And This Is Just the Beginning. *TIME*. Elérhető: https://time.com/6257790/ai-chatbots-love/ (2024. 05. 10.)

[265] Harari a gép-ember közötti kapcsolat kialakulásának társadalmi hatásain kívül felhívja a figyelmet az ilyen rendszerek kultúraformáló erejére is. A Nagy Nyelvi Modellek immár képessé váltak arra, hogy az emberi képzeletet is gyökeresen átformálják azáltal, hogy egészen új kulturális narratívákat, szimbólumrendszereket, törvényeket és hiedelemvilágokat teremtsenek számunkra. A meggyőzés erejével könnyen átvehetik a kultúratermelés feletti irányítást is. Lásd bővebben: Harari, Y. N. (2023): Yuval Noah Harari argues that AI has

(3) A GenMI megjelenése jelentős – egyúttal diszruptív – hatással lehet számos ma ismert szakmára és munkakörre. Bár az évezred elején még az volt az uralkodó vélekedés[266], hogy az MI és a robotizáció elterjedése elsősorban a fizikai munkások megélhetését fogja veszélyeztetni, a 2020-as évek elejére kijelenthető, hogy a GenMI *fehérgallérossá* vált.[267] Csak néhány olyan területet és foglalkozást kiemelve, melyeket veszélyeztet – vagy alapjaiban alakíthat át – a technológia; a szöveges tartalomgyártás területén dolgozók, mint az újságírók, szövegírók, bloggerek és forgatókönyvírók szembesülhettek elsőként komoly kihívásokkal, minthogy a Nagy Nyelvi Modellek már most is képesek gyorsan és nagy mennyiségben létrehozni emberi kreativitási szintet elérő szövegeket. Általánosságban véve a nyelvi készségeket igénylő munkakörök egészen biztosan nem maradnak érintetlenek, így a fordítók és tolmácsok foglalkoztatottsága is csökkenni fog, ahogy az LLM fordítóprogramok egyre pontosabbá válnak (de ide értendő például az irodalmi lektorok és szerkesztők munkája is). Ezzel párhuzamosan a kreatív-vizuális területeken is érezhető lesz a technológia kiszorító hatása: a grafikai tervezők, illusztrátorok, logótervezők és web-designerek munkája is átalakul a képgeneráló GenMI fejlődésével.

Az ügyfélszolgálati szektorban is komoly átrendeződés várható, mivel a chatbotok és virtuális asszisztensek fokozatosan képesek lesznek átvenni munkaköri feladatok meghatározó részét. A jogi területen a jogi asszisztensek és kutatók munkája részben automatizálható lehet, különösen az információfeldolgozás és -elemzés terén. Az adatkezeléssel foglalkozó szakmák

---

hacked the operating system of human civilisation. The Economist. https://www.economist.com/byinvitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-systemof-human-civilisation (2023. 04. 28.). E jelenséggel összefüggésben, Harari legújabb könyvében a szabad demokráciák előtt álló kihívásokról és a jövőbeli autoriter rendszerek felemelkedésének veszélyéről is értekezik: Lásd bővebben: Harari, Yuval Noah (2024): Nexus: A Brief History of Information Networks from the Stone Age to AI. New York, Random House. 305-361.o.

[266] Például Martin Ford is erről ír *A robotok kora* c. könyvében. Érdekesség, hogy ugyan legtöbben egyetértettek Ford álláspontjával a robotikai fejlesztések munkaerőpiacra gyakorolt hatásával kapcsolatosan, az un. *Moravec paradox* mégis már az 1980-as években viszonylag pontos előrejelzést adott az MI technológiai fejlődésének ívéről. E kifejezés a robotika és az MI kutatás azon meglepő következtetésére utal, mely értelmében látszólag egyszerű szenzomotoros tevékenységek (tárgyak megragadása, akadálytalan mozgás, gépi térlátás) végrehajtása lényegesen nehezebb, összetetteb feladat, ráadásul sokkal több számítási erőforrást is igényel annál, mint a rendszereket intelligens képességekkel történő ellátása (absztrakt logikai gondolkodás, magas szintű fordítás, integrálás, sakkjáték szabályainak elsajátítása, tételbizonyítás stb.). Hans Moravec kanadai-amerikai tudós, 1988-ban közzétett könyvében úgy fogalmaz, hogy viszonylag könnyű olyan szintre fejleszteni az MI rendszereket, hogy azok „*a felnőttekéhez mérhető szintű eredményeket érjenek el intelligenciateszteken vagy a dámajátékban, de a lehetetlent súrolóan nehéz eljuttatni őket az egyéves gyerekek szintjére az észlelés és manőverező képesség területén*". Moravec, Hans (1988): Mind Children. Cambridge, Harvard University Press. 15-16.o.

[267] A kifejezést Tilesch György MI-kutató használta 2023 szeptember 11-én a Magyar Zene Házában megrendezett az Economx AI Summit konferencián. Lásd: Németh Tamás (2023): Tilesch György: Sokan még mindig a Transformers-filmek szintjén kezelik a mesterséges intelligenciát. EconomX. Elérhető: https://www.economx.hu/gazdasag/ai-summit-2023-tilesch-gyorgy-mesterseges-intelligencia-kkv-k-munkaltatok.777160.html (2023. 09. 11.)

munkája is hamar veszélybe kerülhet, de a szoftverfejlesztés és programozás területén is – főleg az egyszerűbb kódolási feladatok esetében, melyek már ma is automatizálhatóak[268] – változást hoz a következő néhány év. Szintén idesorolható még a digitális marketing területe is, minthogy a közösségi média menedzserek és SEO szakértők munkája nagyrészt kiválthatóvá válik az MI-alapú fejlett tartalomoptimalizálás révén.

(4) A Google DeepMind társalapítója, Mustafa Suleyman úgy véli, hogy a GenMI csak rövid, átmeneti szakaszt jelent a technológiaicsalád fejlődésében. A következő hullám az autonóm és célirányos cselekvésekre képes interaktív MI lesz. Amíg a jelenlegi generatív modellek bemeneti adatokból új tartalom előállítására képesek, az interaktív MI ezen túl azt is lehetővé teszi majd a felhasználói számára, hogy különböző online feladatok elvégzését is rábízzák.[269] Egy személyre szabott digitális asszisztenshez hasonlóan képes lesz a felhasználók eszközein másokkal interakcióba lépni, és különböző munkafolyamatokat végrehajtani (útiterv-készítése, repülőjegy-foglalás, éltelrendelés stb.).

Az előbbiekben ismertetett négy lehetséges irányvonal – valamint az azokhoz kapcsolódó kockázatok és szabályozási stratégiák – feltérképezése egyenként is fontos kutatási témával szolgálnának. Az elmúlt másfél évben azonban konszenzus látszik kialakulni abban, hogy a GenMI modellek legégetőbb – rövid távú – kihívását az jelenti, hogy az általuk generált megközelítőleg korlátlan mennyiségű, olcsón előállítható szintetikus szöveges és audiovizuális tartalmakat a dezinformációs- és spamkampányoktól kezdve, a pénzügyi

[268] A programkód-generálás képességével összefüggésben érdemes megemlíteni, hogy az OpenAI 2024 szeptemberében teszi majd nyilvánosan is hozzáférhetővé az o1 nevezetű legújabb GPT modelljét. Ennek különlegessége, hogy megerősítéses tanulást és „*chain-of-thought*" (gondolatmenet-lánc) módszert alkalmaz, ami lehetővé teszi, hogy összetett problémákat lépésről lépésre oldjon meg, fokozatosan bontsa le a bonyolult problémákat kisebb, könnyebben kezelhető lépésekre, és ezen keresztül fejlessze a saját megoldási stratégiáit. A modell képes önállóan felismerni és javítani a hibákat anélkül, hogy emberi beavatkozásra lenne szüksége. A megerősítéses tanulásnak köszönhetően az o1 modell folyamatosan finomítja érvelési technikáit, ami javítja az olyan összetett feladatokban nyújtott teljesítményét, ahol a több lépéses érvelés és a logikai megközelítések központi szerepet játszanak (például a matematikai és kódolási folyamatokban). Lásd: OpenAI (2024): Introducing OpenAI o1 Preview. Elérhető: https://openai.com/index/introducing-openai-o1-preview. (2024. 08. 20.); Kim, Eugene (2024): In a leaked recording, Amazon cloud chief tells employees that most developers could stop coding soon as AI takes over. *Business Insider.* Elérhető: https://www.businessinsider.com/aws-ceo-developers-stop-coding-ai-takes-over-2024-8 (2024. 08. 20.)

[269] Az interaktív MI/autonóm ágensek olyan rendszerek, amelyek képesek önállóan döntéseket hozni és cselekedni anélkül, hogy közvetlen emberi beavatkozásra lenne szükségük Bővebben: Suleyman, Mustafa, Michael Bhaskar (2023): A következő hullám. Mesterséges intelligencia, technológia, hatalom és a 21. század legnagyobb kihívása. (ford. Farkas Veronika). Budapest, Magnólia kiadó; Will Douglas Heaven (2023): DeepMind's cofounder: Generative AI is just a phase. What's next is interactive AI. *MIT Technology Review.* Elérhető:https://www.technologyreview.com/2023/09/15/1079624/deepmind-inflection-generative-ai-whats-next-mustafa-suleyman/ (2023. 09. 15.)

csaláson és személyazonosság-lopáson át,[270] a politikai- és választási manipulációig bezárólag számtalan célra fel lehet használni.[271] Teljesen nyilvánvaló az is, hogy a GenMI modellek által létrehozott szintetikus tartalmak példátlan gyorsaságú elterjedése hosszútávu hatással bír majd a körülöttünk lévő online és offline információs közegre. Ennek feltérképezésére vállalkozik a következő fejezet. A folyamat megértéséhez egészen az internet létrejöttének kezdetéig érdemes visszatekinteni.

[270] Annak egyik lehetséges következményét, hogy gyakorlatilag mindenki számára lehetségessé vált, a magas minőségű szöveges tartalmak nagy mennyiségben történő előállítása jól érzékelteti, hogy a ChatGPT 2022 novemberi megjelenése óta több, mint 1.260%-kal nőtt a hitelesnek tűnő, tehát kevesebb nyelvtani hibát tartalmazó és koherens szövegszerkezettel bíró (ezáltal hatékonyabb és eredményesebb) adathalász e-mailek száma. Lásd: Nelson, Jason (2023): Email Phishing Attacks Up 1,265% Since ChatGPT Launched. EMERGE. Elérhető: https://decrypt.co/203564/since-chatgpt-launch-phishing-emails-are-up-1265-slashnext

[271] 2023 januárjában számolt be arról a New York Times, hogy az OpenAI felfüggesztette egy olyan felhasználójának fiokját, aki a demokrata elnökjelölt-aspiráns Dean Phillips képviselőt megszemélyesítő automatizált botba integrálta a cég nagy nyelvi modelljét. A *Dean Bot*, amely valós időben tudott kommunikálni a szavazókkal egy weboldalon keresztül, ugyan tartalmazott figyelmeztetést arra vonatkozóan, hogy nem a valódi Dean Phillips, az OpenAI mégsem engedte annak alkalmazását. A cég azzal indokolta a döntést, hogy belső szabályzatuk tiltja a technológiájuk politikai kampányokban történő olyan jellegű felhasználását, amely során az érintett beleegyezése nélkül visszaélnek annak a személyazonosságával. Dwoskin, Elizabeth (2024): OpenAI suspends bot developer for presidential hopeful Dean Phillips. *The Washington Post*. Elérhető: https://www.washingtonpost.com/technology/2024/01/20/openai-dean-phillips-ban-chatgpt/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=sam-altman-s-secret-ai-chip-venture (2024. 01. 20.)

6 Szép új (online) világ

6.1 Orwell vagy Athén?[272]

Az internet térnyerését kezdetben széles körű lelkesedés és optimizmus kísérte. E pozitív hangulat abból a reményből fakadt, hogy a világháló 1990-es évektől kezdődő[273] terjedésével megindulhat majd az információ és tudás egyetemes demokratizálódása; a hagyományos információáramlási struktúrák átalakulásával a történelem során felhalmozott ismeretek széles körben elérhetővé és hozzáférhetővé válhatnak. A világháló új csatornákat biztosított az információ megosztására, lebontva azon időbeli, térbeli és anyagi korlátokat, amelyek korábban akadályozták a tudáshoz való egyetemes hozzáférést. Az országok, kormányok, állampolgárok közötti kommunikációs korlátok csökkenésével az emberek a világ legtávolabb pontjain is könnyen kapcsolatba léphettek egymással, elősegítve ezzel a kulturális információs cserét, és a nemzetközi együttműködést.[274]

Egy rövid ideig úgy tűnt, hogy az információ demokratizálódásán túl az internet aktív szerepet játszhat a politikai demokratizálódási folyamatokban is. A különböző online platformok és a közösségi média – mint *felszabadító technológia* (liberation technology)[275] – lehetőséget teremtettek az állampolgároknak arra, hogy gyakorolják az alapvető, első

[272] A kérdés „*Orwell vagy Athén*?" az informatikai fejlesztések és a demokrácia jövője közötti dilemmát hivatott érzékeltetni. A gondolatmenet egy 1995-ös tanulmányból származik, amely a demokrácia működését és az információs technológiák hatását elemzi. Az alapvető kérdés az, hogy a digitális technológiák hozzájárulnak-e a demokratikus részvétel növeléséhez és a nyitott társadalmak erősítéséhez (Athén, mint szimbólum), vagy inkább az autoriter megfigyelés és a központi irányítás világát teremtik meg (Orwell, utalva az „1984" című regényre). Lásd bővebben: van de Donk, W. B. H. J., & Tops, P. W. (1995): Orwell or Athens? Informatization and the Future of Democracy. In W. B. H. J. van Donk, I. T. M. Snellen, & P. W. Tops (szerk.): Orwell in Athens: a Perspective on Informatization and Democracy. IOS Press. 13-32.o.

[273] Tim Berners-Lee 1989-ben hozta létre a máig leghíresebb „World Wide Web" globális információs rendszert, amely a különböző dokumentumokat un. hiperhivatkozással kapcsolta össze. Ugyan a rendszert 1991-ben már használták kutatók, a szélesebb közönség számára csak a 90-es évek első felében, a webböngészők elterjedésével vált ismertté. (ezek közül a leghíresebb az 1993 április 22.én megjelent Mosaic böngésző).

[274] Az internet kezdeti fejlődéséről lásd bővebben: Castells, Manuel. (2002): Az Internet-galaxis. Gondolatok internetről, üzletről és társadalomról. Budapest, Network Twenty One Hungary kiadó.

[275] A liberation technology kifejezés népszerűsítése Larry Diamond amerikai politikai szociológushoz köthető. Diamond a kifejezést annak leírására alkalmazta, hogy miként tudják az információs és kommunikációs technológiák (mint például a közösségi média) elősegíteni a demokratikus folyamatokat, az emberi jogok védelmét és a politikai aktivizmust az autoritás és elnyomás elleni harcban. Bővebben lásd: Diamond, Larry (2010): Liberation Technology. *Journal of Democracy*, vol. 21, no. 3, 69-83.o.

generációs polgári és politikai jogaikat, önként szerveződjenek és szabadon fejezzék ki véleményüket ott, ahol a hagyományos média korlátozott vagy szoros állami ellenőrzés alatt állt. Ennek jelentősége különösen az arab tavasz eseménysorozata során vált nyilvánvalóvá, amikor is a Facebook és a Twitter központi szerepet játszott – az esetek többségében alulról szerveződő – reformpárti tüntetések és politikai mozgalmak koordinálásában és lebonyolításában.[276]

Azonban az online tér iránt érzett kezdeti lelkesedés korán alábbhagyott. Az idő előrehaladtával az internet egyre kevésbé testesítette meg a születésekor vizionált szabad információáramlást biztosító, korlátozásoktól és beavatkozásoktól mentes demokratikus környezetet. Napjainkban az online térben történő információgyűjtést, a digitális tartalmakhoz való hozzáférést kormányok vagy vállalatok által működtetett algoritmus alapú döntések szabályozzák, amelyek működését – olykor a kelleténél kevesebb átláthatóság és közösségi felügyelet mellett – a hatalom kiteljesítésének vagy a profit maximalizálásának szándéka motiválja. Az információs közeg ahelyett, hogy tájékozottabbá tette volna a felhasználókat sok esetben kételyt ébreszt és dezinformál, ahelyett, hogy közelebb hozta volna az embereket, inkább ellentéteket gerjeszt és polarizál. Ahhoz, hogy ennek az okait és következményeit megértsük, a jelenleg ismert online környezet rövid fejlődéstörténetét és működésének sajátosságait szükséges áttekinteni.

6.2 Zajos terek, kétes információk

Az internet megjelenésének kezdetén – a Web 1.0 korszakában – a weboldalak statikus jellegűek voltak, melyek főként egyszerű HTML forráskódból álltak, így a felhasználók számára kizárólag előre megírt szöveges és képi tartalmakat jeleníthettek meg. E webhelyek

[276] A fentebb említett online platformok és közösségi médiumok (pl.: Facebook, Twitter) nagyban segítették a tüntetéshullámban érintett országokban a kollektív fellépés sikeres megszervezését. Ez köszönhető volt többek között annak, hogy gyorsan és valós időben voltak képesek tömegek számára információt terjeszteni a szervezett tüntetésekről, a hatóságok közbeavatkozásáról, a várható rendőri jelenlétről stb. Segítségükkel széles közönséghez juthattak el az adott ügyekkel kapcsolatos hírek, és olyan embereket is mozgósíthattak, akiket hagyományos eszközökkel talán nem értek volna el (a helyi és regionális résztvevőkön túl, a nemzetközi sajtót, megfigyelőket és egyéb támogatókra is ideértve). A témáról lásd bővebben: Howard, Philip N. és Muzammil M. Hussain (2013): Democracy's Fourth Wave? Digital Media and the Arab Spring. Oxford University Press; Tufekci, Zeynep. (2017): Twitter and Tear Gas: The Power and Fragility of Networked Protest. Yale University Press

csupán egyirányú kommunikációra nyújtottak lehetőséget: a tartalmat a weboldal üzemeltetői helyezték el, míg az oda látogatók pusztán fogyasztói lehettek a közreadott információknak, aktív interakciós vagy hozzászólási lehetőség nélkül. A felhasználó tehát hozzáférhetett az információhoz, de nem oszthatta meg saját tartalmát. A ma ismert keresőmotorok még gyerekcipőben voltak, a közösségi média pedig nem is létezett. A kétezres évek elején jelent meg az internet új generációja – a Web 2.0 – ami radikálisan átformálta a digitális világot. A Web 2.0 már egy dinamikus, felhasználóközpontú digitális környezet, ahol a tartalom nem csak olvasható, hanem szerkeszthető, bővíthető és módosítható is egyben. A kommunikációs csatornák kétirányúvá váltak a blogokon, fórumokon és komment szekciókon keresztül. Megjelentek a könnyen kezelhető tartalomkezelő rendszerek (CMS), amelyek lehetővé tették a felhasználók számára a tartalom egyszerű közzétételét és szerkesztését, az egyszerű szövegalapú HTML weboldalak mellett elterjedtek a multimédiás tartalmak, (pl. audiok, GIF-ek videók), a keresőmotorok pedig hatékony eszközzé váltak a webes tér navigálásához.[277] A közösségi média megjelenése forradalmasította az emberek közötti interakciók formáit, intenzitását és gyakoriságát. Ez az új korszak egy sokkal interaktívabb környezetet teremtett, melyben a blogok, közösségi hálózatok, és videómegosztó oldalak lehetővé tették, hogy bárki saját tartalmat hozzon létre és osszon meg.[278] Ennek eredményeképpen összemosódtak a határok a tartalmak fogyasztása és előállítása között, a felhasználó egyidejűleg tartalomfogyasztóvá és tartalomgyártóvá is vált.[279]

A Web 2.0 elterjedésével a hagyományos kapuőri[280] szerep is gyökeres változáson ment keresztül, melynek köszönhetően egy rövid ideig szinte bárki közzé tehette az általa

[277] O'Reilly, T. (2005): What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, No.65 (1), 17-37.o.

[278] Szőke Gergely László (2012): Az Európai adatvédelmi jog megújítása. Tendenciák és lehetőségek az önszabályozás területén. PTE-ÁJK, 57-58.o.

[279] E jelenség leírására gyakran alkalmazzák az Alvin Toffler amerikai író és jövőkutató által 1980-ban bevezetett *prosumer* kifejezést, amely a „producer" (gyártó) és „consumer" (fogyasztó) szavak összevonásából keletkezett. A kifejezés a fogyasztói szerepkör megváltozását hivatott tükrözni, akik immár a passzív fogyasztáson túl, aktívan részt vehetnek a termékek és szolgáltatások létrehozásában, formálásában is Lásd: Toffler, Alvin (1980): The Third Wave. New York, Bantam Books. A későbbiekben még utalás szintjén szó lesz arról is, miként segíti a felhasználók kettős szerepe a dezinformáció, az álhírek, az áltudományos teóriák és konspirációs elméletek terjedését.

[280] A kapuőr (gatekeeper) fogalma a kommunikációtudományok területén olyan személyt, jogi személyt takar, aki vagy amely dönt arról, hogy mely információk jutnak el a szélesebb közönséghez. A kapuőri szerep a hagyományos médiában – nyomtatott sajtóban, televízióban, rádióban – még leggyakrabban az újságírókhoz, kiadványszerkesztőkhöz vagy hírigazgatókhoz kötődött. Ők döntöttek arról, hogy mely hírek, cikkek vagy jelentések érdemelnek figyelmet, és hogyan kerülnek azok bemutatásra. Ez a szerep az internet 2.0 kezdetekor egy rövid ideig demokratizálódni látszott, majd irányítását átvették a piacvezető tech. vállalatok által működtetett közvetítő szolgáltatások (pl.: keresőmotorok, közösségi média platformok).

létrehozott tartalmat anélkül, hogy ezt hagyományos szerkesztői folyamatok szűrték volna.[281] Kétségtelen, hogy ez az új online környezet jelentősen bővítette az elérhető információk körét, de egyben komoly kihívást is állított a digitális tartalmak hitelességének és minőségének megőrzése kapcsán.

A kétezres évek elején a korlátozásoktól mentes – de számottevő információs zajtól terhes – digitális környezetet fokozatosan átalakította azon *online közvetítők* megjelenése, amelyek a tartalomszolgáltatók és a felhasználók közötti kapocsként egyre jelentősebb szerepet játszottak annak eldöntésében, hogy kihez, miként és milyen információ jutott el az online térben (illetve, hogy melyek maradjanak rejtve vagy váljanak nehezen hozzáférhetővé).

Ezek a közvetítők, mint például a közösségi média platformok és a keresőmotorok jellemzően algoritmusok segítségével szűrik, rangsorolják és szabják személyre a tartalmakat. Az ilyen ajánlóalgoritmusok döntéseikben nagyban támaszkodnak az online személyiségprofilokra, amelyeket a fokozatosan *digitalizálódó egyén* internetes aktivitásából származó adatokból építenek fel.[282] Az, hogy milyen módon működik az algoritmus, amely dönt a felhasználó számára hozzáférhetővé tett tartalmakról, kizárólag e rendszerek üzemeltetőin múlott. A felhasználók ebből eredő kiszolgáltatottságát pedig a piaci környezetre jellemző koncentrálódás csak tovább növelte.[283]

Ez az új online ökoszisztéma egy olajozottan működő digitális gazdasági modellé fejlődött, ahol a felhasználók információs szabadságának korlátozása nem szándékos cél, hanem inkább egy elkerülhetetlen következmény. A rendszer legfőbb célja, hogy a felhasználók vélt vagy

[281] Ezzel kapcsolatban Török Bernát így fogalmaz; „*A közösségi média – működésének első időszakában – megteremtette a társadalmi párbeszéd kapuőrök nélküli szféráját, ahol nem lap- vagy tévétulajdonosokon, nem is szerkesztőségeken, de még csak nem is újságírókon, hanem mindenekelőtt a megszólalón múlik, hozzászól-e közügyeinkhez*." Lásd: Török Bernát (2022): A szólásszabadság a közösségi platformokon és a Digital Services Act. In: Török Bernát és Ződi Zsolt (szerk.): Az internetes platformok kora. Budapest, Ludovika Egyetemi Kiadó, 197.o.

[282] E felhasználói profilok ahogy egyre növekvő adatforrásból táplálkoznak úgy válnak lényegesen összetettebbé is. Lásd bővebben: Pataki Gábor – Szőke Gergely László (2017): Az online személyiségprofilok jelentősége – régi és új kihívások. *Infokommunikáció és jog*, 2017/2., 63–70.o.

[283] Ezt többek között a digitális gazdaság kapcsán gyakran emlegetett *hálózati hatás* jelensége magyarázza: A ma ismert tech. vállalatok jelentős része azáltal tudta idejében leuralni az online piacot, hogy ők rendelkeztek elsőként számottevő felhasználói bázissal. A domináns helyzetük abból fakadt, hogy a felhasználók nem kerestek alternatív szolgáltatásokat, hiszen a már kiválasztott platformjuk vonzerejét éppen az adta, hogy az ismerőseik közül sokan szintén azt használták. Ezt erősíti az un. *hólabda hatás* is, mely szerint, ha egy szolgáltatásnak elegendő felhasználója van sokkal könnyebben válik számára terméke minőségének javítása, ezzel létrehozva egy öngerjesztő folyamatot, amelyben a javuló minőség még több felhasználót vonz, ami további fejlesztést tesz lehetővé és így tovább. Az Uber esetében például minél több sofőr és utas csatlakozik a platformhoz, annál gyorsabb és hatékonyabb a szolgáltatás (rövidebb várakozási idő/több elérhető autó), az eBay online aukciós oldal esetében minél több eladó és vevő használja az oldalt, annál nagyobb a választék és a vásárlási lehetőségek.

valós, és néha mesterségesen befolyásolt preferenciáit kiszolgálja, majd az ebből fakadó online aktivitást monetizálja.

6.3 A figyelmed eladó

A figyelemgazdaság korában – ahogy ötletes elnevezése is sugallja – a szolgáltatók közötti verseny a felhasználók figyelmének megragadásáért és megtartásáért zajlik. Ebben a környezetben – ahol a figyelem értékes árucikknek számít – a tartalomszűrő algoritmusok azon tartalmakat rangsorolják előre és teszik láthatóvá, amellyel leginkább képesek lekötni a felhasználót. Itt az „ingyenes" szolgáltatásokért és tartalmakért a felhasználó figyelmével, interakcióival és az azokból kinyerhető adatokkal fizet. Minél hosszabb idejű és változatosabb a felhasználói aktivitás, annál nagyobb a potenciálisan realizálható profit.[284] Ez a működési logika szorosan összefügg az olyan népszerű közösségi médiaplatformok kialakításával is, mint a Facebook, a YouTube vagy a TikTok. Ezek a platformok szándékosan aknázzák ki az emberi agy dopamin-válaszát[285] annak érdekében, hogy fokozzák a felhasználói érdeklődést és interakciókat.[286] Amikor a felhasználó pozitív visszajelzést kap a közösségi médiában – legyen az egy *kedvelés* a bejegyzése alatt vagy egy új, automatikusan feldobott videó a TikTokon –, az agyban dopamin szabadul fel, ami kellemes érzést vált ki. Ez a folyamat egyfajta jutalomként működik, ami arra motiválja a felhasználót, hogy folytassa azon tevékenységeket, amelyek ezt a pozitív érzést előidézik. Ezzel kapcsolatban három pszichológiai hatás kiemelése indokolt;

(1) A felhasználók egyre több időt töltenek az ilyen közösségi média felületeken, folyamatosan keresve a jutalmazó interakciókat, ami könnyen függőséghez, depresszióhoz és

[284] Tim Wu (2016): The Attention Merchants: The Epic Scramble to Get Inside Our Heads. New York, Knopf.
[285] A dopamin egy a központi idegrendszerben termelődő neurotranszmitter, amely kiemelt szerepet játszik az agy jutalmazási és motivációs rendszereiben. Rendelkezésre állása (vagy nem állása) nagyban befolyásolja az örömérzetet, az elkötelezettséget, és a motivációt. Mivel dopamin rendszer aktiválása jutalmazó érzetet, örömöt vagy elégedettséget okoz, hozzájárulhat bizonyos viselkedések megerősítéséhez és ismétléséhez. A modern társadalmakra (kiváltképp a mai fiatalokra) jellemző dopaminéhségről és dopaminfüggőségről tudományos részleteséggel, de olvasmányos formában értekezik 2021-ben megjelent könyvében Anna Lembke, amerikai pszichiáter. A könyv idén magyar fordításban is megjelent. Lásd: Lembke Anna (2024): Dopaminkorszak – Hogyan találjunk egyensúlyt a függőségekre épülő világban. (Ford.: Bujdosó István). Budapest, Libri Könyvkiadó Kft.
[286] Ilyen technikák például az értesítések gyakori megjelenítése, a végtelen görgetés, vagy az automatikusan lejátszódó videók.

szorongás kialakulásához vezethet;[287] (2) A dinamikus és ingergazdag online környezetben a folyamatosan felkínált új tartalmak – mint dopamin stimulánsok – hozzájárulnak ahhoz, hogy a digitális bennszülöttek figyelemküszöbe drasztikusan csökkent. Bár ez kedvez a rövid, egyszerű tartalmak népszerűségének (nem véletlen, hogy a legfiatalabb generáció körében már a TikTok a második leglátogatottabb platform)[288], egyúttal korlátozza a hosszabb, összetett, komplex tartalmak megértésének és értelmezésének képességét;[289] (3) Az új tartalmak folyamatos, automatikus felkínálása révén a felhasználók akaratukon kívül könnyen un. virtuális „*nyúlüregekbe*" (rabbit hole) találhatják magukat. E kifejezés egy olyan automatizált információ-torzító folyamatra utal, mely során a felhasználó egy meghatározott témával kapcsolatban kezd el információkat keresni vagy tartalmat fogyasztani, de fokozatosan, lépésről-lépésre a kezdeti témától egyre távolabb eső – és ahhoz egyre kevésbé kapcsolódó – tartalmakhoz jut el az algoritmusok ajánlásai révén. Ez gyakran észrevétlenül történik, és mire a felhasználó feleszmél, már teljesen más témájú tartalmak kerülnek elé, mint amire eredetileg kíváncsi volt. Ez egyfajta akaratlan eltávolodást jelent a kezdeti érdeklődési körtől azáltal, hogy az algoritmusok révén folyamatosan felkínált, de apránként eltérő ajánlások elcsábítják a felhasználót új irányokba. Ezzel összefüggésben említést érdemel az algoritmikus radikalizáció jelensége is, ami azt a folyamatot írja körbe, hogy az

[287] Jonathan Haidt, amerikai pszichológus, a New York University professzora szerint a közösségi média jelentős szerepet játszik a tinédzserek mentális egészségének romlásában. Kutatásai azt mutatják, hogy az elmúlt évtizedben jelentősen megnőtt a depresszió, szorongás, önsértés és az öngyilkossági kísérletek aránya a serdülők körében. Ezek szintje kimutathatóan magasabb azok körében, akik naponta több mint két órát töltenek a közösségi média felületeken. Haidt, Jonathan, Nick Allen (2020): Scrutinizing the effects of digital technology on mental health. *Nature*. Vol 578, 226-227o.;

[288] A Pew Research Center 2023-as felmérése szerint a YouTube, TikTok és az Instagram a legnépszerűbb platformok a tinédzserek körében. A felmérés alapján a YouTube-ot használja a legtöbb fiatal, míg a TikTokot 63%, az Instagramot pedig 59% használja. A fiatalok körében a közösségi média használatának gyakorisága is magas, sokan „*szinte folyamatos*" (almost constant use) használatról számoltak be, különösen a TikTok esetében (16-17%). A Facebook és Twitter (X) használata visszaesett, a Facebook felhasználók aránya 2014-2015 óta 71%-ról 33%-ra csökkent. Lásd: Pew Research Center (2023): Teens, Social Media and Technology 2023. 5-6.o. Elérhető: https://www.pewresearch.org/wp-content/uploads/sites/20/2023/12/PI_2023.12.11-Teens-Social-Media-Tech_FINAL.pdf

[289] Az *információs túlterhelés* (information overload) valamint az elmélyedést és alaposabb megértést akadályozó tartalomfogyasztási szokások (kiegészítve a csökkenő figyelemfenntartási képeségekkel) könnyen az un. Dunning-Kruger hatás felerősödéséhez vezethetnek. A kifejezés arra a pszichológiai jelenségre utal, miszerint egy adott témában alacsonyabb szintű tudással, gyakorlattal rendelkező személyek rendszerint túlértékelik saját képességeiket vagy esetleges teljesítményüket a területen. A szerzőpáros által lebonyolított kísérlet szerint ez arra vezethető vissza, hogy e személyek a témához kötődő koherens tudásuk hiányában képtelenek voltak olyan önreflexióra, mellyel felismerhetnék az adott témához kapcsolódó tudásbéli, teljesítménybéli korlátiakat. A dolgozat témájához kötődve fontos kiemelni azt is, hogy a tudásdeficit nem csak saját magukkal szembeni pozitív elfogultságot okozott, hanem egyúttal azzal is járt, hogy képtelenek voltak felismerni és megkülönböztetni azt, ha egy a témában valóban jártas szakértő gondolatait hallották. Bővebben a kísérletről: Kruger, Justin & Dunning, David. (1999): Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology,* 77(6), 1121–1134.o; Az információs túlterhelésről részletesebben: Koltay Tibor (2017): Egy „örökzöld téma": az információs túlterhelés. *Információs Társadalom*, XVII. évf. 3. szám, 39–54.o.

ajánló algoritmusok miként hajlamosak fokozatosan egyre szélsőségesebb tartalmak felé taszítani a felhasználókat. Egyre több tanulmány hívja fel arra a figyelmet, hogy e folyamat különösen veszélyezteti a fiatal férfi felhasználókat, akiket sokszor ártalmatlannak tűnő tartalmakkal, például mentális és fizikai egészséghez kapcsolódó tippekkel, vagy gazdasági tudatosságra és pénzszerzésre vonatkozó tanácsokkal vonzanak be a magas követőszámmal rendelkező influenszerek. Ebben a nyúlüregben a közösségi médiában népszerű – a fiatal identitáskeresés nehézségeire összpontosító hasznos életmódtanácsokat adó – tartalmakat rövid idő alatt részben vagy egészben felváltják a nemek közötti egyenlőség eszményét és a szexuális kisebbségek jogait támadó, sok esetben szélsőjobboldali, soviniszta vagy antiszemita ideológiákat hirdető videók. Idővel ezek a tartalmak még inkább szélsőségesebb világnézetek felé vezetik a felhasználót, kihasználva az identitáskeresés nehézségeit és a társadalmi frusztrációkat.[290]

6.4 A jövőd (is) eladó?

Shoshana Zuboff amerikai szociálpszichológus 2019-ben megjelent könyvében a figyelemgazdasághoz hasonló gazdasági modellről, a *megfigyelési kapitalizmus*ról ír részletesen. Zuboff arra a következtetésre jut, hogy az a figyelemgazdaság korára még érvényes mondás, miszerint, *ha valamilyen szolgáltatás ingyenesen veszel igénybe, akkor te magad vagy a termék*, már nem fedi le a teljes valóságot. A tényleges helyzet ennél sötétebb képet mutat, hiszen a felhasználók ebből az eleve kiszolgáltatott termék pozícióból mára leértékelődtek egyszerű nyersanyaggá, melynek elsődleges szerepe, hogy a belőle kinyert adatok bemenetként szolgáljanak a jövőbeli magatartások előrejelzésére programozott algoritmusoknak. Ezzel párhuzamosan a valódi termékekké a felhasználók jövőbeli magatartásának előrejelzései váltak. Ezeket az *előrejelzési termékeket* (prediction production) különböző piaci szereplőknek értékesítik tovább, akik ezáltal hatékonyabban tudják befolyásolni a fogyasztót.[291] Ha a cégek elegendő viselkedési adattal rendelkeznek lehetővé

[290] A témával kapcsolatos új trendekről – főként a Tik-Tok videó és a youtube short tartalmakra fokuszálva - lásd bővebben a Dublini egyetem 2024-es tanulmányát: Baker, Catherine, Debbie Ging, Maja Brandt Andreasen (2024): Recommending Toxicity: The role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male supremacist influencers. DCU Centre, Dublin City University. 2-33o. Elérhető: https://antibullyingcentre.ie/wp-content/uploads/2024/04/DCU-Toxicity-Full-Report.pdf

[291] Zuboff, Shoshana (2019): The Age of Surveillance Capitalism. New York, PublicAffairs. 14.o.

válik az olyan online magatartások ösztönzése is, mint egy adott termékhirdetésre kattintás, videó megnyitása, vagy egy meghatározott hírfolyam követése. Zuboff szerint bár a megfigyelési kapitalizmus alig pár éve kezdte bontogatni szárnyait, már most is rengeteg ember döntéseire bír komoly befolyással (sok esetben az ő tudtuk nélkül). Ahogy egyre több okosinfrastruktúra kapcsolódik a digitalizált ökoszisztémához, az értékes adathalmazok, az adatelemzés által kínált lehetőségek, valamint a fogyasztói magatartások előrejelzésének és befolyásolásának eszköztárai is növekedni fognak a jövőben.[292] Valószínűleg hamarabb elérkezik az idő, mint azt sokan gondolnák, amikor már közel lehetetlenné válik kiszabadulni a folyamatosan bővülő *megfigyelési* hálóból.[293]

Ebben az új online környezetben a felhasználók nemcsak a piaci szereplők manipulatív gyakorlataival szemben váltak kiszolgáltatottá, de a politikai indíttatású befolyásolási technikákkal szemben is. Ennek kifejtése előtt azonban szükséges egy, az algoritmikus tartalomajánláshoz kapcsolódó másik jelenségre is kitérni, mely fontos tényezőként játszott szerepet az információs környezet (és egyéni információfeldolgozás) napjainkban tapasztalható eltorzulásában.

6.5 Információs buborékok, személyre szabott valóságok

Eli Pariser 2011-ben megjelent könyvében ismertette meg a médiatudományokban kevésbé jártas felhasználókkal a *szűrőbuborék* (filter bubble) fogalmát. A kifejezés arra a jelenségre utal, amkor az online platformokon alkalmazott ajánlóalgoritmusok az internetes aktivitást és hírfogyasztási szokásokat figyelembe véve főként személyre szabott tartalmakat jelenítenek meg a felhasználóknak, ezzel egyfajta információs „buborékba" zárva őket. Minél inkább az egyéni preferenciák alapján szűrik és értékelik az információt, annál inkább valószínű, hogy a felhasználók olyan tartalmakhoz férnek hozzá főként, amelyek illeszkednek a meglévő attitűdjeikhez, ismereteikhez és álláspontjukhoz, ezáltal olyan *hiedelemkamrákba*[294] szorulnak, melyek megnehezítik számukra, hogy a saját nézeteiktől eltérő információokhoz

[292] Diósi Szabolcs, Barcsi Tamás (2021): The legacy of disciplinary society – how relevant is Foucault's theory today? *Évkönyv - Újvidéki egyetem magyar tannyelvű tanítóképző,* XVI. évfolyam, 1. szám, 10-33.o.
[293] Zuboff, Shoshana (2019): The Age of Surveillance Capitalism. New York, PublicAffairs. 310.o.
[294] Veszelszki Ágnes (2021): deepFAKEnews: Az információmanipuláció új módszerei. In Balázs László (szerk.): Digitális Kommunikáció és Tudatosság. Budapest, Hungarovox kiadó. 96.o.

jussanak.[295] Az ilyen szűrők azon túl, hogy meghatározott irányban formálhatják a felhasználó világképét, hozzájárulhatnak az un. *önbeteljesítő identitások* (self-fulfilling identites) kialakulásához is.[296] Mivel a felhasználók folyamatosan olyan tartalmakat kapnak, amelyek a korábbi érdeklődési körüket tükrözik, hamar olyan helyzetbe kerülhetnek, ahol a múltbeli preferenciáik kezdik irányítani – és befolyásolni – a jövőbeli döntéseik, érdeklődésük és identitásuk alakulását. Ez az önbeteljesítő hatás korlátozza a felhasználók autonómiáját, mivel nehezebbé teszi számukra, hogy szabadon fedezzenek fel új vagy kihívást jelentő gondolatokat. A szűrőbuborékok ezáltal nem csupán a sokszínű információ szabad áramlását akadályozzák meg, de egyben gátolják a sokoldalú személyiség kifejlődésének lehetőségét is.

További kihívást jelent, hogy az online környezetben a felhasználók gyakran nem tudatosan választják a szűrőbuborékokat; az algoritmusok automatikusan szelektálják a tartalmakat és formálják az információáramlást. Ez új jelenségként tűnik fel a hagyományos médiafogyasztási szokásokhoz képest, ahol az egyének tudatosan választhatnak újságokat vagy tévécsatornákat (például annak alapján, hogy az adott platform milyen ideológiai keretezéssel tálalja a híreket). Ettől eltérően az online térben könnyen előfordulhat, hogy a felhasználó nincs is tudatában annak, hogy algoritmusok által szűrt információs környezetben tájékozódik,[297] amiből egyenesen következik, hogy nem is a saját döntéséből lépett be az információs buborékba (arról nem is beszélve, hogy az abból való kilépés is egyre nehezebbé válik).

6.5.1 Az információfeldolgozás torzulása

A szűrőbuborékok és az ajánlóalgoritmusok társadalmi hatásait vizsgálva fontos megemlíteni az amerikai jogtudós, Cass Sunstein által már 2001-ben feltárt *visszhangkamrákhoz* (echo

[295] Bizonyos értelemben veszélyeztetve ezáltal az Európai Unió Alapjogi Chartájának 11. cikke, (1) bekezdésében, illetve az az Emberi Jogok Európai Egyezményének 10. cikkében foglalt információk és eszmék megismerését garantáló szabadságjogok korlátoktól mentes gyakorlását. (*...az információk, eszmék megismerésének és közlésének szabadságát országhatárokra tekintet nélkül…*).
[296] Pariser, Eli (2021): The Filter Bubble. What the Intenet is Hiding from Yout New York. The Penguin Press. 112.-113.o.
[297] Uo., 9-10.o.

chambers) köthető lehetséges kockázatokat is.[298] Visszhangkamra-hatásként szokás utalni arra az – online és offline környezetben egyaránt megfigyelhető – jelenségre, amikor az emberek hasonló gondolkodású közösségekbe csoportosulnak, melynek következtében főként olyan információkkal és véleményekkel találkoznak, amelyek megerősítik saját előítéleteiket és nézeteiket. Egy 2009-ben megjelent könyvében Sunstein már arról ír, hogy az online felhasználókat azok aktív közreműködésük nélkül is visszhangkamrákba taszíthatják a közösségi média algoritmusai, ami nagyban hozzájárul az un *digitális enklávék* létrejöttéhez.[299] Az olyan csoportok pedig, akik tartósan nem kommunikálnak a sajátjuktól eltérő véleménnyel bíró közösségekkel, hamar elszigeteltté válnak az online térben, ami hosszútávon az online közösségek ideológiai vagy érdeklődési kör alapján történő fragmentálódásához, un. *kiberbalkanizációhoz* vezethet.[300]

E jelenséghez szorosan kapcsolódik az a folyamat is, ahogyan az internet elterjedése tovább torzította az emberek észlelését arról, hogy bizonyos álláspontok és nézetek mennyire elterjedtek (népszerűek vagy elutasítottak) a valóságban. Míg az offline világban az egyén korlátozott számú, többnyire a közelében élő személyekkel való interakció alapján szerzi információt arról, hogyan gondolkodnak mások, az online világban ezek a fizikai határok megszűnnek, és hozzáférés nyílik egy információban ugyan gazdag, de azok hitelességét és reprezentativitását tekintve szélsőségesen aránytalan környezethez.[301] A közösségi médiában a

[298] A visszhangkamrákról bővebben: Sunstein, Cass R. (2001): Republic.com Princeton NJ: Princeton University Press

[299] Sunstein hangsúlyozza, hogy az ilyen visszhangkamrák és digitális enklávék elősegíthetik a politikai és társadalmi csoportpolarizációt, ami a vélemények és attitűdök még szélsőségesebbé válásával súlyos akadálya lehet az érdemi, konstruktív párbeszédnek és a nézetek ütköztetésének Lásd: Sunstein, Cass (2009): Republic.com 2.0. New Jersey, Princeton Univerity Press. 11.o.; Sunstein (2017): #Republic: Divided Democracy in the Age of Social Media. Princeton, Princeton University Press.

[300] Az ilyen online környezetben a különböző csoportok saját információs univerzumokban működnek, eltérő tényekkel és narratívákkal. Közöttük kevés az interakció, nem hallják meg egymás álláspontját. A csoportidentitás megerősödik, az ellenvélemények elutasítása fokozódik. Megnő az előítéletek, sztereotípiák és szélsőséges nézetek terjedésének esélye. Általánosságban elmondható, hogy a társadalmi kohézió gyengül. L ásd: Sunstein, Cass (2009): Republic.com 2.0. New Jersey, Princeton Univerity Press. 79-80.o.

[301] Természetesen már a hagyományos média is – legyen az nyomtatott újság, televízió vagy rádió – példátlan befolyással bírt, és bír a mai napig a közvélemény és közvélekedés alakításában. Ennek bemutatására és feltérképezésére számos mára alapvetőnek számító írás tett kísérletet. Walter Lippmann Közvélemény című könyvének egyik fő állítása például, hogy a média az a hatalom, mely képes olyan pszichikai környezetet (psychic environment) teremteni, amelyben az emberek a valóságról alkotott alapvető képüket formálják. Lippmann is kiemeli, hogy a média által közvetített üzenetek nemcsak informálják, de gyakran el is torzítják a közvéleményt, hiszen az emberek csak közvetett (irányított) képet kapnak a világról. Lásd: Lippmann, Walter (1971): A közvélemény I-II. Budapest, Tömegkommunikációs Kutatóközpont. 215-281.o.; Noam Chomsky és Edward S. Herman pedig *Az egyetértés gépeze*t (Manufacturing Consent) című könyvükben mutatnak rá arra, hogy a média alapvetően az elit érdekeit szolgálja, és a hatalom képviselői által kívánatosnak tartott narratívákat erősíti meg, ezzel befolyásolva a társadalom gondolkodását. A médiát Chomsky propagandamodellként írja le, amelyben a tartalmakat szűrőkön keresztül tálalják, így alakítva ki az emberek egyetértését bizonyos politikai és gazdasági kérdésekben. Bővebben: Chomsky, Noam - Edward S. Herman (2016): Az Egyetértés-gépezet - A

marginális csoportok is látszólag széles körben elterjedtnek tűnhetnek, ha kellően aktív és zajos közösségeket alkotnak.[302] Ez könnyen azt a téves benyomást keltheti, hogy az extrém vélemények is széles körben elfogadottak és sokak által támogatottak. Míg a valós életben nehéz olyan emberekkel találkozni, akik abban hisznek, hogy a Föld lapos, az interneten és a közösségi médiában, ahol több milliárd aktív felhasználó van jelen, viszonylag könnyen találkozhatnak egymással azok, akik ezt a nézetet vallják.[303] Az ilyen aktív és hangos véleményközösségek legitimációt biztosítanak egy személy hitének, ami még inkább nehezebbé teszi az elgondolásuk megváltoztatását. Ezt a folyamatot egy sor az emberi információfeldolgozást érintő heurisztikai és kognitív torzítás is erősít, melyek közül a legfontosabbak (1) a hamis konszenzus hatás; (2) a megerősítéses torzítás és a kognitív disszonancia, valamint; (3) a hitbéli meggyőződés állandósága (belief perseverance).

(1) A „hamis konszenzus hatás" (false consensus effect) arra a pszichológiai jelenségre utal, amely során az emberek hajlamosak feltételezni, hogy saját nézeteik, hiedelmeik és értékrendjük általánosabbak és elterjedtebbek, mint valójában; (2) A megerősítési torzítás okán az emberek hajlamosak azokat az információkat keresni, előtérbe helyezni és mélyebben az emlékezetükbe vésni, amelyek összhangban vannak és megerősítik előzetes meggyőződéseiket, nézeteiket és értékrendszerüket.[304] Amikor pedig olyan információkkal találkoznak, amelyek ellentmondanak a meglévő hiedelmeiknek un. kognitív disszonanciát élhetnek át. [305] Ennek feloldására hajlamosak lehetnek figyelmen kívül hagyni vagy elutasítani

---

tömegmédia politikai gazdaságtana. (Ford.: Konok Péter). Budapest, L'Harmattan Kiadó. E hatásokkal szemben azonban, a törzsszöveg elsősorban arra utal, hogy az internet megjelenésével már nem csak a valóságról alkotott képünk, hanem a különböző álláspontokat valló emberek és közösségek nagyságával és reprezentativitásával kapcsolatos benyomásunk is sérül.

[302] Leviston, Z., I. Walker, and S. Morwinski (2013): Your opinion on climate change might not be as common as you think. *Nature Climate Change*, 3: 334-337.o.

[303] A Föld lakosságának megközelítőleg 62%, kb. 5 milliárd ember használt már valamilyen típusú közösségi médiát (csak 2022-ben 266 millió új felhasználó csatlakozott először közösségi platformra. Bővebben: Meltwater & We Are Social (2024): Digital 2024 Global Overview Report. Elérhető: https://datareportal.com/reports/digital-2024-global-overview-report

[304] Nickerson, Raymond S. (1998): Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2 (2), 175–220.o.

[305] A kognitív disszonancia – melynek elméletét Leon Festinger amerikai szociálpszichológus dolgozta ki 1957-ben – egyfajta mentális feszültséget vagy kellemetlen érzést takar, amely akkor lép fel, amikor egy személy egyszerre tart fenn egymással összeegyeztethetetlen gondolatokat, attitűdöket vagy cselekedeteket. Festinger elmélete alapján az emberek folyton törekednek a belső kognitív egyensúly elérésére. Amikor információk, vélemények vagy cselekedetek ütköznek egymással, feszültség keletkezik, és az egyén igyekszik csökkenteni ezt a feszültséget. A jelenségről először a *When Prophecy Fails* c. 1956-ban megjelent könyvében írt, Henry Riecken és Stanley Schachter társszerzőkkel. A könyv egy amerikai UFO-szekta csoportot követ nyomon, amely azt hirdette, hogy a világ hamarosan véget ér, és csak őket menti meg egy földönkívüli civilizáció. Festinger és társai beépültek a csoportba, és megfigyelték a tagok viselkedését, különösen azt, hogyan reagáltak, amikor a jóslatuk nem teljesült. A csoport tagjai nem vetették el hiedelmeiket, hanem új magyarázatokat kerestek, hogy fenntarthassák meggyőződéseiket, ami példázza a kognitív disszonancia csökkentésére irányuló törekvéseket. A

a tényeket, és ragaszkodni a saját meggyőződéseikhez, még akkor is, ha azok nem állják ki a valóság próbáját; (3) A viselkedéspszichológiában megfigyelt jelenség, hogy az emberek egy jelentékeny hányada gyakran akkor sem változtatja meg kialakult hiedelmeit, ha új, azoknak ellentmondó információkkal szembesül (belief perseverance). Ez azzal magyarázható, hogy ítéletalkotásuk során ok-okozati összefüggéseket feltételeznek a világ eseményei között, és az okozati következtetéseket mélyen elraktározzák az emlékezetükben. Ha később ki is derül, hogy a kialakult következtetésük alapjául szolgáló információ hamis volt, az abból eredő következtetés továbbra is befolyásolja a gondolkodásukat (az eredeti információ cáfolata, nem írja automatikusan felül az abból kialakított és már rögzült következtetést vagy világmagyarázatot).[306]

Összegezve a fejezetben elhangzott állításokat; az elmúlt bő három évtizedben kialakult új online környezet – amely a figyelemgazdaságra és a megfigyelési kapitalizmusra épül – pszichológiai és technológiai (az online tér működési logikája) szempontból egyaránt sebezhetővé tette a felhasználókat az információfeldolgozás terén. Pszichológiai értelemben többek között olyan hatások figyelhetők meg, mint a csökkenő figyelemküszöb, a dopaminéhség- és függőség, a hamis konszenzus hatás, a megerősítési torzítás, a kognitív disszonancia és a csoportpolarizáció. Míg technológiai szempontból a személyre szabott ajánlóalgoritmusok, a szűrőbuborékok, visszhangkamrák és a kiberbalkanizáció járulnak hozzá a felhasználók kiszolgáltatottságához.

Ebben a sérülékeny és kiszolgáltatott közegben érkeztek meg a 21. század társadalmai a dezinformációtól terhes – és konspirációs elméletekkel telített – post-truth (posztigazság) korszakába.

---

témáról lásd bővebben: Festinger, L., Riecken, H.W., & Schachter, S. (1956): When Prophecy Fails. University of Minnesota Press; Festinger, L. (1957): A Theory of Cognitive Dissonance. Stanford University Press.

[306] E jelenség mutatja, hogy nem elég egyszerűen cáfolni a hamis információkat. Új, hiteles magyarázatokat is kell adni az eseményekre, hogy az emberek le tudják cserélni a téves következtetéseiket. Például, ha valaki azt a téves információt hallja, hogy az oltások autizmust okozhatnak, melyre alapozva arra a következtetésre juthat, hogy az oltások veszélyesek, és nem akarja beoltatni a gyermekeit. Hiába szembesül később hiteles cáfolatokkal, amelyek egyértelműen bizonyítják, hogy az oltások nem okoznak autizmust, az eredeti hiedelme megmaradhat, és továbbra is befolyásolhatja a döntéseit. Ahhoz, hogy megváltoztassa a véleményét, nem elég pusztán cáfolni a hamis információt. Szüksége van egy új, hiteles magyarázatra is, amely meggyőzően bemutatja, hogy miért biztonságosak és fontosak az oltások (az oltások szigorú biztonsági és hatékonysági teszteken mennek át, az oltást elutasítók gyermekei sokkal nagyobb eséllyel kaphatnak el súlyos betegségeket stb. Lásd: Brendan Nyhan, Jason Reifler (2015): Displacing Misinformation about Events: An Experimental Test of Causal Corrections, *Journal of Experimental Political Science,* Volume 2, Issue 1, 81-93.o.

7 Post-truth: az igazság hanyatlása és a tényeken túli valóság

A „post-truth" kifejezés egy olyan társadalmi és politikai környezetre utal, ahol az érzelmek és a személyes meggyőződések nagyobb szerepet játszanak a közvélemény formálásában, mint az objektív igazság. Ebben a környezetben az emberek döntéseit inkább az érzéseik és saját hiedelmeik vezérlik, semmint a valós, ellenőrizhető tények. Krekó Péter a „Tömegparanoia 2.0" című könyvében ezt úgy foglalja össze, hogy: *„A post-truth mágikus világában elmosódik a tények és a vélemények közti különbség, és ugyanarról a jelenségről több, egymást kizáró „tény" is létezhet".* [307]

A kifejezés a 2010-es évek második felében vált széles körben ismertté, különösen a 2016-os Brexit népszavazás és az amerikai elnökválasztási kampány(ok) kapcsán.[308] E két esemény során feltűnő volt, hogy a politikusok a szokásosnál is gyakrabban terjesztettek olyan érzelmekre ható, valóságot leegyszerűsítő, félrevezető vagy egyenesen hamis üzeneteket, melyeket a célközönség hajlamos volt igaznak elfogadni, függetlenül azok valóságtartalmától. A post-truth korszakra jellemző politikai kommunikációt jól szemlélteti Kellyanne Conway – Donald Trump tanácsadójának – 2017. január 22-én a *Meet the Press* televíziós műsorban adott interjúja; ebben Conway megvédte Sean Spicer sajtótitkár azon állítását, hogy Trump beiktatásán rekordméretű tömeg volt, annak ellenére, hogy az eseményről készült fotók és videós beszámolok ennek jól láthatóan ellentmondtak. Conway azt mondta, hogy Spicer csak *alternatív tényeket* (alternative facts) közölt, amikor azt állította, hogy ez volt a legnagyobb egybegyűlt közönség, amely valaha is jelen volt egy elnöki beiktatáson.[309]

---

[307] Krekó Péter (2021): Tömegparanoia 2.0 – Összeesküvés-elméletek, álhírek és dezinformáció. Budapest, Athenaeum Kiadó.

[308] Jól érzékelteti, hogy ezekben az években mekkora figyelem összpontosult a jelenségre, az is, hogy az Oxford Dictionaries 2016-ban a post-truth kifejezést választotta az év szavának. A meghatározás szerint a post truth olyan helyzetekre utal, amikor *a tárgyilagos tények kevésbé hatnak a közvéleményre, mint az érzelmeken, személyes hiten alapuló érvek* ("relating to or denoting circumstances in which objective facts are less influential in shaping

[309] Chuck Todd, a műsor házigazdája, erre úgy reagált, hogy "az alternatív tények nem tények, hanem valótlanságok (alternative facts are not facts, they are falsehood). Conway megpróbálta elterelni a beszélgetést más témákra, mint például az egészségügyi biztosítás kérdése, anélkül, hogy közvetlenül válaszolt volna a hamis állításra vonatkozó kérdésekre Lásd: Gajanan, Mahita (2017): Kellyanne Conway Defends White House's Falsehoods as 'Alternative Facts'. *TIME*. Elérhető: https://time.com/4642689/kellyanne-conway-sean-spicer-donald-trump-alternative-facts/ (2017. 01. 12.)

Kétségtelen, hogy az *igazság hanyatlásának* (truth decay)[310] folyamata nem a 2010-es években indult meg. A történelem számos példát szolgál arra nézve, hogy különböző totalitárius rendszerek miként igyekeztek az igazságot relativizálni és a valóságot elferdíteni a hatalmuk biztosítása érdekében.[311]

A modern digitális környezetében mégis fontos részletesebben is tárgyalni e témakört, hisz a korábban vázolt technológiai eredetű információs torzulás, kiegészülve az emberre jellemző pszichológiai sajátosságokkal különösen sebezhetővé tették az információs társadalmat. Ezekre a sebezhetőségekre vezethető vissza az is, hogy napjainkban gazdag táptalajra találtak, és példátlan mértékben elszaporodtak a legkülönfélébb – első hallásra a valóságtól és racionális belátástól teljesen elrugaszkodottnak tűnő – összeesküvés elméletek (egyebek mellett azok a mélyállamról szóló konspirációk is, amelyek szerepet játszottak az Egyesült Államok Capitoliumának 2021 január 6-án bekövetkezett ostromában).[312]

---

[310] Az *igazság hanyatlása* (truth decay) kifejezést Kavanagh és Michael Rich a RAND amerikai nonprofit politikai kutatóintézet tanácsadó megbízásából készített 2018-as jelentésében használták, arra, hogy leírják az objektív tények szerepének csökkenését a közbeszédben és a politikai viták során. A jelenséget négy trenddel jellemezték: (1) A tényekbe vetett bizalom csökkenése; (2) A vélemények és a tények közötti különbségek elmosódása; (3) A személyes vélemény felértékelődése a tényekkel szemben; (4) A korábban hitelesnek elismert információforrásokba vetett bizalom csökkenése. Lásd: Kavanagh, Jennifer and Michael D. Rich (2018): Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life. Santa Monica, CA: RAND Corporation. 21-38.o. Elérhető: https://www.rand.org/pubs/research_reports/RR2314.html.

[311] Az autoriter rendszerek sajátosságait kutató Hannah Arendt is mélyrehatóan vizsgálta az igazság jelentésének mibenlétét. Munkásságában különbséget tett a *tényigazságok* (factual truth) és az *észigazságok* (racional truth) között. Értelmezése szerint az észigazság filozófiai, tudományos, matematikai igazságokat takar, melyekhez az elme útján, gondolkodás révén lehetséges eljutni. Ezek állandóak és biztosak (logikai kényszerítő erővel bírnak) ellentétpárjuk a tévedés, a tudatlanság vagy az illúzió. A tényigazságok ezzel szemben egy adott helyzetre történésre, eseményre vonatkoznak, ami természetükből adódóan - különösen a politikai szférában- könnyen torzíthatóak, manipulálhatóak (például átalakulhatnak véleményekké, amikor politikai diskurzusok tárgyává válnak). A tényigazság ellentétpárját viszont nem a tévedés vagy a tudatlanság adja, hanem a szándékos hazugság. A rendszerszinten gyakorolt szándékos hazugság a totalitárius berendezkedések sajátja. Lásd: Arendt, Hannah (1995): Igazság és politika. In: Múlt és jövő között. Nyolc gyakorlat a politikai gondolkodás terén. (ford: Módos Magdolna). Budapest, Osiris–Readers International, 233-251.o.; E témához kapcsolódva érdemes szót ejteni arról, hogy Arendt már az 1951 ben megjelent *The Origins of Totalitarianism* (A totalitarizmus gyökerei) c. könyvében hangsúlyozta, hogy a propaganda a totalitárius rendszerek egyik központi eszközeként jelenik meg., és célja - tömegek irányításán túl –, hogy eltorzítsa a valóságot oly módon, hogy a polgárok elveszítsék a képességüket arra, hogy megkülönböztessék a hazugságot az igazságtól. A propaganda eszközével a rendszer képes elérni, hogy a polgárok elfogadják a legképtelenebb hazugságokat is, ha azok következetesen ismétlődnek. Bővebben: Arendt, Hannah (1992): A totalitarizmus gyökerei. (Ford.: Berényi Gábor, Erős Ferenc, Seres Iván, Braun Róbert). Budapest, Európa Könyvkiadó. 577-602.o.

[312] Még a 2016-os amerikai elnökválasztási kampány utolsó heteiben terjedt el a *Pizzagate* néven elhíresült konspirációs elmélet, mely szerint Hillary Clinton kampányfőnöke, John Podesta és más prominenst demokrata politikusok gyermek szexkereskedelmi hálózatot üzemeltetnek egy washingtoni pizzéria (Comet Ping Pong) alagsorában. Bár erre semmilyen konkrét bizonyíték nem volt, a Wikileaks által korábban kiszivárogtatott Cilnton - Podesta privát e-mailek szövegeiben az elmélet hívei különböző rejtett utalásokat véltek felfedezni az állításaik alátámasztására. Erre az elméletre épült később a QAnon összeesküvés-elmélet, mely szerint egy rejtélyes "Q" nevű kormányzati hírszerző információkkal rendelkezik egy globális elit csoportról, akik pedofil hálózatokat üzemeltetnek. Donald Trump, az Amerikai Egyesült Államok 45. elnöke, azonban harcot indított ezen elit csoport ellen, és titokban küzd az ő globális befolyásuk csökkentéséért. Az elmélet hívői a 2020-as

A helyzetet tovább bonyolítja, hogy ebben a kiszolgáltatott közegben indultak meg azon jól felépített, tudatosan tervezett dezinformációs kampányok is, melyek a közvélemény befolyásolását és az igazság relativizálását tűzték ki céljukként. Jól illik ebbe az új információs környezetbe a modern dezinformáció természete, amely sok esetben nem is a meggyőzésre, hanem elbizonytalanításra törekszik. [313]

7.1 Félretájékoztatás, dezinformáció

Mára széles körű egyetértés látszik kialakulni azzal kapcsolatban, hogy a 2016-os amerikai elnökválasztásban az álhírek és a dezinformációs kampányok kiemelt szerepet játszottak. Ebben az időszakban az orosz kormánnyal közvetlen kapcsolatba hozható *trollfarmok* nagy mennyiségben terjesztettek megosztó és félrevezető tartalmakat a közösségi médiában, hogy befolyásolják a választások kimenetelét és mélyítsék a politikai polarizációt az USA-ban.[314] Az erre szakosodott műhelyek közül a legismertebb az azóta *váratlan* repülőbalesetben elhunyt, ex-Wagner vezér Jevgenyij Prigozsin tulajdonában álló szentpétervári *Internet Research Agency* (IRA) volt. A hibrid hadviselés[315] részét képző dezinformációs művelet

elnökválasztáson Trump vereségét úgy értelmezték, mint a szabadságharcosok bukását a korrupt háttérhatalom ellen. Ebben a feltüzelt helyzetben kérdőjelezte meg Trump az elnökválasztás eredményét és vádolta csalással a demokratákat. A választást követő két hónapban egyre szélesebbé és hangosabbá váltak azok a visszhangkamrák, amik a korrupt és romlott demokrata politikusok felelősségét kiemelve a választási eredmények rendszerszerű meghamisítását hangoztatták. A leköszönő elnök január 6-ai beszédét követően a reményvesztett tömeg Capitoliumhoz vonult, áttörték a kordonokat, összecsaptak a rendőrökkel, betörtek az épületbe, amit több órán át megszállva tartottak. Az ostromban számos QAnon követő is részt vett, a leghíresebb közülük Jake Angeli volt, akit a médiában csak „QAnon sámán"-ként emlegettek.

[313] Krekó Péter, Molnár Csaba (2023): Tényrelativizmus és a hírforrásokkal szembeni elbizonytalanodás a magyar közvéleményben. Lakmusz-HDMO. 5.o. Elérhető: https://politicalcapital.hu/pc-admin/source/documents/hdmo _pc_tanulmany_2_tenyrelativizmus_kozvelemeny_20231130.pdf (2023.11.30.)

[314] Az orosz beavatkozást az amerikai hírszerző szervek is megerősítették, kiemelve, hogy Oroszország célja a választási folyamat befolyásolása és az amerikai politikai rendszer meggyengítése volt. Egy az Egyesült Államok szenátusának hírszerzési bizottsága által közzétett több száz oldalas jelentés megállapítja, hogy bár Moszkva korábban is szivárogtatott ki politikailag érzékeny információkat, illetve 2014 óta egyre nagyobb mennyiségben tett elérhetővé harmadik fél közreműködésével (pl.: WikiLeaks) illegálisan szerzett iratanyagokat, a 2016-os amerikai elnökválasztás során tanúsított aktivitása egyértelműen szintlépésnek tekinthető. Lásd bővebben: U.S. Government Publishing Office (2019): Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election. Senate Report, II. Volume, 116–290.o. Elérhető: www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures

[315] A hibrid hadviselés olyan katonai stratégia, amely egyesíti a hagyományos katonai műveleteket a nem hagyományos módszerekkel, mint a kibertámadások, dezinformációs kampányok, gazdasági nyomásgyakorlás, valamint politikai destabilizáció. Ezen módszerek célja a közvélemény befolyásolása, a politikai stabilitás aláásása és a döntéshozatali folyamatok megzavarása.

kiterjedtségét mutatja, hogy a Facebook az amerikai szenátus igazságügyi bizottsága előtt elismerte, hogy az Oroszország által támogatott digitális tartalmak a 2016-os amerikai elnökválasztás alatt és után körülbelül 126 millió amerikait felhasználót értek el a közösségi platformjukon.[316] Ebben az időszakban az Oroszországhoz köthető oldalak több mint 80 ezer bejegyzést hoztak létre, amelyek közvetlenül 29 millió amerikai felhasználóhoz jutottak el. Azonban, a másodlagos megosztásokat és interakciókat is számításba véve a bejegyzések – eredeti címzetteken túl, a felhasználók barátainak és követőinek hálózatába terjeszkedve – valójában közel ötször annyi emberhez jutottak el. A Twitter esetében még kiterjedtebb kampányról volt szó, melyben 50 ezer automatizált felhasználói fiók (bot) hozzávetőlegesen 1,4 millió bejegyzést tett közzé, közel 288 millió interakciót generálva.[317]

Bár az Internet Research Agency amerikai tevékenysége tekinthető talán a legismertebb esetnek, a hibrid hadviselés keretein belül folytatott félretájékoztatási kampányok nem korlátozódnak egyetlen országra vagy szereplőre. A világ különböző állami és nem állami szereplői hasonló tevékenységet folytatnak politikai, ideológiai vagy gazdasági haszonszerzés céljából.[318] Például az elmúlt években az Európai Unióra összpontosuló dezinformációs hibrid hadműveletek olyan világnézeteket és politikai erőket karoltak fel – és törekedtek propagandájuk által felerősíteni –, amelyek az európai integráció és transzatlanti együttműködés ellen foglalnak állást, és a status quo felborítását célozzák. Tevékenységük során igyekeztek kiélezni a belső megosztottságot a célországokban és felnagyítani a szélsőséges hangokat olyan érzékeny társadalmi ügyek kapcsán, mint a migráció, szexuális kisebbségek jogai, eutanázia, terhességmegszakítás vagy drogliberalizáció. Az oltásokkal kapcsolatos álhíreket és összeesküvés-elméleteket felkapva igyekeztek aláásni a bizalmat a tudományos intézményekben és szakértőkben, ezzel gyengítve a járványkezelési erőfeszítéseket a nyugati országokban. Geopolitikai válsághelyzeteket, mint a brexit, a katalán

[316] Scola, Nancy - Ashley Gold (2017): Facebook: Up to 126 million people saw Russian-planted posts. POLITICO. Elérhető: https://www.politico.eu/article/facebook-up-to-126-million-people-saw-russian-planted-posts/ (2022. 10. 15.)

[317] Jack Dorsey a twitter akkori vezérigazgatója utólag vallotta csak be, hogy ez idő tájt milyen nagy számban voltak jelen az Internet Research Agencyhez köthető hamis fiókok a felületükön. Az elnökválasztásról folyamatosan tweetelő automatizált fiókok (botok) számát kezdetben 36.000-re becsülték, amely számot később korrigáltak 50.000-re. Bővebben: Romm, Tony (2018): Twitter has notified at least 1.4 million users that they saw Russian propaganda during the election. *VOX*. Elérhető: https://www.vox.com/2018/1/31/16956958/twitter-jack-dorsey-russia-trolls-election-us-trump-clinton-propaganda (2022. 10. 15.)

[318] A nemzetközi politikára gyakorolt hatását mutatja, hogy a Világgazdasági Fórum a 2024-es Global Risk Report jelentése az elkövetkező két év kiemelt és már rövid-távon is meghatározó kockázatait a következő módon rangsorolja: (1) dezinformáció és félretájékoztatás globális elterjedése; (2) extrém időjárási események; (3) társadalmi polarizáció; (4) kiberbiztonsági fenyegetések; (5) államközi fegyveres konfliktusok; (6) gazdasági lehetőségek hiánya; (7) infláció; (8) kényszerű migráció; (9) gazdasági visszaesés; (10) környezetszennyezés. Lásd: World Economic Forum (2024): The Global Risk Report 2024 - 19th Edition. 14-37.o.

függetlenségi népszavazás, vagy egyes nagyszabású terrortámadások utóhatásait meglovagolva terjesztettek zavarkeltő álhíreket, fokozva a bizonytalanság és a széthúzás légkörét.[319]

7.2 A dezinformáció szolgálatában – Mikrocélzás, állhírek és botok

Az Európai Bizottság meghatározása szerint a dezinformáció „*olyan igazolhatóan hamis vagy félrevezető információ, amelyet gazdasági haszonszerzés vagy szándékos megtévesztés céljából hoznak létre, hoznak nyilvánosságra és terjesztenek, és amely kárt okozhat a közérdeknek”*.[320]

Fontos különbséget tenni a félretájékoztatás és a dezinformáció esetei között. A *félretájékoztatás* (misinformation) fogalma alatt olyan pontatlan vagy hamis információk közlését és terjesztését értjük, amelyek nem tudatos félrevezetési szándékból erednek. [321] Gyakran előfordul, hogy egy személy – vagy szervezet – tévedésből oszt meg hibás információkat, anélkül, hogy rossz szándék vezérelné; a félretájékoztatás tipikus esetei közé sorolhatók azon helyzeteket, amikor valaki olyan információt terjeszt, amelyet tévesen megbízhatónak vélt, illetve amikor egy forrás helytelen értelmezéséből von le téves következtetéseket. Fontos tehát, hogy a terjesztésnek nem a károkozás a célja (egyszerűen csak olyan információ közlése/megosztása, amelyről a tévesen úgy hiszik, hogy igaz). Valós példaként hozható fel, az az eset, amikor a *The Lancet* orvosi szaklap 1998-ban közölt egy tanulmányt, amely összefüggést sugallt az MMR vakcina és az autizmus között. Bár a tanulmány módszertana hibás volt, és következtetései tévesnek bizonyultak, a folyóirat nem szándékosan terjesztett hamis információt. Ez a publikáció azonban évekig tartó félretájékoztatáshoz vezetett az oltások biztonságosságával kapcsolatban.[322] Ezzel szemben a *dezinformáció* (disinformation) – ahogy a Bizottság definíciója is hangsúlyozza – hamis információk szándékos létrehozását vagy terjesztését jelenti, amelynek célja mások félrevezetése vagy károkozás. A dezinformáció készítői és terjesztői tisztában vannak azzal,

[319] Lásd részletesebben: EU DisinfoLab (2023): Connecting the disinformation dots: insights, lessons, and guidance from 20 EU Member States. 1-12.o. Elérhető: https://www.disinfo.eu/wp-content/uploads/2023/12/20231204_Connecting-disformation-dots_comparative-study-1.pdf

[320] De nem tartoznak e kategóriába a jelentéstételi hibák, a szatírák és paródiák, illetve az egyértelműen azonosítható módon pártokhoz köthető hírek és kommentárok. Lásd: Európai Bizottság (2018): Európai megközelítés az online félretájékoztatás kezelésére. COM(2018) 236 final, 2.1. pont

[321] Országgyűlés Hivatala (2023): Infodémia – A dezinformációs járvány hatásai és a védekezési lehetőségek. *Infojegyzet, 2023*. Elérhető: https://www.parlament.hu/documents/10181/64399821/Infojegyzet_2023_21_infodemia.pdf/93d64698-1928-84ac-df43-eaaba42124eb?t=1688048577522 (2023. 08. 10.)

[322] Európai Bizottság (2020): Az Európai Demokráciára vonatkozó cselekvési tervről. COM(2020) 790 final

hogy az információ hamis, de mégis megosztják, hogy zavart keltsenek vagy valamilyen módon előnyhöz juthassanak. [323]

A dezinformációk tartalmáról – általánosságban – elmondható, hogy gyakran leegyszerűsített narratívákat közvetítenek, amelyek az összetett kérdések megértése helyett arra ösztönzik a célszemélyeket, hogy figyelmen kívül hagyják egy adott ügy árnyaltabb részleteit. Emellett előszeretettel használnak olyan erős érzelmi elemeket, mint a félelem, a düh és a bizonytalanság, amely nem csak nagyobb elérést és gyorsabb terjedést biztosít a figyelemgazdaság korában, de egyúttal erőteljesebben is képes befolyásolni az emberi ítélőképességet, ezáltal fogékonyabbá téve a fogyasztót a manipulációra.[324] Érzékelhető tehát, hogy a dezinformáció terjedésének gyorsaságában szerepet játszik az előzőkben említett információfeldolgozási torzulás mindkét formája: (1) Technológiai torzítás: A személyre szabott tartalomszűrő algoritmusok a magas aktivitást kiváltó tartalmakat helyezik előtérbe, és információs buborékokat hoznak létre; (2) Pszichológiai torzítás: az emberek erőteljesebben reagálnak az érzelmekre ható negatív üzenetekre, elfogultak előzetes tudásukat illetően, túlbecsülik saját ismeretüket egy adott témában, illetve a figyelemküszöb csökkenése okán a rövid és könnyen emészthető tartalmakat részesítik előnyben. Ennek eredményeképpen a szenzációhajhász dezinformációs narratívák gyakran vírusként terjednek, míg a pontos és tényszerűén ellenőrzött (terjedelmesebb, szárazabb, érzelemmentes, nehezebben befogadható) információk elterjedéséhez több idő szükséges. Ráadásul az is súlyosbítja a problémát, hogy a közösségi médiákban a felhasználó fogyasztó és tartalom-előállító is egyben (prosumer), így a megosztások, kedvelések és hozzászólások által könnyen maga is dezinformációs kampányok propagálójává válik.

Céljaik tekintetében a dezinformációs kampányok rendkívül változatosak lehetnek: a közvélemény szándékos félrevezetésétől kezdve, egy adott kérdésben kialakított többségi vélemény formálásán át, egészen a szabad és tisztességes választási folyamatok befolyásolásáig terjedhetnek. Ezek elérésére különböző megoldások állnak rendelkezésre,

[323] Szintén nem tartoznak e kategóriába a jelentéstételi hibák, a szatírák és paródiák, illetve az egyértelműen azonosítható módon pártokhoz köthető hírek és kommentárok. Lásd: Európai Bizottság (2018): Európai megközelítés az online félretájékoztatás kezelésére. COM(2018) 236 final, 2.1. pont.

[324] A negatív hírek erőteljesebb érzelmi reakciókat váltanak ki, mint a pozitív vagy semleges hírek. Ennek oka, hogy. az emberi agy hajlamosabb nagyobb figyelmet fordítani a negatív információkra, ami az evolúciós múltból ered, ahol a negatív információk, mint a veszély vagy fenyegetés, fontos túlélési jelzések voltak (az olyan erős érzelmek, mint a düh vagy a félelem pedig, mélyebb emlékezetei rögzülést is eredményezhetnek). Bővebben: Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023): Negativity drives online news consumption. *Nature human behaviour, 7(5),* 812–822.o.

melyek közül ehelyütt három elterjedt gyakorlat kerül rövid kifejtésre: (1) mikrocélzás; (2) (állhírek); (3) automatizált botok.

(1) A kereskedelmi gyakorlatokban elterjedt online hirdetési módszerek, különösen a mikrocélzás (micro-targeting) technikák nem csak a termékek reklámozásában hasznosak, hanem a félrevezető információk terjesztésében is. A mikrocélzás lényege, hogy a hirdetők a felhasználókról gyűjtött személyes adatok alapján pontosan meg tudják határozni, kik tartoznak a célcsoportjukba, így az üzeneteiket aprólékosan személyre szabhatják, növelve ezzel azok hatékonyságát.[325] A politikai mikrocélzás legveszélyesebb formái kifinomult pszichográfiai profilalkotási módszereket alkalmazva az emberek személyes sebezhetőségeit használják ki.[326] A félretájékoztatási műveletek esetében az ilyen mikrocélzás abban segíthet, hogy a megtévesztő vagy hamis információkat célzottan juttassák el azokhoz a csoportokhoz, amelyek a legfogékonyabbak rájuk. Például, ha valaki összeesküvés-elméleteket akar terjeszteni, a mikrocélzással azonosíthatja azokat a felhasználókat, akik korábban hasonló tartalmakkal léptek interakcióba, majd célzottan nekik küldheti az üzeneteket.

(2) A közvélemény befolyásolásának szándékával vagy – az esetek nem elhanyagolható számában – egyszerű haszonrealizálás reményében[327] terjesztett *álhírek* (fake news) is fontos eszközei lehetnek a dezinformációnak. Az ilyen tartalmak a szenzációhajhász cím mellett előszeretettel használnak internetes mém vagy infografika formátumokat, amivel az összetett kérdéseket végletekig leegyszerűsítve, vizuálisan is vonzó, könnyen befogadható módon közvetítik.[328] Itt is érvényes az a szabály, hogy az érzelmekre ható tartalmak (állhírek) különösen hatékonyak, mivel könnyen felkeltik az olvasók érdeklődését, egyúttal ösztönzik a

[325] Talán a legismertebb példa erre, a Cambridge Analytica politikai tanácsadó cég esete, amely mintegy 87 millió Facebook felhasználó adataihoz fért hozzá a felhasználók beleegyezése nélkül. A cég a 2016-os amerikai választásokon ezeket az adatokat pszichológiai profilozásra és a szavazók politikai hirdetésekkel való célzott megszólítására használta. Bővebben: Isaak, J., & Hanna, M. (2018): User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51, 56-59.o. https://doi.org/10.1109/MC.2018.3191268.

[326] . Európai Bizottság (2021): A dezinformáció elleni gyakorlati kódex megerősítésére vonatkozó iránymutatás. COM(2021) 262 final, 11-12.o.

[327] Az Észak-Macedóniai Veles városában például több tucat, főleg fiatalok által működtetett weboldal specializálódott ennek kiaknázására. A városban működő weboldalak egyszerű, de annál hatékonyabb gazdasági modellje az álhírek gyártásán és terjesztésén, főként amerikai olvasók célzott bevonzásán, majd a hirdetések megjelenítése révén a nyereség realizálásán alapult (a Google AdSense szolgáltatásának segítségével). Kihasználva az amerikai politikai polarizációt és az online hírfogyasztási szokásokat a Veles-i weboldalak tipikusan amerikai politikával kapcsolatos tartalmakat gyártottak - például Hillary Clinton állítólagos korrupciós bűncselekményeiről vagy Barack Obama „eltitkolt, valódi" születési helyéről. Ezek a weboldalak aztán a közösségi médiában, különösen a Facebookon keresztül, terjesztették ezeket a tartalmakat. Bővebben: Hughes, H., & Waismel-Manor, I. (2020): The Macedonian Fake News Industry and the 2016 US Election. PS: Political Science & Politics, 54, 19 -23.o. https://doi.org/10.1017/S1049096520000992.

[328] Erről részletesebben: Veszelszki Ágnes (2017): Az álhírek extra- és intralingvális jellemzői. *Századvég*, 84: 51–82.o.

megosztásokat. A megtévesztés érdekében ezek forrásaiként sok esetben olyan weboldalak szolgálnak, amik felépítésükben, elrendezésükben és domain nevükben is hasonlóak lehetnek, mint a legitim, ismert hírügynökségek. [329]

(3) A dezinformáció terjedésének fokozását nagyban segítheti a hamis profilokat irányító botok bevetése. Ezek olyan automatizált szoftverek, amiket olyan egyszerű online művelet elvégzésére hoztak létre, mint a bejegyzések megosztása, komment kedvelése, hashtagek felfuttatása stb. Ezek egyik altípusát, a *politikai botokat* gyakran használják arra, hogy a választási kampányokban egy jelölt vagy politikai párt üzeneteit terjesszék, de akár agresszív kampányolásra is programozhatják, mely során úgy képesek manipulálni – egyúttal torzítani – a nyilvános diskurzust, hogy propagandát terjesztenek, újságírókat támadnak, vagy éppen politikai vezetőket járatnak le.[330] Samuel Woolley a politikai botokkal kapcsolatban megjegyzi, hogy elterjedt félreértés, hogy azokat úgy tervezték, hogy a felhasználókkal kommunikáljanak és befolyásolják azok nézeteiket. Ezek elsődleges feladata valójában az, hogy a közösségi média platformok algoritmusait manipulálják, mivel azok döntik el, hogy milyen tartalmak jelennek meg hangsúlyosan a felhasználóknak. A botok oly módon igyekeznek befolyásolni ezt a folyamatot, hogy mesterségesen felduzzasztják bizonyos tartalmak népszerűségét, például sok lájkkal, megosztással, hozzászólással, így segítve azt, hogy a lehető legtöbb felhasználóhoz jussanak el.[331] Fontos, hogy minél többször minél több helyen jelenjenek meg az ilyen tartalmak, hiszen a dezinformációs stratégiák módszertanához tartozik az ismétlés általi megerősítés (ahol a hamis információkat gyakran és különböző csatornákon ismétlik, hogy legitimitást és hitelességet sugalljanak).[332]

---

[329] Ilyenek például azok a kérészéletű, Magyarországon is rendszerint az országgyűlési választásokat megelőző hónapokban aktivizálódó, tisztán pártpolitikai érdekek mentén működő, impresszum nélküli online kiadványok, melyek nincsenek bejegyezve az NMHH-nál, nem minősülnek sajtóterméknek és amelyekkel szemben sajtó-helyreigazítási per sem indítható. Ezek a weboldalak néhány hónapon keresztül – főleg a kampányidőszakban – osztanak meg cikkeket, melyek sok esetben erős keretezéssel, torzítással közöltek – fő célok között a dezinformáció, karaktergyilkosság és lejáratás áll. A választások végével ezen weboldalak aktivitása és közösségi média hirdetései is elapadnak. Lásd bővebben: Ambrus Balázs (2023): Megtévesztő híroldalakkal építik magukat a pártok. *Index*. Elérhető: https://index.hu/belfold/2023/11/07/impresszum-lapok-kommunikacio-helyi-hirek-propaganda/

[330] Howard, P. N., Woolley, S., & Calo, R. (2018): Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of information technology & politics*, 15(2), 81-93.o.

[331] Wooley, Samuel C. (2020): Bots and Computational Propaganda: Automation for Communication and Control. In Social Media and Democracy: State of the Field. (Szerk:) Persily, Nathaniel and Tucker, Joshua A.Cambridge, Cambridge University Press, 89-110.o.

[332] A botok hasznosságát szemlélteti a Kínai Kommunista Párthoz (KKP) köthető „*Spamouflage"* online dezinformációs művelet, amely aktuális nemzetközi konfliktusokat, globális eseményeket felhasználva igyekszik társadalmi és politikai polarizációt előidézni, valamint Amerika-ellenes vagy a KKP-t támogató narratívákat terjeszteni. Stratégiájuk, hogy fényképek, videók és hírcikkek képernyőképeit használják (más esetekben GenMI

7.3 GenMI a dezinformáció szolgálatában

A 2010-es években indított dezinformációs műveletek sikeres végrehajtása még több korlátba ütközött. Ezek közül kiemelendő a rendelkezésre álló (anyagi és humán) erőforrások korlátozottsága, az üzenetek és közvetített tartalmak minőségének alacsony színvonala[333], valamint a *rejtett* műveletek és akciók könnyű észlelhetősége. Mivel ezek a gyenge minőségű üzenetek könnyen észlelhetőek és felismerhetőek voltak, korlátozott volt a kampányok hatékonysága is (a célközönség gyakrabban utasította el a hamis információkat).

A Nagy Nyelvi Modellek megjelenésével és szabadon hozzáférhetővé válásával ezen korlátok egy jelentős része áthidalhatóvá vált. Az LLM segítségével könnyen, gyorsan, minimális anyagi költségekkel és minimális emberi erőforrás felhasználásával lehet nagy mennyiségű, hitelesnek tűnő dezinformációt előállítani. Továbbá a felhasználók egyéni preferenciáihoz igazodó célzott dezinformáció is pontosabb, hatékonyabb lehet segítségével. Több kutatás is arra a következtetésre jutott, hogy a legújabb Nagy Nyelvi Modellek már kreatívabbnak[334] és

---

által létrehozott képeket), melyek látványos, figyelemfelkeltő, szarkasztikus grafikai illusztrációkként szolgálnak fő narratívák képi közléséhez. A kampány gerincét azonban nem a tartalomgyártás, hanem azok a botfiókok alkotják, melyek egy viszonylag kisebb csoportja a tartalom közzétételére, míg egy nagyobb csoport, a bejegyzések kedvelésével, megosztásával és kommentálásával foglalkozik (annak érdekében, hogy felnagyítsák a bejegyzések láthatóságát és hatását). Ez a kampány több mint 50 különböző platformon és fórumon volt aktív, beleértve a Facebookot, Instagramot, TikTokot, YouTube-ot és a Twittert (X). A Meta jelentése szerint több mint 7,700 Facebook-fiókot távolítottak el, amelyek részt vettek ebben a kiterjedt online „spam" műveletben Lásd: Josh Taylor (2023): Meta closes nearly 9,000 Facebook and Instagram accounts linked to Chinese 'Spamouflage' foreign influence campaign. The Guardian. Elérhető: https://www.theguardian.com/australia-news/2023/aug/30/meta-facebook-instagram-shuts-down-spamouflage-network-china-foreign-influence (2023. 08. 30.)

[333] Például visszatérő jelenség volt, hogy a dezinformációs kampányok sok esetben angol nyelvterületű országot céloztak meg, azonban a szöveges tartalmak előállítására nem angol anyanyelvű, alacsony képzettségű, olcsó munkaerőt alkalmaztak, akiknek a helyesírási és nyelvtani hibái hamar leleplezték az üzenetek kétes eredetét.

[334] Hubert és társai egy 2024-es tanulmányukban emberi résztvevők (N=151) és a GPT-4 GenMI válaszait hasonlították össze a következő három divergens gondolkodási feladat megoldásában: (1) Alternatív felhasználási asszociáció (Alternative Uses Task, AUT), ahol résztvevőknek egy hétköznapi tárgy, újszerű kreatív felhasználási módjait kellett kitalálniuk; (2) Lehetséges következmények feladat (Consequences Task, CT), mely során résztvevőknek képzeletbeli forgatókönyvek lehetséges következményeit kellett megfogalmazniuk; (3) Divergens asszociációs feladat (DAT): A résztvevőknek 10 olyan főnevet kellett megadniuk, amelyek kinézetük, tartalmuk, funkciójuk stb. tekintetében a lehető leginkább különbözőek. Mindhárom vizsgált feladatra igaz volt, hogy az LLM válaszai egyrészt eredetibbek voltak, (azaz egyedibbek, szokatlanabbak, „out-of-the-box" jellegűek) és kidolgozottabbak is (vagyis részletesebbek, hosszabbak, alaposabbak) mint az emberi résztvevők válaszai. Az LLM-ek esetében feltűnő volt azonban, hogy nagyobb gyakorisággal használtak újra és újra ismétlődő szavakat válaszaik generálása során. Bővebben: Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024): The current state of artificial intelligence generative language models is

meggyőzőbbnek[335] bizonyulnak a szöveges tartalomgyártásban terén, mint az átlagemberek.[336] Bai és társai például egy 2023-as tanulmányukban számoltak be három különböző kísérletről, melyet 2022 októbere és 2022 decembere között végeztek, összesen 4836 résztvevővel. (1203, 2023, 1610 fő). A kísérletek során különféle politikai üzenetekkel próbálták megváltoztatni a részvevők nézeteit olyan témákban, mint a dohányzással kapcsolatos szabályozások szigorítása, fegyverviselési jogok korlátozása vagy a széndioxid adó bevezetése. A kísérletben három üzenetkategóriát használtak: (1) MI által generált érvek (AI Condition); (2) természetes személyek által létrehozott üzenetek (human condition); (3) öt MI által írt szöveg, melyek közül emberi döntés alapján került kiválasztásra egy (human in the loop condition). A résztvevők mindegyik esetben véletlen besorolás alapján kapták a politika tartalmú üzeneteket, így azok emberi vagy szintetikus jellegéről nem volt tudomásuk. Mindhárom kísérletben az MI által generált üzenetek következetesen meggyőzőnek bizonyultak a részvevők számára (minimum 2-4 pontos változás a 101 pontos összesített attitűdskálán). Bár a különböző kategóriák közötti a meggyőzés hatékonysága statisztikailag elhanyagolható volt, a résztvevők értékelése szerint az MI által létrehozott szövegek inkább tűntek objektívnek, tényszerűnek, következetesnek és jól követhető, racionális érvekből felépítettnek (*better informed, more logical, less angry*).[337]

Jól érzékelteti e gyakorlatok való életben történő előfordulását is, hogy az OpenAI egy 2023-as jelentésében arról írt, hogy az orosz Doppelgänger nevű csoport a ChatGPT-t használta fel olyan álhírek terjesztésére, amelyek célja az Ukrajna iránti szolidaritás és bizalom aláásása volt. A csoport több nyelven terjesztett olyan hamis információkat (jól felépített és

more creative than humans on divergent thinking tasks. *Scientific reports*, 14(1), 3440. https://doi.org/10.1038/s41598-024-53303-w

[335] Goldstein A. Josh, Jason Chao, Shelby Grossman, Alex Stamos, Michael Tomz (2024): How persuasive is AI-generated propaganda? PNAS Nexus. Volume 3, No 2, 1-7o.; Vykopal, I., Pikuliak, M., Srba, I., Móro, R., Macko, D., & Bieliková, M. (2023). Disinformation Capabilities of Large Language Models. ArXiv, abs/2311.08838. https://doi.org/10.48550/arXiv.2311.08838.

[336] Sőt egy 2024-es tanulmány megállapítása szerint a GPT 3.5 LLM már nem csak kreatívabb, de humorosabb is (200 a kísérletben résztvevő emberi bíráló értékelése alapján). Lásd: Gorenz D, Schwarz N (2024): How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PLoS ONE* 19(7), 2-10.o. https://doi.org/10.1371/journal.pone.0305364

[337] Ezzel szemben az emberek által írt üzenetek inkább személyes narratívákra és történeti elbeszélésekre épültek (pl: tapasztalatok, élmények, képi-leíró elemek használata stb.). A tanulmány azt a következtetést állapítja meg, hogy a GPT-3as modell által generált kimenetek, csakúgy, mint a profi marketingesek által létrehozott politikai üzenetek, képesek arra, hogy érdemben befolyásolják az azt olvasó, fogyasztó személyek előzetesen kialakult nézetét. Lásd: Bai, Hui, Jan G. Voelkel, johannes C. Eichstaedt, and Robb Willer. (2023): Artificial Intelligence Can Persuade Humans on Political Issues. OSF Preprints. 7.o. https://doi.org/10.31219/osf.io/stakv

megfogalmazott Facebook-bejegyzések, cikkek, hozzászólások formájában), amelyek oroszpárti, valamint ukrán és nyugat ellenes üzeneteket közvetítettek.[338]

A modern dezinformációs műveletek azonban már nem csak szöveges üzenetekre korlátozódnak; a meggyőzőerejük fokozása érdekében a kifinomult multimédiás tartalmakra is nagyban építkeznek.

7.4 Szintetikus hang- és videótartalmak

Az audiovizuális tartalmak előre meghatározott céllal történő szándékos manipulációja nem új keletű jelenség, ilyen gyakorlatok már a fotózás és a filmkészítés kezdetei óta léteznek.[339] Lényeges különbség azonban, hogy korábban e manipulációs technikák alkalmazásához technikai tudásra és szakmai tapasztalatra volt szükség, mely jelentős anyagi és humánerőforrást igényelt. Napjainkra jelentős változások mentek végbe ezen a területen, a modern, könnyen hozzáférhető, felhasználóbarát digitális technológiák lehetővé teszik a fotók és videók – magas minőségű – szerkesztését és manipulálását akár néhány egyszerű kattintással. Ezzel a vizuális tartalmak hamisítása olyan szintre jutott, ami korábban elképzelhetetlen volt. Az amatőr és professzionális felhasználók számára egyaránt elérhetővé váltak azok az eszközök, amelyek lehetővé teszik a képek és mozgóképek meggyőző hamisítását. [340]

[338] Például hamis híroldalakat hoztak létre és különböző hírességek nevével fiktív idézeteket generáltak, melyeken keresztül olyan hamis információkat közvetítettek, mint hogy a nyugati országok az Ukrajnának nyújtott támogatást Izraelbe irányítják át. Lásd bővebben: OpenAI (2023): AI and Covert Influence Operations: Latest Trends. REPORT. 17-23.o. Elérhető: https://downloads.ctfassets.net/kftzwdyauwt9/5IMxzTmU clSOAcWUXbk VrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf (2023. 10. 22.)

[339] Már a Szovjetunióban is előszeretettel fordultak fényképek manipulációjához (retusálásához) a történelem újraírása érdekében. Ez különösen jellemző volt a sztálini időszakra, amikor a politikai tisztogatások során sok korábbi szövetségest és közismert személyt eltávolítottak a hivatalos történelemből. Az egyik leghíresebb példa erre, amikor Lev Trotszkijt, aki korábban Sztálin egyik legfőbb riválisa volt, eltávolították a hivatalos fotókról és dokumentumokból, miután kegyvesztett lett. De a második világháború idején is készültek olyan szerkesztett (hamisított) propaganda filmek is, amelyek azt voltak hivatottak bemutatni, milyen tisztességesen és emberségesen bánnak a náci katonák a zsidó származású foglyokkal. Lásd: Margry Karel (1992): Theresienstadt (1944–1945), the Nazi propaganda film depicting the concentration camp as paradise. Historical Journal of Film, Radio and Television. Vol (12) No. 2,1 45–162.o. doi: 10.1080/01439689200260091.

[340] Guld Ádám (2023): A deepfake és CGI-technológia az influencer marketing szolgálatában: így formálják át a digitális karakterek az ismertségipar működését. In: Aczél, Petra; Veszelszki, Ágnes (szerk.) Deepfake: a valótlan valóság. Budapest, Gondolat Kiadó. 189-190.o.

A GenMI modellek segítségével létrehozott vagy manipulált audiovizuális tartalmak (deepfake tartalmak)[341] járványszerű elterjedését valójában évtizedes technológiai folyamat előzte meg, melynek fontos állomását képezte az un. antagonisztikus hálózatok (Generative Adversarial Networks, GAN) 2014-es megjelenése. A GAN a gépi tanulás innovatív területét képezi, ahol az új tartalom létrehozása két különálló neurális hálózat együttműködésén alapul; Ezek közül a *generátor hálózat* (generator) felelős az új adatok generálásáért, célja, hogy olyan adatokat hozzon létre, amelyek valóságosnak tűnnek (audio, kép, videó). A *diszkriminátor hálózat* (discriminator) pedig értékeli, hogy a generátor által létrehozott adatok valóságosak-e vagy sem, célja, hogy különbséget tegyen a valódi és a mesterséges adatok között. A GAN működése során a generátor és a diszkriminátor egyfajta (antagonisztikus) versenyben van egymással, ahol a generatív hálózat egyre jobban igyekszik utánozni a valóságot, hogy becsapja a diszkriminatív hálózatot, amely viszont egyre hatékonyabban próbálja felismerni és kiszűrni a generált adatokat. Ez a folyamatos versengés hozzájárul a generatív modell szüntelen fejlődéséhez és finomításához, így a generált adatok egyre nehezebben különböztethetők meg a valós adatoktól. Ennek köszönhetően a GAN-ok számos területen váltak alkalmazhatóvá, például képfeldolgozásban, művészeti alkotások generálásában, vagy akár realisztikus szimulációk létrehozásában is.[342]

Több kutatás is alátámasztja, hogy az MI által mesterségesen előállított vagy manipulált audiovizuális tartalmak képesek kihasználni az emberi agy vizuális információkra való fogékonyságát. Hameleers és munkatársai kutatása szerint a téves információk sokkal meggyőzőbbnek tűnhetnek, ha audiovizuális formában jelennek meg, szemben a szöveges tartalommal. Ez a jelenség abból adódik, hogy az audiovizuális tartalmak – amelyek egyidejűleg használják a látványt és a hangot – erőteljesebb érzelmi hatást képesek kiváltani, így nagyobb befolyást gyakorolhatnak a nézők vagy hallgatók véleményére, gondolataira, hiedelmeire.[343]

[341] Olyan digitális médiatartalom (elsősorban audiovizuális formátumú, de lehet például kép, animáció vagy 3D-modell), amelyet részben vagy egészben MI technológiák felhasználásával hoztak létre, illetve olyan már létező tartalom, amelyet MI-alapú eszközökkel módosítottak vagy manipuláltak. A rendelet 3. cikk (60) fogalommeghatározásában: az MI által generált vagy manipulált kép, audio- vagy videotartalom, amely hasonlít létező személyekre, tárgyakra, helyekre, entitásokra vagy eseményekre, és amely egy személy számára megtévesztő módon autentikusnak vagy valóságosnak tűnne.

[342] Goodfellow, Ian, Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014): Generative Adversarial Nets. 2-8.o. arXiv:1406.2661

[343] Hameleers, M, T. E. Powell, T. G. Van Der Meer, L. Bos. (2020): A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, 37(2):281–301.o.

7.4.1 Megtévesztő szándékkal készített deepfake – esettanulmányok a közelmúltból

A megtévesztő szándékkal készített deepfake tartalmak hatásai szerteágazóak lehetnek. Danielle Citron és szerzőtársa arra hívják fel a figyelmet, hogy az olyan (idő)érzékeny területeken bevetett félrevezető vizuális tartalmak, mint a nemzetközi kapcsolatok, a diplomácia vagy a tőzsde kiemelt kockázatot képviselnek.[344] Egy olyan hamisított videó, amelyben egy politikus provokatív vagy kompromittáló kijelentéseket tesz, súlyos nemzetközi konfliktusokat idézhet elő, hasonlóképpen a tőzsdén egy megtévesztő hamis nyilatkozat azonnal nemvárt árfolyammozgásokhoz vagy árfolyameséshez vezethet.

A szintetikus kép és videótartalmak elterjedt változatát képviselik azon politikai deepfakek, melyeknek a közvélemény manipulálásán túl, a politikai polarizáció növelése vagy konkrét választások befolyásolása is célja.[345] Ezek illusztrálására – a közelmúltból is – több esettanulmány áll rendelkezésre:

(1) Az orosz-ukrán háború kirobbanását követően, az orosz erők aktívan használták a deepfake technológiát abból a célból, hogy megtörjék az ukrán nép egységét és háborús elszántságát, illetve, hogy megrendítsék a bizalmat a katonák, civilek és a nemzetközi

[344] Különösen veszélyes lehet egy strategikusan időzített deepfake tartalom megjelenése, például egy nemzetközi csúcstalálkozó alatt. Az ilyen manipulált tartalom képes lehet annyira felkorbácsolni a közvéleményt, hogy átmenetileg ellehetetleníti a tárgyalások folytatását vagy egy küszöbön álló megállapodás megkötését. Mindez komoly kockázatot jelent mind a nemzetbiztonságra, mind a globális stabilitásra nézve. Lásd: Chesney, Bobby, Danielle Citron (2019): Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, Vol. 107, 1775-1786.o.

[345] Az audio deepfake tartalmak előállításának közvetlen politikai kockázatai miatt döntött úgy az OpenAI hogy a saját Voice Engine nevű GenMI modelljét egyelőre nem teszi szabadon elérhetővé. A modell képes bármely 15 másodperces hangmintából a bementi forrással teljesen megegyező szintetikus hangot generálni. A hang klónozását követően a felhasználó beírhat szöveget a Voice Engine-be, és megkapja a GenMI által generált hangkimenetet (text-to-audio model). Bár rengeteg pozitív felhasználási módja lehet a jövőben – például a beszédkárosodással küzdők számára is lehetőséget nyújthat a saját hangjuk visszanyerésére – az OpenAI úgy döntött, hogy a hangklónozásból eredő potenciális veszélyekre tekintettel – kiváltképp a 2024-es globális választások évében – nem teszi még elérhetővé a modellt. Hivatalos oldalukon azt írják: Tudatában vagyunk annak, hogy az emberi beszédet imitáló tartalmak előállításának komoly kockázatai vannak, kiváltképp a választási évben. A Voice Engine-t ma tesztelő partnerek elfogadták a felhasználási irányelveinket, amelyek tiltják létező személyek és szervezetek személyazonosságának engedély nélküli vagy jogosulatlan utánzását. (We recognize that generating speech that resembles people's voices has serious risks, which are especially top of mind in an election year. The partners testing Voice Engine today have agreed to our usage policies, which prohibit the impersonation of another individual or organization without consent or legal right.) Edwards, Benj (2024): OpenAI Can Re-Create Human Voices—but Won't Release the Tech Yet. *WIRED*. Elérhető: https://www.wired.com/story/openai-voice-engine-artificial-intelligence-release (2024. 04. 03.); OpenAI (2024): Navigating the Challenges and Opportunities of Synthetic Voices. Blog, Product Announcements. Elérhető: https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices (2024. 04. 02.)

közvélemény körében, ezáltal is erősítve Moszkva tárgyalási pozícióit. A legjelentősebb ilyen próbálkozás 2022 márciusában történt, amikor egy olyan hamis videó terjedt el az interneten, amelyben az ukrán elnök, Volodimir Zelenszkij felszólította katonáit, hogy tegyék le a fegyvert. A videót széles körben megosztották, számos platformon, köztük ukrán honlapokon, a Facebookon, a Twitteren, a YouTube-on és több orosz csatornán is megjelent, de a vizsgálatok hamar megerősítették, hogy MI által generált deepfake volt.[346]

2023 novemberében egy újabb hasonló hamis videó látott napvilágot, ezúttal az ukrán fegyveres erők főparancsnokát, Valerij Zaluzsnijt ábrázoló manipulált felvételen, amelyben Zelenszkijhez hasonlóan ő is a fegyverletételre szólított fel. Más hamis videókban Zelenszkijt szvasztika viselete közben, illetve egy LMBTQ-felvonuláson lépve mutatták be. Noha a hivatalos orosz álláspont elutasítja a videók orosz eredetét, joggal feltételezhető, hogy ezek az Oroszország által kezdeményezett és terjesztett deepfake-ek a kiterjedt információs hadviselés és befolyásolás eszközeként szolgáltak.[347]

(2) A 2023-es szlovák parlamenti választások alatt több deepfake hangfelvétel is elterjedt a közösségi médiában, ezekben a Progresszív Szlovákia (Progresívne Slovensko) EU-párti, liberális párt elnöke, Michal Šimečka és Monika Tódová nevű újságíró között meghamisított párbeszéd volt hallható. A beszélgetés során többek között olyan témákat érintettek, mint a választási eredmények meghamisításának terve vagy a sör árának felemelése. A hanganyagokat a szavazás kezdetét megelőző 48 órás kampánycsend időszaka alatt tették közzé, emiatt – a szlovák választási szabályok értelmében – a posztok cáfolására tett kísérletek, illetve az azokkal kapcsolatos politikai kommunikáció is akadályozott volt. Nem véletlen a célzott támadás a párt és a politikusok ellen, hiszen az akkori legfrissebb közvélemény-kutatások szerint a Progresszív Szlovákia felzárkózott az addigi vezető párthoz, a Smerhez. A választásnak nemzetközi politika jelentősége is volt, mivel a Robert Fico által vezetett SMER a választási kampány során nyíltan EU ellenes retorikát folytatott, illetve megválasztása esetén Ukrajna katonai segítségek leállítását is ígérte. A videókat végül eltávolították a YouTube-ról, de a Facebookon továbbra is elérhetők maradtak.[348] (A Meta

[346] Allyn, Bobby (2022): Deepfake Video of Zelenskyy Could Be 'Tip of the Iceberg' in Info War, Experts Warn. *NPR*. Elérhető: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-expertswar-manipulation-ukraine-russia (2023. 06. 14.)

[347] Byman, Daniel, Linna Jr., D. W., Subrahmanian, V. S. (2024): Government Use of Deepfakes. Center for *Strategic & International Studies*, 5-23.o.

[348] Meaker, Morgan (2023): Slovakia's Election Deepfakes Show AI Is a Danger to Democracy. WIRED. Elérhető: https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/?redirectURL=https%3A%2F%2F (2023. 11. 05.)

szabályzata jelenleg nem tiltja a megtévesztő információt közvetítő audio, tehát kizárólag hangalapú deepfakek megosztását).[349]

(3) 2024 januárjában a New Hampshire-i előválasztások előtt deepfake Biden-robothívásokkal igyekeztek eltántorítani választókat szavazatuk leadásától. A robothívás – Biden elnök hangját imitálva – arra szólította fel a szavazókat, hogy ne szavazzanak, és azt a megtévesztő üzenet hangsúlyozta, hogy a szavazatukat a novemberi választásokra kell tartalékolniuk.[350]

(4) 2023. május 22-én reggel rövid időre elterjedt a közösségi médiában egy hamis, GenMI által generált kép az amerikai Pentagon épülete előtti robbanásról. Ugyan a védelmi minisztérium hivatalos közleményében hamar cáfolta a kép valóságtartalmát, a robbantásról szóló hamis hírek egy rövid időre így is megingatták a főbb részvénypiaci indexeket. Külön nehezítette a helyzetet, hogy a híresztelést terjesztő Twitter-fiókok többségén kék pipajel volt, ami korábban a hitelesített fiókokat jelezte, de Elon Musk felvásárlását követő változtatások nyomán mára bárki megvásárolhatja a Twitter Blue előfizetéssel. A hamis Pentagon-képet megosztó fiókok között volt egy a Bloomberg hírügynökséget utánzó fiók is, valamint a valódi, Kreml-közeli orosz RT hírszolgálat fiókja. Az RT később törölte a bejegyzését, míg a hamis Bloomberg-fiókot felfüggesztette a Twitter.[351]

7.4.2 Explicit deepfake tartalmak

[349] A vállalat várható jövőbeli tartalommoderálási politikája szerint az MI által generált videókat, képeket és hanganyagokat címkézni fogják, hogy a felhasználók tudják, hogy ezek mesterségesen lettek létrehozva vagy módosítva. Az oversight board, amely Meta tartalompolitikai felülvizsgálatait végzi, javasolta, hogy a cég bővítse a közvélemény megtévesztésére szánt video deepfakek tiltását az audió deepfake-ekre is. Hozzáteszik, hogy a manipulált tartalmak eltávolítását csak a legmagasabb kockázatú esetekre kell korlátozni, ahol a tartalom egyértelműen kárt okoz. Frissítés: A Meta 2024 április 5-én jelentette be, hogy jelentős változtatásokat fog eszközölni a digitálisan létrehozott és módosított tartalmakra vonatkozó szabályzatában. A vállalat májustól Made by AI (Mesterséges intelligenciával készült) címkéket fog alkalmazni a Facebookon és az Instagramon közzétett, GenMI-vel generált videókra, képekre és hanganyagokra is, lényegesen kiterjesztve korábbi szabályzatát, mely csak a manipulált videók egy szűk szeletére vonatkozott. Lásd: Meta, Bickert, Monika (2024): Our Approach to Labeling AI-Generated Content and Manipulated Media. META Newsroom. Elérhető: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/(2024. 04. 05.)

[350] Eliott, Vittoria (2024): The Biden Deepfake Robocall is Only the Beggining. *Wired*. Elérhető: https://www.wired.com/story/biden-robocall-deepfake-danger/ (2024. 01. 31.)

[351] Brian Bushard (2023): Fake Image Of Explosion Near Pentagon Went Viral—Even Though It Never Happened. *Forbes*. Elérhető: https://www.forbes.com/sites/brianbushard/2023/05/22/fake-image-of-explosion-near-penta gon-went-viral-even-though-it-never-happened/

Egy ritkábban tárgyalt, de szintén (fokozódó) aggodalomra okot adó trend a GenMI eszközök segítségével létrehozott szexuális tartalmú deepfake képek és videók rohamos terjedéséhez köthető. Az efféle tartalmak előállítása szintén bármiféle grafikus tudás vagy videószerkesztői előismeret nélkül lehetséges a lényegében bárki által hozzáférhető és nyílt forráskódú technológiák (pl. a CivitAI szoftver) segítségével.[352] A GenMI eszközök felbukkanásával drámai növekedés volt megfigyelhető az explicit deepfake videók számában; míg 2022-ben az interneten körülbelül 3725 ilyen tartalom volt elérhető, ez a szám 2023-ban 21019-re emelkedett. A New York Times egy 2024-es cikkében egyenesen járványnak nevezte a jelenséget. [353]

Az előző alfejezethez kötődően megemlítendő, hogy akárcsak a politikai deepfakek esetében, az explicit tartalmak létrehozásának és terjesztésének is áldozatául eshetnek ismert közéleti személyek, politikusok. A legismertebb eset Giorgia Meloni-hez, Olaszország miniszterelnökéhez kötődik, aki 100 000 euró kártérítést követelt, miután deepfake pornográf videókat készítettek és tettek közzé róla az interneten (értelemszerűen tudta és hozzájárulása nélkül). A videókban Meloni arcát egy másik személy testére illesztették digitálisan egy MI eszköz segítségével, majd feltöltötték azokat egy amerikai weboldalra.[354]

7.4.3 A megtévesztésen túl

[352] A CivitAI platformján több ezer szintetikus színész/színésznő érhető el (ezek gyakran valós emberek képeiből készülnek, amelyeket sok esetben engedély nélkül gyűjtenek össze az internet különböző szegleteiről) és bár a platform ezt nyilvánosan nem reklámozza, amennyiben a szolgáltatást igénybe vevő ilyen célra kívánja azt használni, pornográf tartalmak előállítása is könnyen lehetségessé válik. Sőt a felhasználó a saját eszközén tárolt, általa ismert személyről készült képet is felhasználhatja hasonló célra. A témáról, és a CivitAI platformján elérhető erősen szexuális tartalmú beágyazott szerkesztőfilterekről bővebben: Maiberg Emanuel (2023): Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale. *404media.* Elérhető: https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/

[353] A szexuális tartalmú deepfakek túlnyomórészt nőket céloznak meg, egy felmérés szerint az ilyen videókban szereplő alanyok 99%-a nő.

[354]A hatóságok videók elkészítésével egy 40 éves férfit és 73 éves apját gyanúsítják, akiket mobileszköz-nyomkövetéssel azonosítottak Meloni ügyvédje, Maria Giulia Marongiu azt nyilatkozta, hogy amennyiben a bíróság megítéli ügyfele számára a 100 000 eurós kártérítési összeget a miniszterelnök olyan nők támogatására kívánja fordítani, akik férfiak általi erőszak áldozatai lettek. Ez az összeg szimbolikus jelentőségű, és célja, hogy üzenetet küldjön a hasonló visszaélések áldozatainak, ne féljenek feljelentést tenni. Gozzi, Laura (2024): Giorgia Meloni: Italian PM seeks damages over deepfake porn videos. *BBC.* Elérhető: https://www.bbc.com/news/world-europe-68615474 (2024. 03. 21.); Starcevic, Seb (2024): Italy's Giorgia Meloni called to testify in deepfake porn case. Elérhető: https://www.politico.eu/article/italian-pm-giorgia-meloni-called-to-testify-in-deepfake-porn-case/ (2024. 03. 21)

A deepfake tartalmak hatékonysága nagyban függ attól, hogy mennyire tudnak már meglévő kétségekre vagy bizonytalanságokra építeni. Legnagyobb hatásukat akkor érik el, amikor egy olyan környezetben jelennek meg, ahol már eleve jelen van valamiféle gyanú vagy bizonytalanság.[355] Ilyenkor a videó és a háttérben lévő információk – vagy éppen a kellő információ hiánya – egymást erősítik: a videó alátámasztja a korábbi feltételezéseket, miközben a meglévő információk – vagy azoknak a hiánya – hihetőbbé teszik a videón látottakat. Ez a kölcsönös erősítés különösen veszélyes lehet, ha társadalmi megosztottságokat céloznak meg vele; márpedig a mai demokráciákban a politikai nézetek és meggyőződések fokozódó polarizációja tapasztalható (amit még a korábban tárgyalt társadalmi-pszichológiai folyamatok is tovább erősítenek). Ennek is köszönhető, hogy a deepfake tartalmak egyre gyakrabban képesek kijátszani a tartalomfogyasztók kritikai érzékét és gondolkodását.[356]

Egy az Egyesült Királyságban végzett nagymintás felmérés (N=2005), amely a hamisított digitális tartalmakkal kapcsolatos állampolgári attitűdöket vizsgálta, kimutatta, hogy a deepfake technológiák jó eséllyel növelik az általános bizonytalanságot, szkepticizmust és cinizmust a lakosság körében. A felmérés megállapította, hogy az internethasználók egy része hiába volt képes felismerni a deepfake videókat, ezeknek a hamisítványoknak a puszta jelenléte számottevően csökkentette az online hírekbe vetett általános bizalmukat. A tanulmány írói hangsúlyozták, hogy az elbizonytalanodás érzése nem korlátozódik csupán az adott deepfake videót tartalmazó híradásra, hanem általánosságban véve is csökkenti az online megjelenő hírekbe vetett bizalmat (kiváltképp igaz volt ez a közösségi médiában megjelenő hírekre).[357]

A deepfake-ek megjelenése valójában csak tovább fokoz egy olyan trendet, melynek a jelei már korábban is megjelentek. A híradásokba vetett bizalom az elmúlt egy évben további 2

[355] Herke Csongor (2023): Deepfake: Áldás vagy átok? Jogi szabályozási szempontok. *Pro Futuro - A Jövő nemzedékek joga*, 13(1), 169.o.

[356] A kialakult polarizált közhangulatban nem is várható el reálisan, hogy a választópolgárok a videós tartalmak hitelességét vagy hamisságát vizsgálják; ehelyett sokan inkább azt mérlegelik, hogy azok összhangban vannak-e saját politikai nézeteikkel vagy sem. Bővebben: Mráz Attila (2022): Deepfake, demokrácia, kampány, szólásszabadság. In: Török Bernát–Ződi Zsolt (szerk.): A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi Kiadó, 254.o.

[357] Vaccari, C., & Chadwick, A. (2020): Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1) https://doi.org/10.1177/2056305120903408.

százalékponttal csökkent. A teljes mintánkból tíz emberből átlagosan csak négyen (40%) mondják azt, hogy a legtöbbször megbíznak a hírekben.[358]

A növekvő társadalmi bizalmatlanság egyik sajátos veszélye, hogy könnyen önmagát erősítő folyamattá válhat. Chesney és Citron még 2018-ban alkották meg a *hazugság osztaléka* (liar's dividend) fogalmát, mely azt a jelenséget írja le, amikor valaki a szintetikus tartalmak (jelen esetben a deepfake-ek) létezésére hivatkozva kétségbe vonja a valós bizonyítékok hitelességét, így rombolva az emberek általános bizalmatlanságát a média és az objektív bizonyítékok iránt. A hazugság osztalék különösen hasznos lehet a politikusok számára, akik valós botrányokat próbálhatnak meg elhárítani azáltal, hogy azokat hamis híreknek vagy deepfake-nek állítják be.[359]

A 2023-as évben tapasztalt innovációshullám bizonyítékul szolgált arra is, hogy a jövőben a jelenleginél még inkább fejlettebb, intuitívabb és mindenki számára könnyen hozzáférhető modellek, szolgáltatások jelennek majd meg az online térben. Ahogy a GenMI technológiák fejlődnek, a deepfake-ek egyre valósághűbbek és nehezebben észlelhetőek lesznek. Ez növeli az előbb ismertetett politikai és társadalmi kockázatokat, valamint szükségessé teszi mind a kifinomult felismerési és szűrési technológiák fejlesztését, mind pedig a kockázatokat adekvát módon kezelni képes jogi keretrendszerek megalkotását.

7.5 Az Európai Unió dezinformációs stratégiája

Az Európai Unió az elmúlt években jelentős erőfeszítéseket tett a dezinformáció elleni küzdelem terén, és számos olyan intézkedést hozott, amelyek célja a digitális tér biztonságának növelése és a demokratikus folyamatok védelme volt. Az Unió ezirányú törekvései 2015 márciusáig vezethetők vissza, ekkor bízta meg az Európai Tanács az Unió külügyi és biztonságpolitikai főképviselőjét, hogy dolgozzon ki egy *stratégiai kommunikációra vonatkozó cselekvési tervet* Oroszország folyamatos dezinformációs kampányainak kezelésére. E cselekvési terv eredményeként hoztak létre egy új stratégiai kommunikációs részleget (StratCom) az Európai Külügyi Szolgálaton belül. A StratCom első munkacsoportjának – az úgynevezett keleti stratégiai kommunikációs munkacsoportnak – feladatává vált az Unión kívülről (elsősorban Oroszországból) származó dezinformáció elleni

[358] Reuters Institute (2023): Digital News Report. 9-10.o. Elérhető: https://reutersinstitute.politics.ox.ac.uk /digital-news-report/2023

[359] Lásd: Chesney, Bobby, Danielle Citron (2019): Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, Vol. 107, 1775-1786.o.

fellépés, valamint pozitív stratégiai kommunikáció kialakítása és terjesztése az Unió keleti szomszédságában.[360]

A Bizottság 2017 végén hozta létre az *Álhírekkel és dezinformációval foglalkozó magas szintű szakértői csoportot* (High-Level Expert Group on Fake News and Disinformation), melynek legfőbb feladatául szabták, hogy konkrét, a gyakorlatban alkalmazható tanácsokkal szolgáljon a dezinformáció kezeléséhez. A csoport a rákövetkező év márciusában nyújtotta be jelentését, amelyben többek között az átláthatóság növelésével, a felhasználók és hírolvasók médiatudatosságának elősegítésével, illetve a platformszolgáltatókkal közös, hatékony együttműködéssel kapcsolatban fogalmaztak meg javaslatokat.[361]

A jelentés fontos alapjául szolgált a Bizottság által ugyanazon év áprilisában publikált „*megközelítés az online félretájékoztatás kezelésére*" című bizottsági közleménynek.[362] E megközelítésben a Bizottság megállapítja, hogy a *szigorú szakmai előírásokon alapuló, jól működő, szabad és plurális információs környezet* elengedhetetlen feltétele az egészséges demokratikus viták létrejöttének, egyúttal pedig a demokratikusan működő társadalmak hosszútávú fenntartásának. Ennek előmozdítása érdekében ajánlásokat fogalmazott meg a *magas színvonalú információk előnyben részesítésén alapuló* digitális ökoszisztémák megteremtésére.[363] A közlemény ezenfelül szorgalmazza egy európai gyakorlati kódex kidolgozását is a dezinformáció visszaszorítása céljából.

---

[360] A Kelet-StratCom munkacsoport működteti például az EUvsDisinfo weboldalt is, melynek elsőszámú célja, hogy felfedje az orosz forrásból származó hamis és megtévesztő állításokat. A megfogalmazott hosszútávú célok között szerepel ezenfelül, a Kreml információs műveleteivel kapcsolatos állampolgári tudatosság növelése, valamint a polgárok megfelelő tudással történő felvértezése a megtévesztő digitális információkkal és a média manipulációjával szemben. A weboldal az alábbi linken érhető el: https://euvsdisinfo.eu/

[361] A javaslat az átláthatóság növelése érdekében szorgalmazta például a tájékoztatásnyújtási kötelezettség kiterjesztését a digitális médiaplatformokon megjelenő hirdetések forrásai és finanszírozási módjaival kapcsolatban. A médiatudatosság növelése érdekében a javaslat az újságírói függetlenség biztosításán, valamint a tényellenőrző szervek támogatásán túl különböző oktatási programok indítását is javasolta, melyek arra tanítanák a hírfogyasztó állampolgárokat, hogyan értékeljék kritikusan az információkat, hogyan ismerjék fel a dezinformációt, és hogyan navigáljanak sikeresen az újonnan létrejövő online környezetben A digitális platformokkal való együttműködéshez kötődően olyan intézkedések közös kidolgozását és folyamatos monitorozásával javasolták, melyek nagyobb felelősségvállalásra köteleznék online platformokat a dezinformációs tartalmak terjedésének csökkentésében. Erről bővebben: European Commission (2018): A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation. Luxembourg, Publications Office of the European Union, 22-31o.

[362] Európai Bizottság (2018): Európai megközelítés az online félretájékoztatás kezelésére. COM(2018) 236 final

[363]A megfogalmazott ajánlások az alábbi öt területbe sorolhatóak: (1) Átláthatóbb, megbízhatóbb és elszámoltathatóbb online környezet kialakítása; (2) Biztonságos és ellenállóképes választási rendszerek és folyamatok garantálása; (3) Az oktatás és a médiaműveltség előmozdítása; (4) A színvonalas és minőségi újságírás támogatása; (5) A belső és külső információs fenyegetések elleni kommunikációs stratégiák kidolgozása. Lásd bővebben: Európai Bizottság (2018): Európai megközelítés az online félretájékoztatás kezelésére. COM(2018) 236 final 7-18.o.

Az uniós erőfeszítések a kezdetektől fogva kiemelt jelentőséget szenteltek a dezinformáció visszaszorítását célzó, önszabályozó gyakorlati kódex megalkotására. Az EU és a *nagy online platformok* között létrejött önkéntes megállapodás legfőbb célkitűzése, hogy átláthatóbbá és elszámoltathatóbb tegye a legnagyobb befolyással bíró online platformokat, illetve, hogy egyfajta innovatív keretként szolgáljon ezen platformok dezinformációval kapcsolatos politikáinak nyomon követéséhez és javításához. Az önszabályozási megközelítésen alapuló kódexben[364] a 16 aláíró fél összesen 21 kötelezettséget vállalt többek között olyan intézkedésekkel kapcsolatban mint a; reklámelhelyezések ellenőrzése, politikai és a közügyekkel kapcsolatos hirdetések átláthatósága, szolgáltatások integritása, tudatos fogyasztói magatartás előmozdítása, tényellenőrzők és a kutatók szerepvállalásának erősítése, kódex eredményességének mérése.[365]

Az Európai Tanács a 2018 júniusi következtetéseiben kérte fel a Bizottságot arra, hogy „*terjesszenek elő konkrét javaslatokat a félretájékoztatás jelentette problémára adandó koordinált uniós válaszintézkedésekre vonatkozóan",* valamint határozzák meg az e célból biztosítandó szükséges mértékű és elégséges erőforrásokat.[366]

Az Európai Bizottság 2018 decemberében egy cselekvési tervet is közzétett a dezinformáció elleni küzdelem elősegítésére, melyben az Unió által megfogalmazott ajánlások az alábbi négy fő célra összpontosítottak: (1) Az uniós intézmények azon képességeinek javítása, hogy felderítsék, elemezzék és leleplezzék a dezinformációt; (2) A dezinformációval kapcsolatos koordinált és együttes válaszlépések erősítése; (3) A magánszektor bevonása és mozgósítása a dezinformáció elleni küzdelembe; (4) A dezinformációból eredő problémák tudatosítása és a társadalom ellenálló képességének javítása.[367]

A dezinformáció elleni küzdelem és a biztonságos online környezet megteremtéséhez köthető erőfeszítésekhez szorosan kapcsolódik, hogy az Európai Bizottság 2020. december 15-én

[364] Európai Bizottság (2018): A dezinformáció visszaszorítását célzó uniós gyakorlati kódex. 4-10.o. A dokumentum magyar nyelven is letölthető: https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation

[365] E kódexet 2022-ben megerősítették acélból, hogy szigorúbb intézkedésekkel küzdjön a dezinformáció ellen. 34 aláíró vállalt 44 kötelezettséget, beleértve a dezinformáció bevételi forrásainak elzárását, a politikai hirdetések átláthatóságának növelését, a manipulált tartalmak azonosításának és jelentésének egyszerűsítését, valamint a tényellenőrzés bővítését. Ezen felül létrehoztak egy átláthatósági központot, szilárd nyomonkövetési rendszert és egy állandó munkacsoportot a kódex fejlesztésére és kiigazítására, valamint a kutatók számára az adatokhoz való hozzáférés javítására. Lásd: European Commission (2022): Strengthened Code of Practice on Disinformation. 18-34.o.

[366] Európai Tanács (2018): Következtetések. EUCO 9/18 (2018. június 28.) II. Cím (Biztonság és védelem), 10. bekezdés, 5.o.

[367] European Commission (2018): Action Plan against Disinformation. JOIN(2018) 36 final, 5-11.o.

javaslatot tett a digitális tér teljes „ambiciózus” reformjára, melynek keretében egy új, átfogó, az összes digitális szolgáltatásra (közösségi média, online piacok, az Európai Unióban működő egyéb online platformok) kiterjedő szabályrendszert terjesztett elő. Thierry Breton, belső piacért felelős biztos nyilatkozata alapján ennek szükségszerűségét az adta, hogy ”*sok online platform központi szerepet játszik az EU polgárainak és vállalkozásainak életében, sőt társadalmunkban és a demokráciában is... e javaslatokkal a következő évtizedek digitális terét szervezzük meg”.* [368]

7.6 Digitális Szolgáltatásokról Szóló Rendelet

Az Európai Unió új Digitális Szolgáltatásokról szóló rendelete (DSA)[369] harmonizált jogszabályokat állapít meg az Unió állampolgárinak[370] életére közvetlen hatással bíró online közvetítő szolgáltatók számára. A rendelet célja egy biztonságos, kiszámítható és megbízható online környezetet megteremtése, amely többek között hatékonyan képes kezelni a jogellenes online tartalmak, a személyre szabott információs terek és a dezinformációs kampányok terjedéséből eredő társadalmi kockázatokat. [371]

Az Európai Bizottság – egy átfogó konzultációs és hatásvizsgálati folyamatot követően – 2020 decemberében terjesztette elő a DSA tervezetét. A tervezetet ezután az Európai Parlament és az EU Tanácsa tárgyalta meg és módosította a rendes jogalkotási eljárás keretében. Hosszas egyeztetések után 2022. október 19-én hivatalosan is elfogadták és kihirdették a végleges szöveget mely ezáltal formálisan is az EU jogrendjének részévé vált. A szabályozás nagy része, beleértve az összes online platformra vonatkozó kötelezettségeket, 2024. február 17-től vált kötelezően alkalmazandóvá, 15 hónapos felkészülési időt biztosítva a vállalkozásoknak, hogy hozzáigazítsák működésüket az új szabályokhoz. A nagyon nagy

[368] Európai Bizottság (2020): Sajtóközlemény - A digitális korra felkészült Európa: A Bizottság új szabályokat javasol a digitális platformokra. 1-3.o.
[369] Európai Bizottság: Az Európai Parlament és a Tanács rendelete a digitális szolgáltatások egységes piacáról (digitális szolgáltatásokról szóló jogszabály) és a 2000/ 31/EK irányelv módosításáról. COM(2020)825 final
[370] A DSA követelményeinek alkalmazását az igénybe vevők lakó- vagy tartózkodási helyéhez köti, nem pedig a közvetítő szolgáltató letelepedési helyéhez vagy székhelyéhez. A 2. cikk (1) bekezdés kimondja, hogy „*a rendelet az olyan igénybe vevők részére kínált közvetítő szolgáltatásokra vonatkozik, amelyeknek, illetve akiknek a letelepedési helye az Unióban található vagy akik, illetve amelyek az Unióban találhatók, függetlenül az említett közvetítő szolgáltatást biztosító szolgáltatók letelepedési helyétől”.*
[371] DSA 9. preambulum

online platformokra (Very Large Online Platforms, VLOPs) és a nagyon nagy online keresőprogramokra (Very Large Online Searc Engine, VLOSEs) vonatkozó szigorúbb kötelezettségek már 2023. augusztus 25-től érvénybe léptek.

A DSA egy többszintű és differenciált szabályozási modellt vezetett be, amely annak alapján különbözteti meg az egyes szolgáltatókat, hogy azok mekkora felhasználói bázissal rendelkeznek. Ez a modell tükrözi azt a felismerést, hogy a nagyobb platformok jelentősebb hatást gyakorolnak az alapvető jogok érvényesülésére, a politikai diskurzusra és a demokratikus nyilvánosságra. Ebből következik, hogy nagyobb a felelősségük is a társadalmi hatások kezelése tekintetében. Ennek értelében a rendelet szigorúbb jogi követelményeket állapít meg az ún. online óriásplatformok és nagyon népszerű online keresőprogramok számára. Azon szolgáltató tartoznak e kettő kategóriába, amelyek havonta átlagosan legalább 45 millió aktív igénybe vevővel rendelkeznek az EU-ban (jelentősebb és gyorsan bekövetkező népességváltozás esetén az Unió összlakosságának 10%-a lesz a mértékadó).[372]

Ezekre a platformokra vonatkozó további kötelezettségek közé tartozik például az általuk nyújtott szolgáltatások használata során felmerülő online veszélyekkel kapcsolatos kockázatok átfogó éves értékelése. Az ilyen platformoknak biztosítaniuk kell, hogy rendszeresen felülvizsgálják és mérsékeljék a felhasználókat érintő potenciális kockázatokat (ideértve a felhasználói jogokra, az adatvédelemre és a szólásszabadságra gyakorolt hatásokat). A szolgáltatóknak a rendszerszintű kockázatok négy meghatározott kategóriáját kell részletesen értékelniük: (1) *jogellenes tartalmak terjesztésével (pl.: gyűlöletbeszéd) vagy a jogellenes tevékenységek folytatásával kapcsolatos kockázatok*; (2) *a felhasználók alapvető jogainak gyakorlását érintő tényleges vagy várható negatív hatáso*k (pl.: véleménynyilvánítás és a tájékozódás szabadságát, tömegtájékoztatás szabadságát vagy a magánélethez való jogot érintő korlátozások); (3) *a demokratikus folyamatokra, a polgári közbeszédre, a választási folyamatokra és a közbiztonságra gyakorolt tényleges vagy előrelátható negatív hatások*; (4) *a nemi alapú erőszakkal, a közegészség és a kiskorúak védelmével összefüggő tényleges vagy előre látható negatív hatások, valamint a személyek testi és szellemi jóllétére gyakorolt súlyos*

[372] A Bizottság 2023. április 25-én azonosította azon szolgáltatókat, amelyek e két kategória valamelyikébe tartoznak. Ennek értelmében óriásplatformnak (Very Large Online Platform, VLOP) minősül az: AliExpress / Amazon Store / AppStore / Booking / Facebook / Google Maps / Google Play / Google Shopping / Instagram / LinkedIn / Pinterest / Snapchat / TikTok / Twitter / Wikipedia / YouTube / Zalando. Nagyon népszerű keresőmotor szolgáltatóként (Very Large Searching Engine), VLSE) pedig a Bing és Google Search szolgáltatókat nevezték meg.

negatív következmények (pl.: manipuláción keresztül vagy olyan online felületek kialakítása útján, melyek függőségéhez vezethetnek) .[373]

A továbbiakban – dolgozat témájához kötődően – a DSA jogszabály következő rendelkezéseit indokolt ismertetni:

7.6.1 Átláthatósági és tájékoztatási kötelezettség

Az alapvető jogok védelme tekintetében kiemelkedő jelentőségű a rendelet 14. cikke, amely amellett, hogy szigorú transzparenciakövetelményeket állít, széles körű tájékoztatási kötelezettséget is előír a közvetítő szolgáltatok számára. Ennek alapján *világosan, egyszerűen, érthetően, felhasználóbarát módon* tájékoztatást kell nyújtani a szolgáltatást igénybe vevők számára, többek között a tartalommoderálás[374] céljából alkalmazott valamennyi szabályra, eljárásra, intézkedésre és eszközre (például, hogy milyen módon szűrik vagy blokkolják a tartalmakat), ideértve az algoritmikus döntéshozatalt és az emberi felülvizsgálatot is.[375]

7.6.2 A személyre szabott célzott hirdetés és manipuláció tilalma

A DSA szintén fontos eleme, hogy az online platformot üzemeltető szolgáltatók nem tervezhetik meg, alakíthatják ki vagy üzemeltethetik online interfészeiket oly módon, *amely megtéveszti vagy manipulálja a szolgáltatásaikat igénybe vevőket, vagy gyengíti azok szabad és tájékozott döntéshozatalra való képességét*. [376]

---

[373] DSA 33-35. cikk, 80-84. preambulum.

[374] DSA 3.cikk t) értelmében a tartalommoderálás a közvetítő szolgáltató olyan *„automatizált vagy nem automatizált tevékenysége, amely különösen a szolgáltatás igénybe vevője által közzétett jogellenes tartalom vagy a közvetítő szolgáltató szerződési feltételeivel összeegyeztethetetlen információ észlelésére, azonosítására és kezelésére szolgál, ideértve az ilyen jogellenes tartalom vagy információ elérhetőségét, láthatóságát és hozzáférhetőségét érintő intézkedéseket, például annak hátrasorolását, a demonetizálását, az ahhoz való hozzáférés megszüntetését vagy annak eltávolítását, vagy a szolgáltatás igénybe vevője általi információközlés lehetőségét érintő intézkedéseket, például a fiókja megszüntetését vagy felfüggesztését*"

[375] DSA 14. cikk

[376] DSA 25. cikk (1) bekezdés

A 26. cikk (3) bekezdése kimondja, hogy az online platformokon nem jeleníthetnek meg profilalkotáson alapuló hirdetéseket a szolgáltatások igénybe vevői számára a személyes adatoknak a GDPR-ban említett különleges kategóriáinak felhasználásával. Ide tartoznak a *faji* vagy *etnikai származásra, politikai véleményre, vallási* vagy *világnézeti meggyőződésre* vagy *szakszervezeti tagságra* utaló személyes adatok, valamint a természetes személyek egyedi azonosítását célzó *genetikai* és *biometrikus adatok*, az *egészségügyi adatok* és a természetes személyek *szexuális életére* vagy *szexuális irányultságára* vonatkozó személyes adatok.[377]

A DSA 26. cikke értelmében az online hirdetéseket megjelenítő platformok üzemeltetőinek *világosan, tömören, egyértelműen és valós időben fel kell tüntetniük, amennyiben az általuk közölt információ hirdetésnek minősül*. Emellett fel kell tüntetniük azt is, hogy mely természetes vagy jogi személy nevében került sor a hirdetés megjelenítésére (26. cikk (1) b); a hirdetést mely természetes vagy jogi személy finanszírozta, amennyiben e személy eltér a b) pontban említett természetes vagy jogi személytől (26. cikk (1) c); valamint azt is, hogy mely paraméterek szolgáltak a hirdetéssel megcélzott személyek meghatározására (26. cikk (1) d).[378]

Az online környezetben az egyének számára a legtöbb esetben nincs lehetőség megérteni, hogy hogyan és miért jut el hozzájuk bizonyos tartalom. Ezt az információs aszimmetriát orvosolja a DSA azzal, hogy kötelezi az összes online platformot, hogy közzé tegye tartalomajánló rendszereik[379] paramétereit, ezáltal világos magyarázatot szolgáltatva arra, hogy miért is ajánlják az adott információkat a szolgáltatás igénybe vevője számára.[380]

Ezen túlmenően a 38. cikk előírja, hogy az ajánlórendszereket használó online óriásplatformot vagy nagyon népszerű online keresőprogramot üzemeltető szolgáltatók minden ajánlórendszerük esetében legalább egy olyan lehetőséget biztosítanak, amely nem a GDPR

[377] GDPR 9.cikk (1) bekezdés

[378] A 39. cikk továbbiakban előírja azt is, „hogy az online hirdetéseket megjelenítő óriásplatformot és nagyon népszerű online keresőprogramot üzemeltető szolgáltatók nyilvánosan hozzáférhetővé teszik a hirdetésekkel kapcsolatos további információkat is. 39. cikk (2) a) a hirdetés tartalma, beleértve a termék, a szolgáltatás vagy a márka nevét és a hirdetés tárgyát; b) kinek nevében került sor a hirdetés megjelenítésére; c,) mely természetes vagy jogi személy fizette a hirdetést, amennyiben e személy eltér a b) pontban említett személytől; d) mely időszakban került sor a hirdetés megjelenítésére"

[379] DSA 3. cikk s) az ajánlórendszer „*teljes mértékben vagy részben automatizált rendszer*, amelyet az online platform arra használ, hogy az online interfészén konkrét információkat javasoljon vagy ezeket az információkat rangsorolja a szolgáltatás igénybe vevője számára, többek között a szolgáltatás igénybe vevője által indított keresés alapján *vagy egyéb módon meghatározva a megjelenített információk relatív sorrendjét vagy elsőbbségét*."

[380] DSA 27.cikk (1) és (2) bekezdés

szerint meghatározott profilalkotási[381] eljáráson alapul. A DSA nem teszi kötelezővé a nem profilozáson alapuló ajánlást, de ösztönözheti a platformokat, hogy kínáljanak ilyen lehetőséget.

A rendelet olyan új mechanizmusokat is bevezet, amelyek segítenek a felhasználóknak bejelenteni a jogellenes tartalmakat[382] annak érdekében, hogy a platformok hatékonyabban legyenek képesek azonosítani és eltávolítani azokat. A DSA hangsúlyozza, hogy a platformoknak proaktív intézkedéseket kell tenniük az illegális tartalmak ellen, de nem kötelezi őket általános monitorozásra.[383] Ehelyett a platformoknak együtt kell működniük az un. „*trusted flaggers*" (megbízható bejelentők) rendszerével, hogy hatékonyan kezeljék az illegális tartalmakat.

7.6.3 DSA-val kapcsolatos észrevételek

A tartalomajánló rendszerekkel kapcsolatban feltehetően várható még vita az Európai Unióban. Ezt támasztja alá, hogy 2023. december 20-án 17 európai parlamenti képviselő levelet írt a Bizottságnak, melyben szorgalmazták, hogy a technológiai platformok perszonalizált tartalomajánlórendszereit alapértelmezésben kapcsolják ki. Levelükben úgy fogalmaznak: „*Az interakción alapuló ajánlórendszerek, különösen a kifinomult személyre szabott rendszerek komoly veszélyt jelentenek polgárainkra és társadalmunkra általában, mivel az érzelmekre ható és szélsőséges tartalmakat helyezik előtérbe, és kifejezetten a*

[381] A GDPR 4. cikk 4 pontja úgy határozza meg a profilalkotást, mint a: „*személyes adatok automatizált kezelésének bármely olyan formája, amelynek során a személyes adatokat valamely természetes személyhez fűződő bizonyos személyes jellemzők értékelésére, különösen a munkahelyi teljesítményhez, gazdasági helyzethez, egészségi állapothoz, személyes preferenciákhoz, érdeklődéshez, megbízhatósághoz, viselkedéshez, tartózkodási helyhez vagy mozgáshoz kapcsolódó jellemzők elemzésére vagy előrejelzésére használják"*

[382] Bármely olyan információ, amely *jogellenes a vonatkozó nemzeti vagy európai jogszabályok* szerint. Például olyan információk közzététele, amely mások személyes adatainak jogosulatlan felhasználását jelenti; Minden olyan anyag, amely erőszakos szélsőséges nézeteket népszerűsít, vagy terrorcselekmények elkövetésére ösztönöz; Minden olyan tartalom, amely faj, vallás, etnikai hovatartozás, nem, szexuális irányultság stb. alapján diszkriminál, erőszakra uszít vagy gyűlöletet kelt (gyűlöletbeszéd).

[383] Az általános monitorozás azt jelentené, hogy a platformok folyamatosan figyelnének minden felhasználói tartalmat, hogy illegális vagy káros tartalmakat találjanak. Kritikaként megfogalmazható, hogy ez a követelmény túlzott terhet jelentene a platformok számára, és aggályokat vetne fel a felhasználói magánélet és szólásszabadság védelme szempontjából. Megjegyzendő, ugyanakkor, hogy a digitális platformok már a törvény előtt is széles körben alkalmaztak tartalomfelismerő algoritmusokat annak érdekében, hogy automatikusan felismerjék és blokkolják a jogsértőnek tartott felhasználói hozzászólásokat (ezek többnyire a kulcsszavak és más jellemzők alapján szűrik a tartalmat).

provokációra hajlamos személyeket célozzák meg”. [384] Annak ellenére, hogy ez az ötlet már a DSA tárgyalások során is felmerült, végül nem került be a végleges rendeletbe Ehelyett a jogalkotók arra az előbb ismertetett kompromisszumos megoldásra jutottak, hogy a nagyobb platformoknak minden ajánlórendszerük esetében legalább egy olyan tartalomszolgáltatást kell nyújtaniuk, amely nem profilalkotáson alapul (opt out megközelítés).

A rendelet hiányosságaként felróható Hacker és társai azon kritikája, miszerint a szabályozás érdemi része nem terjed ki a „*privát üzenetküldő szolgáltatásokra*”.[385] Ennek következtében az olyan népszerű közösségi platformok, mint a WhatsApp és Telegram zárt csoportjai, ahol a problémás tartalmak és a dezinformáció különösen elterjedt, nem esnek a DSA szabályozási hatálya alá.[386]

Előfordulhat az is, hogy a jogszabályi kötelezettségek betartásának nehézségeit nem az érintett szolgáltatók jogkövetési hajlandóságának vagy a részletes, következetes szabályozás meglétének hiánya, hanem sokkal inkább a rendelkezésre álló technológiai képesség hiánya okozza. A Nagy Nyelvi Modellek esetében például sok esetben hiába építenek be olyan visszaélés-megelőző funkciókat és biztonsági korlátokat, melyek célja, hogy visszautasítsák vagy megtagadják a nem kívánt utasításokat (például gyűlöletkeltő, erőszakra felbujtó esszék és pamfletek előállítása), a felhasználók mégis megtalálják a módot ezeknek a biztonsági intézkedéseknek a megkerülésére.[387] Egy ilyen megkerülési technika például az un, „prompt injection" támadás is, ahol úgy álcázzák a bemeneti szöveget, mint a felhasználó vagy a fejlesztő utasításait, annak érdekében felülírják a modellekbe épített vagy tanított korlátozásokat.[388]

---

[384] “Interaction-based recommender systems, in particular hyper-personalised systems, pose a severe threat to our citizens and our society at large as they prioritize emotive and extreme content, specifically targeting individuals likely to be provoked,” Lásd bővebben: Natasha Lomas (2023): Big Tech’s divisive ‘personalization’ attracts fresh call for profiling-based content feeds to be off by default in EU. TechCrunch. Elérhető: https://techcrunch.com/2023/12/20/dsa-recommender-systems/ (2023. 12. 20.)

[385] DSA 14. preambulumbekezdés

[386] A DSA mechanizmusainak teljes skálája csak a hagyományos közösségi hálózatokon közzétett GenMI tartalomra alkalmazható. Lásd: Hacker et al. (2023): Regulating ChatGPT and other Large Generative AI Models. *Fairness, Accountability, and Transparency* (FAccT ’23), 1112-1123.o. https://doi.org/10.1145/3593013.3594067

[387] A *tartalomszűrők* (content filter) feladata, hogy észleljék és eltávolítsák vagy blokkolják a meghatározott kategóriák szerinti nemkívánatos tartalmakat. A *biztonsági szűrők* (safety filter) célja, hogy megakadályozzák az ellenséges promptokat és biztosítsák, hogy a modell kimenete biztonságos maradjon különböző támadási forgatókönyvek alatt. Ezek a szűrők összetettebb manipulációkat és ellenséges támadásokat kezelnek, amelyek megpróbálják kijátszani a modell védelmi mechanizmusait

[388] Lásd: Fábio, Perez, Ian Ribeiro (2022): Ignore previous prompt: Attack techniques for language models. 3-6.o. arXiv.2211.09527; Az ilyen gyakorlatokhoz köthetőek az un. jailbreaking támadások is. A kettő közötti különbség abban rejlik, hogy míg a prompt injection egy adott utasítással törekszik tiltott, vagy káros kimenet

8 Web 0.0 – a digitális szimulákrum kora

Az internet – abban a formában, ahogy azt eredetileg létrehozták – már nem létezik. Kezdetben egy nyílt hálózatként jött létre, amely lehetővé tette a felhasználók számára, hogy szabadon hozzáférjenek a megosztott ismeretekhez és közvetlenül kapcsolatba léphessenek egymással. Az utóbbi években lezajlott változások visszalépést jelentettek az eredeti célokhoz és eszményekhez képest; az egykori nyílt és szabad hozzáférés helyét a korlátozások és a zárt rendszerek vették át, fokozatosan ritkul a valódi emberek közötti közvetlen kapcsolat, helyüket botok és szintetikus tartalmak veszik át. Ezt a visszafejlődést hivatott jelezni a címben feltüntetett név, az internet új korszaka: *Web 0.0.*[389]

E korszak legfőbb jellemzője, hogy az emberi jelenlét és az *autentikus tartalmak* fokozatosan háttérbe szorulnak az online térben. A felhasználók – kifejezett tudomásuk nélkül – egyre gyakrabban találkoznak MI által generált szövegekkel, képekkel, videókkal, miközben a valódi emberi interakciók aránya drasztikusan csökken. Jól dokumentált tény, hogy az internetes forgalom jelentős része már ma is botokhoz köthető; 2023-ban az emberi felhasználóktól származó online aktivitás rekord mélységekbe (50,4%-ra) zuhant, mellyel párhuzamosan az internetes forgalom csaknem fele (49,6%) már kizárólag botok által volt generálva. Ez a legmagasabb arány azóta, hogy 2013-ban az Imperva – egy vezető amerikai kibervédelmi vállalat – elkezdte nyomon követni és mérni a webes forgalmat. [390]

Bár a jelenség nem új, az idő előrehaladtával fokozatos romlás volt megfigyelhető e téren. A Youtube videómegosztó felülete a 2010-es évek közepén egy rövid ideig olyan jelentős botforgalommal szembesült, hogy egyes alkalmazottak attól tartottak, az algoritmusaik

---

generálására, addig a jailbreaking végrehajtójának a modell biztonsági mechanizmusainak teljes felszámolása a célja. Bővebben: Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu (2023): Jailbreaking ChatGPT via prompt engineering: An empirical study. 4-9.o. arXiv:2305.13860v2; Anthropic (2024): Many-shot jailbreaking. Elérhető: https://www-cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking__2024_04_02_0936.pdf (2024. 06. 24.)

[389] Legjobb tudomásom szerint nem használták még e kifejezést hasonló kontextusban.

[390] Külön aggasztó, hogy a rosszindulatú botok aránya öt egymást követő évben nőtt, 2023-ban elérve a 32%-ot. A legmodernebb botok már nem csak egyszerű automatizált rendszerekként működnek, képesek emberi viselkedést szimulálni, például egérmozgásokat vagy kattintásokat utánozni, így még nehezebben észlelhetők. A 2024-es tanulmány megállapításairól bővebben: Imperva (2024): Bad Bot Report – 2024. Elérhető: https://community.imperva.com/blogs/percy-smith/2024/05/16/impervas-11th-annual-bad-bot-report-2024 (2024. 05. 16.)

rövidesen a botok által indított interakciókat tekinthetik majd hitelesnek az emberi felhasználók aktivitása helyett.[391] 2018-ban pedig a New York Times nagy visszhangot kiváltó oknyomozó cikke fedte fel, hogy egy amerikai cég (Devumi) mesterséges Twitter-követőket és interakciókat árusított többszázezer felhasználó számára. A cég legalább 3,5 millió automatizált bot fiókot használt, megközelítőleg 200 millió mesterséges követőt és interakciót biztosítva ezáltal ügyfeleinek. [392]

A GenMI megjelenésével e trend új szintre lépett: a mesterségesen előállított tartalmak mennyisége és minősége is jelentősen megnövekedett. Aki figyelmesen szemléli a közösségi médiát láthatja, hogy a felületeit elárasztották a szintetikus képi tartalmak. Az ezek terjesztése és megosztása mögött meghúzódó szándék legtöbb esetben – egyelőre – még csak nem is a manipuláció, hanem a kattintások és interakciók monetizálása.[393]

A szintetikus tartalmak terjedése azonban nem korlátozódik a közösségi média platformjaira, változó tendenciák kezdenek kirajzolódni a hírportáloknál is. Különösen szembetűnő ez a Google News felületén, ahol egyre gyakrabban bukkannak fel GenMI által létrehozott szintetikus cikkek.[394] Ez a trend nem csak a hírszolgáltatás minőségének és megbízhatóságának romlását vetíti előre, de azt is jelzi, hogy a Google szűrőrendszere egyre kevésbé képes – vagy hajlandó – különbséget tenni emberi és gépi alkotások között. Ennek az egyik oka az lehet, hogy a GenMI által generált hírek viszonylag rövid idő alatt eláraszthatnak

[391] A Youtube platformján azonosított botforgalom annyira magas volt, hogy az összes megtekintés közel felét robotok generálták. Az aggodalom tárgyául az szolgált, hogy a YouTube algoritmusai és rendszerei, amelyek a hamis forgalom kiszűrésére készültek, esetleg elkezdhetik a botok által generált forgalmat valódinak tekinteni, míg az emberi tevékenységet hamisnak. Ezt a – legjobb tudásunk szerint – be nem következett elméleti forgatókönyvet nevezték *Inverziónak* (the Inversion). Lásd: Keller, H. Michael (2018): The Flourishing Business of Fake YouTube Views. *The New York Times*. Elérhető: https://www.nytimes.com/interactive/2018/08/11/technology/youtube-fake-view-sellers.html (2023. 07. 21.)

[392] A cikk arra is rámutatott, hogy a Devumi szolgáltatásait politikai célokra is használták., például Kína állami hírügynöksége és Ecuador elnöke is vásárolt követőket. Egyes szakértők szerint a Twitternek üzleti érdeke fűződhetett ahhoz, hogy ne lépjen fel túl agresszíven a botok ellen, mivel ez befolyásolhatja a felhasználói növekedési mutatóit. A Twitter végül felfüggesztette a Devumi fiókját, miután a botokkal kapcsolatos problémák napvilágra kerültek. Lásd: Confessore, Nicholas, Gabriel J.X. Dance, Richard Harris, Mark Hansen (2018): The Follower Factory: Everyone Wants to Be Popular Online. Some Even Pay for It. Inside Social Media's Black Market. *The New York Times*. Elérhető: https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html (2022. 03. 19.)

[393] A – főként vizuális – szintetikus tartalmakat nagy számban posztolók (spammerek) jellemzően clickbait taktikákat alkalmaznak, hogy a felhasználókat más platformokra vagy alacsony minőségű, hirdetésekkel teli oldalakra irányítsák, így közvetlen pénzügyi nyereséget generálva a látogatásokból és az interakciókból. Lásd bőveben: DiResta, R. - Goldstein, J.A. (2024): How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. 2-17.o. ArXiv, abs/2403.12838.

[394] Cox, Joseph (2024): Google News Is Boosting Garbage AI-Generated Articles. *404media*. Elérhető: https://www.404media.co/google-news-is-boosting-garbage-ai-generated-articles/ (2024. 01. 18.)

bármilyen online felületet, mivel ezek a tartalmak rendkívül gyorsan és nagy mennyiségben készülnek.[395]

A Web 0.0 korszakának egyik legnyugtalanítóbb következménye az, hogy az emberek bizalma az online, majd fokozatosan az offline információforrásokban is egyre inkább erodálódik. Hamar eljöhet az idő, és az emberek nem tudják többé megkülönböztetni a valós tényeket a szintetikus, mesterségesen generált információktól. Amikor már nem lehet tudni, hogy egy online hír, egy Wikipedia bejegyzés vagy egy YouTube videó valós eseményeket tükröz-e, a valóság is elveszíti jelentőségét. Amikor minden tartalom – legyen az kép, szöveg vagy hír – olyan mértékben manipulálható és mesterségesen előállítható, hogy a valósággal egyenértékűnek tűnik. (Képzeljünk el, egy reaktív Wikipedia oldalt[396], ami a felhasználó kattintásának pillanatában hoz létre tökéletesen szerkesztett cikkeket nem létező eseményekről, vagy olyan videómegosztó oldalt, ahol a keresőkifejezés leütése pillanatában jön létre egy szintetikus dokumentumfilm *a hírhedt* 1526-os mohácsi magyar győzelemről.)

A kialakuló bizalomvesztés tovább mélyíti a már meglévő – dezinformációs kampányokhoz, konspirációs elméletekhez, alternatív tényekhez kötődő – krízistüneteket. Ebben a kontextusban a Web 0.0 nem csupán technológiai visszalépés, hanem egy társadalmi és filozófiai válság előhírnöke is egyben.

8.1 A valóság talaján. Stratégiák és javaslatok az episztemológiai összeomlás megelőzésére

A Web 0.0 korszaka egyre összetettebb kihívások elé állítja a társadalmakat az információszerzés és a kollektív valóságértelmezés terén. Bár léteznek különböző stratégiák e problémák kezelésére, hatékonyságuk gyakran korlátozott. A megoldás minden bizonnyal több megközelítés együttes alkalmazását igényli, ideértve: (1) a digitális kompetenciák fejlesztését és az oktatási rendszerek modernizálását; (2) hatékony technológiai megoldások fejlesztését, valamint; (3) előremutató jogi keretrendszerek kidolgozását.

[395] Csak szemléltetésképpen a NewsGuard-Tech kutatása alapján a mesterséges hírtartalmakat közlő World-Today-News.com naponta átlagosan 1200 cikket tesz elérhetővé, míg ezzel egyidőben a The New York Times naponta körülbelül 150 eredeti cikket publikál Lásd: Brewster, Jack, Zack Fishman, Elisa Xu (2023): Misinformation Monitor: June 2023. Funding the Next Generation of Content Farms. Elérhető: https://www.newsguardtech.com/misinformation-monitor/june-2023/ (2024. 01. 20.)

[396] Olyan online megjelenített tartalom, ami valós időben képes reagálni a felhasználói interakciókra (képes valós időben létrehozni és módosítani a felületet, alkalmazkodva a felhasználók egyedi preferenciáihoz vagy az adott kontextushoz).

8.1.1 A digitális kompetenciák fejlesztése és az oktatási rendszerek modernizálása

A dezinformációs kampányok és a hibrid hadviselés visszaszorításáért tett erőfeszítések számos olyan értékes tapasztalattal és tanulsággal szolgáltak az elmúlt években, melyek beépíthetők lehetnek a digitális kompetenciák fejlesztését célzó oktatási programokba. Jó példaként szolgálhat erre az Európai Unió átfogó dezinformációs stratégiája, mely jelentős mértékben támaszkodik a tényellenőrző platformok támogatására[397] és a digitális írástudás előmozdítására. Az Uniós stratégia hangsúlyozza, hogy a digitális jártasság elsajátítása elengedhetetlen annak érdekében, hogy az uniós polgárok maximálisan kihasználhassák az online világ által nyújtott előnyöket és lehetőségeket, miközben ellenállók maradnak annak ártalmas hatásaival szemben. Ez magába foglalja többek között, hogy az interneten tájékozódó egyén: (1) *kritikusan tudja értékelni* az információforrás és a digitális tartalom hitelességét és megbízhatóságát; (2) tudatában van annak, hogy az internetes keresőmotorok, a közösségi média és a tartalomplatformok gyakran alkalmaznak MI-algoritmusokat annak érdekében, hogy olyan reakciókat generáljanak, illetve tartalmakat jelenítsenek meg, amelyeket az adott *felhasználó preferenciáihoz igazítanak, vagy azok alapján határoznak meg*.[398]

A modern információs korban már magától értetődőnek tűnik, hogy a tradicionális ismeretátadás mellett egyre nagyobb hangsúlyt kell fektetni az információfeldolgozási készségek fejlesztésére is. A felhasználóknak meg kell tanulniuk, hogyan értékeljék a források megbízhatóságát, hogyan különböztessék meg a tényeket a véleményektől, és hogyan ismerjék fel az érvelési hibákat és a manipulációt. A megoldás (egyik) kulcsa tehát a *digitális*

[397] Ez a törekvés tetten érhető abban is, hogy 2022-ben megalapították az *Európai Tényellenőrző Hálózatot* (European Fact-Checking Standards Network, EFCSN), amely a különböző európai országok tényellenőrző (fact-checking) szervezeteit hivatott egyesíteni. A hálózat célja a magasabb szintű tényellenőrzési szabványok előmozdítása és terjesztése, illetve, hogy tagjai munkáján keresztül biztosítsa a pontos és megbízható információkhoz való hozzáférést. Jelenleg több mint 45 szervezet tartozik hozzá Európa különböző országaiból, többek között a brit Full Fact, a lengyel Demagog, és a magyar Lakmusz. A hálózathoz olyan szervezetek csatlakozhatnak, akik vállalják, hogy munkájuk során betartják az EFCSN által megfogalmazott szakmai, etikai és módszertani szabályokat. Ez a szabályzat olyan kötelezettségeket ír elő, mint például: (1) tényeken alapuló és pontos tájékoztatás; (2) a tényellenőrzés módszertanának és a szervezet működésének átláthatósága; (3) a hibajavítások transzparens kezelése. Lásd: Zöldi Blanka (2023): A Lakmusz megkapta az Európai Tényellenőrző Hálózat tanúsítványát. *LAKMUSZ*. Elérhető: https://www.lakmusz.hu/a-lakmusz-megkapta-az-europai-tenyellenorzo-halozat-tanusitvanyat (2023. 12. 18.)

[398] Európai Bizottság (2022): Iránymutatások tanárok és oktatók számára a dezinformáció kezeléséről és a digitális jártasság oktatás és képzés révén történő előmozdításáról. Luxembourg, Az Európai Unió Kiadóhivatala, 20-23.o.

kompetenciák / írástudás (digital literacy / competence) és a *médiaműveltség* (media literacy) fejlesztése; ezek révén a felhasználó a megtanulja, hogy különböző helyzetekben milyen típusú információkra van szüksége, hogyan juthat hozzá azokhoz, miként értékelje kritikusan a megtalálat információt (nem utolsó sorban képes kritikusan kezelni a médiafogyasztás során megszerzett élményeket, benyomásokat, mindig szem előtt tartva a tartalmak konstruált jellegét).[399]

Ezen felül szükséges annak tudatosítása is, hogy a 21. században az egyén információfeldolgozási képességét két jelentős tényező befolyásolja és torzítja, melyek együttes hatása különösen sebezhetővé teszik a hamis, megtévesztő, szintetikus tartalmakkal szemben. Elengedhetetlen, hogy a felhasználók megértsék, hogy mind az (1) online tér belső működési logikája, mind pedig (2) az emberi psziché sajátosságai befolyásolják az információfeldolgozásuk és valóságalkotásuk folyamatait. Csak e kettős tudás birtokában válhatnak képessé arra, hogy tudatosan és kritikusan kezeljék az őket érő információáradatot.

A digitális írástudás és kompetenciák oktatásának tehát ki kell terjednie egyrészt (1) a tartalomajánló algoritmusok, a pszichografikus profilozás és a microcélzás működésének alapszintű megértésére, a figyelemgazdaság és a megfigyelőkapitalizmus működési mechanizmusának átlátására, a vélemény-és szűrőbuborékok, valamint a visszhangkamrák kialakulásának tudatosítására is. Másrészt magában kellene foglalnia (2) az olyan kognitív sajátosságok, sebezhetőségek ismertetését is, mint az információs túlterheltség, a csökkenő figyelemküszöb és Dunning-Kuger hatás, a dopaminéhség- és függőség, a hamis konszenzus hatás, a csoportpolarizáció, a kognitív disszonancia és a megerősítési torzítás. Ezek bemutatására a 6. fejezet tett kísérletet.

További észrevételként megemlítendő az is, hogy bár fontos feladat a kritikai gondolkodás és a digitális írástudás folyamatos fejlesztése, a Web 0.0 korszakában talán már nem elegendő csak állhírek és dezinformáció elterjedésével, az információfeldolgozás torzulásával, vagy az online platformok működési modelljével tisztában lenni. Ennél is tovább kell bővíteni az oktatási keretrendszert: elengedhetetlen a GenMI rendszerek technológiai alapjainak, ezek korlátainak és lehetőségeinek alapos megértése is.[400] A Nagy Nyelvi Modellek működését

[399] Koltay Tibor (2010): Az új média és az írástudás formái. *Magyar Pedagógia*, 110.évf. (4), 301-309.o.
[400] Alapvető, hogy bármely jövőbeli reform esetén: (1) a digitális írástudás oktatásának kötelező tantárggyá kell válnia az iskolákban (már az alapfokú oktatásban); (2) a tananyagot interaktív tanulási módszerek alkalmazásával, gyakorlati projektalapú megközelítéssel kell leadni, annak érdekében, hogy a célközönség megismerje (kiismerje) a technológiák működését; (3) Emellett sürgős feladat lenne az idősek számára

illetően fontos lenne megérteni, hogyan működik az autoregresszív nyelvi modellezés (és miért hívják azt olykor sztochasztikus papagájnak), illetve azt is, hogy mik azok a gépi hallucinációk, amelyekkel ezek a modellek küzdenek. A technológia iránt nyitottabbak – arra fogékonyabbak – számára hasznos lehet tisztában lenni olyan fogalmakkal is, mint a modell összeomlás, a pozíciós elfogultság, vagy a *lost in the middle* (közép elvesztésének) problémája. Közvetíteni kellene a társadalom felé az ilyen rendszerek fejlődő képességeivel kapcsolatos legújabb tudományos ismereteket is (az 5. fejezetben hivatkozott tanulmányok következtetéseit az LLM-ek kreativitásával, logikai képességeivel és meggyőzőerejével kapcsolatban).[401] Tudatosítani kell az emberekben, hogy minden esetben fennáll annak az esélye, hogy amit online látnak, az mesterségesen létrehozott tartalom, legyen szó közösségi média profilokról, megosztásokról, kommentekről, képi illusztrációkról, vagy akár dokumentumfilmekről.

A bizalomvesztés megakadályozása érdekében szükséges minden segítséget megadni az embereknek a szintetikus tartalmak felismeréséhez. Ennek az alapja, hogy minden azonosított szintetikus tartalmat jelölni kell. Azoknál a tartalmaknál, amelyek jelenleg nehezebben azonosíthatók, a technológia és a jog eszköztárával kell biztosítani e nehézségek orvoslását.

8.1.2 Hatékony technológiai megoldások fejlesztése

A szintetikus tartalmak azonosítása összetett kihívást jelent, amely az etikai, társadalompolitikai és jogalkotási dimenziókon túl, jelentős technológiai akadályokat is felvet. Jelenleg a GenMI által létrehozott tartalmak felismerésére és megjelölésére leggyakrabban

kialakított speciális tanfolyamok indítása is. A legtöbb felmérés alapján ők azok, akik a legnehezebben boldogulnak a digitális világban, és ők a legkiszolgáltatottabbak az online átverések és dezinformációk terén.

[401] Ahogy azt sem ildomos elfelejteni, hogy a GenMI velünk marad a következő évtizedben is. E rendszerek fejlődésének kezdeti szakaszában vagyunk. Ahogy a számítási kapacitások folyamatosan növekednek, a jövőbeli képességeik várhatóan felülmúlják a jelenlegi rendszerekét. Gyorsabb és pontosabb válaszokat adhatnak majd. Komplexebb feladatok megoldására lesznek alkalmasak. Kreatívabb és eredetibb tartalmakat hozhatnak létre. Jobban megérthetik majd az emberi nyelv árnyalatait. Fejlettebb problémamegoldó képességekkel rendelkezhetnek. Szélesebb körű ismeretekkel bírhatnak különböző területeken. Jobban alkalmazkodhatnak a felhasználók egyedi igényeihez.

alkalmazott módszerek a (1) Metaadatokok (metadata); (2) Vízjelek (watermark); (3) Ujjlenyomat-technika (fingerprint) és a; (4) Gépi észlelés, detektálás (detection).[402]

(1) Az első esetben információkat (metaadatokat) ágyaznak be egy fájlba, amelyek segítenek az adott tartalom eredetiségének, hitelességének vagy szerzői jogi státuszának ellenőrzésében. Metaadatban fellelhető információ lehet például a tartalom létrehozójának neve, a készítés időpontja és helye, a felhasznált eszközök és szoftverek, valamint az esetleges módosítások részletei. Mivel ezek az adatok viszonylag könnyen módosíthatók vagy eltávolíthatók, nem minden esetben biztosítanak megbízható azonosítást; (2) A vízjelek lényegében olyan digitális információkat tartalmaznak, amiket különféle digitális tartalmakba ágyaznak annak érdekében, hogy biztosítsák a tartalom hitelességét és nyomon követhetőségét. Általában láthatatlanok vagy alig észrevehetők[403] az emberi szem számára, de megfelelő technikai eszközökkel azonosíthatók. Mivel eltávolításuk a tartalom jelentős minőségromlásával járhat, megbízható eszközként szolgálnak a hamisítás elleni védekezésben; (3) Az újlenyomat technika egyedi azonosítót generál az adott tartalomhoz (ezt ujjlenyomatnak vagy *hash-nek* nevezik). Ez a kód később egy külső adatbázisban kerül tárolásra, amely alapján bármikor visszakereshető és összehasonlítható a tartalom az eredetivel. Az ujjlenyomat-technika hatékony eszköz lehet, hátránya, hogy a külső adatbázisok fenntartása és kezelése további erőforrást igényel (adatbiztonság fenntartása, tűzfalak, rendszerfrissítés stb.); (4) A detektálás olyan MI-alapú eszközök felhasználásával történik, melyek gépi tanulási *osztályozási technikákat* (classification techniques) alkalmaznak, hogy megkülönböztessék az ember által készített és mesterségesen létrehozott tartalmakat. Ezen módszerek megbízhatósága – kiváltképp a szöveges tartalmak esetén – rendkívül alacsony. Egyre gyakrabban jelennek meg hírek azzal kapcsolatban, hogy az emberek által készített „valódi" tartalmakat tévesen mesterségesen előállított, szintetikus tartalomként jelölik meg.[404] E problémát jól érzékelteti,

[402] Lásd bővebben: Hamon, R., Sanchez, I., Fernandez Llorca, D. and Gomez, E. (2024): Generative AI Transparency: Identification of Machine-Generated content. European Commission, Joint Research Centre, 2-5.o.

[403] Az explicit vízjelek általában jól láthatóak a tartalomban, például egy fényképen vagy videóban elhelyezett szöveg vagy logó formájában. Implicit vízjelek viszont úgy vannak kialakítva, hogy rejtve maradjanak az emberi szem elől, és elsősorban erre kifejlesztett algoritmusok segítségével azonosíthatók. Bővebben: Grinbaum, A. and Adomaitis, L. (2022): The Ethical Need for Watermarks in Machine- Generated Language. 2-6.o. arXiv preprint arXiv:2209.03118

[404] Amióta a Nagy Nyelvi Modellek mindenki számára szabadon elérhetővé váltak – kiváltképp az új technológiák iránt nyitott diákok számára – megsokszorozódott mind a technológiát használók, mind pedig a használattal tévesen megvádolt esetek száma. Jelenleg úgy tűnik, hogy az oktatási intézmények által leggyakrabban használt szoftverek (DeepTrace / Sensity / Fakespot /Turnitin) egyre kevésbé megbízhatók a mesterséges tartalmak azonosításában, elsősorban a növekvő falspozitív értékelések száma miatt. A szoftverek hiányosságairól bővebben: Sadasivan, V., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S. (2023): Can AI-

hogy az OpenAI hivatalos weboldalán már azt javasolta felhasználóinak, hogy az általuk 2023-ban elérhetővé tett tartalomkategorizáló szoftver következtetéseit főszabályként csak kiegészítő információként értelmezzék, és ne támaszkodjanak kizárólag ezekre az eredményekre az MI által generált tartalmak azonosításához.[405]

Összeségében kijelenthető, hogy a technológiai megoldások többsége hasznos ugyan, de az átlagos felhasználó számára nem képes teljes körű védelmet nyújtani. Könnyen orvosolhatná ezt az állapotot, ha létrehoznának egy olyan *böngészőben futó modult,* ami automatikusan ellenőriz minden megjelenített tartalmat*, és valamilyen formában értesítést küld,* amennyiben szintetikus tartalmat érzékel*. „Ez azonban jelenleg még inkább science-fiction*" [406]

Pozitív fejleményként értékelhető azonban, hogy egyre több példát látni iparági önszabályozásra. Növekszik azon vállalatok száma, melyek támogatják a metaadatok használatát a mesterségesen létrehozott tartalmaik megjelölésére. Példaként említhető az Adobe vállalat által indított *Content Authenticity Initiative* (CAI) kezdeményezés, amely arra ösztönzi a tartalomgyártókat, hogy metaadatokkal lássák el szintetikus alkotásaikat, annak érdekében, hogy azok mindinkább nyomon követhetők és átláthatók legyenek.[407] Ezzel együtt a legbiztosabb és legkívánatosabb megoldás mégis valamilyen nemzetközi jogi norma, vagy standard kialakítása lenne. Az Európai Unió MI-rendelete ugyan már előírja szintetikus tartalmak metaadatokkal való ellátását, a szabályozás szigorítása mégis indokolt lenne e tartalmak megjelölése, címkézése és a felhasználók tájékoztatása tekintetében.

Generated Text be Reliably Detected?. ArXiv, abs/2303.11156; Subramaniam, R. (2023): Identifying Text Classification Failures in Multilingual AI-Generated Content. *International Journal of Artificial Intelligence & Applications,* Vol.14(5), 57-63.o. https://doi.org/10.5121/ijaia.2023.14505

[405] Kirchner JH, Ahmad L, Aaronson S, Leike J (2023): OpenAI: New AI classifier for indicating AI-written text. Elérhető: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text (2023. 05. 20.). 2023. július 20-án frissítették az oldalt, a következő szöveget hozzáillesztve. *A pontosság és megbízhatóság alacsony szintje miatt az LLM szövegazonosító a továbbiakban nem elérhető. A visszajelzéseket feldolgozva jelenleg hatékonyabb tartalommeghatározási technikákat kutatunk a szintetikus szövegek számára. Elköteleztük magunkat amellett, hogy a jövőben kifejlesztünk és bevezetünk olyan mechanizmusokat, amelyek lehetővé teszik a felhasználók számára, hogy megértsék, hogy egy audio vagy vizuális tartalom MI által generált-e.* (saját fordítás)

[406] Üveges István (2024): A mesterséges intelligencia által generált tartalmak vízjelezése: megoldás vagy látszatmegoldás? *Jogászvilág*. Elérhető: https://jogaszvilag.hu/a-jovo-jogasza/a-mesterseges-intelligencia-altal-generalt-tartalmak-vizjelezese-megoldas-vagy-latszatmegoldas (2024. 03. 27.)

[407] Lásd: ADOBE (2024): What are Content Credentials? Elérhető: https://helpx.adobe.com/creative-cloud/help/content-credentials.html (2024. 07. 21.)

8.1.3 Előremutató jogi keretrendszerek kialakítása

Az elmúlt évek során a szintetikus tartalmak szabályozásakor az Unió főként a döntésbefolyásolásra alkalmas manipulációs gyakorlatokkal, a dezinformációval és félretájékoztatással, valamint a demokratikus intézmények és szabad választások elleni támadásokkal kapcsolatos kockázatokra összpontosított. E veszélyek visszaszorítása érdekében a korábban bemutatott DSA és az MI-rendelet is olyan szabályozási megközelítést alkalmaz, ami a modellek / online platformok működésének átláthatóságát és elszámoltathatóságát kívánja előmozdítani. Ez az elv tükröződik, mind a DSA előírásaiban (pl.: a felhasználók tájékoztatására vonatkozó kötelezettségek garantálása révén), mind az MI-rendelet rendelkezéseiben.

Az MI rendelet 50. cikk (1) bekezdése például kimondja, hogy a szolgáltatóknak biztosítaniuk kell, hogy a *természetes személyekkel való közvetlen interakcióra szánt MI-rendszereket* úgy tervezzék meg és fejlesszék ki, hogy az érintett természetes személyek tájékoztatást kapjanak arról, hogy egy MI-rendszerrel állnak interakcióban.[408] Hasonlóan az átláthatóság és a tájékoztatási kötelezettség fontosságát tükrözi az is, hogy *a szintetikus hang-, kép-, video- vagy szöveges tartalmat létrehozó MI-rendszerek* – köztük az általános célú MI-rendszerek – szolgáltatóinak „*biztosítaniuk kell, hogy az MI rendszer kimeneteit géppel olvasható formátumban jelöljék meg, és azok mesterségesen létrehozottként vagy manipuláltként észlelhetők legyenek”* [409]

Ugyanezen elv mentén írja elő a rendelet, hogy az olyan MI-rendszerek alkalmazóinak, amelyek *eredetinek vagy valóságosnak tűnő („deepfake”) kép-, hang- vagy videotartalmat hoznak létre vagy manipulálnak*, közölniük kell, hogy a tartalmat mesterségesen hozták létre vagy manipulálták. Azonban – feltehetően – annak okán, hogy a szintetikus tartalmak esetén a jogalkotók figyelme elsősorban a közvélemény manipulálásából, választások befolyásolásából, és politikai polarizáció kiéleződéséből fakadó kockázatokra összpontosult, a bekezdés második fele enyhít ezen a kötelezettségen, és azt írja*: amennyiben a tartalom nyilvánvalóan művészeti, kreatív, szatirikus, fiktív vagy hasonló mű vagy program részét képezi, az e bekezdésben meghatározott átláthatósági kötelezettségek az ilyen létrehozott vagy*

[408] Kivéve, ha ez a körülményekre és a felhasználási kontextusra figyelemmel, egy észszerűen jól tájékozott, figyelmes és körültekintő természetes személy szempontjából nyilvánvaló tény.
[409] MI-rendelet 50. cikk (1) bekezdés

manipulált tartalom meglétének megfelelő, a mű megjelenítését vagy élvezetét nem akadályozó közlésére korlátozódnak. [410]

Azon túl, hogy a kelleténél tágabban értelmezhető az a kikötés, hogy „*meglétének megfelelő, vagy az élvezetet nem akadályozó*" módon jelöljék a mesterséges tartalom voltát,[411] a szabályozás valódi hiányossága, hogy nem lép fel kellő erővel a szintetikus és az autentikus tartalmak feltűnésmentes keveredése ellen.

Ezen hiányossághoz kötődik, hogy az 50. cikk (4) bekezdése főszabályként azt írja elő, hogy a „*nyilvánosság közérdekű ügyekről való tájékoztatása céljából közzétett szöveget generáló vagy manipuláló MI-rendszer alkalmazóinak közölniük kell, hogy a szöveget mesterségesen hozták létre vagy manipulálták".*[412] Például egy hírportál, vagy egy blogoldal, amely Nagy Nyelvi Modellekkel gyártott automatizált cikkeket közöl (folyamatosan frissülő választási eredmények, vagy tőzsdei hírek formájában ilyen már ma is létezik) főszabályként köteles feltüntetni, hogy szintetikus tartalmat olvas a néző. Azonban a bekezdés második része enyhít ezen a követelményen úgy, „*hogy a kötelezettség nem alkalmazandó abban az esetben, ha a közzétett tartalmon emberi felülvizsgálatra vagy szerkesztési ellenőrzésre került sor, és amennyiben a tartalom közzétételéért természetes vagy jogi személy szerkesztői felelősséget visel."*[413] Itt is megfigyelhető az információ tartalmára összpontosító szabályozási megközelítés (nem pedig az információ típusára: mesterséges, szintetikus / ember által létrehozott, autentikus), hiszen amennyiben a természetes személy felelősséget vállal a közzétételért, a rendelet nem látja szükségességét annak, hogy az olvasó számára biztosítsa a jogot arra, hogy tájékoztassák a tartalom mesterséges eredetéről.

---

[410] MI-rendelet 50. cikk (4) bekezdés; E szabály – részletesebben kifejtve – a 134. preambulumkezdés második felében is olvasható: az ... „*átláthatósági kötelezettségnek való megfelelés nem értelmezhető úgy, hogy az azt jelzi, hogy az MI-rendszer vagy annak kimenete akadályozza a véleménynyilvánítás szabadságához való jogot, valamint a művészet és a tudomány szabadságához való, a Chartában garantált jogot, különösen, ha a tartalom egy nyilvánvalóan kreatív, szatirikus, művészeti, fiktív vagy hasonló munka vagy program részét képezi, a harmadik felek jogaira és szabadságaira vonatkozó megfelelő biztosítékok mellett. Az említett esetekben az e rendeletben meghatározott, a deepfake tartalmakra vonatkozó átláthatósági kötelezettség az ilyen, előállított vagy manipulált tartalom olyan megfelelő módon történő felfedésére korlátozódik, amely nem akadályozza a mű megjelenítését vagy élvezetét, beleértve annak rendes kiaknázását és használatát, a mű hasznosságának és minőségének fenntartása mellett. Emellett helyénvaló hasonló felfedési kötelezettséget előírányozni az MI által előállított vagy manipulált szöveg tekintetében is, amennyiben azt a nyilvánosság közérdekű ügyekről való tájékoztatása céljából teszik közzé, kivéve, ha az MI által létrehozott tartalom emberi felülvizsgálaton vagy szerkesztői ellenőrzésen esett át, és a szerkesztői felelősséget egy természetes vagy jogi személy viseli a tartalom közzétételéért."*

[411] Az (5) bekezdés ezt annyival pontosítja csak, hogy „*az említett tájékoztatást legkésőbb az első interakció vagy kitettség alkalmával, egyértelmű és jól megkülönböztethető módon kell az érintett természetes személyek számára nyújtani."*

[412] MI rendelet 50. cikk (4) bekezdés

[413] Uo.

E megközelítés helyett arra volna szükség, hogy minden tartalmat – legyen az komment, művészeti videó, dokumentumfilm-illusztráció vagy szórakoztató rövid videó – kötelezően jelöljenek, amennyiben az szintetikus eredetű. Az ilyen címkézésnek (1) sohasem szabad félrevezetőnek lenni (másképpen megfogalmazva; legyen egyértelmű) és; (2) sohasem szabad rejtettnek lennie (másképpen megfogalmazva; azonnal észrevehetőnek kell lennie, még akkor is, ha ez csökkenti a digitális tartalom fogyasztásának élvezetét); (3) Emellett további elvárásként megfogalmazható valamilyen egyetemes szabványosított címkézési rendszer kidolgozása is. Jelenleg a szintetikus tartalmak címkézése platformonként eltérő lehet, ami zavart és bizonytalanságot okozhat a felhasználók számára.

E három címkézéssel kapcsolatos követelményt következetesen be kell(ene) tartani. Nem pusztán azért, mert a szintetikus tartalmak manipulálhatnak, megtéveszthetnek, dezinformációt terjeszthetnek, vagy a döntéshozatali folyamatok befolyásolására használhatóak. A felsorolt kockázatkon túl fontos annak felismerése is, hogy ezen kimenetek feltűnésmentes elvegyülése végérvényesen elmoshatja a határvonalat az emberi kreativitás, intuíció és tapasztalás által vezérelt *autentikus tartalmak*, valamint az algoritmusok és programkódok által generált szintetikus tartalmak között.

A korábban tárgyalt kognitív és technológiai információfeldolgozási torzulások már így is túlságosan sebezhetővé tették a társadalmat a szintetikus tartalmak elterjedésével szemben. Ha a mindenki számára szabadon elérhető GenMI modellek korábban nem tapasztalt sebességgel és intenzitással kezdik előállítani a szintetikus tartalmakat, ez a folyamat elkerülhetetlenül tovább fog eszkalálódni. A (túlzott) óvatosság ebben az esetben indokoltnak tekinthető, hiszen a cél az *episztemológiai összeomlás* elkerülése. Ha a társadalom tagjai elveszítik bizalmukat az általuk olvasott hírekben, tudományos tényekben és korábban konszenzust élvező történelmi eseményekben, akkor az a demokratikus intézmények iránti bizalom végzetes megingását eredményezheti. Csak a tényeken alapuló közmegegyezés és a demokratikus intézményekbe vetett bizalom képes biztosítani a társadalmak szükséges alkalmazkodóképességét az előttük álló globális átalakulások sikeres navigálásához.

9 English Summary

The doctoral thesis is structured into eight chapters. At the end, there is an English summary, a bibliography, and a list of the author's publications.

Following the (Introduction), the second chapter briefly presents the history of AI development and introduces basic concepts related to the technology family. Special attention is given to mapping the machine learning patterns characteristic of modern AI. Its significance is non-negligible, as while traditional computer systems operated strictly according to pre-written instructions by programmers, dictating step-by-step what the systems must do to execute a task, machine learning has enabled AI to learn and evolve independently (iteratively) based on patterns in data and past experiences without needing further instructions.

The third chapter presents various case studies of AI applications—predominantly used in the public sector—that can significantly affect citizens' rights and obligations (e.g., systems used by courts, tax authorities, law enforcement). The selection of practices is not arbitrary; they facilitate an understanding of the risks and challenges identified by EU legislative strategies initiated in 2016 concerning these systems. The second part of the chapter discusses the most significant regulatory and societal challenges arising from practical applications, notably: (1) the lack of transparency, accountability, and predictability in AI operations trained through machine learning; (2) the risk of erroneous, biased, or discriminatory decision-making driven by AI; (3) the harmful impact of AI systems on fundamental human freedoms, decision-making autonomy, and democratic institutions.

The fourth chapter discusses the legislative process of the AI Act, focusing on the area of high and unacceptable risk AI applications and the establishment of new regulatory frameworks for general-purpose AI models. The chapter chronologically guides the reader through the process of creating the AI Regulation, starting with the Commission's draft regulation published on April 21, 2021, followed by the Council's common position in December 2022 and the Parliament's compromise proposal in May 2023. It then covers the tripartite negotiations in the second half of 2023 and the circumstances of the final adoption of the legislation on March 13, 2024.

The fifth chapter undertakes to map out the peculiarities of the widely recognized GenAI models at the end of 2022. The essence of these systems lies in their ability to generate new content based on user prompts (instructions), and they are characterized by numerous additional application possibilities. While early models were limited to a single modality—understanding and generating only textual content—the latest multimodal models can process and generate content in various formats. A notable example of such GenAIs is the Luma Dream Machine model, freely available to all users since June 2024, which can produce videos and animations reminiscent of professional film studios in about two minutes based on image or textual inputs. The chapter also addresses the most relevant challenges arising from the widespread application of Large Language Models like ChatGPT, such as machine hallucination, intentional deception, copyright conflicts related to training data, training on synthetic data, information homogenization, and model collapse. The concluding part of the chapter outlines the newest directions in the modern discourse related to GenAI systems. Among the most relevant issues are (1) the increasing—and in the long run, unsustainable—energy needs and environmental damage stemming from training these models; (2) the radical transformation of human relationships and human-machine interactions; (3) the disruptive impact of GenAI models on the labor market; (4) the potential consequences of the emergence of new AI agents (InteractiveAI).

The sixth, seventh, and eighth chapters explore the short- and long-term risks associated with the proliferation of synthetic content generated by GenAI models. In order to fully map out the process, the sixth chapter looks back in time to the early 2000s—the emergence of Web2.0—and from this point forward, illustrates how the structure of the public sphere has changed to the present day. It highlights the contradiction that while the internet was initially the embodiment of information democratization, it often only sows doubt and disinformation instead of informing users, and instead of bringing people closer, it tends to generate divisions and polarize.

The seventh chapter examines the process of the decline of truth, accompanied by the spread of conspiracy theories, fake news, and disinformation campaigns. The chapter aims to show how GenAI—especially Large Language Models and Deepfake technologies—contribute to the poisoning of truth and the erosion of societal trust. As technology evolves, it becomes increasingly difficult to recognize manipulated content, further exacerbating the crisis of social trust. The thesis also highlights how the proliferation of such technologies can threaten the stability of democratic systems and interpersonal relationships in the long term. The

second part of the chapter attempts to describe the EU's disinformation strategy and the Digital Services Act (DSA). The legislation—aimed at creating a safe, predictable, and reliable online environment for all Union citizens—is closely linked to the challenges discussed in the thesis, as it outlines service provider obligations to ensure transparency in automatic decision-making and algorithmic content filtering, and to limit manipulative practices that can affect individual decision-making freedom. The chapter concludes that although the DSA is a constructive and necessary addition to the existing legal materials addressing the potential risks of AI, none of the currently applicable laws provide comprehensive protection against the long-term threat posed by the indistinguishable blending of synthetic content with human-generated authentic content.

The eighth chapter undertakes to present the newest era of the internet (Web 0.0). A key characteristic of this era is that users, often without their knowledge, increasingly encounter AI-generated texts, images, and interactions online, while the proportion of real human connections and authentic content dramatically decreases. One of the most worrying consequences of the Web 0.0 era is that users' trust in information—initially online, later offline—gradually erodes. People can no longer distinguish real facts from artificially generated information.

9.1 Irodalomjegyzék

- Access Now (2021): Access Now’s submission to the European Commission’s adoption consultation on the AI Act. 15.o. Elérhető: https://www.accessnow.org/wp-content/uploads/2021/08/Submission-to-the-European-Commissions-Consultation-on-the-Artificial-Intelligence-Act.pdf (2022. 05. 09.)

- Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A., Babaei, H.R., LeJeune, D., Siahkoohi, A., & Baraniuk, R. (2023): Self-Consuming Generative Models Go MAD. 7-15.o arXiv:2307.01850

- Alignment Research Center (2023): Update on arc’s recent eval efforts. Elérhető: https://metr.org/blog/2023-03-18-update-on-recent-evals (2023. 08. 21.)

- Allyn, Bobby (2022): Deepfake Video of Zelenskyy Could Be ‘Tip of the Iceberg’ in Info War, Experts Warn. *NPR*. Elérhető: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-expertswar-manipulation-ukraine-russia (2022. 03. 16.)

- Anthropic (2024): Many-shot jailbreaking. Elérhető: https://www.cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking__2024_04_02_0936.pdf (2024. 06. 24.)

- Arendt, Hannah (1992): A totalitarizmus gyökerei. (Ford.: Berényi Gábor, Erős Ferenc, Seres Iván,Braun Róbert). Budapest, Európa Könyvkiadó. 577-602.o.

- Arendt, Hannah (1995): Igazság és politika. In: Múlt és jövő között. Nyolc gyakorlat a politikai gondolkodás terén. (ford: Módos Magdolna). Budapest, Osiris–Readers International, 233-251.o.

- Bai, Hui, Jan G. Voelkel, johannes C. Eichstaedt, and Robb Willer. (2023): Artificial Intelligence Can Persuade Humans on Political Issues. OSF Preprints. https://doi.org/10.31219/osf.io/stakv

- Baker, Catherine, Debbie Ging, Maja Brandt Andreasen (2024): Recommending Toxicity: The role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male

supremacist influencers. DCU Centre, Dublin City University. 2-33o. Elérhető: https://antibullyingcentre.ie/wp-content/uploads/2024/04/DCU-Toxicity-Full-Report.pdf

- Banasik, J., Crook, J., & Thomas, L. (2003): Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8), 822–832.o. https://doi.org/10.1057/palgrave.jors.2601578

- Bara Zoltán (2017): Dinamikus árazás az online kereskedelemben – Hogyan lehet hátrányos a fogyasztónak a dinamikus árazás? In: *Versenytükör*, XIII. évfolyam (2), 4-19.o.

- Barcsi Tamás - Diósi Szabolcs (2022): Hasznos test, dividuum, nyersanyag. A felügyeleti társadalomtól az (ön)ellenőrző és megfigyelési kapitalizmusig. Magyar filozófiai szemle 66(3), 190-191.o.

- Barocas, Solon – Selbst, Andrew D. (2016): Big Data's Disparate Impact. *California Law Review,* 104. 671–732.o.

- Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021): On the dangers of stochastic parrots: can language models be too big? *Proceedings of FAccT '21: Conference on Fairness, Accountability, and Transparency*. 610–623.o.

- Bertuzzi, Luca (2023): AI Act: EU countries mull options on fundamental rights, sustainability, workplace use. Elérhető: https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-countries-mull-options-on-fundamental-rights-sustainability-workplace-use/ (2024. 02. 25.)

- Bertuzzi, Luca (2023): EU's AI Act negotiations hit the brakes over foundation models. EURAVTIV. Elérhető: https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/ (2024. 01. 24.)

- Bertuzzi, Luca (2023): France, Germany, Italy push for 'mandatory self-regulation' for foundation models in EU's AI law. EURACTIV. Elérhető: https://www.euractiv.com/section/artificial-intelligence/news/france-germany-italy-push-for-mandatory-self-regulation-for-foundation-models-in-eus-ai-law (2024. 01. 20.)

- Blattner, L., Nelson, S. (2021): How Costly is Noise? Data and Disparities in Consumer Credit. 27-30.o. ArXiv, abs/2105.07554

- Bostrom, Nick (2015): Szuperintelligencia. Utak, veszélyek, stratégiák. (ford. Hidy Mátyás). Budapest, Ad Astra. 89-96.o.
- Bobadilla, J., Ortega, F. Hernando, A. and Gutierrez, A. (2013): Recommender Systems Survey. *Knowledge-Based System.* Vol(46), 112-113.o.
- Bradford, Anu (2020): The Brussels Effect: How the European Union Rules the World. New York, Oxford University Press
- Brendan Nyhan, Jason Reifler (2015): Displacing Misinformation about Events: An Experimental Test of Causal Corrections, *Journal of Experimental Political Science.* Volume 2, Issue 1, 81-93.o.
- Brewer, Jordan, Dhru Patel, Dennie Kim, Alex Murray (2023): Navigating the challenges of generative technologies: Proposing the integration of artificial intelligence and blockchain. *Business Horizons Journal Pre-proof*, 4-7.o.
- Brewster, Jack, Zack Fishman, Elisa Xu (2023): Misinformation Monitor: June 2023. Funding the Next Generation of Content Farms. Elérhető: https://www.newsguardtech.com/misinformation-monitor/june-2023/ (2024. 01. 20.)
- Bsharat Sondos Mahmoud, Aidar Myrzakhan, Zhiqiang Shen (2023): Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. 5-6.o.
- Burrell, Jenna (2016): How the Machine 'Thinks: Understanding Opacity in Machine Learning Algorithms. *Big Data and Society*, 3/1. 3-6.o.
- Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002): Deep Blue. *Artificial Intelligence*, *134(1-2*), 59-60.o.
- Castells, Manuel. (2002): Az Internet-galaxis (Gondolatok internetről, üzletről és társadalomról). Budapest, Network Twenty One Hungary kiadó.
- Cheney-Lippold, John (2017): We Are Data: Algorithms and the Making of Our Digital Selves. New York: NYU Press. 88-92.o.
- Chesney, Bobby, Danielle Citron (2019): Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, Vol. 107, 1775-1786.o.

- Chomsky, Noam - Edward S. Herman (2016): Az Egyetértés-gépezet - A tömegmédia politikai gazdaságtana. (Ford.: Konok Péter). Budapest, L'Harmattan Kiadó

- Chronowski, Nóra, Kálmán, Kinga,Szentgáli-Tóth, Boldizsár (2022): Régi keretek, új kihívások: a mesterséges intelligencia prudens bevonása a bírósági munkába és ennek hatása a tisztességes eljáráshoz való jogra. *Glossa Iuridica*, 8(4), 7-38.o.

- Confessore, Nicholas, Gabriel J.X. Dance, Richard Harris, Mark Hansen (2018): The Follower Factory: Everyone Wants to Be Popular Online. Some Even Pay for It. Inside Social Media's Black Market. *The New York Times*. Elérhető: https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html (2022. 03. 19.)

- Cox, Joseph (2024): Google News Is Boosting Garbage AI-Generated Articles. *404media*. Elérhető: https://www.404media.co/google-news-is-boosting-garbage-ai-generated-articles/ (2024. 01. 18.)

- Fantoly Zsanett, Herke Csongor and Szabó Barbara (2023): The role of AI-based systems in negotiated proceedings. *e-Revue Internationale de Droit Pénal*, 2023, A-18, 4.o.

- Chow, R. Andrew (2023): AI-Human Romances Are Flourishing—And This Is Just the Beginning. *TIME*. Elérhető: https://time.com/6257790/ai-chatbots-love/ (2024. 05. 10.)

- Collingridge, David (1980): The Social Control of Technology. New York. St. Martin's Press

- Cox, Joseph (2023): GPT-4 Hired Unwitting TaskRabbit Worker By Pretending to Be 'Vision-Impaired' Human. VICE. Elérhető: https://www.vice.com/en/article/jg5ew4/gpt4-hired-unwitting-taskrabbit-worker (2023. 05. 10.)

- Curreli, Eleonora (2023): ChatGPT Case: How the Italian Data Protection Authority Is Trying To Address AI Risks. *MONDAQ*. Elérhető: https://www.mondaq.com/italy/privacy-protection/1317670/chatgpt-case-how-the-italian-data-protection-authority-is-trying-to-address-ai-risks (2023. 05. 08.)

- Csepeli György (2020): Ember 2.0 – A mesterséges intelligencia gazdasági és társadalmi hatásai. Budapest, Kossuth Kiadó

- Danaher J. (2016): The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*. 29 (3), 245-268.o.

- Danaher, J. (2019): The ethics of algorithmic outsourcing in everyday life. In K. Yeung & M. Lodge (szerk.) Algorithmic Regulation (91–118o.). Oxford University Press. 107.o. https:// doi.org/ 10. 1093/ oso/ 97801 98838 494. 003. 0005

- Diamond, Larry (2010): Liberation Technology. *Journal of Democracy*, vol. 21, no. 3, 69-83.o.

- Diósi Szabolcs (2022): Behaviorally informed regulations- an emerging trend in modern public policymaking. In Bendes Ákos L. et al. (szerk): III. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar Doktori Iskola. 125-142.o.

- Diósi Szabolcs (2021): The rise of algorithmic decision-making in the age of Big Data. In Bendes Ákos L. et al. (szerk): I-II. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar Doktori Iskola. 150-165.o.

- Diósi Szabolcs (2023): Tiltott gyakorlatok, "elfogadhatatlan kockázat": Az Európai Bizottság rendelettervezete a diszruptív Mesterséges-Intelligencia-technológiák megfékezésére. *Európai Jog*, 23(3), 17-24.o.

- Diósi Szabolcs, Barcsi Tamás (2021): The legacy of disciplinary society – how relevant is Foucault's theory today? *Évkönyv - Újvidéki egyetem magyar tannyelvű tanítóképző*, XVI. évfolyam, 1. szám, 10-33.o.

- DiResta, R. - Goldstein, J.A. (2024): How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. 9.o. ArXiv, abs/2403.12838.

- Dupré H. Maggie (2023): AI Loses Its Mind After Being Trained on AI-Generated Data. *Futurism.* Elérhető: https://futurism.com/ai-trained-ai-generated-data (2023. 12. 07.)

- Edwards, Benj (2024): OpenAI Can Re-Create Human Voices—but Won't Release the Tech Yet. *WIRED*. Elérhető: https://www.wired.com/story/openai-voice-engine-artificial-intelligence-release (2024. 04. 03.)

- El-Gayar, M.M., Abouhawwash, M., Askar, S.S. et al. (2024): A novel approach for detecting deep fake videos using graph neural network. *J Big Data*, 11, 22.o. https://doi.org/10.1186/s40537-024-00884-y

- Eli Pariser (2021): The Filter Bubble. What the Intenet is Hiding from Yout New York. The Penguin Press. 112-113.o.

- EU DisinfoLab (2023): Connecting the disinformation dots: insights, lessons, and guidance from 20 EU Member States. 1-12.o. Elérhető: https://www.disinfo.eu/wp-content/uploads/2023/12/20231204_Connecting-disformation-dots_comparative-study-1.pdf

- Eubanks, Virginia (2018): Automating Inequality:How High-Tech Tools Profile, Police, and Punish the Poor. New York, St Martin's Publishing. 14-39.o.

- Európai Adatvédelmi Testület (2021): Az Európai Adatvédelmi Testület és az európai adatvédelmi biztos 5/2021. sz. közös véleménye. 13-16.o.

- Európai Bizottság (2018): A mesterséges intelligenciáról szóló összehangolt terv COM(2018) 795 final.

- Európai Bizottság (2019): Az emberközpontú mesterséges intelligencia iránti bizalom növelése COM(2019) 168 final

- Európai Bizottság (2020): Az Európai Demokráciára vonatkozó cselekvési tervről. COM(2020) 790 final

- Európai Bizottság: Az Európai Parlament és a Tanács rendelete a digitális szolgáltatások egységes piacáról (digitális szolgáltatásokról szóló jogszabály) és a 2000/ 31/EK irányelv módosításáról. COM(2020)825 final

- Európai Bizottság (2021): A dezinformáció elleni gyakorlati kódex megerősítésére vonatkozó iránymutatás. COM(2021) 262 final,

- Európai Bizottság (2018): A dezinformáció visszaszorítását célzó uniós gyakorlati kódex. 4-10.o.

- Európai Bizottság (2020): Európa digitális jövőjének megtervezése. A Bizottság közleménye az Európai Parlamentnek, a Tanácsnak, az Európai Gazdasági és Szociális Bizottságnak és a Régiók Bizottságának, COM/2020/67, final

- Európai Bizottság (2018): Európai megközelítés az online félretájékoztatás kezelésére. COM(2018) 236 final, 2.1. pont.

- Európai Bizottság (2020): Fehér könyv a mesterséges intelligenciáról – A kiválóság és a bizalom európai megközelítése COM(2020) 65 final. 3-4.o.

- Európai Bizottság (2022): Iránymutatások tanárok és oktatók számára a dezinformáció kezeléséről és a digitális jártasság oktatás és képzés révén történő előmozdításáról. Luxembourg, Az Európai Unió Kiadóhivatala, 20-23.o.

- Európai Bizottság (2018): Mesterséges intelligencia Európa számára (A közös európai adattér felé) COM (2018) 237 final. 7-20.o.

- Európai Bizottság (2020): Sajtóközlemény - A digitális korra felkészült Európa: A Bizottság új szabályokat javasol a digitális platformokra. 1-3.o.

- Európai Bizottság, Kommunikációs Főigazgatóság, Leyen, U. (2019): Ambiciózusabb Unió – Programom Európa számár: politikai iránymutatás a hivatalba lépő következő Európai Bizottság számára (2019–2024). Elérhető: https://data.europa.eu/doi/10.2775/56352

- Európai Bizottság, Mesterséges intelligenciával foglalkozó magas szintű szakértői csoport (2019): Etikai iránymutatás a megbízható mesterséges intelligenciára vonatkozóan, 19-24.o.

- Európai Gazdasági és Szociális Bizottság (2021): Vélemény - Javaslat európai parlamenti és tanácsi rendeletre a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról INT/940, 4-5.o

- Európai Parlament és a Tanács (EU) 2016/679 rendelete a természetes személyek védelméről a személyes adatok kezelése során és az ilyen adatok szabad áramlásáról, valamint a 95/46/EK irányelv hatályon kívül helyezéséről (Általános Adatvédelmi Rendelet)

- Európai Tanács, Az Európai Tanács ülése (2018. június 28.) – Következtetések, EUCO 9/18 2018, 8.o.

- Európai Tanács (2017): Következtetések, EUCO 14/17, 7.o.
- Európai Tanács 2018: Következtetések. EUCO 9/18, 5. o
- Európai Unió Tanácsa (2022): Javaslat – Az Európai Parlament és a Tanács rendelete a mesterséges intelligenciára vonatkozó harmonizált szabályok (a mesterséges intelligenciáról szóló jogszabály) megállapításáról és egyes uniós jogalkotási aktusok módosításáról – Általános megközelítés. 14954/22
- European Commission (2018): A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation. Luxembourg, Publications Office of the European Union, 22-31o.
- European Commission (2018): Action Plan against Disinformation. JOIN(2018) 36 final, 5-11.o.
- European Commission (2022): Strengthened Code of Practice on Disinformation. 18-34.o.
- Fábio Perez, Ian Ribeiro (2022): Ignore previous prompt: Attack techniques for language models. 3-6.o. arXiv.2211.09527
- Fair Trials (2021): Automating Injustice - The Use of Artificial Intelligence & Automated Decision- Making Systems in Crimninal Jutice. 8-21.o
- Fantoly Zsanett, Herke Csongor, Szabó Barbara (2023): The role of AI-based systems in negotiated proceedings. *e-Revue Internationale de Droit Pénal*. 2023, A-18, 4.o.
- Festinger, L., Riecken, H.W., & Schachter, S. (1956): When Prophecy Fails. University of Minnesota Press
- Festinger, L. (1957): A Theory of Cognitive Dissonance. Stanford University Press.
- Fejes Erzsébet - Futó Iván (2021): Mesterséges intelligencia a közigazgatásban – az érdemi ügyintézés támogatása. *Pénzügyi Szemle*. Különszám 2021/1, 44.o.
- Floridi, Luciano (2020): AI and Its New Winter: from Myths to Realities. *Philosophy & Technology*. 33, 1–3.o. https://doi.org/10.1007/s13347-020-00396-6

- FRA Report (2020): European Union Agency for Fundamental Rights (2020): Artificial Intelligence, Big Data and Fundamental Rights Country - Research Estonia 2020.

- Gajanan, Mahita (2017): Kellyanne Conway Defends White House's Falsehoods as 'Alternative Facts'. TIME. Elérhető: https://time.com/4642689/kellyanne-conway-sean-spicer-donald-trump-alternative-facts/ (2017. 01. 12.)

- Gartner (2012): Gartner Research. The Importance of 'Big Data': A Definition. Elérhető: https://www.gartner.com/en/ documents/2057415

- Gálik Mihály (2019): A hálózati hírmédia sajátosságai különös tekintettel a visszhangkamra- és a szűrőbuborék-jelenségre. *In Medias Res*, 8, 2. 330-342.o.

- Goldstein A. Josh, Jason Chao, Shelby Grossman, Alex Stamos, Michael Tomz (2024): How persuasive is AI-generated propaganda? *PNAS Nexus*. Volume 3, No 2, 1-7.o.

- Gombos Katalin, Gyuranecz Franciska Zsófia, Krausz Bernadett, Papp Dorottya (2021): A mesterséges intelligencia jogalkalmazási területen való hasznosíthatóságának alapjogi kérdései. In Török Bernát és Ződi Zsolt (szerk): A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 331-332.o.

- Goodfellow, Ian, Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014): Generative Adversarial Nets. 2-8.o. arXiv:1406.2661

- Google (2023): Environmental Report. 49-55.o.

- Gorenz D, Schwarz N (2024): How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PLoS ONE*, 19(7), 2-10.o. https://doi.org/10.1371/journal.pone.0305364

- Guld Ádám (2023): A deepfake és CGI-technológia az influencer marketing szolgálatában: így formálják át a digitális karakterek az ismertségipar működését. In: Aczél, Petra; Veszelszki, Ágnes (szerk.) Deepfake : a valótlan valóság. Budapest, Gondolat Kiadó. 189-190.o.

- Grynbaum, M Michael, Ryan Mac (2023): The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*. Elérhető: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (2023. 12. 27.)

- Gyires-Tóth Bálint (2020): A mélytanulás múltja, jelene és jövője. *Hiradástechnika*. 75/1., 23-29o.

- Hacker, Philipp, Andreas Engel, Marco Mauer (2023): Regulating ChatGPT and Other Large Generative AI Models. In Proceedings of the Fairness, Accountability, and Transparency (FAccT ’23). New York, ACM. 1115.o. https://doi.org/10.1145/3593013.3594067

- Hagendorf, Thilo (2024): Deception abilities emerged in large language models. *PNAS*, Vol. 121 No. 24, 1-8.o. https://doi.org/10.1073/pnas.2317967121

- Haidt, Jonathan, Nick Allen (2020): Scrutinizing the effects of digital technology on mental health. *Nature*. Vol 578, 226-227.o.

- Hainsdorf, Clara, Tim Hickman, Sylvia Lorenz, Jenna Rennie (2023): Dawn of the EU's AI Act: political agreement reached on world's first comprehensive horizontal AI regulation. White & Case Tech Newsflash. Elérhető: https://www.whitecase.com/insight-alert/dawn-eus-ai-act-political-agreement-reached-worlds-first-comprehensive-horizontal-ai (2024. 01. 30.)

- Hameleers, M, T. E. Powell, T. G. Van Der Meer, L. Bos. (2020): A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, 37(2): 281-301.o.

- Hamon, R., Sanchez, I., Fernandez Llorca, D. and Gomez, E. (2024): Generative AI Transparency: Identification of Machine-Generated content. European Commission, Joint Research Centre, 2-5.o.

- Han, Byung-Chul (2020): Pszichopolitika. (Ford. Csordás Gábor) Budapest, Typotex. 91.o.

- Hassabis, D. (2017): Artificial Intelligence: Chess match of the century. *Nature*, 544, 413-414.o. https://doi.org/10.1038/544413a

- Hassani, K. Bertrand (2020): Societal bias reinforcement through machine learning. *AI Ethics*, Volume 1, 239-247.o. https://doi.org/10.1007/s43681-020-00026-z

- Harari, Yuval Noah (2018): 21 Lecke a 21. századra. (ford: Torma Péter). Budapest, Animus kiadó. 55-60.o.

- Harari, Yuval Noah (2024): Nexus: A Brief History of Information Networks from the Stone Age to AI. New York, Random House. 305-361.o.

- Harari, Y. N. (2023): Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. The Economist. https://www.economist.com/byinvitation/2023/04/28/ yuval-noah-harari-argues-that-ai-has-hacked-the-operating-systemof-human-civilisation (2023. 04. 28.)

- Herke Csongor (2021): A mesterséges intelligencia kriminalisztikai aspektusai. *Belügyi Szemle*. 69/10. 1714–1720.o.

- Herke Csongor (2023): Deepfake: Áldás vagy átok? Jogi szabályozási szempontok. *Pro Futuro - A Jövő nemzedékek joga*. 13(1), 157-178.o.

- Herke Csongor (2023): Mesterséges intelligencia a büntetőjogi döntéshozatalban. *Jogtudományi Közlöny*. 78(4), 165–176.o

- Héder Mihály (2020): Mesterséges Intelligencia. Filozófiai kérdések, gyakorlati válaszok. Budapest, Gondolat Kör Kiadó

- Hohmann Balázs (2021): A mesterséges intelligencia közigazgatási hatósági eljárásban való alkalmazhatósága a tisztességes eljáráshoz való jog tükrében. In Török Bernát és Ződi Zsolt (szerk.) A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 413.o.

- Houy, C., Hamberg, M. & Fettke, P., (2019): Robotic Process Automation in Public Administrations. In: Räckers, M., Halsbenning, S., Rätz, D., Richter, D. & Schweighofer, E. (Hrsg.), Digitalisierung von Staat und Verwaltung. Bonn: Gesellschaft für Informatik. 62-74.o.

- Howard, Philip N. és Muzammil M. Hussain (2013): Democracy's Fourth Wave? Digital Media and the Arab Spring. Oxford University Press

- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 9-10.o. arXiv:2311.05232, https://doi.org/10.48550/arXiv.2311.05232.

- Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024): The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. Scientific reports, 14(1), 3440.

- Hughes, H., & Waismel-Manor, I. (2020): The Macedonian Fake News Industry and the 2016 US Election. PS: Political Science & Politics, 54, 19 – 23.o. https://doi.org/10.1017/S1049096520000992.

- Human Rights Watch (2021): How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net. 25.o.

- Imperva (2024): Bad Bot Report - 2024. Elérhető: https://community.imperva.com/blogs/percy-smith/2024/05/16/impervas-11th-annual-bad-bot-report-2024 (2024. 05. 16.)

- Isaak, J., & Hanna, M. (2018): User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer.* 51, 56-59.o. https://doi.org/10.1109/MC.2018.3191268.

- Italie, Hillel (2023): John Grisham, George R.R. Martin and other authors sue OpenAI for copyright infringement. Los Angeles Times. Elérhető: https://www.latimes.com/world-nation/story/2023-09-20/john-grisham-george-r-r-martin-and-other-authors-sue-openai-for-copyright-infringement (2023. 09. 20.)

- Jeffrey Dastin (2018): Insight - Amazon scraps secret AI recruiting tool that showed bias against women. REUTERS. Elérhető: https://www.reuters.com/article/idUSKCN1MK0AG (2021. 04. 11.)

- Jackie; Raskino, Mark (2008): Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Massachusetts, Harvard Business Publishing

- Josh Taylor (2023): Meta closes nearly 9,000 Facebook and Instagram accounts linked to Chinese 'Spamouflage' foreign influence campaign. The Guardian. Elérhető: https://www.theguardian.com/australia-news/2023/aug/30/meta-facebook-instagram-shuts-down-spamouflage-network-china-foreign-influence (2023. 08. 30.)

- Kavanagh, Jennifer and Michael D. Rich (2018): Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life. Santa Monica, CA: RAND Corporation. 21-38.o

- Keller, H. Michael (2018): The Flourishing Business of Fake YouTube Views. The New York Times. Elérhető: https://www.nytimes.com/interactive/2018/08/11/technology/youtube-fake-view-sellers.html (2023. 07. 21.)

- Kirchner JH, Ahmad L, Aaronson S, Leike J (2023): OpenAI: New AI classifier for indicating AI-written text. Elérhető: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text (2023. 05. 20.).

- Kirk, Robert Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, Roberta Raileanu (2024): Understanding the Effects of RLHF on LLM Generalisation and Diversity. 4-10.o. arXiv:2310.06452

- Klein Daniel (2018): Mighty mouse. *MIT Technology Review.* Elérhető: https://www.technologyreview.com/2018 /12/19/138508/ mighty-mouse (2021. 05. 23.)

- Kollár Csaba (2020): Kína és a társadalmi kredit rendszere. *Hadtudomány: A magyar hadtudományi társaság folyóirata*. 30:2, 79-97.o.

- Koltay András (2021): Előszó. In Török Bernát és Ződi Zsolt (szerk) A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi kiadó. 11.o.

- Koltay Tibor (2010): Az új média és az írástudás formái. *Magyar Pedagógia*, 110.évf.(4) 301-309.o.

- Koltay Tibor (2017): Egy „örökzöld téma”: az információs túlterhelés. *Információs Társadalom*, XVII. évf. 3. szám, 39–54.o.

- Krekó Péter, Molnár Csaba (2023): Tényrelativizmus és a hírforrásokkal szembeni elbizonytalanodás a magyar közvéleményben. Lakmusz-HDMO. 5.o. Elérhető: https://politicalcapital.hu/pc-admin/source/documents/hdmo _pc_tanulmany_2_tenyrelativizmus_kozvelemeny_20231130.pdf

- Krekó Péter (2021): Tömegparanoia 2.0 - Összeesküvés-elméletek, álhírek és dezinformáció. Budapest, Athenaeum Kiadó. 22.o.

- Kreps, Sarah & Doug Kriner (2023): How AI Threatens Democracy. *Journal of Democracy*. Volume 34(4), 123-125.o.

- Kruger, Justin & Dunning, David. (1999): Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology,* 77(6), 1121–1134.o.

- LeCun, Yann, Yoshua Bengio és Geoffrey Hinton (2015): Deep learning. *Nature 521 (75539),* 436-444.o.

- Lembke, Anna (2024): Dopaminkorszak – Hogyan találjunk egyensúlyt a függőségekre épülő világban. (Ford.: Bujdosó István). Budapest, Libri Könyvkiadó Kft.

- Lemley, A. Mark - Bryan Casey (2020): Fair learning. *Texas Law Review.* 99, 743.o.

- Leviston, Z., I. Walker, and S. Morwinski (2013): Your opinion on climate change might not be as common as you think. Nature Climate Change, 3:334–337.o.

- Lim, N., Kuan, M., Pu, M., Lim, M., & Chong, C. (2022). Metamorphic Testing-based Adversarial Attack to Fool Deepfake Detectors. *2022 26th International Conference on Pattern Recognition* (ICPR), 2503. https://doi.org/10.1109/ICPR56361.2022.9956543.

- Liu, Chuncheng (2019): Multiple Social Credit Systems in China. Economic Sociology: *The European Electronic Newsletter*, Vol 21 (1), 22–32.o.

- Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023): Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.o.

- Lippmann, Walter (1971): A közvélemény I-II. Budapest, Tömegkommunikációs Kutatóközpont. 215-281.o

- Lucas, Jessica (2024): The teens making friends with AI chatbots. The Verge. Elérhető: https://www.theverge.com/2024/5/4/24144763/ai-chatbot-friends-character-teens (2024. 05. 10.)

- Luccioni, A. S., Jernite, Y., Strubell, E. (2023): Power hungry processing: Watts driving the cost of ai deployment? 13-14o. arXiv preprint arXiv:2311.16863

- Mezei Kitti, Bán-Forgács Nóra (2022): Discrimination in the Age of Algorithms. In: Human Rights as a Guarantee of Smart, Sustainable and Inclusive Growth. Budapest; Józefów, Milton Friedman University; Alcide De Gasperi University of Euroregional Economy, 73-80.o.

- Maciejewski, Mariusz (2016): To do more, better, faster and more cheaply: using big data in public administration. *International Review of Administrative Sciences*. 83, 120–135.o.

- Margry Karel (1992): Theresienstadt (1944–1945), the Nazi propaganda film depicting the concentration camp as paradise. Historical Journal of Film, Radio and Television. Vol(12) No. 2,1 45–162.o. doi: 10.1080/01439689200260091.

- Mayer, J., & Mitchell, J. C. (2012): Third-Party Web Tracking: Policy and Technology. *IEEE Symposium on Security and Privacy*

- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955): *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Az eredeti szöveg angol nyelven elérhető: https://www.formal.stanford.edu/jmc/history/ dartmouth/dartmouth.html

- McNulty, Eileen (2014): Understanding Big Data: The Seven V's. DataConomy. Elérhető: https://dataconomy.com/2014/05/22/ seven-vs-big-data (2014. 05. 22.)

- Meltwater & We Are Social (2024): Digital 2024 Global Overview Report. Elérhető: https://datareportal.com/reports/digital -2024-global-overview-report

- Meta, Bickert, Monika (2024): Our Approach to Labeling AI-Generated Content and Manipulated Media. *META Newsroom*. Elérhető: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/(2024. 04. 05.)

- Michal Kosinski, D. Stillwell, and T. Graepel (2013): Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*. 110:5802–5805.o.

- Misuraca, Gianluca., - van Noordt, C. (2020): Overview of the use and impact of AI in public services in the EU. EUR 30255 EN, Publications Office of the European Union, Luxembourg 45-46.o.

- Mnih, Volodymyr, Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A.,

Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S.,& Hassabis, D., (2015): Human-level control through deep reinforcement learning. *Nature,* 529-533.o.

- Mok, Aaron (2023): A disturbing AI phenomenon could completely upend the internet as we know it. *Business Insider*. Elérhető: https://www.businessinsider.com/ai-model-collapse-threatens-to-break-internet-2023-8 (2023. 08. 30.)

- Molnar, Christoph (2020): Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub. Elérhető: https://christophm.github.io/interpretable-ml-book/

- Moravec, Hans (1988): Mind Children. Cambridge, Harvard University Press. 15-16.o.

- Mráz Attila (2022): Deepfake, demokrácia, kampány, szólásszabadság. In: Török Bernát–Ződi Zsolt (szerk.): A mesterséges intelligencia szabályozási kihívásai. Budapest, Ludovika Egyetemi Kiadó

- Natasha Lomas (2023): Big Tech's divisive 'personalization' attracts fresh call for profiling-based content feeds to be off by default in EU. TechCrunch. Elérhető: https://techcrunch.com/2023/12/20/dsa-recommender-systems/

- N. C. Köbis and L. Mossink (2021): Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, vol.114(2):106553 doi: 10.1016/j.chb.2020.106553

- Németh Tamás (2023): Tilesch György: Sokan még mindig a Transformers-filmek szintjén kezelik a mesterséges intelligenciát. EconomX. Elérhető: https://www.economx.hu/gazdasag/ai-summit-2023-tilesch-gyorgy-mesterseges-intelligencia-kkv-k-munkaltatok.777160.html (2023. 09. 11.)

- Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, Florian Tramèr (2023): Poisoning web-scale training datasets is practical. 4 -13.o. arXiv:2302.10149.

- Nickerson, Raymond S. (1998): Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2 (2), 175-220.o.

- Nils J. Nilsson (2009): The Quest for Artificial Intelligence. Cambridge University Press. 13.o.

- O’Hara, K. (2020): Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review,* Vol.39. 3-6.o.

- OpenAI (2023): AI and Covert Influence Operations: Latest Trends. REPORT. 17-23.o. Elérhető:https://downloads.ctfassets.net/kftzwdyauwt9/5IMxzTmUclSOAcWUXbkVrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf (2023. 10. 22.)

- OpenAI (2024): Navigating the Challenges and Opportunities of Synthetic Voices. Blog, Product Announcements. Elérhető: https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices (2024. 04. 02.)

- OpenAI (2023): GPT-4 Technical Report. 55.o. arXiv:2303.08774v6

- O'Reilly, T. (2005): What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, No.65 (1), 17-37.

- Országgyűlés Hivatala (2023): Infodémia – A dezinformációs járvány hatásai és a védekezési lehetőségek. *Infojegyzet,* 2023. Elérhető: https://www.parlament.hu/documents/10181/64399821/Infojegyzet_2023_21_infodemia.pdf/93d64698-1928-84ac-df43-eaaba42124eb?t=1688048577522 (2023. 08. 10.)

- Oswald et al (2018): Algorithmic risk assessment policing models: lessons from the Durham HART model and ‘Experimental’ proportionality. *Information & Communications Technology Law.* 27(2), 223-250.o.

- Pataki Gábor – Szőke Gergely László (2017): Az online személyiségprofilok jelentősége – régi és új kihívások. *Infokommunikáció és jog.* 2017/2., 63-70.o.

- Park A. Lee (2019): Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing. *UCLA Law Review*. Elérhető: https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/ (2021. 07. 22.)

- Pasquale, Frank. (2015): The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

- Peterson, Jake (2024): AI Companies Are Running Out of Internet. LifeHacker. Elérhető: https://lifehacker.com/tech/ai-is-running-out-of-internet (2024. 04. 03.)

- Pew Research Center (2023): Teens, Social Media and Technology - 2023. 5-6.o. Elérhető: https://www.pewresearch.org/wp-content/uploads/sites/20/2023/12/PI_2023.12.11-Teens-Social-Media-Tech_FINAL.pdf

- Pődör Lea (2021): Leibniz and His Possible Effect on AI – Some Thoughts on the Legal System and Judicial Decision-Making in the Age of AI. In *12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021)*, 853–858.o.

- ProPublica (2016): How We Analyzed the COMPAS Recidivism Algorithm. Elérhető: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (2020. 05. 14.)

- Ranerup, Agneta & Henriksen , Z. H (2019): Value positions viewed through the lens of automated decision-making: The case of social services. Government Information Quarterly. Volume 36(4) 2-3.o.

- Ranerup, Agneta, Henriksen, Z. H. (2022): Digital Discretion: Unpacking Human and Technological Agency in Automated Decision Making in Sweden's Social Services. Social Science Computer Review. 40(2), 445-461.o.

- Régiók Európai Bizottsága (2021): Vélemény. A mesterséges intelligenciával kapcsolatos európai megközelítés – A mesterséges intelligenciáról szóló jogszabály. SEDEC-VII/022, 13.o.

- Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023): Negativity drives online news consumption. *Nature human behaviour, 7(5)*, 812-822.o.

- Romm, Tony (2018): Twitter has notified at least 1.4 million users that they saw Russian propaganda during the election. *VOX*. Elérhető: https://www.vox.com/2018/1/31/16956958/twitter-jack-dorsey-russia-trolls-election-us-trump-clinton-propaganda (2022. 10. 15.)

- Rossen, Jake (2023): Please Tell Me Your Problem': Remembering ELIZA, the Pioneering '60s Chatbot. Mental Floss. Elérhető: https://www.mentalfloss.com/posts/eliza-chatbot-history (2023. 02. 14.)

- Rudin, C. (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. (1), 206–215.o

- Russel, Josh (2023): Sanctions ordered for lawyers who relied on ChatGPT artificial intelligence to prepare court brief.Courthouse News Service. Elérhető: https://www.courthousenews.com/sanctions-ordered-for-lawyers-who-relied-on-chatgpt-artificial-intelligence-to-prepare-court-brief (2023. 06. 22.)

- Russell, Stuart J. – Norvig, Peter (2010): Artificial Intelligence: A Modern Approach. Third Edition. Essex, Pearson. 2-16.o.

- Sadasivan, V., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023): Can AI-Generated Text be Reliably Detected? ArXiv, abs/2303.11156

- Sarkar, Soumyadeep (2024): OpenAI to get access to Reddit data to train its AI models. The Tech Portal. Elérhető: https://thetechportal.com/2024/05/17/openai-gpt-reddit-ai-model-training (2024. 05. 20.)

- Satariano, Adam (2024): France Fines Google Amid A.I. Dispute With News Media. The New York Times. Elérhető: https://www.nytimes.com/2024/03/20/business/france-google-fine.htmlutm_source=www.mindstream.news&utm_medium=newsletter&utm_campaign=france-sues-google (2024. 03. 20.)

- Scheurer, J., Balesni, M. and Hobbhahn, M., (2023): Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. 2-6.o. arXiv:2311.07590

- Schmidhuber, J. (2015): Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.o.

- Scola, Nancy - Ashley Gold (2017): Facebook: Up to 126 million people saw Russian-planted posts. *POLITICO*. Elérhető: https://www.politico.eu/article/facebook-up-to-126-million-people-saw-russian-planted-posts/ (2022. 10. 15.)

- Selbst, A. (2017): Disparate Impact in Big Data Policing. *Georgia law review.* 52, 3373.o. https://doi.org/10.2139/SSRN.2819182.

- Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou (2024): Make Your LLM Fully Utilize the Context. 3-10.o. arXiv:2404.16811

- Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023): The Curse of Recursion: Training on Generated Data Makes Models Forget. 3-6.o. ArXiv, abs/2305.17493.

- Sibai, Noor (2022): OpenAI Chief Scientist Says Advanced AI May Already Be Conscious. *The_BYTE*. Elérhető: https://futurism.com/the-byte/openai-already-sentient (2022. 11. 20.)

- Smuha, Nathalie A., Ahmed-Rengers, Emma Harkens, Adam, Li, Wenlong MacLaren, James Piselli, Riccardo, Yeung, Karen (2021): How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. http://dx.doi.org/10.2139/ssrn.3899991

- Stim, Rich: What Is Fair Use? Stanford Libaries. Elérhető: https://fairuse.stanford.edu/overview/fair-use/what-is-fair-use

- Subramaniam, R. (2023). Identifying Text Classification Failures in Multilingual AI-Generated Content. *International Journal of Artificial Intelligence & Applications*. Vol.14(5), 57-63.o. https://doi.org/10.5121/ijaia.2023.14505

- Sugumaran, D., John, Y., C, J., Joshi, K., Manikandan, G., & Jakka, G. (2023): Cyber Defence Based on Artificial Intelligence and Neural Network Model in Cybersecurity. 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 1-8.o. https://doi.org/10.1109/ICONSTEM56934.2023.10142590.

- Suleyman, Mustafa, Michael Bhaskar (2023): A következő hullám. Mesterséges intelligencia, technológia, hatalom és a 21. század legnagyobb kihívása. (ford. Farkas Veronika). Budapest, Magnólia kiadó. 74-76.o.

- Sunstein, Cass R. (2001): Republic.com Princeton NJ: Princeton University Press

- Sunstein, Cass (2009): Republic.com 2.0. New Jersey, Princeton Univerity Press. 11.o.

- Sunstein (2017): #Republic: Divided Democracy in the Age of Social Media. Princeton, Princeton University Press.

- Susser, D., Roessler, B., & Nissenbaum, H. (2019): Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), 1-21.o.

- Szántó Richárd - Dudás Levente (2017): A döntési helyzetek tudatos tervezésének háttere: A nudge fogalma, módszerei és kritikái. *Vezetéstudomány,* XLVIII. évf. 10. szám, 48-57.o.

- Szőke Gergely László (2012): Az Európai adatvédelmi jog megújítása. Tendenciák és lehetőségek az önszabályozás területén. PTE-ÁJK, 57-58.o.

- Szőke Gergely László (2018): Big Data and Algorithms in the Public Sector and Their Impact on the Transparency of Decision-Making. *Central and Eastern European eDem and eGov Days*, 301-306.o

- Szőke, Gergely László (2021): A közösségi oldalak szabályozási problémái, 105-120., In: Kis Kelemen, Bence; Mohay, Ágoston (szerk.): A technológiai fejlődés jogi kihívásai: Kézikönyv a jogalkotás és jogalkalmazás számára, PTE ÁJK,

- Tegmark, Max (2017): Élet 3.0. Embernek lenni a mesterséges intelligencia korában (Ford.:Weisz Böbe, Garai Attila). Budapest, HVG Könyvek, 69.o

- Thaler, R., Sunstein, C. (2008): Nudge: Improving Decisions About Health, Wealth, and Happiness London, Penguin Books

- Tilesch György – Hatamleh, Omar (2021): Mesterség és Intelligencia. Vegyük a kezünkbe a sorsunkat az MI korában. Budapest, Libri Kiadó

- Tim Wu (2016): The Attention Merchants: The Epic Scramble to Get Inside Our Heads. New York, Knopf

- Tiku, Nitasha (2022): The Google engineer who thinks the company's AI has come to life. The Washington Post. Elérhető: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine (2022.11. 05.)

- Toffler, Alvin (1980): The Third Wave. New York, Bantam Books

- Török Bernát (2022): A szólásszabadság a közösségi platformokon és a Digital Services Act. In: Török Bernát és Ződi Zsolt (szerk): Az internetes platformok kora. Budapest, Ludovika Egyetemi Kiadó, 197.o.

- Tufekci, Zeynep. (2017): Twitter and Tear Gas: The Power and Fragility of Networked Protest. Yale University Press

- Turing, A. M. (1950): Computing Machinery and Intelligence. *Mind, 59(236),* 433 – 460.o.

- U.S. Government Publishing Office (2019): Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election. Senate Report, II. Volume, 116-290.o.

- Üveges István (2024): A mesterséges intelligencia által generált tartalmak vízjelezése: megoldás vagy látszatmegoldás? *Jogászvilág.* Elérhető: https://jogaszvilag.hu/a-jovo-jogasza/a-mesterseges-intelligencia-altal-generalt-tartalmak-vizjelezese-megoldas-vagy-latszatmegoldas (2024. 03. 27.)

- Vaccari, C., & Chadwick, A. (2020): Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society,* 6(1) https://doi.org/10.1177/2056305120903408

- Varga Imre (2020): Algoritmusok és a programozás alapjai. A Debreceni Egyetem Informatikai Karának oktatási tananyaga. Elérhető: https://irh.inf.unideb.hu/~vargai/APA/index.html

- van de Donk, W. B. H. J., & Tops, P. W. (1995): Orwell or Athens? Informatization and the Future of Democracy. In W. B. H. J. van Donk, I. T. M. Snellen, & P. W. Tops (szerk.): Orwell in Athens: a Perspective on Informatization and Democracy. IOS Press. 13-32.o.

- Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017): Attention is All you Need. *Advances in Neural Information Processing Systems*. 5998-6008.o.

- Wells Fargo Bank (2023): Investor Presentation - 2023. 10.o. Elérhető: https://www08.wellsfargomedia.com/ assets/pdf/about/investor-relations/presentations/2023/april-investor-presentation.pdf?utm_source=www.mind stream.news&utm_medium=newsletter&utm_campaign=ai-vs-energy-consumption (2023. 08. 12.)

- Veszelszki Ágnes (2017): Az álhírek extra- és intralingvális jellemzői. Századvég, 84: 51-82.o.

- Veszelszki Ágnes (2021): deepFAKEnews: Az információmanipuláció új módszerei. In Balász László (szerk.): Digitális Kommunikáció és Tudatosság. Budapest, Hungarovox kiadó, 96.o.

- Viktor Mayer-Schönberger - Kenneth Cukier (2014): Big data. Forradalmi módszer, amely megváltoztatja munkánkat, gondolkodásunkat és egész életünket. Ford. Dankó Zsolt. Budapest, HVG Könyvek

- Vincent, James (2023): OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host.*The Verge*. Elérhető: https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit (2023. 06. 09.)

- Von Neumann, J. (1945): First Draft of a Report on the EDVAC. Az eredeti szöveg beszkennelt formában angolul elérhető: https://archive.computerhistory.org/resources/text/Knuth_Don_X4100/PDF_index/k-8-pdf/k-8-u2593-Draft-EDVAC.pdf

- von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C. and Schuecker, J., (2019): Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 35, 614-633.o.

- Voss, Peter (2007): Essentials of General Intelligence: The Direct Path to Artificial General Intelligence. In: Goertzel, B., Pennachin, C. (szerk.): Artificial General Intelligence. Cognitive Technologies. Berlin, Springer Kiadó, 131-157.o. https://doi.org/10.1007/978-3-540-68677-4_4

- Voss, Peter, Jovanovic, Mlađan (2023): Why We Don't Have AGI Yet. ArXiv, abs/2308.03598, 5-7.o. https://doi.org/10.48550/arXiv.2308.03598

- Vörös Zoltán (2019): Kína a digitális korszakban - Kínai internet és a Társadalmi Kreditrendszer. In: Dombi Judit – Rimai Dávid (szerk.): Digitális forradalom világunkban: tanulmánykötet. Pécs, Institutio Kiadó. 165-182.o.

- Vykopal, I., Pikuliak, M., Srba, I., Móro, R., Macko, D., & Bieliková, M. (2023): Disinformation Capabilities of Large Language Models. ArXiv, abs/2311.08838. https://doi.org/10.48550/arXiv.2311.08838.

- Wooley, Samuel C. (2020): Bots and Computational Propaganda: Automation for Communication and Control. In Social Media and Democracy: State of the Field. (Szerk:) Persily, Nathaniel and Tucker, Joshua A.Cambridge, Cambridge University Press, 89–110.o.

- World Economic Forum (2024): The Global Risk Report 2024 - 19th Edition. 14-37o.

- W. Youyou, M. Kosinski, and D. Stillwell (2015): Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112:1036-1040.o.

- Yeung, Karen (2017): Hypernudge: Big data as a mode of regulation by design. *Information, Communication and Society*, Vol.20(1), 118-136.o

- Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu (2023): Jailbreaking ChatGPT via prompt engineering: An empirical study. 4-9.o. arXiv:2305.13860v2

- Yuval Noah Harari (2018): 21 Lecke a 21. századra. (ford: Torma Péter). Budapest, Animus kiadó. 55-60.o.

- Zhang, Yue, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, Shuming Shi (2023): Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. 2-6. o. https://doi.org/10.48550/arXiv.2309.01219

- Zöldi Blanka (2023): A Lakmusz megkapta az Európai Tényellenőrző Hálózat tanúsítványát. *LAKMUSZ.* Elérhető: https://www.lakmusz.hu/a-lakmusz-megkapta-az-europai-tenyellenorzo-halozat-tanusitvanyat (2023. 12. 18.)

- Ződi Zsolt (2023): A Collingridge dilemma működés közben. Hogyan szabályozhatunk valamit, amit nem is ismerünk? *Ludovika*. Elérhető: https://www.ludovika.hu/blogok/itkiblog/2023/07/31/a-collingridge-dilemma-mukodes-kozben/ (2023. 11. 27.)

- Ződi Zsolt (2017): Jog és jogtudomány a Big Data korában. *Állam- és Jogtudomány 2017/1.*

- Ződi Zsolt (2018): Platformok, robotok és a jog. Új szabályozási kihívások az információs társadalomban. Budapest, Gondolat Kiadó. 234.o.

9.2 Publikációs jegyzék

- Diósi Szabolcs (2023): Adatvezérelt döntések, előrejelző algoritmusok, Mesterséges Intelligencia: Technológiai innováció a köz szolgálatában? In. Barcsi, Tamás (szerk.); Csefkó, Ferenc (szerk.); Diósi, Szabolcs (szerk.): HCS 70. Ünnepi írások Horváth Csaba ny. egyetemi docens születésnapjára. Pécs, Jövő Közigazgatásáért Alapítvány, PTE ÁJK, 74-87.o.

- Diósi Szabolcs (2022): Behaviorally informed regulations- an emerging trend in modern public policymaking. In Bendes Ákos L. et al. (szerk): III. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar Doktori Iskola, 125-142.o.

- Diósi Szabolcs (2020): Big Data and mental privacy. In: Barna, Boglárka Johanna; Kovács, Petra; Molnár, Dóra; Pató, Viktória Lilla (szerk.): XXIII. Tavaszi Szél Konferencia 2020. Absztraktkötet: MI és a tudomány jövője. Budapest, DOSZ, 522.o.

- Diósi Szabolcs (2022): Facing 'unacceptable risk'. In: Molnár, Dániel; Molnár, Dóra (szerk.) XXV. Tavaszi Szél Konferencia 2022: Absztraktkötet. Budapest, DOSZ, 126.o.

- Diósi Szabolcs (2022): Great opportunities, greater threats: The unfolding history of data-driven social scoring In: Berek, Patrícia; Fodor, Krisztina Dóra (szerk.): Ab ovo usque ad mala – Selected Studies from the "Destinies and Processes" conference. Pécs, Történelmi Ismeretterjesztő Tudás- és Emléktár Alapítvány, 123-136.o.

- Diósi Szabolcs (2023): Hogyan (ne) pontozzuk állampolgárainkat? Ludovika. Kormányzás és Tudomány Blog. Elérhető: https://www.ludovika.hu/blogok/kormblog/2023/07/14/hogyan-ne-pontozzuk-allampolgarainkat/ (2023. 07.14.)

- Diósi Szabolcs (2019): How algorithm-controlled societies could affect individual autonomy. In: Bódog, Ferenc; Csiszár, Beáta (szerk.): 8th Interdisciplinary Doctoral Conference 2019: Book of Abstract. Pécs, PTE Doktorandusz Önkormányzat. 45.o.

- Diósi Szabolcs (2023): Latest Amendments to the European Union's Artificial Intelligence Act: The Emergence of General-Purpose and Generative AI Technologies. In:Ivanova,

Mariana; Dragica, Odzaklieska; Rasim, Yilmaz (szerk.) XX. International Balkan and Near Eastern Congress Series on Economics, Business and Management: Proceedings. Sofia, University St. Kliment Ohridski, Faculty of Economics and Business Administration, 687.o.

- Diósi Szabolcs (2021): Personalized paternalism - present-day analysis of an age-old idea. In. Kajos, Luca Fanni (szerk.); Bali, Cintia (szerk.); Preisz, Zsolt (szerk.); Polgár, Petra [Polgár, Petra Ibolya (szerk.); Glázer-Kniesz, Adrienn (szerk.); Tislér, Ádám [szerk.]; Szabó, Rebeka (szerk.): 10th Jubilee Interdisciplinary Doctoral Conference : Book of Abstracts 241.o.

- Diósi Szabolcs, Barcsi Tamás (2021): The legacy of disciplinary society – how relevant is Foucault's theory today? Évkönyv - Újvidéki egyetem magyar tannyelvű tanítóképző. XVI. évfolyam, 1. szám, 10-33.o

- Diósi Szabolcs (2021): The rise of algorithmic decision-making in the age of Big Data. In Bendes Ákos L. et al. (szerk): I-II. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar Doktori Iskola. 150-165.o.

- Diósi Szabolcs (2023): Tiltott gyakorlatok, "elfogadhatatlan kockázat": Az Európai Bizottság rendelettervezete a diszruptív Mesterséges-Intelligencia-technológiák megfékezésére. Európai Jog 23(3), 17-24.o.

- Diósi, Szabolcs (2023): Trustworthy AI in Public Administration - the EU perspective on regulating disruptive technologies. In: Ivanova, Mariana; Nikoloski, Dimitar; Yilmaz, Rasim (szerk.): Proceedings of XVII. International Balkan and Near Eastern Social Sciences Congress Series on Economics. Mitrovica, University of Isa Boletini, 687.o.

- Barcsi Tamás - Diósi Szabolcs (2022): Hasznos test, dividuum, nyersanyag. A felügyeleti társadalomtól az (ön)ellenőrző és megfigyelési kapitalizmusig. Magyar filozófiai szemle 66(3), 190-191.o.

- Szabó, Gábor, Diósi, Szabolcs (2022): Ecological debt and sustainable development. In: Simon, Zoltán; Ziegler, Dezső Tamás (szerk.): European Politics - Crises, Fears, and Debates. Paris, L'Harmattan, 89-102.o.

Szerkesztői munkák

- Barcsi Tamás (szerk.); Csefkó, Ferenc (szerk.); Diósi, Szabolcs (szerk.): HCS 70. Ünnepi írások Horváth Csaba ny. egyetemi docens születésnapjára. Pécs, Jövő Közigazgatásáért Alapítvány, PTE ÁJK (2023)

- Barcsi Tamás; Diósi, Szabolcs (szerk.): A válság elméleti vonatkozásai: Tanulmánykötet. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar (2022)

- Barcsi Tamás, Diósi Szabolcs, Mészáros Gábor, Monori Gábor (szerk.): Globális igazságosság, emberi jogok, jogászi etika: Szabó Gábor-emlékkötet. Pécs, Pécsi Tudományegyetem Állam- és Jogtudományi Kar (2023)

- Tóth Dávid; Bendes, Ákos László; Diósi, Szabolcs; Gáspár, Zsolt; Projics, Nárcisz; Serbakov, Márton Tibor; Szívós, Alexander Roland (szerk.) I-II. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, PTE Állam- és Jogtudományi Kar, Doktori Iskola (2021)

- Bendes Ákos László (szerk.); Diósi, Szabolcs (szerk.) ; Gáspár, Zsolt (szerk.) ; Gáti, Balázs (szerk.) ; Projics, Nárcisz (szerk.) ; Tóth, Dávid (szerk.): III. Konferenciakötet: A pécsi jogász doktoranduszoknak szervezett konferencia előadásai. Pécs, PTE Állam- és Jogtudományi Kar, Doktori Iskola (2022)