Iustum Aequum Salutare
XXI. 2025. 3. • 177–196.

A MESTERSÉGES INTELLIGENCIA ALAPÚ TARTALOMSZŰRŐ ALGORITMUSOK TÁRSADALMI HATÁSAI, KÜLÖNÖS TEKINTETTEL A KÖZÖSSÉGIMÉDIA-PLATFORMOKRA

Üveges István

1. Bevezetés

Napjainkra a közösségekben zajló kommunikáció elsődleges tere egyértelműen az online világ lett. A különböző online platformok, közösségimédia-felületek, videómegosztó oldalak stb. az információáramlás domináns médiumát képezik. Az ilyen, gyakran globálisan működő kommunikációs terek azonban fokozottan ki vannak téve a szélsőséges ideológiák, szerzői jogi szempontból aggályos, esetleg más szempontból jogsértő, gyakran felkavaró tartalmak ellenőrizetlen terjedése okozta kockázatnak.

A probléma kezelésére bevett gyakorlat, hogy az ilyen platformokra feltöltött tartalmak ellenőrzésen esnek át, akár automatikusan, közvetlenül a feltöltésük után, akár később, ha egy felhasználó problémásként bejelenti az adott tartalmat. Ennek következménye lehet többféle szankció is, a tartalom törlésétől a felhasználó kizárásán át egészen jogi eljárás megindításáig. Ezt a fajta ellenőrzési mechanizmust szokás összefoglalóan 'moderálásként' hivatkozni.

Azáltal, hogy például a közösségimédia-platformok mögött álló cégek mind saját megoldásokat alkalmaznak a felületeiken felbukkanó tartalmak moderálására, valójában olyan hatalomra tesznek szert, amellyel korábban csak államok és egyházak rendelkezhettek.[1] Vékony a határvonal az intézményesített cenzúra, és a felhasználók érdekeit szem előtt tartó tartalomszűrés között. A kettő összemosódása különösen gyakran kerül szóba annak kapcsán, hogy a legnagyobb közösségimédia szolgáltatók mind jobban támaszkodnak akár mesterséges intelligenciát is alkalmazó, de mindenképpen teljesen automatizált rendszerekre a moderálási folyamat során.

* Tudományos segédmunkatárs, ELTE Társadalomtudományi Kutatóközpont.
ORCID: https://orcid.org/0000-0001-5897-9379

1 Gosztonyi Gergely: *Cenzúra Arisztotelésztől a Facebookig.* Budapest, Gondolat, 2022.

https://doi.org/10.69695/ias.2025.3.11

Az ilyen rendszerek működése során az emberi tényező háttérbe szorul, hiszen a döntéseket sok esetben kizárólag algoritmusok hozzák meg. Ez nem pusztán az ilyen rendszerek pontatlansága miatt lehet problémás, de például gépi tanulás alkalmazása esetében az azok tanítóadataiban jelenlévő torzítások *(bias)* is kockázati faktort jelentenek.

A jelen tanulmány célja éppen ezért, hogy feltárja az automatizált/algoritmikus moderálás elterjedése mögött álló tényezőket, felhívja a figyelmet azok kockázataira és hatásaira a társadalom egészére nézve, valamint, hogy egyfajta helyzetképet adjon az online kommunikációs terek körül kialakult viták aktuális állapotáról, néhány lehetséges megoldást is felvázolva. Ehhez elsőként a tanulmány áttekinti a jogi helyzetet, amely az Egyesült Államokban lehetőséget teremtett az online kommunikáció ma ismert formáinak elterjedésére. Ezt követően ismerteti, hogy milyen (nyilvánosan megismerhető) módszereket alkalmaznak az érintett platformok a náluk megjelenő tartalmak megfelelőségének biztosítására. Végül felvázol néhányat azon problémás pontok közül, ahol ezek a módszerek kihatnak az egyén szólásszabadságára, vagy éppen az egyenlő bánásmódhoz való jogára.

2. A szólásszabadság védőbástyája – és Achilles-sarka

Az online szólásszabadság kérdését a közösségi médiában alapvetően a *Communications Decency Act* (CDA)[2] 230-as szakasza határozza meg az USA esetében. Ez alapvető fontosságú, hiszen a legtöbb ma ismert közösségimédia-platform innen származik. A CDA megszületésének hátteréről több átfogó elemzés[3] is napvilágot látott.[4] Ezek alapján rekonstruálható a törvénybe, főként annak leginkább vitatott 230-as szakaszába foglaltak háttere.

Az online tartalommoderálás szükségességét érintő viták talán egyik első sarkalatos pontját egy 1995-ös USA-beli per, a Stratton Oakmont v. Prodigy Services ügy jelentheti.[5] A per első résztvevője, a felperes, a Stratton Oakmont Inc. volt, egy Long Island-i székhelyű brókercég, amelyet Jordan Belfort alapított 1989-ben. A vállalat elsősorban alacsony áron beszerezhető, tőzsdén kívüli „centes részvényekkel" *(penny stock)* kereskedett, és az 1990-es években már a legnagyobb tőzsdén kívüli részvények kereskedésére szakosodott vállalat volt az USÁ-ban.[6]

---

2 47 U.S.C. § 230 (1996).

3 Például David S. Ardia: Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity under Section 230 of the Communications Decency Act. *Loyola of Los Angeles Law Review,* Vol. 43. (2010) 373–506.; Benjamin Volpe: From Innovation to Abuse: Does the Internet Still Need Section 230 Immunity? *Catholic University Law Review,* Vol. 68. (2019) 597–624.; vagy éppen Jeff Kosseff: A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not). *Berkeley Technology Law Journal,* Vol 37., No. 2. (2022) 757–802. https://doi.org/10.15779/Z38VT1GQ97

4 Jeff Kosseff: *The Twenty-Six Words That Created the Internet*. Ithaca, NY, Cornell University Press, 2019.

5 Stratton Oakmont, Inc. v. Prodigy Services Co., 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

6 Matthew Partridge: Great frauds in history: Jordan Belfort and Stratton Oakmont. *MoneyWeek,* 2019. augusztus 7. https://tinyurl.com/3tv7wcy4

Az alperes egy korai internetes tartalomszolgáltató, a Prodigy Services Co.[7] volt. A cég fő profilja információs szolgáltatások nyújtása volt a feliratkozóik részére, például „hirdetőtáblák" *(bulletin boards)* segítségével. Ezeken harmadik felek információt cserélhettek egymással, nagyjából a mai értelemben vett internetes fórumokhoz hasonló módon. A cég olyan online szolgáltatóként hirdette magát, mint aki a puszta közzétételen túl szerkesztői ellenőrzést is gyakorol az általa üzemeltetett online platformokon megjelenő tartalmak felett. Geoffrey Moore, a Prodigy piaci programokért és kommunikációért felelős igazgatója a pert megelőzően több alkalommal is sajtónyilvánosság előtt hangoztatta, hogy ezáltal ők sokkal inkább a nyomtatott sajtóhoz hasonlítanak, semmint egy ellenőrizetlen online felülethez.[8]

A cég egyik igen népszerű fóruma a „Money Talk" volt, ahol a résztvevők pénzügyi kérdéseket vitathattak meg egymással, és amelyet mai szóhasználattal élve 'moderátorok' is felügyeltek. 1994-ben a fórumon aztán egy olyan bejegyzés jelent meg, amely a Stratton Oakmont-ot bűncselekményekkel és csalással vádolta. A Stratton Oakmont erre válaszul beperelte a Prodigy-t rágalmazásért. A per során a kulcskérdés az volt, hogy a Prodigy a megjelent információk kapcsán 'kiadónak' *(publisher)* vagy csak 'terjesztőnek' *(distributor)* minősült-e. A kérdés azért bírt különös relevanciával, hiszen az USA-beli esetjog hagyományosan erősen elkülöníti egymástól az e két szerepkörrel járó felelősséget.

E szerint a megkülönböztetés szerint egy *kiadótól* elvárható, hogy tudatában legyen az általa közzétett anyag tartalmának, jellegének, valódiságának vagy valóságtartalmának, így tehát felelősségre vonható az általa közzétett jogellenes tartalmakért. Ezzel szemben egy *terjesztőnek* valószínűleg nincs tudomása az általa elérhetővé tett anyagok tartalmáról, így pedig mentesülhet az azzal összefüggő jogi felelősség alól is.[9]

Ahogyan például a Doe v. America Online esetben, a Zeran v. America Online esetből idézve az Egyesült Államok Negyedik Kerületi Fellebbviteli Bírósága ekkor megállapította, a rágalmazási perekben alapvető elem a 'publikálás' (~kiadás) fogalma, ami azt jelenti, hogy a közölt anyagok tartalmáért csak az vonható felelősségre, aki azokat publikálja. A publikálás itt nem csak az információk közzétételére vonatkozik, hanem a rágalmazó állítás gondatlan közlésére vagy a mások által elsőként közölt állítás eltávolításának elmulasztására is.[10]

A Prodigy szempontjából a döntés végül hátrányosan alakult, hiszen a bíróság a cég fentebb említett nyilatkozatai, és az alkalmazott moderálási elvei miatt *kiadóként* tekintett rá, ennek megfelelően pedig felelősnek is tekintette az általa gondozott oldalakon megjelent állításokért. Ez egyébként részben szembement egy korábbi, 1991-es

---

7 Prodigy Communications Corporation History. FundingUniverse, én. https://tinyurl.com/mr3zs69v

8 Stratton Oakmont, Inc. v. Prodigy Services Co., No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995), opinion by Justice Stuart L. Ain.

9 Richard J. Hunter Jr. − Hector R. Lozada − John H. Shannon: Distributor vs. Publisher vs. Provider: That Is the High-Tech Question: But is an Extension of Liability the Answer? *International Journal of Education and Social Science,* Vol. 8., No. 1. (2021) 28–36.

10 Doe v. America Online, Inc. (2001). 783 So. 2d 1010 (Florida Supreme Court); Zeran v. America Online, Inc. (1997). 129 F. 3d 327 (United States Court of Appeals, 4th Circuit).

döntéssel, amikor is a bíróság a CompuServe-et, mint egy oldal üzemeltetőjét csak mint 'terjesztőt' ismerte el.[11]

Az eset kapcsán szárnyra kaptak olyan álláspontok, hogy a Stratton Oakmont kapcsán hozott döntés inkább visszalépés a korábbi gyakorlathoz képest, hiszen ezen állásfoglalásával a bíróság arra biztatja az internetes platformok üzemeltetőit, hogy semmilyen tartalommoderálást ne végezzenek, ezzel elhárítva mindennemű felelősséget az USA-beli jog alapján. Voltaképpen ennek az ellentmondásnak a feloldása lett az, amely később a manapság is híres/hírhedt CDA 230. szakaszában öltött testet.

A *Telecommunications Act of 1996*[12] részeként elfogadott *Communications Decency Act* (CDA)[13] biztosítja a platformok számára a jogi védelmet a felhasználói tartalommal kapcsolatos felelősség alól, miközben lehetővé teszi számukra a tartalom moderálását is. A CDA 230. szakasza kimondja, hogy az online szolgáltatók nem felelősek a felhasználók által közzétett tartalomért,[14] és biztosítja számukra az immunitást a moderációs döntéseikért, feltéve, hogy jóhiszeműen jártak el egy-egy károsnak ítélt tartalom eltávolítása során.[15] E védelem azonban nem terjed ki minden esetre: a szerzői jogi igényeket a *Digital Millennium Copyright Act* (DMCA), a szexuális szolgáltatásokkal kapcsolatos tartalmakat pedig a FOSTA/SESTA törvénycsomag kifejezetten kivonja a 230. § oltalma alól.

Egészen pontosan a 230 (c)(1) bekezdése a következők szerint fogalmaz:

> Egyetlen interaktív számítógépes szolgáltatás szolgáltatója vagy felhasználója sem tekinthető egy másik információs tartalomszolgáltató által nyújtott információ kiadójának *(publisher)* vagy előadójának.

Az érem másik oldalát közvetlenül ezután a (2) bekezdés tárgyalja:

> Egyetlen szolgáltató vagy interaktív számítógépes szolgáltatás felhasználója sem vonható felelősségre az alábbiak miatt:
> (A) bármely, a szolgáltató vagy felhasználó által obszcénnek, kéjesnek, bujának, mocskosnak, túlzottan erőszakosnak, zaklatónak vagy más módon kifogásolhatónak tartott anyaghoz való *hozzáférés* vagy *ilyen tartalom elérhetőségének korlátozása érdekében jóhiszeműen önkéntesen tett intézkedésért, függetlenül attól, hogy az ilyen anyag alkotmányos védelem alatt áll-e vagy sem*; vagy
> (B) bármely olyan intézkedés, amely lehetővé teszi vagy elérhetővé teszi az információs tartalomszolgáltatók, vagy mások számára az (1) bekezdés-

[11] Cubby, Inc. v. CompuServe Inc., 776 F. Supp. 135 (S.D.N.Y. 1991).

[12] Telecommunications Act of 1996, Pub. L. No. 104-104, 110 Stat. 56 (1996).

[13] Ruth Ann Strickland: Telekommunications Act of 1996 (1996). *Free Speech Center,* January 1, 2009. https://firstamendment.mtsu.edu/article/telecommunications-act-of-1996/

[14] Anthony Ciolli: Chilling Effects: The Communications Decency Act and the Online Marketplace of Ideas. *University of Miami Law Review,* Vol. 63. (2008) 137. https://doi.org/10.2139/ssrn.1101910

[15] Richard J. (2021) i. m. 30–32.

ben leírt anyagokhoz való hozzáférés korlátozására szolgáló technikai eszközöket.

Ahogyan azt a Doe v. GTE Corp. ügyben[16] a bíróság kifejtette, a 230. szakasz (c)(1) és (2) bekezdés közötti különbség az, hogy az első „megakadályozza a polgári jogi felelősségre vonást, ha a webtárhelyek és más internetszolgáltatók (ISP-k) tartózkodnak az oldalukon található információk szűrésétől vagy cenzúrázásától", míg az utóbbi biztosítja, hogy a „sértő anyagokat kiszűrő szolgáltató ne legyen felelősségre vonható a cenzúrázott ügyfél által".[17]

Egy másik eltérés, hogy a 230. (c)(2) szakasz védelme csak a „jóhiszeműen tett" intézkedésekre vonatkozik. A „jóhiszeműség követelménye" nem létezik a 230. (c)(1) szakasz tágabb értelemben vett rendelkezéseiben. A bíróságok kifejtették, hogy ez a követelmény a kifogásolható tartalmakat eltávolító vagy az azokat tévesen el nem távolító online szolgáltatások védelmére szolgál, anélkül, hogy azokat is védelemben részesítené, akik versenyellenes vagy más rosszindulatú szándékból (visszaélésszerűen) távolítanak el tartalmakat.[18]

A fentiek miatt végül a CDA 230. szakasza lett az internetes szólásszabadság védőbástyája, és egyben Achilles-sarka is. Azáltal, hogy a rendelkezés megvédi az online szolgáltatásokat a harmadik fél által készített tartalmakért való felelősségre vonástól, egyben meg is nyitotta az utat a különböző, ma már széles körben elterjedt üzleti modellek előtt, amelyek a felhasználók által generált tartalmakra támaszkodnak, átalakítva az ekkor még csak formálódóban levő online gazdaságot is. Lényegében ez tette lehetővé, hogy olyan tudásmegosztó oldalak, mint a Wikipédia, szabadon gyakorolhassanak tartalommoderálást a felhasználói bejegyzések felett. De ugyanígy ez teremtett lehetőséget például a kis webshopok üzemeltetői számára, hogy a forgalmazott termékeikről szóló vásárlói véleményeket tehessenek közzé a gyártók retorzióitól való félelem nélkül.

3. Az automatikus moderáláshoz vezető út

A CDA 230. szakasza által nyújtott szabadság tehát hatalmas üzleti lehetőségeket nyitott, egyúttal azonban számos probléma kialakulásának is megágyazott. Az egyik sarkalatos pont például, amely körül a viták az USÁ-ban újra és újra fellobbannak, a terrorizmussal kapcsolatos, pontosabban a terrorista szervezetek által közzétett on-

---

16 Doe v. GTE Corp., 347 F.3d 655 (7th Cir. 2003).

17 A terület hazai irodalmával kapcsolatban lásd különösen: Imre Melinda: Az internet-szolgáltatók felelősségének szabályozása a szerzői jogot sértő tartalmak tekintetében – Az amerikai, a közösségi és a magyar szabályozás bemutatása. *Iustum Aequum Salutare,* 2006/1–2. 213–227.; Muraközi Gergely: A szerzői jog és az internet – Az internet technikai megvalósítása a szerzői jog tükrében. *Jogi Fórum,* én. https://www.jogiforum.hu/files/publikaciok/drMurakozi-A_szerzoi_jog_es_az_internet(jf).pdf; Gosztonyi Gergely: A platformszolgáltatók felelősségének új szabályozása az európai uniós digitális szolgáltatásokról szóló rendelet alapján. *Pro Futuro,* 2023/3. 3–26.

18 Ash Johnson − Daniel Castro: Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved. *Information Technology & Innovation Foundation,* 2021. február 22. https://tinyurl.com/mr2w953n

line tartalmak kezelése. A leggyakoribb téma ennek kapcsán a szólásszabadság és a nemzetbiztonság közötti egyensúly kérdése, amely napjainkban is élénk jogi és társadalmi viták tárgyát képezi.[19] Az Egyesült Államok alkotmánya, különösen az Első Alkotmánykiegészítés[20] garantálja a szólásszabadságot, amely magában foglalja az online tartalmak szabad megosztását is. Ugyanakkor, ha ezek a tartalmak terrorizmust támogató vagy erőszakra buzdító üzeneteket tartalmaznak, a hatóságok és a közvélemény egy része azt érzi, hogy a nemzetbiztonsági kockázatok miatt szükséges ezen tartalmak kiemelten hatékony korlátozása vagy eltávolítása. A fő kérdés ennek kapcsán a moderálás hatékonyságának állami kikényszeríthetősége alapvetően magántulajdonban levő vállalatokra vonatkozóan.

A helyzetet bonyolítja, hogy az Egyesült Államokban a kezdeti időszakban a kormányzat az újra és újra kiélesedő viták ellenére elsősorban nem közvetlen jogszabályi úton kívánta rendezni az olyan kérdéseket, amelyek hosszabb távon problémásnak bizonyultak.[21] Ennek hátterében nem kizárólag a közvélemény megosztottsága állt, hanem az is, hogy a platformok felhasználói bázisa, valamint az általuk kezelt tartalmak mennyisége ekkor még lényegesen kisebb volt, mint napjainkban. A jogalkotók ezért elsősorban kevésbé formális, együttműködésen alapuló megoldásokat részesítettek előnyben, például köz- és magánszféra közötti partnerségeket, figyelemfelkeltő kampányokat, egyetértési megállapodásokat, a rendelkezésre álló szakértelem megosztását, valamint a felelősségek elhatárolását az amerikai kormány és a nagy közösségimédia-vállalatok között. Ezt az időszakot a szakirodalom gyakran „társszabályozásként" *(co-regulation)* jellemzi, amely a szabályozói szándék és a piaci szereplők önszabályozása közötti egyensúly kialakítására törekedett.[22]

Persze emellett az sem elhanyagolható szempont, hogy például a nagy közösségimédia-platformok mindegyike a tőzsdén is bejegyzett vállalat. A céggel kapcsolatos negatív hírek sok esetben direkten megmutatkoznak a részvényárfolyamok csökkenésében is, ami pedig közvetetten, a befektetőkön keresztül erős nyomást gyakorol az érintett cégek döntéshozóira is.[23] Mindezek mellé adódik még a szociális normák szerepe, vagyis, hogy sok esetben maga a szolgáltatást használó közösség az, aki értékrendjét az online platformokon megjelenő tartalmakban is szeretné viszont látni, illetve ellenkező esetben ezt az értéktranszfert az üzemeltető irányában kikényszeríteni.

Le kell azt is szögezni, hogy a közösségimédia-szolgáltatók is felelősek az olyan tartalmak korlátozásáért, amelyek törvénybe ütköznek, például éppen az *Anti-Terrorism*

---

19 Gosztonyi Gergely – Lendvai Gergely Ferenc: Twitter kontra Taamneh és Gonzalez kontra Google, avagy ki a felelős az online platformokra feltöltött tartalomért? *Magyar Jog,* 2023/10. 587–593.

20 U.S. Const. amend. I.

21 Alexander Tsesis: Social Media Accountability for Terrorist Propaganda. *Fordham Law Review,* Vol. 86. (2017) 605–631.

22 Koltay András – Nyakas Levente (szerk.): *Magyar és európai médiajog*. Budapest, Wolters Kluwer, 2017.; Lawrence Lessig: What Things Regulate Speech. In*: Code: And Other Laws of Cyberspace*. New York, Basic Books, 1999. 86–88.

23 Nicole Softness: Terrorist Communications: Are Facebook, Twitter, and Google Responsible for the Islamic State's Actions? *SIPA Journal of International Affairs,* Vol. 70. No. 1. (2017) 201–215.

Act[24] hatálya alá tartozó esetekben, vagy szerzői jogot sértő tartalmak esetében. Részben emiatt, részben pedig a törvényileg nem szabályozott, de a társadalom értékrendjét esetlegesen sértő esetek szabályozása miatt a legtöbb platform a felhasználási feltételeiben (*Terms of Service,* a továbbiakban: ToS) rendelkezik arról, hogy milyen tartalmakat töröl, vagy mely esetekben szankcionálja a tartalom közzétevőjét. Ez a szankció a legtöbb esetben az érintett tartalom eltávolításán keresztül valósul meg, amelyet a szolgáltatók – például a folyamatos társadalmi nyomásra reagálva – igyekeznek hatékonyan, és legfőképpen, gyorsan végrehajtani. Ugyanakkor nem kizárólag a tartalom törlése jelenthet szankciót: előfordulhat a felhasználói fiók ideiglenes vagy végleges felfüggesztése, a tartalom láthatóságának korlátozása (pl. *shadow banning*), földrajzi alapú korlátozása *(geoblocking)*, vagy az elérés algoritmikus háttérbe szorítása is. Ezek a módszerek gyakran kevésbé átláthatók, és a felhasználók számára nehezebben követhetők nyomon, ami külön kérdéseket vet fel az elszámoltathatóság és a transzparencia kapcsán.

Az elmúlt éveket jellemzi, hogy az online tartalmakra vonatkozó szabályozások fokozatosan egyre inkább szigorodtak. A ma legnépszerűbb közösségimédia-platformok mögött álló szolgáltatók többsége az USÁ-ban van bejegyezve, ugyanakkor, tekintettel arra, hogy tevékenységük a világ számos országára kiterjed, így érvényesek rájuk például az Európai Unió területén annak vonatkozó jogszabályai is. Ezek közül talán a legfontosabb az EU digitális szolgáltatásokról szóló jogszabálya (*Digital Services Act,* DSA)[25], amely 2024. február 17-én lépett hatályba teljeskörűen. Ennek értelmében új kötelezettségek hárulnak azokra az online platformokra, amelyeknek az EU-ban is vannak felhasználói, azzal a céllal, hogy a felhasználók, illetve a felhasználók jogainak védelme hatékonyabbá válhasson.[26] A DSA rendelkezései több tucatnyi új előírást tartalmaznak, ám a tartalommoderálás és felhasználói jogvédelem szempontjából négy olyan pillér emelkedik ki, amely közvetlenül alakítja a platformok és a felhasználók mindennapi viszonyát.

– *„Notice-flag"* rendszer: ez az első, gyakorlatban már rövid távon érezhető újítás, mert kifejezetten a moderációs folyamat elejét – a problémás tartalom bejelentését – szabványosítja, így a felhasználók ténylegesen élhetnek a jogsértő bejegyzések gyors jelzésének lehetőségével.
– Kiskorúaknak célzott hirdetések tilalma: tiltja a kiskorúak profilalkotáson vagy személyes adataikon alapuló hirdetésekkel való megcélzását (pl. mikrotargetálás[27]). A DSA ezzel a legvédtelenebb csoportot célozza, ahol a profilalkotásból fakadó károk – például egészség vagy testképtorzító üzenetek – a legnagyobbak.

---

[24] Anti-Terrorism Act (ATA), 18 U.S.C. § 2333.

[25] Az Európai Parlament és a Tanács 2022/2065/EU rendelete (2022. október 19.) a digitális szolgáltatásokról és a 2000/31/EK irányelv módosításáról (digitális szolgáltatásokról szóló rendelet), PE/30/2022/REV/1, HL L 277., 2022.10.27., 1–102. o.

[26] New rules to protect your rights and activity online in the EU. European Commission, 2024. február 16. https://tinyurl.com/yc44sre3

[27] Anja Prummer: Micro-targeting and polarization. *Journal of Public Economics,* Vol. 188. (2020) 104210. https://doi.org/10.1016/j.jpubeco.2020.104210

– Reklám-átláthatósági kötelezettség: a platform-gazdaság üzleti modellje hirdetés-vezérelt, így az információs aszimmetria csökkentése (miért pont ezt látom, ki fizetett érte?) kulcslépés a felhasználói autonómia megerősítéséhez.
– Érzékeny adatokon alapuló célzás teljes tiltása: teljes egészében betiltja az olyan hirdetéseket, amelyek érzékeny adatok, például politikai vagy vallási meggyőződés, szexuális preferenciák stb. alapján célozzák meg felhasználók egy csoportját.

A DSA hatálybalépése után a szabályok, amelyek 2023 óta már eddig is vonatkoztak néhány kiugróan nagy platformra és keresőmotorra, immár minden platformra és tárhelyszolgáltatásra érvényesek. Nem meglepő tehát, hogy különösen a közösségimédia szolgáltatók esetében, amelyek profitjuk jelentős részét adatkereskedelemből és hirdetések értékesítéséből nyerik, kiemelten fontossá vált ezek betartása, illetve betartatása.[28]

Ahhoz, hogy a jogsértő tartalmakat gyorsan és hatékonyan szűrhessék ki, a platformok hagyományosan moderátorokat alkalmaznak, vagyis olyan embereket, akik feladata az oldalakra kikerülő tartalmak ellenőrzése, a beérkező bejelentések felülvizsgálata, illetőleg a szükséges szankciók kiszabása volt. Különböző algoritmikus megoldások is léteztek már egészen a kezdetektől, de a domináns irányvonal sokáig a *human-in-the-loop* megközelítés volt. Ez azt jelentette, hogy a legtöbb esetben a döntések legalább minimális emberi felülvizsgálat után születhettek csak meg.[29] Az utóbbi években azonban az automatizált előszűrés került túlsúlyba, amit a *Digital Services Act* (DSA) is elismer, ugyanakkor egyértelművé tesz, hogy az eltávolítási vagy korlátozási döntések nem alapulhatnak kizárólag automatizált megoldásokra. A szabályozás további újdonsága az átláthatóság követelménye. Ennek értelmében a legnagyobb szolgáltatóknak éves jelentésekben kell feltárniuk moderációs gyakorlataikat, és kutatók számára hozzáférést biztosítaniuk a releváns adatokhoz. Bár az első egységes DSA-jelentések csak 2025 tavaszán válnak elérhetővé, jelenleg a közölt információk inkább összesítő jellegűek, így a konkrét döntési logikák továbbra is nagyrészt rejtve maradnak.

A moderálási eljárást érő kiritkák már az automatizált módszerek bevezetése előtt is jellemzőek voltak. A New York University Stern, Centre for Business and Human Rights 2020 júniusában elkészített jelentése például a Facebookra, mint a legnagyobb közösségimédia platformra vonatkozóan az alkalmazott moderátorok kiszervezését, és a moderátorok számának elégtelenségét kritizálja.[30] A moderálás miként aztán a COVID

---

[28] Üveges István: A mesterséges intelligencia közösségi médiában történő alkalmazásának társadalmi és politikai következményei. In: Kovács Zoltán (szerk.): *A mesterséges intelligencia és egyéb felforgató technológiák hatásainak átfogó vizsgálata.* Budapest, Katonai Nemzetbiztonsági Szolgálat, 2023. 301–327.

[29] Tom Lymn – Jessica Bancroft: The use of algorithms in the content moderation process. *Responsible Technology Adoption Unit Blog,* 2021. augusztus 5. https://tinyurl.com/yt9x75a6

[30] Paul M. Barrett: Who Moderates the Social Media Giants? A Call to End Outsourcing. NYU Stern Center for Business and Human Rights. *NYU STERN,* 2020 június. https://tinyurl.com/4t2zuwfs

hatására drasztikusan megváltozott.[31] A pandémia alatt a közösségimédia platformok jelentős mértékben támaszkodtak algoritmikus tartalom moderációra, mivel a járvány következtében az emberi moderátorok munkája erősen korlátozottá vált. Az automatizált rendszerek bevezetése ekkoriban igencsak vegyes eredményeket hozott. Noha hatékonyak voltak a nagy mennyiségű tartalom kezelésében, az olyan problémákkal, mint a kontextus megértése és a különböző nyelveken való moderáció nehézségei, nem tudtak hatékonyan megbirkózni. A dezinformáció elleni küzdelem során a platformok gyakran szembesültek a politikai és tudományos viták bonyolultságával is, amely sokszor jelentősen megnehezíti a moderációs döntéseket. A tapasztalatok egyértelműen rávilágítottak az algoritmikus rendszerek korlátaira és az emberi felügyelet fontosságára.

A tanulmány eddigi részében a moderálás fogalma a felhasználói tartalmak feletti döntéshozatalra korlátozódott. Egyes szerzők meghatározásában azonban a *tartalommoderáció* ennél kiterjesztőbb értelmezést nyer, olyan irányítási mechanizmusok gyűjtőneveként, amelyek strukturálják a közösségben való részvételt az együttműködés megkönnyítése és a visszaélések megelőzése érdekében.[32] Fontos, hogy ebben az értelmezésben tehát a moderációs tevékenység részei természetesen a tartalom eltávolításáért felelős adminisztrátorok vagy moderátorok is, de ugyanígy beletartoznak azok a tervezési döntések is, amelyek alapjaiban megszervezik és befolyásolják, hogy egy adott közösség tagjai hogyan lépnek/léphetnek kapcsolatba egymással. Egy ilyen tervezési döntés lehet a moderálási folyamat mind inkább automatizmusokra történő felépítése is.[33]

4. Az algoritmikus tartalom moderáció formái

Annak felderítése, hogy az egyes közösségimédia-platformok moderációs tevékenysége mögött pontosan milyen technikai megoldások húzódnak meg, közel sem triviális feladat. Ennek oka főként az, hogy kevés a nyilvánosan elérhető, megbízható információ a témában, hiszen az érintett vállalatok ezeket a legtöbb esetben üzleti titokként kezelik. Éppen ezért a helyzet feltáráshoz elengedhetetlen például a nyilvánosan elérhető jelentések, vagy éppen az olyan nem konvencionális források használata, mint az oknyomozó újságírás. Ezek felhasználása egyébként a vonatkozó szakirodalomban sem példa nélküli.[34]

Az algoritmikus tartalommoderálás a közösségi média platformokon olyan technikák és eljárások összessége, amelyek célja a felhasználók által generált tartalom automatikus szűrése és felügyelete. Voltaképpen a már ismertetett *moderálási folyamat* au-

---

[31] Marc Faddoul: COVID-19 is triggering a massive experiment in algorithmic content moderation. *Brookings,* 2020. április 28. https://tinyurl.com/2p732haw

[32] James Grimmelmann: The Virtues of Moderation. *Yale Journal of Law & Technology,* Vol. 17. (2015) 42–109. https://doi.org/10.31228/osf.io/qwxf5

[33] Gergely Gosztonyi: Human and Technical Aspects of Content Regulation. *Erdélyi Jogélet,* Vol. 2., No. 4. (2022) 7–17.

[34] Robert Gorwa – Reuben Binns – Christian Katzenbach: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society,* Vol. 7. No. 1. (2020) https://doi.org/10.1177/2053951719897945

tomatizálása ez, amelynek során a döntések adatbázisokkal való teljes, vagy részleges egyezések, esetleg gépi tanuló modellek által végzett osztályozás alapján történnek.

Napjainkban a legnagyobb közösségimédia platformok felhasználói bázisa bőven a milliárdos nagyságrendben van,[35] ezen felhasználók pedig ugyanekkora léptékben állítanak elő tartalmakat minden nap. Nem meglepő tehát, hogy a kizárólag moderátorok által végzett, vagy minden esetben általuk felülvizsgált ellenőrzési modell közelít a fenntarthatatlanhoz. A tartalommoderálás automatizálása ezzel szemben egy praktikusnak tűnő, jól skálázható megoldást jelent a közösségimédia-platformok számára.[36]

Az alkalmazott megoldások rendkívül széles skálán mozognak mind összetettségben, mind pedig hatékonysági szempontok szerint. Közöttük talán a legfontosabb két kategória, a *hash* alapon működő, valamint a gépi tanulásra alapozott megoldások csoportja. Léteznek persze ennél történetileg régebbi, például kulcsszó alapon kereső eljárások is, de ezek képességei mai szemmel már annyira korlátozottak, hogy néhány speciális alkalmazási esetet leszámítva relevanciájukat is szinte teljes mértékben elvesztették.

2019-ben a Facebook nyilvánosan elérhetővé tette a PDQ, valamint a TMK + PDQF nevű, *hash* alapon működő megoldásait,[37] amelyek forráskódja mind a mai napig nyilvánosan elérhető.[38] A folyamat kulcsai a *hash* függvények. Ezek lényegében olyan algoritmusok, amelyek előállítják a fix hosszú reprezentációt olyan módon, hogy tetszőleges hosszúságú bitsort egy fix hosszúságúra képeznek le. Ilyen esetben kulcsfontosságú az, hogy amennyiben az eredeti bitsorozatban – például egy közösségimédia poszt szövegében – akár egyetlen karakter is megváltozik, akkor a belőle képzett, „hashelt" érték is módosuljon. Ezen felül azt is biztosítani kell, hogy nagyon alacsony legyen annak a valószínűsége, hogy két különböző szöveg hashelt értéke véletlenül megegyezzen.

Egy-egy tartalom *hash*ének kiszámítása után arról el kell dönteni, hogy az eredeti szöveg, kép, videó stb. megvalósít-e valamiféle szabály- vagy törvénysértést. Ilyen szabálysértés lehet bármi, ami a platform használatakor elfogadott ToS szerint ellentétes a közösség irányelveivel. Törvénysértő tartalmakra jó példa szerzői jogilag védett tartalmak ismételt feltöltése a felhasználó által sajátként. Ennek megállapításához léteznek olyan *hash*-adatbázisok, amelyek ismerten problémás tartalmak *hash* értékeit tartalmazzák, és amelyeket az egyes szolgáltatók meg is oszthatnak egymással *(Shared Industry Hash Database).* Mivel a *hash* kiszámítása, és az ilyen jellegű adatbázisokban való keresés rendkívül gyors, ezért a módszer jól skálázódik még a közösségi médiára jellemző adatmennyiségek esetében is.

---

[35] Dave Chaffey: Global social media statistics research summary. *Smart Insights,* 2024. május 1. https://tinyurl.com/sm74a5za

[36] Joseph Seering − Tian Wang − Joon Youn − Geoff Kaufman: Moderator engagement and community development in the age of algorithms. *New Media & Society,* Vol. 21., No. 7. (2019) 1417–1443. https://doi.org/10.1177/1461444818821316

[37] Antigone Davis – Guy Rosen: Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer. *Meta*, 2019. augusztus 1. https://tinyurl.com/4p8e2d2f

[38] Az informatikában, főleg a kriptográfiában a hashelés egy olyan eljárást jelent, amely során egy változó méretű bemenetből – például szövegekből vagy képekből – egy jellemzően rövidebb, fix hosszúságú reprezentációt állítanak elő.

A problémát itt a pontos egyezés kérdése jelentheti. Elvégre, például egy kép esetében elég csak egyetlen pixel színét megváltoztatni, és a kapott *hash* érték teljesen más lesz, mint amire az adatbázisban való keresés találatot jelezne. Egy pixel eltéréstől azonban egy eredetileg jogsértő tartalom jó eséllyel nem lesz kevéssé jogsértő; a jogszabályok ráadásul a platformot sok esetben arra is kötelezik, hogy az „azonos vagy egyenértékű" (lényegileg azonos) változatokat is eltávolítsa, amint arról az Európai Unió Bírósága döntött a Glawischnig-Piesczek kontra Facebook ügyben (C-18/18, ECLI:EU:C:2019:821).

Erre kínálnak megoldást az olyan *hash* függvények és adatbázisok, amelyek nem csak a pontos egyezéseket képesek előre jelezni, hanem az ún. lényegi egyezéseket is. A perceptuális, avagy „észlelési" *hash* függvények különbözősége abban áll, hogy két olyan tartalom esetében, amelyet az emberek hasonlónak *észlelnek*, ott ezek perceptuális *hash* értéke is azonos, vagy legalábbis nagyon hasonló lesz.[39] A módszer gyakorlatilag az emberi észleléssel nagyjából azonosnak mutatkozó objektumokat (például képeket) hivatott ugyancsak azonosként, vagy *lényegileg azonosként* felismerni. Éppen ezért az ilyen algoritmusok alkalmasak arra, hogy felismerjék az olyan tartalmakat, amelyekre esetleg csak egy plusz vízjel került képek esetében, vagy detektálják, ha egy korábban gyűlöletkeltőnek talált szöveget valaki ismét feltölt néhány karakter megváltoztatásával. Valószínűsíthetően a Facebook fent említett megoldásai is ebbe a kategóriába tartoznak.[40]

Egy másik fontos irányvonal a gépi tanuló algoritmusok, illetve a velük készített *modellek* alkalmazása. Ilyen esetben az ellenőrizendő tartalmak esetében nem korábbi, problémásként jelölt esetekkel való teljes vagy részleges egyezések felfedése a cél, hanem soha nem látott példák feletti döntések meghozatala (*prediktív* megközelítés). Ezt előre címkézett tanítóadatok segítségével lehetséges megvalósítani, amely a *felügyelt gépi tanulás*[41] alapvető módszertana. Ilyenkor a gépi tanuló modell feladata az, hogy a tanítóadatokban látott mintázatok alapján olyan generalizálási képességre tegyen szert, amellyel tetszőleges új példákról jó hatékonysággal tudja eldönteni, hogy azok problémásak-e, vagy sem. Cél lehet ilyen esetben például az offenzív nyelvezet, a terrorizmust népszerűsítő, vagy éppen a pornográf tartalmak felismerése is. Utóbbira a Facebook is tett már dokumentált kísérletet.[42]

Emellett léteznek ma már olyan kutatások is, amelyek a legmodernebb szabadon elérhető neurális hálózatokat alapul véve keresnek automatizált megoldást moderálási

---

[39] Hany Farid: An Overview of Perceptual Hashing. *Journal of Online Trust and Safety,* Vol. 1., No. 1. (2021) https://doi.org/10.54501/jots.v1i1.24; Moses Datar– Nicole Immorlica – Piotr Indyk et al.: Locality-sensitive hashing scheme based on p-stable distributions. In: *Proceedings of the Twentieth Annual Symposium on Computational Geometry.* New York, NY, ACM, 2004. 253–262. https://doi.org/10.1145/997817.997857

[40] Davis–Rosen (2019) i. m.

[41] Vladimir Nasteski: An Overview of the Supervised Machine Learning Methods. *HORIZONS.B,* Vol. 4. (2017) 51–62. https://www.doi.org/10.56726/IRJMETS51366

[42] Tarleton Gillespie: *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT, Yale University Press, 2019. https://doi.org/10.12987/9780300235029

problémákra. Egy ilyen módszer lehet például a BERT modellek[43] betanítása a moderálással összefüggő feladatokra, például Reddit posztok kapcsán.[44] Ennek háttere az, hogy a Redditen számos közösség elkezdte a moderálási feladatokat és a feldolgozási logikát algoritmusokba beágyazni, ez pedig lehetőséget nyitott célzott tanítóadatok beszerzésére.[45] De ugyanígy ismerünk működő példákat a Wikipédia kapcsán is, ahol teljesen autonóm rendszerek ugyanúgy működnek, mint a moderátorok támogatására szolgálók.[46] Szintén a kutatási szférából ismert olyan kísérlet is, amely ugyancsak pornográf tartalmak szűrésére többlépéses, neurális hálózat alapú architektúra bevezetését javasolja.[47]

5. A szólásszabadság és az információs ökoszisztéma

A modern közösségimédia-platformok nem csupán technológiai vállalkozások, hanem az elmúlt évtizedben ezek váltak a társadalom alapvető kommunikációs infrastruktúrájának gerincévé is. Ezzel a szereppel azonban együtt jár az a képesség, hogy formálhatják és korlátozhatják a felhasználói interakciókat, diskurzusokat, és bizonyos esetekben a politikai részvételt vagy társadalmi szerepvállalást is.[48] Részben a leközölt tartalmak puszta volumene, részben pedig törvényi kötelezettségeik miatt a már említetteknek megfelelően ma már nem támaszkodhatnak pusztán moderátorok tevékenységére a náluk megjelenő tartalmak szűrése során. Tevékenységükben az automatizált (algoritmikus) megoldások súlya éppen ezért folyamatosan növekszik.[49]

---

[43] Jacob Devlin – Ming-Wei Chang – Kenton Lee – Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota, Association for Computational Linguistics, 2019. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[44] Qinglai He – Yili Hong – T. S. Raghu: Platform Governance with Algorithm-Based Content Moderation: An Empirical Study on Reddit. *Information Systems Research,* Vol. 36., No. 2. (2024) https://doi.org/10.1287/isre.2021.0036

[45] Eshwar Chandrasekharan – Savvas Mattia – Shagun Jhaver – Hannah Charvat – Amy Bruckman – Cliff Lampe – Jacob Eisenstein – Eric Gilbert: The internet's hidden rules: an empirical study of Reddit norm violations at micro, meso, and macro scales. In: *Proceedings of the ACM on Human-Computer Interaction, 2(CSCW)* (2018) Article 32. https://doi.org/10.1145/3274301

[46] Aaron Halfaker – John Riedl: Bots and cyborgs: Wikipedia's immune system. *Computer,* Vol. 45., No. 3. (2012) 79–82. https://doi.org/10.1109/MC.2012.82

[47] Dogus Karabulut – Cagri Ozcinar – Gholamreza Anbarjafari: Automatic content moderation on social media. *Multimedia Tools and Applications,* Vol. 82. (2023) 4439–4463. https://doi.org/10.1007/s11042-022-11968-3

[48] Sylvie Delacroix: Beware of "algorithmic regulation". *SSRN Electronic Journal,* February 1, 2019. https://doi.org/10.2139/ssrn.3327191; Papp János Tamás: *A közösségi média szabályozása a demokratikus nyilvánosság védelmében*. Budapest, Wolters Kluwer, 2022.

[49] Delacroix i. m. 5–9.

5.1. Szólásszabadság

Az olyan platformok, mint például a Facebook vagy a Google tehát tagadhatatlanul jelentős hatalmat gyakorolnak a nyilvános kommunikáció felett. Ez a hatalom főként abban nyilvánul meg, hogy a platformok dönthetnek arról, milyen tartalmakat jelenítenek meg, milyen algoritmusokkal rendezik ezeket, és hogy sok esetben a platform mögött álló cégek döntik el, mi minősül elfogadhatónak vagy elfogadhatatlannak. Ezzel azonban szükségszerűen befolyásolják azt is, hogy milyen információk érhetők el a társadalom egésze számára, mely szereplők hallathatják a hangjukat a nyilvánosságban, illetve, hogy a szólásszabadság mikor és milyen mértékben korlátozódik.

Ez a fajta befolyás nem pusztán technikai vagy üzleti jellegű, hanem jelentős politikai következményei is vannak. A közösségimédia-platformok hatalma a kommunikáció felett azt jelenti, hogy ezek a cégek szabályozhatják a politikai diskurzust, elősegíthetik vagy akadályozhatják a politikai mozgalmakat, és akár közvetlenül befolyásolhatják a választási eredményeket is azáltal, hogy bizonyos információkat előnyben részesítenek – azaz több felhasználóhoz juttatnak el –, másokat pedig háttérbe szorítanak.[50]

Fontos megjegyezni, hogy az ilyen platformok döntései kereskedelmi megfontolások alapján is történhetnek, amelyek gyakran nem esnek egybe a közérdekkel vagy a felhasználók érdekeivel. Ez felveti a kérdést, hogy szükséges lenne-e velük kapcsolatban szigorúbb szabályozásokat alkalmazni. Különösen fontos ez akkor, amikor az egy-egy ilyen szereplő által gyakorolt hatalom olyan mértékű, amelynek birtokában képesek alapvetően megváltoztatni a társadalmi kommunikáció dinamikáját. Ilyen típusú hatalommal a digitalizált világ elterjedése előtt *csakis államok, illetve egyházak rendelkeztek*. Jelenleg azonban a véleményformálásnak ezt a minden korábbinál hatékonyabb és átfogóbb képességét mindössze néhány nagy technológiai cég birtokolja, amelyek felhasználói bázisa országhatárokon túlnyúlva, lényegében az egész modern világra kiterjed.

A szólásszabadságra az algoritmusok esetleges tévedései is komoly hatással vannak. A gépi tanulási algoritmusok nagy mennyiségű adaton tanulnak, és céljuk, hogy felismerjék a mintázatokat, amelyek segítségével eldönthetik, hogy egy adott tartalom megfelel-e egy-egy platform szabályainak. Ilyenkor elkerülhetetlenül előfordulnak *fals pozitív*, és *fals negatív* esetek is. A szólásszabadságra gyakorolt hatás szempontjából a fals pozitív találatok különösen károsak lehetnek. Ilyen esetben az algoritmus tévesen távolít el egy egyébként ártalmatlan tartalmat, ezzel ok nélkül korlátozhatva a felhasználók véleménynyilvánítási jogát. Az ilyen típusú hibák különösen aggasztóak lehetnek a politikai vagy társadalmi diskurzusokban, ahol a szigorúbb moderálás a kritikus vélemények elhallgattatásához vezethet. Például, ha egy algoritmus valós ok nélkül töröl egy politikai tiltakozással kapcsolatos bejegyzést, azzal egyfajta cenzorként működik, és hozzájárul a véleménynyilvánítás korlátozásához.[51]

---

[50] Tomas Apodaca – Natasha Uzcátegui-Liggett: How Automated Content Moderation Works (Even When It Doesn't). *The Markup,* 2024. március 2. https://tinyurl.com/f6f8j9v6

[51] Emma Llansó – Joris van Hoboken – Paddy Leerssen – Jaron Harambam: Artificial Intelligence, Content Moderation, and Freedom of Expression. *Transatlantic Working Group on Content Moderation Online*

Ezzel éppen ellentétesen, a fals negatív találatok egy eltérő kockázatot jelentenek. Amikor az algoritmus nem ismeri fel a szabálysértő tartalmat, azzal a káros információk, például dezinformáció, gyűlöletbeszéd vagy a szélsőséges, destruktív ideológiák szabadon történő terjedéséhez járul hozzá. Ez közvetlen veszélyt jelenthet a felhasználókra, és súlyosan torzíthatja a közvéleményt. Ráadásul az ilyen hibák alááshatják a közösségi média platformok hitelességét és a felhasználók beléjük vetett bizalmát is. A szólásszabadságra gyakorolt hatás persze a tartalomajánló rendszereken keresztül is érvényesül.

5.2. Véleménybuborékok és politikai polarizáció

A szólásszabadság formális keretei mellett azt is vizsgálni kell, hogyan torzulhat a nyilvános diskurzus az algoritmusok tartalom-rangsorolási logikája következtében.

A felhasználók elé kerülő tartalmak moderálása összetett folyamat, amely jóval túlmutat azon az egyszerű binárison, hogy egy tartalmat megtartanak vagy törölnek. A moderálás mára egy skálaszerű rendszerként működik, ahol a szankciók és beavatkozások változatos formában jelenhetnek meg. A döntés szólhat tartalomeltávolításról, de emellett számos olyan eszköz is létezik, amely láthatósági korlátozást alkalmaz. Ilyenek például a *shadow banning* (amikor a tartalom a közönség számára láthatatlan, de a közzétevő számára nem), az algoritmikus háttérbeszorítás *(downranking)*, a földrajzi alapú elérhetetlenné tétel *(geoblocking)*, vagy a személyes hírfolyamból való kiszorítás.

Emellett az is előfordulhat, hogy nem a tartalom, hanem a tartalomhoz kapcsolódó interakciók – például kommentelés vagy megosztás – kerülnek korlátozás alá, ami szintén befolyásolja annak elérhetőségét és társadalmi hatását. Ezek a moderációs gyakorlatok mind hozzájárulnak ahhoz, hogy milyen mértékben láthatók vagy észlelhetők egyes nézetek, és ez közvetlenül hat arra is, hogy a felhasználók milyen nézőpontokkal, politikai vagy kulturális álláspontokkal találkoznak online környezetükben.

Az algoritmikus tartalommoderálás és az ajánlási rendszerek együttes működése jelentős hatással lehet a társadalmi megosztottságra, különösen az úgynevezett visszhangkamrák *(echo chambers)* vagy szűrőbuborékok, esetleg véleménybuborékok *(filter bubbles, opinion bubbles)* kialakulásának elősegítésével.[52] Ez a több néven is hivatkozott jelenség olyankor fordul elő, amikor a felhasználók főként olyan tartalmakkal találkoznak, amelyek megerősítik meglévő nézeteiket, miközben lényegében nem találkoznak ellentétes véleményekkel. Ilyen helyzet legegyszerűbben úgy jöhet létre, hogy például a közösségimédia-platformok tartalomajánló rendszerei hajlamosak kizárólagosan és személyre szabottan az adott felhasználó érdeklődésének, preferenciáinak megfelelő tartalmakat mutatni számára. Megjegyzendő, hogy a felhasználói preferenciákkal ellentétes nézetek nem teljes egészében válnak láthatatlanná. Gyakoriságuk azonban jelentősen csökken a hírfolyamokban és gyakran kedvezőtlen érzelmi

and Freedom of Expression, 2020. február 26. https://tinyurl.com/4tcdtfpr

[52] Eli Pariser: *The Filter Bubble: What the Internet Is Hiding from You*. New York, Penguin Press, 2011.; Papp János Tamás: Ajánlórendszerek és szűrőbuborékok. In: Koltay András (szerk.): *A vadnyugat vége? Tanulmányok az Európai Unió platformszabályozásáról*. Budapest, Wolters Kluwer, 2024. 231–259.

keretezésben jutnak el a közönséghez. Más szavakkal: a buborék nem hermetikusan zár, hanem torzítja a találkozások arányát és módját.[53]

Ez a tartalom-rangsorolás aztán felerősíti egymást a megerősítési torzítással *(confirmation bias[54])* is. Ez arra a pszichológiai jelenségre utal, miszerint az emberek hajlamosak olyan információkat keresni, előnyben részesíteni és alapul venni, amelyek megerősítik a már elve meglévő vélekedéseiket, ellenben figyelmen kívül hagyni az ezzel ellentétes álláspontokat. Ez a kettő együttesen jelentősen befolyásolja a döntéshozatalt és a véleményalkotást is. A téves bizonyosság ezenfelül közvetlenül hozzájárul a társadalmi megosztottság fokozódásához is.

A véleménybuborékok[55] létezése és a politikai polarizáció felerősödése szorosan összefügg, hiszen előbbiek közvetlenül elősegítik a politikai nézetek szélsőségesebbé válását és a társadalmi megosztottság növekedését.[56] A véleménybuborékok következtében a politikai diskurzus egyre inkább polarizálódik, mivel a különböző politikai csoportok egymástól elszigetelődnek, és kevésbé nyitottak az ellentétes nézetek meghallgatására vagy megértésére. Emellett a véleménybuborékokban a nézetek hajlamosak a szélsőségek felé tolódni, ami egyúttal azt is jelenti, hogy a kompromisszum keresése egyre inkább kiveszik, helyét pedig a egyre radikálisabb nézetek propagálása veszi át.

E mögött persze az is fellelhető, hogy a közösségi médiában a megosztó (akár szélsőséges) tartalmak jellemzően több interakciót generálnak, mint a konszolidáltak. Az interakciók nagyobb figyelmet is eredményeznek, ami miatt pedig a figyelmi idő maximalizálásra törekvő vállalatoknak paradox módon érdekük, hogy az ilyen tartalmakat népszerűsítsék. Ennek hátterében az áll, hogy a figyelemalapú gazdaságban[57] a szolgáltatók célja, hogy a felhasználóknak a platformjukon töltött idejét maximalizálják.

6. Az automatizált moderáció közvetlen hatásai a felhasználókra

6.1. Dezinformáció terjedése

A buborékhatás következményeként jelentkezhet a dezinformációk erősödése is, különösen, ha az algoritmusok olyan tartalmakat részesítenek előnyben, amelyek fokozzák

---

53 Eli Pariser: Did Facebook's Big Study Kill My Filter Bubble Thesis? *Wired,* 2015. május 7. https://www.wired.com/2015/05/did-facebooks-big-study-kill-my-filter-bubble-thesis/

54 Joshua Klayman: Varieties of Confirmation Bias. *Psychology of Learning and Motivation,* Vol. 32. (1995) 385–418. https://doi.org/10.1016/S0079-7421(08)60315-1

55 A. K. Saxena: Beyond the Filter Bubble: A Critical Examination of Search Personalization and Information Ecosystems. *International Journal of Intelligent Automation and Computing,* Vol. 2., No. 1. (2019) 52–63.

56 Emily Kubin – Christian von Sikorski: The Role of (Social) Media in Political Polarization: A Systematic Review. *Annals of the International Communication Association,* Vol. 45., No. 3. (2021) 188–206. https://doi.org/10.1080/23808985.2021.1976070; David Hartmann – Sonja Wang – Lena Pohlmann – Bettina Berendt: A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. *Journal of Computational Social Science,* Vol. 8., Article n. 52. (2025) https://doi.org/10.1007/s42001-025-00381-z

57 Thomas H. Davenport – John C. Beck: The Attention Economy. *Ubiquity,* May 2001. https://doi.org/10.1145/376625.376626

a figyelmet vagy felháborodást. A dezinformáció egyik kitüntetett esete a hamis hírek *(fake news)*, valamit a mostanában feltörekvőben lévő – a generatív mesterséges intelligenciára is nagyban támaszkodó – *deepfake*-ek terjedése. Az *álhírek* olyan tudatosan terjesztett félrevezető információk, amelyek megjelenhetnek gazdasági haszonszerzés, kattintásvadászat, szórakoztató vagy éppen szándékolt károkozás *(hoax)* motivációjával is, miközben a politikai vagy geopolitikai manipulációt célzó dezinformáció egy különösen veszélyes formájuknak számít.[58] A politikai dezinformáció szándékos terjesztése a pszichológiai hadviselés egyik eszköze is lehet. A mindennapi beszédben álhírként emlegetünk minden olyan hamis információt, amely széles körben eljut a nyilvánossághoz. A *deepfake* egy olyan technológia, amely mesterséges intelligencia, konkrétabban gépi tanulás segítségével képes hamisított képeket, videókat vagy hangfelvételeket készíteni, amelyek azonban valósághűnek tűnnek.[59]

A tartalommoderálás, és az erre alkalmazott algoritmusok egyik fő célja éppen az ilyen és ehhez hasonló dezinformáció terjedésének megakadályozása. Gyakran előfordul azonban, hogy ezek az automatizált rendszerek nem képesek időben felismerni vagy eltávolítani a hamis(ított) információkat. Ez különösen problematikus a gyorsan terjedő tartalmak esetében *(viral content)*, mivel az algoritmusok, amelyek az elköteleződés (lájkok, megosztások, kommentek) alapján rangsorolják a tartalmakat, hajlamosak az ilyen típusú információkat gyorsan és széles körben terjeszteni. Ezért szélsőséges esetben még azelőtt felhasználók tömegeihez juthat el egy-egy problémás hír, vagy videó, hogy a vele kapcsolatos moderálási döntés megszületne. Ennek hátterében az áll, hogy noha a moderálás maga nagyrészt ma már automatizált, még mindig nagyban támaszkodik a felhasználóktól érkező bejelentésekre is, amelyekhez azonban sokszor idő kell. Automatikus törlés vagy korlátozás hiányában pedig a dezinformáció sokkal gyorsabban elérheti a közönséget, mint ahogy a moderátorok beavatkozhatnának a folyamatba.

A mesterséges intelligencia rendszerek szabályozását rendező uniós rendelet *(AI Act)* kimondja,[60] hogy a szintetikusan létrehozott vagy *„deepfake"* módon jelentősen módosított kép-, hang-, videó- és szövegtartalmakat egyértelmű, jól látható jelöléssel kell ellátni, hogy a felhasználók tisztában legyenek a látott, hallott vagy olvasott információ valóságtartalmával. Előírja továbbá, hogy a jelölés „géppel olvasható", azaz automatikusan detektálható módon is megtörténjen. Ezt a gyakorlatban érvényesíteni viszont közel sem triviális feladat. Különösen, ha figyelembe vesszük, hogy a dezinformáció terjesztésében érdekelteknek magától értetődően nem érdeke az ilyesfajta együttműködés, inkább ennek megkerülése.

---

[58] Howard Tumber – Silvio Waisbord (szerk.): *The Routledge Companion to Media Disinformation and Populism. 1. edition*. London, Routledge, 2021.

[59] Anupama Chadha – Vaibhav Kumar – Sonu Kashyap – Mayank Gupta: Deepfake: An Overview. In*: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security.* Springer, Singapore, 2021. 557–566. https://doi.org/10.1007/978-981-16-0733-2_39; Aczél Petra – Veszelszki Ágnes (szerk.): *Deepfake: a valótlan valóság.* Budapest, Gondolat, 2023.; Gosztonyi Gergely – Lendvai Gergely: Deepfake és dezinformáció – Mit tehet a jog a mélyhamisítással készített álhírek ellen? *Médiakutató,* 2024/1. 41–49. https://doi.org/10.55395/MK.2024.1.3

[60] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence *(AI Act)*, HL L 295, 2024.09.12., Art. 50 (2)–(4).

További probléma, hogy az algoritmusok gyakran nem képesek megfelelően értelmezni a tartalom kontextusát, ami azt eredményezheti, hogy a dezinformációval szembeni küzdelem során valós és releváns információk is eltávolításra kerülnek.[61]

A kontextusértelmezés persze nemcsak a gépi, hanem az emberi moderálásnak is tartós Achilles-sarka. A gyűlöletbeszéd-szabályozásban például éppen ezért született meg az ENSZ által támogatott Rabat-akcióterv *(Rabat Plan of Action)* hatpontos tesztje, amely a szándék, a tartalom, a terjesztés közege, a beszélő helyzete, a célcsoport kiszolgáltatottsága és a beszéd valószínű hatása alapján mérlegeli, mikor indokolt a korlátozás. A gyakorlat azonban azt mutatja, hogy még e keretrendszer alkalmazása mellett is gyakran előfordulnak téves eltávolítások, mert a kontextus (pl. idézés, művészi cél vagy ellenbeszéd) nem mindig dönthető el a rendelkezésre álló információkból. Ez jól illusztrálja: ha az emberi szakértők is vitatkoznak egy tartalom besorolásán, akkor a tisztán algoritmikus döntéshozás szükségszerűen még nagyobb hibaaránnyal működik.

Mindez különösen igaz az olyan helyzetekben, amikor a hamis információk és a valós tények között csak árnyalatnyi különbségek vannak, vagy ahol szarkazmus, irónia, vagy kulturálisan meghatározott speciális kifejezések vagy értelmezési lehetőségek jelennek meg.

6.2. Egyenlő bánásmód és algoritmikus torzítás

A technológia fejlődése során nyilvánvalóvá vált, hogy az automatizált rendszerek gyakran diszkriminatívan járnak el, különösen a marginalizált közösségekkel szemben. Ezek a torzítások sok esetben az algoritmusok fejlesztéséhez használt tanítóadatokban gyökereznek, amelyek implicit módon magukban hordozhatják a társadalomban már meglévő előítéleteket és sztereotípiákat.

Bármilyen gépi tanulásra alapozott megoldás esetében igaz, hogy az algoritmus hatékonysága nagymértékben függ a tanítóadatoktól. Ha ezek az adatok torzítanak *(biased)* vagy hiányosak, az algoritmusok is ezekkel a hibákkal fognak működni. Az adatokban jelenlévő torzítások lehetnek explicitek és implicitek is, mely utóbbiak korrigálása sokkal körülményesebb, ha megoldható egyáltalán.[62] Például, ha egy moderálási algoritmus betanítása során aránytalanul sok esetben kerülnek eltávolításra egy adott kisebbséghez tartozó felhasználók bejegyzései, akkor az algoritmus később is nagyobb valószínűséggel fogja eltávolítani az ilyen bejegyzéseket, még akkor is, ha azok nem sértik a közösségi irányelveket. Ezzel szemben más közösségek hasonló tartalmai érintetlenek maradnak, ami pedig közvetlenül egyfajta egyenlőtlen bánásmódhoz vezet. A közelmúlt empirikus és áttekintő kutatásai szisztematikusan rámutatnak arra, hogy a gépi tanuláson alapuló rendszerek a fejlesztés minden szakaszában (az adatok gyűjtésétől a modellek finomhangolásáig) hordozhatnak és felnagyíthatnak társadalmi előítéleteket. 2022-ben az Európai Unió Alapjogi Ügynöksége (FRA) részletesen dokumentálta, hogy

---

61 Delacroix (2019) i. m. 3–6.

62 Ferry Hoitsma – Guillermo Nápoles – Çiğdem Güven et al.: Mitigating implicit and explicit bias in structured data without sacrificing accuracy in pattern classification. *AI & Society,* Vol. 40. (2024) https://doi.org/10.1007/s00146-024-02003-0

a pontatlan vagy torz tanítóadatok „jelentős és tartós diszkriminációt" idézhetnek elő, ha a szolgáltatók nem végeznek előzetes célzott ellenőrzéseket és utólagos validálást is.[63] Több, az egészségügyben alkalmazott gépi tanulási modellekre fókuszáló összegzés is hasonló következtetésre jutott.[64] Az egészségügyi rendszereket vizsgáló metaanalízis szerint, amely 30 tanulmányt összegez, a klinikai gépi tanulási modellek több mint felében kedvezőtlenebb predikciók várhatók fekete vagy alacsony jövedelmű páciensekre, amelyek az algoritmikus torzításon kívül máshogyan nem indokolhatók.[65]

A fentiekre válaszul a fejlesztők egyre gyakrabban alkalmaznak olyan adat- vagy modell-szintű technikákat (például előzetes újrasúlyozás, *„fair"* veszteségfüggvény, utólagos küszöbhangolás), amelyek célja a csoportok közötti statisztikai különbségek mérséklése *(fairness intervention)*. E módszerek azonban kompromisszumot követelnek pontosság, stabilitás és jogi megfelelés között. Ez lényegében azt jelenti, hogy a rendelkezésre álló tanítóadatok mesterséges manipulálása a torzítások kiküszöbölése érdekében csak akkor lehet valóban hatékony, ha a hatások elemzése maga is többlépcsős, külön fejlesztési folyamat. Ennek hiányában a torzítások kiküszöbölése nem garantált: lényegében csak a modellek általános teljesítménye csökken, de a torzítások nem szűnnek meg. Ezt jól illusztrálja egy nemrégiben megjelent szisztematikus áttekintés, amely 50 pénzügyi, HR- és igazságügyi rendszert elemzett. Az adatokból látható, hogy az elnagyolt, nem kellőképpen auditált beavatkozások a modellek tanítóadataiba az esetek harmadában csak tovább rontották a kisebbségi csoportokra vonatkozó modell-predikciókat, negyedüknél egyáltalán nem javították azt, és csak kevesebb mint a felüknél értek el valódi torzításmentes állapotot.[66] A probléma érdemi kezeléséhez a technológiai cégeknek rendszeres, független adatauditokat kellene bevezetniük, amelyek nemcsak a tanítóadatok sokszínűségét és pontosságát vizsgálják, hanem a főbb metszetekre kiterjedő érzékenységi teszteléseket is elvégzik, azaz a különböző változók (bőrszín, nem stb.) együttes hatását is vizsgálják. Emellett fontos, hogy a tartalommoderálás során az emberi moderátorok jobban kiegészítsék és korrigálják az algoritmusok által hozott döntéseket (pl. *Reinforcement Learning with*

---

63 Christiane Wendehorst: *Bias in Algorithms: Artificial Intelligence and Discrimination*. Luxembourg, Publications Office of the European Union, 2022. https://tinyurl.com/5t5evc6h

64 Michael Colacci – Yu Qing Huang – Gemma Postill – Pavel Zhelnov – Orna Fennelly – Amol Verma – Sharon Straus – Andrea C. Tricco: Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology,* Vol. 178., 111606. (2025) https://doi.org/10.1016/j.jclinepi.2024.111606

65 Syed Ali Haider – Sahar Borna – Cesar A. Gomez-Cabello – Sophia M. Pressman – Clifton R. Haider – Antonio Jorge Forte: The Algorithmic Divide: A Systematic Review on AI-Driven Racial Disparities in Healthcare. *Journal of Racial and Ethnic Health Disparities,* 2024. 12. 18., online ahead of print. https://doi.org/10.1007/s40615-024-02237-0

66 Feng Chen – Liqin Wang – Julie Hong – Jiaqi Jiang – Li Zhou: Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association,* Vol. 31. No. 5. (2024), 1172–1183. https://doi.org/10.1093/jamia/ocae060

Human Feedback – RLHF[67]). Ez különösen fontos olyan esetekben, ahol a kontextus meghatározó egy-egy döntés jogossága kapcsán. Emellett persze a nagyobb átláthatóság és elszámoltathatóság is elengedhetetlen lenne a felhasználói bizalom megszilárdítása érdekében. A legtöbb közösségimédia-platform ma üzleti titokként kezeli a moderálásra alkalmazott algoritmusokat, amely gyakorlat azonban – még ha üzletileg védhető is – jelentősen hátráltatja a velük szembeni bizalom kiépítését, és tekintettel az ilyen cégek szólásszabadságra gyakorolt hatására, több szempontból aggályos gyakorlat is.[68]

6.3. Mentális és pszichológiai hatások

Az algoritmikus tartalommoderálás ténye és megnyilvánulási formái komoly negatív pszichológiai hatást fejthet ki a felhasználókra, különösen akkor, ha azok gyakran szembesülnek indokolatlan hibaüzenetekkel (például egy bejelentés során), vagy nem megfelelően kommunikált moderációs döntésekkel.

Amikor a felhasználók azt tapasztalják, hogy tartalmaikat eltávolítják vagy blokkolják anélkül, hogy erre világos magyarázatot kapnának, az értelemszerűen frusztráció kialakulásához fog vezetni (vö. fals pozitív esetek). A felhasználók számára átláthatatlan és kiszámíthatatlan lehet a fellebbezési folyamat is, amely ráadásul gyakran időigényes és bonyolult, és nem is mindig vezet kielégítő eredményre.[69]

A fentiekkel kapcsolatban érdemes persze megemlíteni, hogy a DSA 17. cikke immár kifejezetten kötelezi a platformokat, hogy minden eltávolításról vagy korlátozásról szóló döntést indokolással küldjenek a felhasználónak, és biztosítsanak könnyen elérhető, pártatlan fellebbezési csatornát is. Ennek célja éppen az, hogy csökkentse a nehezen átlátható moderációs folyamatok, döntések okozta frusztrációt.

A szorongás szintén jelentős pszichológiai következmény lehet, különösen azok számára, akik rendszeresen használják az ilyen platformokat kommunikációra, önkifejezésre vagy akár üzleti célokra. A tartalom moderálásával kapcsolatos állandó bizonytalanság, valamint az attól való félelem, hogy egy fontos bejegyzést vagy fiókot eltávolítanak, komoly stresszfaktorként is jelentkezhet. A tartalmak rendszeres eltávolítása vagy a fiókok blokkolása súlyosan befolyásolhatja a felhasználók mentális egészségét, különösen azoknál, akik online közösségek részei, vagy akiknek online jelenléte alapvető fontosságú a munkájukhoz vagy társadalmi kapcsolataik kiépítéséhez, esetleg fenntartásához.

Mindezek fényében kulcsfontosságú, hogy a platformok a következő években ne pusztán jogi megfelelési feladatként tekintsék a DSA-kötelezettségekre, hanem – a felhasználók mentális jóllétét is szem előtt tartva – transzparens, gyors és empatikus

---

67 Guangli Li − Randy Gomez − Keisuke Nakamura − Bo He: Human-Centered Reinforcement Learning: A Survey. *IEEE Transactions on Human-Machine Systems,* Vol. 49., No. 4. (2019) 337–349. https://doi.org/10.1109/THMS.2019.2912447

68 Ződi Zsolt: *Platformjog*. Budapest, Ludovika, 2023.

69 Kris Holt: Meta's Oversight Board Raises Concerns Over Automated Moderation of Hate Speech. *Engadget,* January 23, 2024. https://tinyurl.com/b39t8uuk

kommunikációs protokollokat alakítsanak ki. E protokolloknak egyaránt tartalmazniuk kell(ene) a döntések laikusok számára is érthető indoklását, a hibás szankciók gyors visszafordítását, valamint az érintettek pszichológiai támogatására szolgáló tájékoztató anyagokat. Ha a moderációs lánc minden szintjén megjelenik ez a szemlélet, az nemcsak a felhasználói bizalom helyreállítását segítheti, hanem hosszú távon az online közösségi terek egészségesebb, kevésbé polarizált működéséhez is hozzájárulhat.

7. Konklúzió

A mesterséges intelligencia alapú tartalomszűrő algoritmusok elterjedése jelentős mérföldkőnek számított a közösségimédia-platformok működésében, ám alkalmazásuk számos kihívással jár. Az ilyen rendszerek számos előnyös tulajdonsággal bírnak, hiszen képesek hatalmas mennyiségű tartalmat kezelni rövid idő alatt, ami a platformok növekvő felhasználói bázisát tekintve szükséges is. Ezek az algoritmusok azonban gyakran nem képesek megfelelően kezelni a tartalom kontextusát, ami számos téves döntéshez vezethet, és ez komolyan veszélyezteti a felhasználók jogait, különösen a szólásszabadságot.

Az algoritmusok működése szorosan összefügg a felhasznált adatok minőségével és sokszínűségével. Amennyiben a gépi tanulási modellek nem megfelelően kiegyensúlyozott adatokon alapulnak, az eredmények diszkriminatívak lehetnek bizonyos társadalmi csoportokkal szemben. Ez különösen problematikus a marginalizált közösségeket illetően, akik emiatt gyakran szembesülhetnek az algoritmusok okozta torzításokkal. A technológiai cégek felelőssége, hogy olyan adatokat és algoritmusokat alkalmazzanak, amelyek biztosítják a tartalom moderálásában az egyenlő bánásmód érvényesülését.

További aggodalomra ad okot a véleménybuborékok kialakulása, amelyek az algoritmusok személyre szabott tartalomajánló rendszereinek melléktermékei. Ezek a rendszerek megerősítik a felhasználók meglévő nézeteit, miközben kizárják az alternatív álláspontokat. A folyamat elősegíti a politikai és társadalmi polarizációt, mivel a felhasználók egyre inkább csak olyan tartalmakkal találkoznak, amelyek megerősítik saját világnézetüket. A szűkített perspektíva veszélye abban rejlik, hogy a társadalmi párbeszéd sokszínűsége elvész, ami súlyos következményekkel járhat a demokratikus folyamatokra nézve.

Ahhoz, hogy ezek a kihívások kezelhetők legyenek, elengedhetetlen a tartalommoderálás szabályozásának korszerűsítése, az ilyen rendszerek transzparenciája, és az emberi tényező hangsúlyosabbá tétele a folyamatban. Az algoritmusok mellett a moderátorok szerepe sem veszíthet jelentőségéből, hiszen az emberi ítélőképesség és az empátia továbbra is nélkülözhetetlen a szürke zónák kezelése során. A szabályozók, a felhasználói civil/társadalmi szereplők és a technológiai cégek közötti együttműködés javítása is kulcsfontosságú, hogy az algoritmusokkal való moderálás fenntartható és igazságos legyen, miközben megőrzi az online terek sokszínűségét és az általános szólásszabadságot.