MTA Law Working Papers
2024/2

# Az online gyűlöletbeszéd detektálásának kérdései, különös tekintettel az algoritmikus rendszerekre

**Bukor Liza – Träger Anikó**

**ISSN 2064-4515**
http://jog.tk.mta.hu/mtalwp
*Társadalomtudományi Kutatóközpont – MTA Kiválósági Kutatóhely*
*HUN-REN Centre for Social Sciences – MTA Centre of Excellence*

**Bukor Liza[1] – Träger Anikó[2]**

# Az online gyűlöletbeszéd detektálásának kérdései, különös tekintettel az algoritmikus rendszerekre[3]

## Absztrakt

A gyűlöletbeszéd kérdése már az offline térben is jelentett problémákat mind az egységes definiálhatóság hiánya mind pedig a másokat veszélyeztető tartalma miatt. Az online tér pedig az egyéb információk gyors terjedése mellett a gyűlölködő tartalmaknak is új teret nyitott. Az egyes ilyen jellegű tartalmak sokkal szélesebb körhöz juthatnak el, valamint azoknak a visszaszorítása is az online platformok, közösségi szolgáltatók működési körébe került. Jelen tanulmány szűkebb témáját ez a szűrés jelenti, amelyben egyszerre jelennek meg a platformokat érintő szabályozási, önszabályozási kérdések, valamint az algoritmusok és a mesterséges intelligencia kérdése.

## Bevezetés

Az online platformok a mindennapi életünk részévé váltak, könnyebbé, gyorsabbá teszik a kommunikációt, információk elérését, terjedését. Az új technológiák térnyerése azonban a pozitív hozadékai mellett számos új problémás területet, megválaszolandó kérdést is magával hozott. Azáltal, hogy bárki közzé teheti ezeken az oldalakon tartalmait, véleményét, általa fontosnak gondolt híreket, információkat, és azok szó szerint egy pillanat alatt juthatnak el akár nagy nyilvánossághoz, szükségessé vált ezeknek a tartalmaknak a szűrése, moderálása. A tartalmak kapcsán felmerülő probléma vagy jogsértés számos különböző módon jelenhet meg, pl.: szerzői jogot sértő tartalmak, álhírek, hirdetések, hamis profilok, haszonszerzési célú átverések, csalások, egyéb megtévesztő tartalmak. Jelen tanulmány a platformokon megjelenő káros tartalmak egy speciális területére, az online gyűlölet terjesztésére és annak visszaszorítási lehetőségeire fókuszál. A gyűlöletkelés és az azzal szembeni fellépés nem kizárólag az online tér problémája, azonban – tekintettel annak a már említett sajátosságaira is – sajátos megoldások kidolgozását is megkívánja, amennyiben ez a platformok felületén történik.

Az online tér társadalmi hasznossága, az emberek közötti könnyű kapcsolatteremtése mellett az internetnek ez a „sötét oldala” is ismert, amely alkalmas a gyűlölet terjesztésére, a sértegetésre, a bántalmazásra és a károkozásra. Rengeteg új tartalom kerül fel a platformokra, amelynek következtében a tartalmak mennyiségével és sebességével nehéz lépést tartani, a moderálás túlterhelő az online platformok számára. Ezért, valamint az online gyűlölet megítélésének komplexitása okán[4] a humán moderátorok, valamint az algoritmusok is

[1] PhD-hallgató, ELTE Állam- és Jogtudományi Kar, Állam- és Jogtudományi Doktori Iskola, Alkotmányjogi Tanszék

[2] Egyetemi tanársegéd, Üzleti Jog Tanszék, Gazdaság- és Társadalomtudományi Kar, Budapesti Műszaki és Gazdaságtudományi Egyetem, Műegyetem rkp. 3., H-1111 Budapest, Projektkutató (Mesterséges Intelligencia Nemzeti Laboratórium), HUN-REN Társadalomtudományi Kutatóközpont 1097 Budapest, Tóth Kálmán utca 4.

[3] A kutatás a Társadalomtudományi Kutatóközpontban valósult meg. A tanulmány a Mesterséges Intelligencia Nemzeti Laboratórium, valamint a 138965. számú NKFIH pályázat keretében készült, az Európai Unió RFF-2.3.1-21-2022-00004 projekt, valamint az Innovációs és Technológiai Minisztérium, a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatásával.

[4] A gyűlölet megítéléséhez az emberi méltósághoz való jog, valamint a véleménynyilvánítás szabadságához való jog kollízióját szükséges feloldani, amely feladatot megnehezíti annak eltérő értelmezése, hogy milyen megnyilvánulás számít gyűlöletnek, hol szabhatóak meg a véleménynyilvánítás szabadságának határai. Az online térben mindez még nagyobb kihívást jelent az online tér sajátosságaira figyelemmel, így többek között a globális, azonnali terjedésre, a különböző nyelvekre, a kontextus és a szándék felismerésének, értelmezésének nehézségeire.

hajlamosak a hibákra, amely – különösen az online tér sajátosságaira figyelemmel – aggodalomra adhat okot. Ez a kihívás olyan megoldásokat követel, amelyek alkalmasak arra, hogy megvédjék az embereket és biztonságos online környezetet tartsanak fenn.

Az online vállalatok szerepe és felelősségvállalása kulcsfontosságú az online gyűlöletbeszéd elleni hatékony küzdelemben annak érdekében, hogy az online tér biztonságos legyen. A technológiai platformokon egyre növekszik a nyomás, hogy megoldást találjanak a gyűlöletbeszéd kezelésének problematikájára, megelőzze és csökkentse a már felkerült ilyen tartalmak terjedését, amelyhez az elvárások szerint algoritmikus moderálásra is szükség lehet, ugyan kérdés, hogy az automatizált moderálás valóban megoldást jelent-e. Algoritmikus moderálásként határozhatóak meg az automatizált hash-egyeztető, valamint a prediktív gépi tanulási eszközök, amelyeket egyre gyakrabban alkalmaznak az olyan platformok, mint a Facebook, a YouTube vagy az X (korábban Twitter), amely az átláthatósági jelentésekből is kitűnik. Az algoritmikus moderálás akár hash egyezés, akár előrejelzés alapján osztályozza a felhasználók által közzétett tartalmakat, az osztályozás pedig valamilyen döntést von maga után. A humán moderálás mellett az automatizált rendszereket szükségessé teszik a jogellenes tartalmak rövid időn belüli eltávolítását előíró szabályozások is, így például a német NetzDG, az Európai Bizottság Jogellenes Online Gyűlöletbeszéddel Szembeni Fellépést Szolgáló Magatartási Kódexe (a továbbiakban: Magatartási Kódex), vagy a DSA rendelet.[5] Az automatizált rendszerek, a mesterséges intelligencia alkalmazása az egyes tartalmak detektálásának és osztályozásának számos kérdést vet fel, többek között a véleménynyilvánítás szabadságának érvényesülésére vonatkozóan. A DSA 14. cikke kötelezettséget ír elő a tartalommoderálás céljából alkalmazott szabályok, eljárások érthető, megismerhető módon történő közzétételére a platformok felhasználási feltételeiben. A DSA előírja azt is, hogy a platformok a felhasználók alapvető jogainak figyelembevételével kell, hogy cselekedjenek a moderálási tevékenységük során, amelyhez támpontokat is ad.

A véleménynyilvánítás szabadságához való jog érvényesülésének kérdéseihez hozzájárul az is, hogy az online gyűlöletbeszéd meghatározására nincsen egységes definíció, ami különösen megnehezíti a gyűlöletbeszéd detektálását. A gyűlöletbeszéd, gyalázkodó beszéd, egyéb sértő beszéd célpontjai elsősorban a kisebbségi csoportok, célja az adott csoporthoz tartozó egyének vagy csoportok ellen irányuló negatív kifejezés, sértés, fenyegetés. A mindennapok során a közösségi médiában gyakran találkozunk olyan tartalmakkal, amelyeket online gyűlöletbeszédként értékelhetünk. Ugyan a gyűlöletbeszéd fogalmi kérdéseinek teljeskörű elemzése kívül esik jele tanulmány keretein, fontosnak tartjuk rögzíteni, hogy a gyűlölködő megnyilvánulások széles spektrumon mozognak, egészen az enyhébb fokú, sértő tartalmaktól a komoly következményekkel járó, gyűlöletkeltő, akár erőszakra uszító tartalmakig. Az internet térhódításával a gyűlöletbeszéd egyfajta összefoglaló kifejezésként a gyűlölködő beszéd széles skáláját fedheti le. Az online térben a gyűlöletbeszéd számos megnyilvánulási formában jelenik meg, így szöveges tartalmakban, képekben (akár mémekben), videókban és ezeknek a kombinációjában, a felhasználók pedig névtelenség vagy álnevek mögé bújva bátrabban fejezik ki gyűlölködő nézeteiket, amelyek gyorsan terjednek. A gyors terjedéshez akár az algoritmikus rendszerek is hozzájárulhatnak, hiszen ezek végzik a tartalmak előresorolását, megjelenítését az egyes felhasználók számára.

A definíciós hiányosságok ellenére az online gyűlöletbeszéd kezeléséhez támpontokat ad a 2012-es Rabati Akciótervben foglalt hat lépéses teszt, amely szerint a gyűlöletbeszéd korlátozásához, a megfelelő küszöbérték megállapításához vizsgálni kell a kontextust, a

[5] Robert Gorwa, Reuben Binns, & Christian Katzenbach (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1)., 1-3. o., az Európai Parlament és a Tanács (EU) 2022/2065 rendelete (2022. október 19.) a digitális szolgáltatások egységes piacáról és a 2000/31/EK irányelv módosításáról (digitális szolgáltatásokról szóló rendelet)

tartalom közlőjét, a szándékát, magát a tartalmat, a beszéd nyilvánosságát, és a beszéd által bekövetkezhető kárt. Így mérhető fel kellőképpen a gyűlölet súlyossága. A kontextus figyelembevétele során vizsgálni kell az aktuális társadalmi és politikai kontextust is. Szükséges tanulmányozni azt is, hogy ki közölte a tartalmat, milyen szerepet tölt be a társadalomban, emellett magát a közönséget is.[6] Tehát a gyűlöletbeszéd detektálása komplex szempontrendszer alapján kell, hogy történjen, amely kihívásokat okoz. Olyan társadalmi kontextust kell figyelembe venni, hogy ki a közlő, mi a közlő nyelvjárása, egyes csoportokon belül elfogadhatóan használt kifejezések csoporton kívüliek által történő használata elfogadható-e vagy sem.[7] Az online platformokon számos típusú és témájú tartalom jelenik meg, amely eltérő kontextusban eltérően értelmezendő, akár például irodalmi, művészeti alkotás lehet vagy oktató célzatú tartalom, az eltérő értelmezésre pedig a detektálási módszereknek képesnek kellene lenniük a véleménynyilvánítás szabadságának érvényesülése érdekében. Jelen tanulmány az online gyűlöletbeszéd detektálásának kérdéseit vizsgálja, különösen az automatizált rendszerek szerepére és esetleges hatékonyságára.

## 2. A felhasználók és a platformok dinamikájának szerepe az online gyűlöletbeszéd detektálásában a DSA szabályozási koncepciójának tükrében

Az online platformok esetében általában is meghatározó sajátosságnak tekinthető, hogy azok – szemben a hagyományos média szolgáltatásaival – nem előzetesen ellenőrzött, szerkesztett tartalmakat kínálnak. Ezzel szemben a felületeiken megjelenő tartalom alakításában meghatározó a felhasználók aktivitása mind a tartalmak feltöltésében, megosztásában, további terjedésében, mind pedig azok elérhetetlenné tételében is.

Ezzel összefüggésben az online gyűlöletbeszéd elleni küzdelemben, az online tartalommoderációban a felhasználói jelentési mechanizmusok szerepe és a felhasználók moderációs folyamatban betöltött szerepe kulcsfontosságú, tekintettel arra is, hogy az online gyűlöletbeszédnek teret biztosító közösségi médiaplatformok alapvetően értesítési-eltávolítási eljárás alapján végzik moderációs tevékenységüket. Az egyes tagállamok, így például Németország a NetzDG-vel, valamint az EU szervei a tartalmak moderálásának feladatát a platformok kezében hagyta, gondoljunk csak a Magatartási Kódexre[8], vagy az eKer irányelvre[9], amely a *notice-and-take down* (értesítési-eltávolítási) eljárást intézményesítette, privatizálva a tartalmak jogellenességéről való döntést[10]. A platformok azonban alapvetően a saját szabályzataik alapján járnak el, nekik van meg az eszközük egy tartalom eltávolítására vagy visszaállítására, illetve akár egy felhasználó hozzáférésének tiltására vagy korlátozásaira. Ehhez kapcsolódva pedig az egyes jogsértésekkel kapcsolatos eljárások is a platform rendszerén belül történnek, amelyek esetében az átláthatóság, pártatlanság szintén vetett fel kérdéseket. Jellemző probléma ugyanis, hogy a vita a felhasználó és a platform között keletkezik (pl. felhasználó profilját felfüggesztik, letiltják), a felhasználó vitatja ezt. Egy ilyen esetben valójában a platform az egyik fél a vitás kérdésben, de a platform – még ha

[6] Rabati Akcióterv http://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf
[7] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith (2019). The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678., Florence, Italy. Association for Computational Linguistics., 1668. o. A tanulmány példaként a „n*gga" és a „queer" kifejezéseket jeleníti meg.
[8] A jogellenes gyűlöletbeszéddel szembeni fellépést szolgáló magatartási kódex https://ec.europa.eu/commission/presscorner/detail/hu/IP_17_1471
[9] Európai Parlament és a Tanács 2018/1808 irányelve
[10] Ez tartalmában azt jelenti, hogy a platform előzetesen nem szűri a tartalmakat, azonban, ha arra vonatkozóan olyan jelzés (notice) érkezik, akkor viszont vizsgálja a jelölt tartalmat és szükség esetén el is távolítja (take down).

másik részleg is foglalkozik ezzel a kérdéssel – jár el a vita eldöntőjeként is. Mivel jelenleg nehezen biztosítható hatékonyan, hogy ezeken a területeken (felhasználók, illetve tartalmaik korlátozása, feloldása, azok felülvizsgálata) ne a platform járjon el, a megoldást az jelentheti, hogy a „szabályozást kell szabályozni". Ennek értelmében továbbra is a platformnál marad saját szabályzatainak elfogadása, illetve a felhasználóihoz kapcsolódó eljárások lefolytatása, azonban ezek legfontosabb elemeinél megjelenik a DSA mint az egységes, rendeleti szabályozás igénye.

Ennek keretében a DSA 16. cikke szerint a tárhelyszolgáltatóknak olyan bejelentési és cselekvési mechanizmusokat kell alkalmazniuk, amelyek lehetővé teszik és megkönnyítik a jogellenesnek vélt tartalmak bejelentését, kellő indokolási lehetőséggel és pontossággal. A DSA preambulum (50) bekezdése alapján ez a kötelezettség minden tárhelyszolgáltatóra vonatkozik annak méretétől függetlenül, a mechanizmust pedig könnyen hozzáférhető és felhasználóbarát, egyértelműen azonosítható módon kell működtetni. A DSA 6. cikke értelmében pedig a tudomásszerzést követően haladéktalanul intézkednie kell a jogellenes tartalom eltávolításáról vagy az ahhoz való hozzáférés megszüntetéséről. Így a DSA alapján is csupán akkor van a platformoknak cselekvési kötelezettsége az egyes tartalmak kapcsán, ha azokat bejelentették a felületnek.

Jól látható tehát, hogy a DSA koncepciója nem megszüntetni akarja a jelenleg is működő mechanizmusokat, hanem azokat egységesíteni, garanciális keretek közé igyekszik szorítani, jobban biztosítva ezáltal a felhasználók védelmét, és az eljárások átláthatóságát, vagyis a platformok aktív részesei maradnak a szabályozás kialakításának és megvalósításának. Az ismertetett koncepciónak a DSA-ban történő alkalmazásával kapcsolatban kritikaként merülhet fel, hogy lesz-e kapacitás a platformok nagy mennyiségű szabályanyagának áttekintésére, valódi ellenőrzésére. Továbbá felvetődik az is, hogy a platformok valóban tevőlegesen részt fognak-e venni ebben a közös szabályozásban, vagy csak látszólag az elveknek megfelelő szabályokat, intézkedéseket fogadnak el, azonban érdemben nem tesznek azért, hogy a szabályozás célja megvalósuljon, valóban együttműködjenek a felhasználók védelme és a DSA-ban megfogalmazott további célok elősegítése érdekében.[11]

A platformok tehát – saját jogszerű működésük biztosítása érdekében is – lehetővé teszik a felhasználók számára, hogy a gyűlöletbeszédnek vélt tartalmakról bejelentést tegyenek, amint ők maguk szembesülnek valamilyen jogsértő, aggályos tartalommal, részletesen leírva az észlelt problémát, ezzel értesítve arról a vállalatot. A felhasználók ezzel felelősséget vállalnak a biztonságos online tér megteremtésének feladatában, a platformokra vonatkozó szabályok érvényesítésében és ezáltal a szabályok esetleges későbbi módosításaiban, kialakításaiban. A felhasználók akár az aggályos tartalom közzétételét és észlelését követően rövid időn belül megtehetik a bejelentést, ráadásul a globális jellegből fakadóan eltérő kulturális és társadalmi kontextusokat ismerhetnek, valamint a nyelvezetet is felismerhetik. Így az összetett tartalmak értelmezésében segíthetnek.[12]

Ugyanakkor fennáll a lehetősége annak is, hogy visszaélésszerűen túlzott bejelentést tegyenek. Sok esetben az algoritmusok vizsgálják a bejelentett tartalmakat, értékelik azokat és automatikusan döntenek annak esetleges eltávolításáról, humán moderátorok eljárása nélkül. Azonban a felhasználók általi bejelentések segítenek a platformoknak és a moderátoroknak gyorsabban reagálni a felmerülő kihívásokra, aggályos tartalmakra, ezzel támogatva a

[11] Nicolo Zingales: The DSA as a paradigm shift for online intermediaries' due diligence https://verfassungsblog.de/dsa-meta-regulation/?fbclid=IwAR1sNE_oloTqQh3LADS2GyBAqRrDqDhwTUskTNS_Flxt_AlMAOuHoN3n_5c

[12] Justin Cheng, Cristian Danescu-Niculescu-Mizil, Jure Leskovec, Michael Bernstein (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing

biztonságos(abb) online tér megteremtését. Ugyanakkor az algoritmikus rendszerek is tanulnak a felhasználók bejelentései alapján, hiszen a gépi tanulás módszerén alapulnak, így a felhasználók segítenek abban is, hogy az algoritmikus rendszerek a jövőben hatékonyabban és pontosabban tudják értékelni az egyes tartalmakat, figyelembe véve az új, éppen aktuális kihívásokat is, ám fontos, hogy a különböző felhasználók megítélési módja eltérő lehet.[13] Az felhasználó gyakran nem értenek egyet bizonyos kifejezések sértő szintjét illetően. A tartalmat értékelők különböző jellemzői alapján változik a tartalom sértőként való megítélése, amely az adatok valódi hasznosságát kérdésessé teszi.[14]

Az online platformot üzemeltető szolgáltatókra alkalmazandó további DSA-ban foglalt kötelezettség az értesítési-cselekvési mechanizmusokhoz kapcsolódóan a megbízható bejelentők általi bejelentések biztosítása és azon bejelentések kiemelt kezelése, indokolatlan késedelem nélküli feldolgozása és az azokról történő indokolatlan késedelem nélküli döntés meghozatala, amely szintén a felhasználói bejelentésekhez hasonló intézmény.[15] A megbízható bejelentővé váláshoz olyan feltételeknek kell megfelelni, mint például a szakértelem és hozzáértés a jogellenes tartalmak észlelése, azonosítása terén vagy a szolgáltatóktól való függetlenség. A megbízható bejelentők olyan személyek vagy szervezetek, akik hozzájárulnak az aggályos tartalmak jelentéséhez és moderálásához, segítve a moderációs folyamatok hatékonyságát, tekintettel arra, hogy rendszeresen figyelik az online platformokat, prokatívan járnak el és szakértelemmel rendelkeznek az aggályos tartalmak felismeréséhez, megtalálva az aggályos tartalmak jelentése és a véleménynyilvánítás szabadságához való jog érvényesülése közötti határt.

Továbbá a különböző társadalmi párbeszédek is elősegítik az online tartalommoderálást és a gyűlölködő tartalmak azonosítását, így például egy nyilvános vita után a Facebook javította a mianmari nyelvű gyűlöletbeszéd osztályozására szolgáló algoritmikus rendszereit, ami az automatizált rendszerek jelzéséből származó eltávolítások számának 39%-os növekedését eredményezte mindössze hat hónap alatt.[16]

Mindezek alapján a felhasználók aggályos tartalmak detektálásában, online tartalommoderálásában betöltött szerepe kulcsfontosságú, hiszen számos ponton hozzájárulnak a biztonságos(abb) online tér és a jövőbeni hatékonyabb moderálás megteremtéséhez, fejlesztéséhez, beleértve az algoritmikus rendszerek hatékonyabb működését is.

Továbbá a platformok közös önszabályozást is támogatja a DSA, kiegészíti az egyes területeken ezzel a rendeletben megfogalmazott célok elérését, és magatartási kódexek elfogadásával igyekszik a legfontosabbnak tartott területeken a platformokat is bevonva ellenőrzött megoldásokat kialakítani. A DSA megfogalmazásában „A Bizottságnak és a Testületnek elő kell mozdítania az önkéntes magatartási kódexek kidolgozását, valamint e kódexek előírásainak végrehajtását, hogy hozzájáruljanak e rendelet alkalmazásához. A Bizottságnak és a Testületnek törekednie kell arra, hogy a magatartási kódexek egyértelműen meghatározzák a kitűzött közérdekű célok jellegét, tartalmazzanak az e célkitűzések

[13] Bővebben ld. például Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th Internatinal Conference on Web and Social Media

[14] European Union Agency For Fundamental Rights (2022). Bias In Algorithms. Artificial Intelligence and Discrimination. 54. o., Reuben Binns, Michael Veale, Max Van Kleek, Nigel Shadbolt, (2017). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In: Giovanni Luca Ciampaglia, Afra Mashhadi,Taha Yasseri, (eds) Social Informatics. SocInfo 2017. Lecture Notes in Computer Science, vol 10540. Springer, Cham.

[15] DSA 22. cikk

[16] Gorwa (2020),2.o.

elérésének független értékelésére szolgáló mechanizmusokat, és egyértelműen meghatározzák az érintett hatóságok szerepét."[17]

Az ön- és társszabályozás alkalmazásának egy kiemelt lépése volt – a jelen tanulmány szűkebb témáját is képező – jogellenes gyűlöletbeszédre vonatkozó magatartási kódex,[18] amelyet az Európai Bizottság és négy nagy tech vállalat - a Facebook, Twitter, YouTube és Microsoft - 2016. májusában fogadtak el.[19] A vállalatok vállalták, hogy elkötelezik magukat a gyűlöletbeszéd felszámolása és visszaszorítása mellett, és megteszik a szükséges lépéseket az online platformjaikon található jogellenes tartalmak eltávolítása és a felhasználókra vonatkozó szigorúbb szabályok bevezetése érdekében.[20]

A kódex kötelezi a vállalatokat a gyorsabb és hatékonyabb tartalmi moderálásra, azaz a jogellenes tartalmak azonosítására és eltávolítására.[21] Ez magában foglalja a bejegyzések, kommentek, videók és más tartalmak gyorsabb észlelését és azokhoz való hozzáférés korlátozását. A vállalatoknak együttműködést kell vállalniuk a hatóságokkal és a civil szervezetekkel, hogy hatékonyabban tudjanak fellépni a gyűlöletbeszéd ellen.[22]

Fontos megjegyezni, hogy a kódexhez való csatlakozás önkéntes, és a vállalatok, platformok dönthetnek arról, hogy részt vesznek-e a kezdeményezésben. Azonban a kódex elfogadása és betartása egy pozitív lépésnek tekinthető a gyűlöletbeszéd visszaszorítása és az online tér biztonságosabbá tétele felé.

Az Európai Bizottság a Kódex 2016-os elfogadása óta folyamatosan monitorozza az abban foglalt célok megvalósulását.[23]Az összes jelentés részletes elemzése túlmutat jelen tanulmány keretein, azonban az megállapítható, hogy a Kódexhez csatlakozók fentebb már hivatkozott bővülése mellett, hogy a kezdetekhez képest magasabb a bejelentések száma, hatékonyabban távolítják el a szolgáltatók a gyűlöletet keltő tartalmakat, bár ez a tendencia nem tisztán emelkedő, egyes szolgáltatóknál kisebb visszaesések megjelennek.[24]

A gyűlöletbeszéd visszaszorításában tehát jelenleg is komoly szerepe van már a platformok részvételének, önszabályozási megoldásainak, valamint közös, társszabályozás keretében történő célok melletti kiállásuknak, amelyeket a DSA tovább kíván ösztönözni.

## 3. Az algoritmikus rendszerek online gyűlöletbeszéd detektálására vonatkozó használata

Ahogyan az az előző fejezetben már említésre került, a gyűlöletbeszéd visszaszorításában meghatározó szerepe van az algoritmusoknak, mesterséges intelligencia alapú szoftvereknek.

---

[17] DSA preambulum(106) bekezdés

[18] A jogellenes gyűlöletbeszéddel szembeni fellépést szolgáló magatartási kódex https://ec.europa.eu/commission/presscorner/detail/hu/IP_17_1471 Ahogyan az az elfogadás időpontjából is látható, a kódex elfogadása jóval megelőzi a DSA hatályba lépését, azonban a magatartási kódexek és az ön- és társszabályozás fenntartása a DSA alkalmazása során is megmarad.

[19] Azóta más tech vállalatok is csatlakoztak a kezdeményezéshez, például a Pinterest, majd 2018-ban az Instagram is csatlakozott a kódexhez, továbbá a kódexnek jelenleg már tagja a Tiktok, a Linkedin, a Viber, a Twich, és a francia JeuxVideo is.

[20] https://ec.europa.eu/commission/presscorner/detail/hu/IP_17_1471

[21] Code of Conduct on Countering Illegal Hate Speech Online https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[22] https://ec.europa.eu/commission/presscorner/detail/hu/IP_17_1471

[23] Az Európai Bizottság ezzel kapcsolatos dokumentumai: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[24] Countering illegal hate speech online 7th evaluation of the Code of Conduct https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

A gyakorlatban jelenleg használt gyűlöletbeszéd-észlelő algoritmusok a fejlett gépi tanulási módszertan és természetes nyelvfeldolgozás kombinációján alapulnak.[25] Az online platformok a felületükön megjelenő tartalmak mennyiségének következtében az online gyűlölet automatikus, proaktív észlelésére törekednek, amelyhez eszközöket fejlesztettek ki. Így például a Facebook (Meta) a PyTorchot, a Google a Perpective-t, amelyek nyílt forráskódú deep-learning keretrendszerek.

Az online platformok mellett általában az algoritmusok, mesterséges intelligencia alkalmazása is egy olyan terület, amely napjainkban új kihívás elé állítja a jogalkotókat. A mesterséges intelligencia szabályozás részletes elemzése jelen tanulmány keretein túlmutat, azonban – a moderálás kérdésével való szoros összefüggésre tekintettel – álláspontunk szerint szükséges röviden kitérni erre a kérdésre. Ezekre vonatkozóan a DSA is tartalmaz kifejezetten az algoritmusok, ajánlórendszerek kockázatértékelésére vonatkozó kötelezettségeket.[26] A DSA ezzel kapcsolatos rendelkezései többféle megfogalmazást alkalmaznak, egyes helyeken kifejezetten megjelölik az ajánlórendszereket, máshol utalnak az egyéb releváns algoritmusokra is. Ezen a különbségtétel azért is nagyon jelentős, mert a másik kiemelkedően fontos terület, ahol a platformok az algoritmusokat rendszerszinten használják, az a tartalmak moderálásának a területe.[27]

A moderálás azért is kiemelten fontos terület, mert ezzel az eszközzel lép fel a platform a felhasználók alapvető jogainak (pl.: szólásszabadság) korlátozására, vagy akár a megélhetésüket jelentő foglalkozás korlátozására (pl.: egy tartalomgyártó fiókjának felfüggesztése, tartalmai monetizálásának kizárása). A DSA a 20. cikk (6) bekezdésében a belső panaszkezelési rendszerrel kapcsolatban utal arra, hogy ezt a tevékenységet nem lehet kizárólag algoritmussal végezni, hanem képesítéssel rendelkező személy felügyelete szükséges hozzá. Ettől eltekintve nem találunk speciális rendelkezést a tartalom moderálását, panaszkezelést végző algoritmusokra vonatkozóan, a rendelet inkább az ajánlórendszerekre tér ki.

Itt szükséges kitérni a mesterséges intelligencia szabályozására vonatkozó rendeletre is, amely nem kifejezetten az online platformokra, hanem általában a mesterséges intelligencia használóira, fejlesztőire vonatkozóan fogalmaz meg megfelelési követelményeket.[28] Jelen tanulmány keretein túlmutat a rendelettervezet teljeskörű elemzése, azonban azt mindenképp szükséges megemlíteni, hogy annak központi elemét jelenti a már hivatkozott „kockázatalapú megközelítés”. Ez a rendelet koncepciójában azt jelenti, hogy az egyes mesterséges intelligencia alapú technológiákat annak megfelelően sorolja be szigorúbb vagy kevésbé szigorú megfelelési csoportokba, hogy azok – a jogalkotó megközelítésében – mekkora kockázatot jelentenek a felhasználókra, illetve az ő alapvető jogaikra nézve.[29]

Az online gyűlölet visszaszorítása egy olyan terület, ahol kiemelten jelentkeznek a fentebb bemutatott aggályok: gyakorlatilag a platform, illetve az általa alkalmazott algoritmus a felhasználó véleménynyilvánítási szabadságának határairól dönt.[30] Erre különböző

[25] A természetes nyelvfeldolgozás a számítástudomány egy területe, végső célja az emberi nyelv által hordozott jelentés megértése, amely „a beszéd témájához illeszkedő alapos háttértudást feltételez, továbbá a szavak egyszerű jelentése mögötti szándékolt közlendők, utalások megértését is kívánja.” Ehhez többek között a mondat szintaktikai szerkezetét vizsgálja és szemantikai elemzést végez. https://gyires.inf.unideb.hu/KMITT/c01/ch04s02.html Utolsó letöltés dátuma: 2023. december 10.

[26] DSA 34-35. és 37. cikkek

[27] Zódi Zsolt: Az európai platformszabályozás jellegzetességei Platformjog és felhasználóvédelem. In Medias Res 2022/1.

[28] Mesterséges intelligencia rendelet: 2. cikk (1) bekezdés

[29] Mezei Kitti: A mesterséges intelligencia jogi szabályozásának aktuális kérdései az Európai Unióban. In Medias Res 2023/1. 61.

[30] A demokrácia egyik esszenciális alapköve a kommunikációs alapjogok anyajogaként a véleménynyilvánítás szabadságához való jog, A véleménynyilvánítás szabadságához való jog kiemelt jelentőségű eszközként szolgál

technológiai megoldásokat alkalmaznak. Ilyen megoldás a hash-illesztés olyan tartalmak proaktív észlelésére, moderálására használható fel, amelyek pontosan megegyeznek egy korábban azonosított tartalommal, tehát ugyan egyértelműen képes beazonosítani az adott tartalmat, de kisebb módosítással közzétett tartalommal könnyen kijátszható módszer, a változó környezetre nem képes reagálni. Egyezés esetén pedig akár azonnali korlátozást, blokkolást is eredményezhet annak ellenére, hogy az nem feltétlenül gyűlöletkeltő, amely a véleménynyilvánítás szabadságának túlkorlátozásához vezethet.[31]

Ehhez képest a gépi tanulás esetén a modelleket arra képezik képzési adatok alapján, hogy megjósolják az eredményeket, így például azt, hogy egy adott tartalom sértő-e vagy sem.

A FRA 2022-es jelentése többnyire a prediktív rendszerekkel foglalkozott és rámutatott arra, hogy ezek az eszközök torz eredményekre juthatnak, tehát az egyik technikai probléma ezekkel a rendszerekkel, hogy elfogultak lehetnek. Az elfogultság és a torzítás mérésének céljából nyilvánosan elérhető adatokon alapuló előrejelző algoritmusokat készítettek és teszteléseket végzetek. Így például a különböző identitáskifejezésekhez kapcsolódó előre jelzett sértő szintje eltér, egyes kifejezések sokkal gyakrabban vezetnek a tartalom sértő előrejelzéséhez, amely a tartalmak téves osztályozásához vezethet, azt eredményezve, hogy a szöveget sértőnek osztályozzák, annak ellenére, hogy valójában nem az. Így kerülhet gyűlöletbeszédként való osztályozásra egy muszlim felhasználó által közzétett olyan tartalom, amely csupán annyit tartalmaz, hogy „Muszlim vagyok." Mindennek az okaként az FRA azt jelölte meg, hogy ezek a kifejezések szerepelnek az algoritmusok létrehozásához, felépítéséhez használt „képzési adatokban" (szöveges adatkészletekben, beleértve a gyűlölet példáit is) rögzített online gyűlölethez.[32] Ugyanakkor a képzéshez kapcsolódóan további probléma, hogy ahhoz nincsenek feltétlenül használható, reprezentatív adatkészletek. Így, ha „ezek az adatkészletek nem tartalmaznak példákat különböző nyelveken és különböző csoportokból vagy közösségekből származó beszédre, az eredményül kapott eszközök nem lesznek felszerelve e csoportok kommunikációjának elemzésére", tehát használhatóságuk limitált. A természetes nyelvfeldolgozáson alapuló eszközök ezért olyan környezetekben teljesítenek a legjobban, amelyek szorosan illeszkednek a betanított képzési adatokhoz.[33]

Az olyan nyelvek esetében, amelyek nyelvtani nemű főneveket használnak, nyelvi elfogultság figyelhető meg, azaz a női és a férfi változatok előrejelzése között is különbség van. Az FRA kutatása során arra a következtetésre jutott, hogy mindez interszekcionális gyűlöletet is tükrözhet, mivel az osztályozás a nemek alapján történik, kombinálva például a vallással, amely a természetes nyelvfeldolgozáson alapuló algoritmusok esetében plusz kihívást jelent. Tehát ezeknek az algoritmusoknak a gyűlöletbeszéd automatikus moderálására való használata nehézségekbe ütközik, mivel a védett tulajdonsághoz kapcsolódó szavakat azonosítják önmagukban a sértő beszédre utaló jelként anélkül, hogy a kontextust figyelembe vennék, különösen az angoltól eltérő nyelvhez használt természetes nyelvfeldolgozáson

---

egyrészt ahhoz, hogy a közösség érdekében felszínre kerülhessen minden vélemény, gondolat. Másrészt pedig individuum önkifejezését is lehetővé teszi. Lásd például Lassányi Tamás: A véleménynyilvánítás szabadsága az Interneten https://mek.oszk.hu/01400/01407/01407.htm#3.1; Nagy Krisztina: A véleménynyilvánítás szabadsága, nyilvánosság, tudatos médiahasználat. In Fundamentum 2017/1-2., 7-8. o.

[31] Gorwa, Binns, Katzenbach (2020), 5. o.

[32] A kutatás szerint például az angol nyelvű modellekben a „zsidó" kifejezést tartalmazó mondatok nagyobb mértékben növelték az előre jelzett sértő szintet, mint a „keresztény" kifejezés. De ez a „muszlim" kifejezésre is igaz lehet. FRA 2022 11-14., 17., 49. o., Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (2016) In Advances in neural information processing systems, 4349-4357, 4351. o.

[33] Emma Llansó, Joris van Hoboken, Paddy Leerssen, Jaron Harambam. Artificial Intelligence, Content Moderation, and Freedom of Expression. (2020) Transatlantic Working Group, Working Paper, 8. o. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

alapuló rendszerek esetében. A különböző nyelveken való detektálás több tényezőn alapul, így a különböző dialektusokat még egy azonos nyelven belül is ismerni kellene, ugyanakkor már az angol nyelvtől eltérő nyelvek esetében sincs az algoritmikus rendszerekhez elérhető megfelelő képzési nyelvi adatmennyiség, amely nehezíti a sértő, gyűlölködő tartalmak hatékony és pontos azonosítását, értékelését, ráadásul a kulturális különbözőségek is előtérbe kerülhetnek, ha más fő nyelvre tanították be az adott rendszert. A nyelvi problémákon túl az elfogultság pedig abból eredhet, hogy a gépeket és a technológiákat emberek fejlesztik, ahol pedig az emberi döntéshozatalban elfogultság merülhet fel, úgy a gépek általi döntéshozatalban is, tehát már eleve elfogultságot tápláltak azokba. Azonban azt maga a jelentés is rögzíti, hogy nem kívánja értékelni a gyakorlatban alkalmazott sértő beszédet detektáló algoritmusokat, valamint azt sem, hogy a célnak megfelelőek-e, ugyanakkor az egyértelmű, hogy azok nagymértékben támaszkodnak bizonyos kifejezések jelenlétére.[34] Ezen technológiák alkalmasak arra, hogy gépi fordítást vagy spamszűrést végezzenek, de például arra nem alkalmasak, hogy a nők elleni diszkriminációt felismerjék.[35] Továbbá a sértő, gyűlölködő beszéd felismerése is komplexebb kihívások elé állítja a rendszereket. A sértés függ a beszéd kontextusától, azonban ilyen értékelést általában nem végeznek a sértő beszédfelismerési algoritmusok, pusztán a szöveg alapján dolgoznak.[36] A gépi tanulási eszközök elemzésébe olyan kontextus lehet beépíthető, mint például a közzétevő felhasználó és a címzett felhasználó, valamint a közöttük lévő kapcsolat, ugyanakkor az adatvédelmi aggályokat vet fel, amelynek további elemzése jelen tanulmánynak nem célja. Ugyanakkor a történelmi, politikai, társadalmi és kulturális kontextusok nehezen észlelhetők egy gépi betanított eszköz számára, amely a gyűlöletbeszéd megítéléséhez elengedhetetlen elem lehet.[37]

Egy 2019-es tanulmányban a faji elfogultságot öt különböző gyűlöletbeszédre és sértő nyelvezetre vonatkozó Twitter (jelenleg X) adatkészletet vizsgáltak és klasszifikálókat képeztek az adatkészletek alapján az afroamerikai és a sztenderd amerikai angol nyelven írt tweetek előrejelzéseit, amely szisztematikus faji elfogultságot mutatott ki mivel az ezekre képzett sértő nyelvészlelő rendszerek hajlamosak voltak azt jósolni, hogy az afroamerikai angol nyelven írt tweetek lényegesen nagyobb arányban sértőek, amely negatív hatással bírhatnak az afroamerikai felhasználóikra, holott gyakran éppen ők a visszaélések célpontjai, így diszkriminálják azokat, akik védelmére tervezték a rendszert. Ez a tanulmány is úgy véli, hogy a torzítás a képzési adatok eredménye. Gépi tanulási modelleket alkalmaznak a gyűlöletbeszéd és a sértő nyelvezet észlelésére a közösségi média platformokon is, ám az elfogultág csökkentheti a pontosságot., ezek a modellek „szisztematikusan elfogultak bizonyos társadalmi csoportokkal, különösen a védett osztályokkal szemben". A tanulmány rámutatott arra is, hogy sértő nyelvezet detektálása nem történhet kontextusfüggetlen megközelítésben, mivel a különböző közösségek eltérő beszédnormákkal rendelkeznek,

[34] A kutatás szerint például az angol nyelvű modellekben a „zsidó" kifejezést tartalmazó mondatok nagyobb mértékben növelték az előre jelzett sértő szintet, mint a „keresztény" kifejezés. De ez a „muszlim" kifejezésre is igaz lehet. FRA 2022 11-14., 17., 49. o., Bolukbasi, Chang, Zou, Saligrama, Kalai (2016), 4351. o.; Anna Schmidt,Michael Wiegand. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10, Valencia, Spain. Association for Computational Linguistics.

[35] European Union Agency For Fundamental Rights (FRA)(2023). Online Content Moderation, Current Challenges In Detecting Hate Speech, 13. o.

[36] FRA (2022) 57. o.

[37] LLansó, Hoboken, Leerssen, Harambam (2020) 7. o.

azonban arról nincs konszenzus arról, hogy hogyan lehet a különböző társadalmi és kulturális kontextusokra érzékeny észlelési rendszereket fejleszteni.[38]

2018-ban Mark Zuckerberg úgy nyilatkozott, hogy a mesterséges intelligencia nem állt készen a gyűlöletbeszéd detektálására és a gyűlöletbeszédet a felhasználóknak kell jelenteniük. Azóta a Facebooknál a gyűlöletbeszéd detektálása kapcsán megnövekedett a mesterséges intelligencia szerepe, az eszközök pro-aktívan (tehát más által történő jelentés előtt) megjelölik ezeket a tartalmakat, amelyről majd a humán moderátorok döntenek és adott esetben manuálisan távolítják el azokat. Úgy vélte, hogy 5-10 éven belül a rendelkezésre fognak állni olyan mesterséges intelligencia eszközök, amelyek a különböző nyelvi apró különbözőségeket is pontosabban ki tudják szűrni a különböző tartalmak vonatkozásában. Tekintettel a kontextusra, valamint a szlengekre is, amelyeket általában nem használnak a mesterséges intelligenciák képzésekor.[39] Az FRA 2022-es jelentése szerint a pro-aktívan eltávolított gyűlöletbeszédként azonosított tartalmak számának korábbi évekhez viszonyított jelentős növekedése a mesterséges intelligencia eszközeinek és az egyes tartalmak az által történő megjelölésének köszönhető, amely alapján a humán moderátorok eljárhattak.[40]

A Facebook szerint ahhoz, hogy a mesterséges intelligencia a gyűlöletbeszéd felismerésének hatékonyabb eszközévé váljon, képesnek kell lennie arra, hogy a tartalmat úgy tudja megérteni, ahogyan az emberek. Amikor például egy mémet néz egy felhasználó, nem gondolkodik egymástól függetlenül a szavakon és a fotón, hanem a mém egyesített jelentését veszi. Mindez rendkívül nagy kihívást jelent a gépek számára, mivel nem elemezhetik csak külön a szöveget és a képet, hanem azt is meg kell érteniük, hogy hogyan változik meg a jelentésük, amikor azok együtt szerepelnek. Ezért a Facebook mesterséges intelligenciája 2020-ban olyan adatkészletet hozott létre és tett nyilvánossá, amely segít olyan rendszerek létrehozásában, amelyek jobban megértik a multimodális gyűlöletbeszédet. A Hateful Memes (Gyűlölködő Mémek) adatkészlet. több mint 10 000 újonnan létrehozott példát tartalmaz multimodális tartalomra. A multimodális jelentés gépek általi megértési nehézségének megoldása érdekében a kutatói közösség olyan eszközök kifejlesztésére összpontosít, amelyek figyelembe veszik az adott tartalomban jelenlévő különböző módozatokat, majd az osztályozási folyamat korai szakaszában összeolvasztják azokat, ezzel a gépek képessé válnak arra, hogy az emberekhez hasonlóan együtt elemezzék a különböző módozatokat.[41]

A közösségi médiavállalatok egyre nagyobb mértékben használnak algoritmikus rendszereket arra, hogy az online gyűlöletet detektálják, a Facebook esetében például jelentősen megnőtt az online gyűlöletbeszéd algoritmikus rendszerek általi automatikus detektálásának mértéke 2022-re, de a YouTube esetében is hasonló tendencia figyelhető meg.[42]

Arról ugyan nincsenek biztos információk, hogy a Facebook, a Twitter és a Google által használt tartalommoderációs rendszerek is a szakirodalomban kimutatott a torzítást mutatják-e, tekintettel arra, hogy ezek a vállalatok nem megismerhető felépítésű, saját tulajdonú algoritmikus rendszereket használnak a moderálási tevékenységük körében és az online gyűlöletbeszéd detektálására. a technológia, amelyet ezek a cégek a tartalom moderálására használnak, saját tulajdonú. Az azonban ismert, hogy a Facebook a Hate Memes adatkészlettel a külső szakemberek segítségére is számítottak a további fejlesztések

[38] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online, 25–35, Florence, Italy. Association for Computational Linguistics., 25-26.o., 33. o. A

[39] https://qz.com/1249273/facebook-ceo-mark-zuckerberg-says-ai-will-detect-hate-speech-in-5-10-years Utolsó letöltés dátuma: 2023. december 15.

[40] FRA (2022) 54.o.

[41] https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/ Utolsó letöltés dátuma: 2023. december 15.

[42] FRA (2023) 67. o.

érdekében. Ehhez képest a 2022 novemberében elfogadott DSA rendelet 14. cikk (1) bekezdése előírja az online platformoknak, hogy a felhasználási feltételeikben tegyenek közzé információkat a tartalmak moderálása céljából használt irányelvekről, eljárásokról, intézkedésekről és eszközökről, beleértve az algoritmikus döntéshozatalt és a humán felülvizsgálatot is.

## 4. Az algoritmikus rendszerek hatása a véleménynyilvánítás szabadságára

Az online platformok moderálási tevékenységük körében amennyiben olyan technológiát alkalmaznak, amelyek a sértő beszédet ismerik fel, a fent kifejtett problémák következtében a tartalmak túl- és az alulkorlátozásának veszélye is felmerül, amelyek egyaránt sérthetik a felhasználók alapvető jogait, így többek között a véleménynyilvánítás szabadságához való jogot. Az automatizált tartalommoderáló rendszerek használatának térhódítása a kutatások eredményei alapján éppen a túlkorlátozásra hívja fel a figyelmet. Tehát az olyan algoritmusok, amelyek célja a gyűlöletkeltő beszéd terjedésének megállítása, valójában felerősíthetik az egyes elfogultságokat, a diszkriminációt, tekintettel az olyan technikai korlátokra, hogy az algoritmusok nem ismerik fel a kontextust, és megjelölik vagy akár blokkolják az adott csoportoktól származó tartalmakat is, ugyanis túlérzékenyek bizonyos csoportazonosító kifejezésekre, azaz potenciálisan elfogultak, amely a hiányos képzési adatkészleteknek és a képzési adatkészletekben felmerülő valós világ torzításainak egyaránt köszönhető lehet. Így például egy afroamerikai felhasználó által *„Wussup, n*gga!"* (Mi a helyzet, n*gga!") tartalommal közzétett, a szakirodalmak szerint nem sértő tweetet a Google Perspective API-ja 90 %-os *„toxicity score"*-ral illet.[43] A Facebook esetében például egy afroamerikai nő „Dear White People" („Kedves Fehér Emberek") tartalmú bejegyzését a vállalat eltávolította, ám a fehér bőrű ismerősei ugyanezzel a tartalommal készült bejegyzését viszont nem.[44] Hasonló lehet a helyzet abban az esetben is, ha egy eszközt olyan alulreprezentált felhasználócsoportra alkalmaznak, amely nem egyezik szorosan a képzési adatokban szereplő csoportokkal, mivel ez is hibás besorolásokhoz vezethet.[45] Tehát az algoritmikus rendszerek számos esetben álpozitív (egy adott tartalmat tévesen minősítenek elfogadhatatlannak) vagy éppen álnegatív (tévesen nem minősíti elfogadhatatlannak) eredményt adnak, amely tartalommoderálás esetén tévesen korlátozhatja a véleménynyilvánítás szabadságához való jogot, vagy akár dermesztő hatást okozhatnak, tekintettel arra, hogy egyes egyének és csoportok online részvételi hajlandósága csökkenhet.[46]

Az is lényeges, hogy az algoritmikus rendszerek gyors megoldásként alkalmazottak, azonban a felmerülő technikai problémák és az azok következtében hozott téves értékelések és az azok alapján automatikusan, humán moderátor felügyeletét nélkülözően alkalmazott következmények szintén sérthetik a véleménynyilvánítás szabadságához való jogot, tekintettel a következmények, döntések elleni fellebbezéssel kapcsolatos aggályokra is. Mindemellett az átláthatósági és elszámoltathatósági kérdések ezen eljárások vonatkozásában még inkább aggályos lehet, mint a humán moderátorok eljárása esetén. Ugyanis az algoritmikus, gépi tanulási rendszerek bonyolultak, nehezen érthetőek, működésük kevéssé megismerhető az egyes eljárások tekintetében.

Ugyanakkor azt is fontos rögzíteni, hogy természetesen az algoritmikus rendszerek alkalmazása pozitív következményekkel is járhat, ugyanis az online platformok működésében

[43] A Perspective egy ingyenes API, amely gépi tanulást használ a kommentek azonosítására. A Pespective API azt a százalékot adja vissza, amely annak valószínűségét jelenti, hogy valaki mérgezőként fogja fel a szöveget. https://perspectiveapi.com/; Sap (2019), 1668. o.

[44] https://revealnews.org/article/how-activists-of-color-lose-battles-against-facebooks-moderator-army/ Utolsó letöltés dátuma: 2023. december 19.

[45] LLansó, Hoboken, Leerssen, Harambam (2020) 8. o.

[46] Uo. 9. o.

olyan központi szerepet játszanak, amelyek az önkifejezést és az információhoz való hozzáférést segítik, feltételeiket megteremtik. Természetesen számos kérdést is felvet azonban arra vonatkozóan, hogy a felhasználók milyen tartalmakat látnak, esetleg befolyásolja-e őket bizonyos esetekben.[47]

**5. A humán moderátorok szerepe**

Az automatizált, algoritmikus rendszerek ugyan gyorsak és nagy mennyiségű tartalom kezelésére képesek, a fentebb kifejtett problémákkal járnak. Ezért az online gyűlöletbeszéd detektálásában a humán felügyelet és moderáció kritikus elem a helyes döntéshozatalban. A DSA is ezt erősíti meg, a 20. cikk (6) bekezdésében amikor a belső panaszkezelési rendszerrel kapcsolatban utal arra, hogy ezt a tevékenységet nem lehet kizárólag algoritmussal végezni, hanem képesítéssel rendelkező személy felügyelete szükséges hozzá. Bár az idézett rendelkezés általában vonatkozik a panaszkezelésre, álláspontunk szerint a gyűlöletkeltő (vagy a platform algoritmusa által annak vélt) tartalmak esetében ennek betartása kiemelten fontos. A tartalmak egy része összetett, a kontextus figyelembe vétele mellett kell értékelni azokat, amelyekre jelenleg az algoritmikus rendszerek önmagukban nem alkalmasak.

A humán moderátorok képesek lehetnek arra, hogy felismerjék a kulturális és társadalmi kontextust, a szóvicceket, a szlenget vagy akár a közlés szándékát, ám ehhez átfogó szakértelemmel kell rendelkezniük. Ugyan a szerepük elengedhetetlen, számos kihívással is szembesülnek az eljárásuk során. Így nem feltétlenül értenek meg különböző kulturális kontextusokat a humán moderátorok sem, tekintettel az eltérő háttérre és az anyanyelvüktől eltérő tartalmak azonosításának feladatára. Az anyanyelvüktől eltérő nyelveken való munka során fennáll annak a veszélye, hogy az apróbb árnyalatbeli különbségeket, szlengeket nem ismerik fel, így téves következtetésre juthatnak.[48] Mindezek érdekében elengedhetetlen fontosságú a humán moderátorok folyamatos képzése.

A már algoritmikus rendszereknél is bemutatott elfogultság a humán moderátoroknál is megjelenhet, a saját értékítéletük és szociológiai hátterük, eltérő nézeteik, értékrendjük befolyásolhatja a moderációs döntéseiket, ugyan kisebb mértékben, mint a gépi algoritmusok alkalmazása során. A humán moderátorok is szembesülnek azzal a nehéz feladattal, hogy meg kell találniuk az egyensúlyt a tartalmak moderálása és a véleménynyilvánítás szabadsága között, amelyhez a megfelelő képzésük elengedhetetlen.

Továbbá számos alkalommal került nyilvánosságra, hogy milyen rövid idő alatt milyen nagy mennyiségű tartalomról kell dönteniük - az érzelmi és a pszichológiai megterhelést nem is említve -, amely túlterheltséget okozhat. a Facebook kiszivárgott belső adatkezelési útmutatói alapján a moderátoroknak csupán nagyjából 10 másodpercük van arra, hogy egy adott tartalom jogellenességéről döntsenek.[49] A megterhelő munkakörülményekkel és a mentálhigiénés kihívások szembesülnek a moderátorok, akik közül sokan kiszervezett vállalkozók a globális délről. Az eljáró moderátorok számos megterhel tartalmat, így például lefejezést tartalmazó videókat, pornográf vagy erőszakos felvételeket látnak és kezelnek. Ugyanakkor kevés adat áll rendelkezésünkre a munkakörülményeikről, valamint eljárásukról.[50]

További aggály lehet a humán detektálás kapcsán, hogy a „szilícium-völgyi elitek” viszonylag homogén csoportja határozza meg a detektáláshoz szükséges globális

[47] Uo. 9. o.

[48] Schmidt, Wiegan (2017)

[49] https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence Utolsó letöltés dátuma: 2023. december 5.

[50] Andrew Arsht, Daniel Etcovitch: The Human Cost of Online Content Moderation. https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation Utolsó letöltés dátuma: 2024. január 20.

beszédszabályokat,[51] amely szintén a társadalmi és kulturális kontextus figyelmen kívül hagyásának veszélyét hordozhatja magában. Mindez pedig hibás értékelésekhez, döntésekhez vezethet, veszélyeztetve a véleménynyilvánítás szabadságának csorbulását.

## 6. Jövőbeli kihívások

Az online világ gyors fejlődése, a felhasználók és a tartalmak növekvő száma, sokszínűsége amellett, hogy a jelenlegi megoldások, eszközök hatékonyságának kérdését vetik fel, újabb és újabb kihívások elé állítja mind a jogalkotókat, mind a technológiai szakembereket a megengedhetetlen tartalmak, így a gyűlöletbeszéd detektálása és moderálása kapcsán. Így továbbra is kihívást jelent a gyűlöletbeszédre való globális jogi válasz, harmonizáció hiánya, valamint a véleménynyilvánításhoz való jog és a moderálás közötti egyensúly megtalálása a hatékonyság figyelembe vétele mellett, illetve a gyorsan változó technológiák és új platformok megjelenéséhez könnyen alkalmazkodó jogi keretek megteremtése, azok hatékony alkalmazása. Tekintettel kell lenni arra is, hogy az egyes országok különböző jogi megközelítéseket alkalmaznak, eltérő jogi hagyományokkal, azonban együttműködéssel egyetemes megoldást szükséges találni a jogi definíciók és keretrendszer tisztázására. A platformoknak, a jogalkotóknak és a témában jártas civil szervezeteknek aktívan együtt kell működniük ennek, valamint a stratégiák és azok végrehajtása érdekében.

A technológia oldalán számos probléma került azonosításra, így az algoritmikus rendszerek beágyazott elfogultsága, az értékelések megfelelőségének kérdése, különösen az online gyűlöletbeszéd komplex és érzékeny megítélésében, ezért az algoritmikus rendszerek fejlesztése során ezen szempontokat és az alkalmas, lehetséges technikákat kell figyelembe venni. Továbbá egyre több tartalom multimediális, tehát képeket vagy videókat kellene hatékonyan detektálniuk és értékelniük a rendszereknek, amely további megoldásra váró kihívást jelent. Ezzel együtt az algoritmikus rendszerek transzparenciájának növelése további feladatként jelölhető meg.

## 7. Összegzés

Ugyan a platformok a rájuk nehezedő szabályozási nyomás következtében az aggályos tartalmak kezelésére, ideértve az online gyűlöletbeszédet, olyan eszközöket alkalmaznak a gyorsaság érdekében a mennyiség következében is, amelyek számos hibalehetőséget rejtenek magukban a pontos detektálás és értékelés vonatkozásában. Mindez számos aggályt vet fel a tartalmak proaktív, valamint bejelentést követő értékelésével összefüggésben.

A kifejtett problémákból kitűnik, hogy az algoritmikus rendszerek még nem használhatóak biztonsággal a gyűlöletbeszéd automatikus moderálására, a jelenleg is fennálló korlátokra számos kutatás és tanulmány rámutatott azzal, hogy önmagában ezen rendszerek jelenleg nem jelenthetnek megoldást.

Az egyik legnagyobb probléma, hogy az algoritmikus rendszerek nem tudják a humán moderáláshoz hasonlóan figyelembe venni a kontextust, ideértve a közlőt és a tartalom címzettjét is, a különböző nyelveket és dialektusokat, továbbá alapvetően hiányoznak azok a reprezentatív adatkészletek is, amelyek az algoritmusok kidolgozásához, tanulásához elengedhetetlenek lennének.

Az algoritmusok használata ahelyett, hogy valódi megoldást kínálna, önmagában, megfelelő biztosítékok, humán moderátorok felügyelete nélkül alkalmazva tovább növeli a tartalmak moderálásának átláthatatlanságát, és a tartalmak túl- vagy éppen alulkorlátozásának

[51] Gorwa (2020) 2. o.

veszélyét, a véleménynyilvánítás szabadságához való jog érvényesülését vagy éppen az egyének, csoportok jogait korlátozva. tovább növelheti a méltányossághoz és az igazságossághoz kapcsolódó kihívásokat.

Ezért a humán moderátorok szerepe elengedhetetlen lenne, ám a tartalmak és felhasználók mennyiségét figyelembe véve erre nincs elegendő kapacitás. Azonban az algoritmikus rendszerek a moderátorok munkáját támogatva fontos elemként szerepelhetnek a platformok tartalomdetektálási és moderálási tevékenységében.

Az Európai Unió esetében egyértelmű törekvést láthatunk az új technológiák jelentette problémák kezelésére, az online tér biztonságosabbá tételére. Az elmúlt években számos olyan szabályozás készült el: már a GDPR esetében is törekvés volt az online óriásvállalatok számára is visszatartó erőt jelentő szabályozás megalkotása, de a DSA-DMA rendeletcsomag vagy a mesterséges intelligencia rendelet egyértelműen idesorolható. Fontos azonban megjegyezni, hogy a DSA-DMA rendeletek esetében a hatályba lépés óta nagyon kevés idő telt el (a mesterséges intelligencia rendelet végleges szövege pedig jelen kézirat lezárásakor még nem is elérhető), így arra csak a gyakorlat fog tudni választ adni, hogy az ön- és társszabályozási megoldások, a szigorúbb auditálási, jelentéstételi kötelezettség mennyiben fog kielégítő megoldást jelenteni a problémákra.

**MTA Law Working Papers**

**Kiadó: Társadalomtudományi Kutatóközpont (MTA Kiválósági Kutatóhely**

**Székhely: 1097 Budapest, Tóth Kálmán utca 4.**

**Felelős kiadó: Boda Zsolt főigazgató**

**Felelős szerkesztő: Kecskés Gábor**

**Szerkesztőség: Hoffmann Tamás, Lux Ágnes, Mezei Kitti**

**Honlap: http://jog.tk.mta.hu/mtalwp**

**E-mail: mta.law-wp@tk.mta.hu**

**ISSN 2064-4515**