Az online platformok diskurzusának moderációja és biztonsági aggályai

Pintér Melinda

Pázmány Péter Katolikus Egyetem, Budapest, Magyarország
E-mail: pinter.melinda@btk.ppke.hu

Beérkezett: 2022. április 30.; Elfogadva: 2022. május 16.; Online megjelent: 2022. július 26.

Összefoglalás

A tanulmány arra a kérdésre keresi a választ, hogy milyen biztonsági aggályokat jelent az, hogy az online platformok diskurzusát egyre nagyobb mértékben mesterséges intelligencia (MI) ellenőrzi. A kérdés megválaszolásához röviden bemutatja a téma szakirodalmának és a releváns empirikus kutatások eredményeit, valamint a legfontosabb értelmezési kereteket és fogalmakat. Megállapítja, hogy az online diskurzusok MI általi moderálásának vonatkozásában meghatározhatók egyrészt belülről kifelé, másrészt pedig kívülről befelé irányuló biztonsági aggályok és fenyegetések. A belülről kifelé irányuló fenyegetések az MI gyakorlati működése, így pedig a diskurzust befolyásoló szerepe okán jelentkeznek, a kívülről befelé irányuló fenyegetések pedig az MI mögött álló emberek döntéseiből adódnak.

Kulcsszavak: online diskurzus, online platformok, mesterséges intelligencia, moderáció, biztonsági aggályok

Moderation and security concerns about discourses on online platforms

Melinda Pintér

Pázmány Péter Catholic University, Budapest, Hungary

Summary

The premise of this study is that online platforms and other public spaces in cyberspace can provide a framework for citizens to discuss issues affecting society as a whole or wider groups in society, contributing to the strengthening of democratic processes. The study seeks to answer the question of what security concerns are posed by the fact that the discourse of online platforms is increasingly moderated by artificial intelligence. To answer this question, the study briefly presents the results of the Hungarian and international literature on the topic and relevant empirical research, as well as reviews the theoretical, conceptual and interpretive frameworks whose presentation and conceptualization of relevant concepts are essential for processing the topic. It introduces the concepts of online platform, online discourse and moderation, and demonstrates applications of artificial intelligence in content moderation, with particular reference to related security concerns and their interpretation. The study then reviews the aspects related to the moderation of the discourse on online platforms by artificial intelligence that may also broaden the horizons of the security interpretation of the issue. The study finds that in connection with the moderation of online discourses by artificial intelligence, threats can be identified in two directions: in the form of "inside out" and in the form of "outside in" security concerns and threats. In the first case, the "inside out" threats are factors which arise from the practical operation of artificial intelligence and thus from its role in influencing discourse. In such way these "inside out" threats can also endanger the world beyond the discourses on online platforms, such as society, politics or even the economy. In the second case, the risks associated with the operation of artificial intelligence in terms of "outside-in" threats are not due to the built-in flaw of the algorithm, i.e. the uniqueness of its operation. In this case, the threats come from the decisions of the owners, developers, managers of artificial intelligence, the people behind artificial intelligence. The study concludes that in the moderation of discourse in artificial intelligence-based decision-making systems, such as online platforms, it is not enough to base accountability on platform conscience and self-restraint. It is absolutely necessary to subject platforms that use artificial intelligence in content moderation to a higher level of external and objective control, and in this connection to create a legal interpretation framework that defines the main guidelines for discourses moderated by artificial intelligence on online platforms and also enforces these guidelines.

Keywords: online discourse, online platforms, artificial intelligence, moderation, security concerns

Bevezetés

Az internet megjelenése nyomán lehetővé vált digitális újgenerációs technológiák alkalmazása, az infokommunikációs eszközök általános elterjedése és a technológiai fejlődés által biztosított lehetőségeknek köszönhetően a hétköznapi tevékenységek kibertérbe való áttétele számos változást és új, korábban nem tapasztalt biztonsági kihívást hozott.

Az infokommunikációs technológiák fejlődése és egyre általánosabb alkalmazása alapjaiban változtatta meg a mindennapokat, amely változások közül jelen tanulmány vizsgálati fókusza szempontjából leginkább a társadalmi nyilvánosság online térben való megjelenése lényeges. Arról a folyamatról van szó, amely során a nyilvánosság mint intézményesített társadalmi tér – amely magában foglalja az összes nem magánjellegű, minden állampolgár és az állampolgárok közössége számára nyilvános információt és jelenséget, és amelyben a polgárok a politikai intézmények feletti ellenőrzést gyakorolják, vagy éppen a társadalom szempontjából fontos kérdéseket vitatnak meg –, egyre inkább az online tér lett. Az információs társadalom elméletei már egészen a XX. század 60-as éveitől kezdődően *(Machlup 1962; Masuda 1980; Toffler 1980; Castells 1996)* hangsúlyozták az információ mint a társadalom alapjává váló tényező, valamint az ebben a kommunikációs folyamatok megkerülhetetlen jelentőségét az új társadalmi-konceptuális keretben, amely később az internet és az infokommunikációs eszközök elterjedésével még nagyobb jelentőséget kapott. Hamarosan megfogalmazódtak olyan kérdések, hogy tényleg megvalósulhat-e a „kiberdemokrácia" *(Poster 1997),* az internet teret adhat-e például a demokratikus gyakorlatok szempontjából elengedhetetlen nyílt diskurzusoknak *(Gimmler 2001),* és hogy az online tér vajon betölthetti-e a nyilvánosság új terének funkcióját *(Papacharissi 2002);* de megjelentek arra vonatkozó kutatások is, hogy az új virtuális közegben milyen a kibernyilvánosság megvalósulása *(Dahlberg 2007)*.

Ezekhez a megközelítésekhez kapcsolódó kihívások immáron nemcsak a földrajzi, illetve fizikai térben jelentkeznek, hanem a kibertérben is, bizonyos esetekben a biztonság koncepciójának teljesen újszerű értelmezését, illetve a biztonságfogalom új megközelítéseinek használatát szükségeltetve. A biztonság mint kizárólagosan a megszerzett értékek (acquired values), azaz például a gazdasági, kulturális és morális javak fenyegetettségének akár objektív, akár szubjektív hiányára *(Wolfers 1952)* vonatkozó koncepció már az információs társadalom 1970-es években történő megjelenésekor is kevés volt a változó és az infokommunikációs eszközök által egyre táguló világban megjelenő biztonsági fenyegetést jelentő tényezők leírására. A Barry Buzan, Ole Waever és Jaap de Wilde által 1998-ban felvázolt, katonai, politikai, gazdasági, társadalmi és környezeti szektorokat magában foglaló biztonságfogalom *(Buzan–Waever–de Wilde 1998)* sem volt elegendő arra, hogy leírja a kibertérben megjelenő új veszélyeket. Az internet megjelenésével és egyre szélesebb körben történő elterjedésével azonban olyan új biztonságfogalmak is a gyakorlati használatba kerültek, mint az információbiztonság, majd a kiberbiztonság *(von Solms–van Niekerk 2013)*, amelyek már reflektáltak a legújabb típusú biztonsági fenyegetésekre. A XXI. századi virtuális térben pedig olyan, a kibertér minden eddiginél szélesebb körű használatából adódó kérdésekkel kell szembenézni, mint például az, hogy az előnyök mellett milyen hátrányokat és biztonsági veszélyeket rejthet magában, ha az online platformok diskurzusának moderációját a mesterséges intelligenciára (MI) bízzuk.

A tanulmány abból a „technopozitivista" jellegű premisszából indul ki, hogy az online platformok és a kibertér egyéb nyilvános terei keretet tudnak biztosítani az egész társadalmat, illetve a társadalom szélesebb csoportjait érintő kérdések megvitatásához, a vélemények cseréjéhez és az egyes vélemények ütköztetéséhez, ezáltal hozzájárulva a demokratikus folyamatok erősítéséhez. Ugyanakkor látni kell, hogy az online térben megvalósuló diskurzusok csak akkor lesznek megfelelőek, ha valamilyen felügyeletük, azaz moderációjuk valósul meg. Egyre több online platform ezt az ellenőrzést a „human in the loop", azaz az ember döntéshozatalban való részvétele mellett az MI segítségével valósítja meg. Ebből kiindulva pedig a tanulmány arra a kérdésre keresi a választ, hogy milyen biztonsági aggályokat jelent az, ha az online platformok diskurzusát egyre nagyobb mértékben az MI ellenőrzi és – adott esetben – hatással is van rá.

Jelen tanulmány elsőként bemutatja a vizsgálat során használt módszert, valamint áttekinti azokat az elméleti és fogalmi, valamint értelmezési kereteket, amelyek bemutatása és vonatkozó fogalmainak konceptualizálása elengedhetetlenül szükséges a téma feldolgozásához, úgymint az online platform, az online diskurzus és a moderáció fogalmai, valamint az MI alkalmazási formáinak bemutatása, különös tekintettel a hozzá kapcsolódó biztonsági aggályokra és ezek értelmezésére. Ezt követően a tanulmány áttekinti azokat a főbb, az online platformokon megvalósuló diskurzus MI általi moderációjához kapcsolódó szempontokat, amelyek szintén a kérdés biztonsági vetületű értelmezésének horizontját tágíthatják. Végül pedig a vizsgálat eredményeinek összegzése kapcsán levonja a következtetéseket, és válaszol a vizsgálat kezdetén felvetett kérdésekre.

Fogalmak és megközelítések

A tanulmány leíró és magyarázó jelleggel dolgozza fel a bevezetőben felvázolt témát. Célja egyrészt az, hogy a releváns szakirodalom és a legfontosabb, vonatkozó kutatási eredmények áttekintésével bemutassa az online platformok diskurzusának moderációja és biztonsági kérdéseinek vizsgálata szempontjából leglényegesebb fogalmakat; másrészt pedig az, hogy mindezek alapján, a vizsgált fogalmak és elméleti megközelítések, valamint empirikus kutatási eredmények alapján feltárja a téma

vizsgálati fókuszának összefüggéseit, és válaszoljon a bevezetőben megfogalmazott kérdésre. Jelen fejezetben a legfontosabb fogalmak bemutatása és a köztük lévő összefüggések rövid áttekintése történik meg.

Online platformok és online diskurzus

A digitális vagy online platformok definiálása nehéz feladat, ugyanis méretük, tevékenységük, illetve a különféle ágazatokhoz való kapcsolódásuk tekintetében annyifélék lehetnek, hogy az egyetlen közös pont talán csak az bennük, hogy mind az interneten, a virtuális térben jelennek meg. Az online platform ugyanis egyszerre jelenthet egy kis látogatottságú webhelyet, vagy egy világvállalat virtuális térben megjelenő felületét, napi többmillió látogatóval. A platformok az általuk kínált szolgáltatásokat tekintve is sokfélék lehetnek: az internetes keresőmotoroktól kezdve az online piactereken, a videómegosztó platformokon és a közösségimédia-oldalakon át egészen a közösségi gazdaság és az online játékok platformjaiig terjednek.

Az OECD 2019-ben kiadott, *An Introduction to Online Platforms and Their Role in the Digital Transformation* című anyagában található definíció szerint az online platform olyan digitális szolgáltatás, amely elősegíti az interakciót két vagy több különálló, de egymástól függő felhasználói csoport között – amelyek lehetnek akár cégek, akár magánszemélyek –, akik a szolgáltatáson keresztül az interneten lépnek kapcsolatba egymással. Az Európai Bizottság állásfoglalása az online platformok lényegét főleg az új piacok létrehozásában és az értékteremtésben látja, ehhez kapcsolódóan pedig 2016-os, *Online platformok és a digitális egységes piac: Lehetőség és kihívás Európa számára* című közleményében kiemeli, hogy az online platformok az adatvezérelt innováció letéteményeseiként működnek, amelyek bővítik a választékot, ezáltal javítva a versenyképességet és emelve a fogyasztói jólétet. A közlemény hangsúlyozza még továbbá azt is, hogy gazdasági előnyeik mellett a platformok lehetővé teszik az információhoz való könnyebb hozzáférés által azt, hogy a polgárok, de különösen a fiatalabb generációk aktívabban vegyenek részt a társadalomban és a demokráciában.

Ződi Zsolt az „óceán metaforát" alkalmazza az online platformok fontosságának az érzékeltetésére. Ennek értelmében szerinte a digitális tér már olyan méretűvé növekedett, hogy a felhasználók már nem tudták átlátni a digitális tér óceánjának végtelenségét, a platformok pedig arra az igényre reflektálva születtek meg, hogy mintegy kapaszkodóként az óceánból kiemelkedve segítsék a digitális térben való eligazodást *(Ződi 2018: 97)*. Ugyanakkor az online platformok definiálására tett számos kísérlet nagy hangsúllyal szerepelteti a fogalommagyarázatban a platformok gazdasági-üzleti-szolgáltató tevékenységét és a digitális piac vonatkozásában betöltött szerepét. Belényesi Pál szerint „[m]indegyik platformra jellemző, hogy a fogyasztókhoz való hozzáférést több csatornán keresztül biztosítja (pl. mobil, fix internet), emellett maga a platform is már egy többoldalú piac, illetve a felhasználók figyelmére építi üzleti modelljét" *(Belényesi 2015: 9)*, Bruno Jullien és Wilfried Sand-Zantman értelmezésében pedig az online platformok olyan kétoldalú („two-sided") piacok, amelyek közvetítőként különböző gazdasági szereplőket kapcsolnak össze *(Jullien–Sand-Zantman 2021)*.

Ebből a néhány példából látható talán az, hogy az online platformoknak nincs olyan egységes, átfogó és minden részletre kiterjedő definíciója, amely a teljesség igényével terelné egy fogalmi meghatározás alá ezen platformok kereskedelmi, kommunikációs vagy éppen szórakoztató céllal létrejött megjelenési formáit. Ezért szükséges egy olyan „munkadefiníció", ami jelen tanulmány értelmezési keretében határozza meg azt, hogy mit ért az online platform fogalma alatt. Jelen tanulmány vizsgálati fókusza az online platformok kommunikatív funkcióira koncentrál. Ezért azokra a digitális keretben létrejövő, ilyen funkcióval is rendelkező terekre koncentrál, amelyek lehetőséget adnak arra, hogy általuk valamilyen diskurzus valósuljon meg online, függetlenül attól, hogy felhasználók vagy valamilyen szolgáltatásnyújtás keretében hozták létre.

Ehhez kapcsolódóan szükséges néhány gondolat erejéig körbejárni az online diskurzus fogalmát is. Online diskurzus alatt egyértelmű módon az online környezetben, online platformokon megvalósuló dialógus értendő, ugyanakkor érdemes lehet specifikálni ezek bizonyos jellegzetességeit. Az online diskurzusok számítógép által közvetített kommunikációként (computer mediated communication – CMC) is értelmezhetők *(Corich-Kinshuk-Lynn 2004)*, amelynek tartalmát egy internethozzáféréssel rendelkező felhasználó bármikor megtekintheti, és ő maga is hozzájárulhat az online diskurzushoz. Az online diskurzusok lehetnek egyidejűek (azaz szinkronban lévők), amikor az online platformon a beszélgetés „real-time" történik, valamint nem egyidejűek (azaz aszinkronban lévők), például amikor egy hozzászólásra akár hónapokkal vagy évekkel később érkezik válasz. Az online diskurzusok nem csak és kizárólag szövegalapúak: tartalmazhatnak hangulatjeleket, képeket, GIF-eket, videókat stb., amelyek a szöveghez hasonlóan integráns részét képezik az online megvalósuló kommunikációnak.

Az online platformokon megvalósuló online diskurzusok ugyanakkor nem minden esetben jók és minőségiek, hiszen olyan hozzászólásokat is tartalmazhatnak, amelyek sértők, károsak, vagy illegális, jogellenes magatartást valósítanak meg. Ezek azért jelentenek különösen nagy gondot, mert ha az online platformok által megvalósított közeget a társadalmi nyilvánosság terepének tekintjük, akkor ezek a problémás tartalmak megakadályozzák azt, hogy a platformokon megvalósuló diskurzus a demokratikus potenciál szempontjából megfelelő legyen, azaz elérje a célját. Emellett pedig a felhasználók által hozzáférhető tartalmakat érintő egyéb olyan tényezők is hatás-

sal lehetnek az online diskurzusokra, mint például a moderáció – vagy ennek végletes és központilag kontrollált formája, az állami (hatósági) cenzúra –, ami háttérbe szorít vagy teljesen kiiktat és láthatatlanná tesz bizonyos tartalmakat, amelyek így nem tudják formálni a diskurzust.

Moderáció és MI

Az online platformokon megvalósuló diskurzusok „mederben tartását" a tartalommoderáció teszi lehetővé, ami gyakorlatilag a diskurzus lefolytatására lehetőséget biztosító online felületeken (internetes oldalak, közösségi média site-ok és egyéb, például hozzászólási lehetőséggel rendelkező weblapok) közétett, a felhasználók által gyártott tartalmak, azaz UGC-k (user generated content) szűrésének szervezett gyakorlatát jelenti az adott webhely által támasztott moderálási elvek alapján.

A moderálás igen változó lehet mind alapelveit, mind pedig gyakorlati megvalósítását tekintve. A moderálás történhet például az állami szabályozás által jogellenesnek ítélt tartalmak (pl. gyűlöletbeszéd) kiszűrése miatt, de a platformok emellett magánszabályozást is alkalmazhatnak például gazdasági érdekeik védelme érdekében, azért, hogy a felhasználóik ne találkozzanak sértő, felháborító vagy zavaró tartalmakkal *(Koltay 2019)*. A tartalom moderálását végezhetik önkéntesek, de a platformok ki is szervezhetik azt magánszemélyek vagy cégek részére, akik díjazásban részesülnek szolgáltatásaikért. Ez utóbbi gyakorlatot kereskedelmi tartalommoderálásnak vagy CCM-nek (commercial content moderation) nevezik *(Roberts 2017)*. A platformok a tartalommoderálás során hagyományosan az emberi erőre támaszkodnak, de a tartalmak mennyiségi növekedésével ez már nem költséghatékony és egyre kevésbé eredményes, ezért az MI-t használó technikák alkalmazása egyre elterjedtebbé vált. Ezek a megoldások, például a gépi tanulás (machine learning – ML) olyan algoritmusok gyakorlati használatát jelentik, amelyek megadott instrukciók, például kulcsszavak alapján automatikusan moderálják a platformon megjelenő tartalmat.

Ugyanakkor ez nem jelenti az emberi moderátorok munkájának végét: a szövegkörnyezet-függő, vagy tartalmi szempontból árnyaltabb, de ugyanolyan sértő jelentéssel bíró megfogalmazásokat ugyanis továbbra is csak az emberi erő tud kiszűrni. A moderálás, azaz az ember és az MI együttműködése az elő- vagy az utómoderálás formájában valósul meg. Az előmoderálás során az MI moderálja a tartalmat a közzététel előtt: a nem káros tartalmakat „átengedi", míg a károsnak ítélt tartalmakat törli, amennyiben pedig kétségei támadnak, úgy megjelöli a tartalmat, amit egy ember felülvizsgál. Az utómoderálás során a felhasználók által ártalmasnak jelölt tartalmakat vizsgálja felül utólag az MI az előmoderálás során megvalósuló folyamatot követve.

Biztonsági aggályok

Az MI mint az egyre növekvő mértékű online UGC-tartalom kezelésére adott válasz már régóta egyfajta aduász a döntéshozók és a technológiai iparág szereplői kezében, akik azt ígérik, hogy az MI-k tartalommoderáló képessége már a közelében jár annak, hogy mindent moderáljon a platformokon megjelenő gyűlöletbeszédtől kezdve a zaklató kommenteken át egészen a terrorista propaganda terjesztéséig. Mark Zuckerberg 2018-ban, az Egyesült Államok Szenátusa ellőtti meghallgatásán például azt mondta, hogy 5-10 év múlva olyan technológia válik elérhetővé, hogy az MI képes legyen már a posztolás előtt felismerni és moderálni a gyűlöletbeszédet tartalmazó hozzászólásokat. Ezek a könnyelmű ígéretek azonban nem számolnak azzal, hogy még a legkifinomultabb MI-k sem lesznek képesek bizonyos jellegű tartalmak pontos azonosítására *(Llansó 2020),* legyenek azok akár károsak, akár ártalmatlanok.

Ezért jelent óriási kihívást egy, az online platformokon megjelenő tartalmak moderálására alkalmasnak tekintett MI kifejlesztése. Már csak azért is, mert rossz vagy nem elégséges működése súlyos negatív következményekkel is járhat, hiszen az interneten, az online platformokon megvalósuló diskurzusok a társadalmi nyilvánosság, ezáltal pedig a közbeszéd részei, tartalmuk moderálása tehát közvetlenül és jelentős mértékben érinti a közérdeket *(Nahmias–Perel 2021)*.

Az MI általi tartalommoderáció az online platformokon zajló diskurzus kapcsán ugyanakkor konkrét biztonsági aggályokat is felvet. Olyan veszélyekről van itt szó, amelyek az MI hibás működéséből kifolyólag befolyásolhatják egyrészt a diskurzus alakulását összességében véve, másrészt pedig – ebből adódóan – akár a diskurzusban részt vevők nézeteit is. Alapvetően kétfélék lehetnek azok a biztonsági vetületű problémák és fenyegetések, amelyek az MI tartalommoderáló tevékenységéből, illetve ennek a tevékenységnek a nem optimális működéséből adódhatnak:

1) ha az MI nem tudja felismerni és kiszűrni a káros, sértő vagy veszélyes tartalmakat;

2) ha olyan tartalmakat jelez károsnak és szűri ki őket, amelyek egyébként nem azok.

Mindkét esetben rövidebb vagy hosszabb távon, de komoly következményekkel és biztonsági kockázatokkal járhat az MI tévedése.

Az első esetben egyértelmű a probléma: ha például a gyűlöletbeszéd nem sértő szavakkal, hanem utalásokkal, hasonlatokkal, metaforákkal valósul meg, amelyek csak a szöveg kontextusának egészét ismerve értelmezhetők, akkor elképzelhető, hogy ezeket az árnyalatnyi különbségeket az MI nem veszi észre, ezért pedig nem jelzi sértőként az adott tartalmat. Ha a gyűlöletbeszéd ilyen formában „egérutat nyer", és az uszító szavak virágnyelven ugyan, de nagyobb tömegeket érnek el, az akár a társadalom bizonyos csoportjai ellen irányuló gyűlölet-bűncse-

lekmények elkövetését alapozhatja meg, vagy radikalizálhat az adott csoport ellen irányuló gyűlölettel addig nem találkozó felhasználókat.

A második esetben talán egy lépéssel korábbról kell kezdeni a probléma értelmezését. A tartalommoderálásban az automatizálás az online platformok diskurzusában minden megnyilvánulást egyfajta előzetes értékelésnek tesz ki, méghozzá olyan módon, hogy az MI az automatizált tartalomszűrés során – ahogyan azt már korábban is említettük – figyelmen kívül hagyja a szöveg nyelvi, társadalmi, történelmi és egyéb releváns összefüggéseit. Mindez pedig nagy kockázatot jelent a véleménynyilvánítás szabadságára nézve *(Llansó et al. 2020)*. Az ilyen típusú tartalomszűrés pedig ütközik egyrészt a véleménynyilvánítás szabadságának a nemzetközi emberi jogi jogban foglalt védelmével *(Llansó 2020)*, de különösen akkor jelent problémát mindez, ha az MI az amúgy nem káros vagy sértő tartalmakat kiszűri az online platformok diskurzusából.

Az ENSZ, az EBESZ, az Amerikai Államok Szervezete és az Afrikai Bizottság különmegbízottjai – ahogyan 1999 óta minden évben – 2011-ben is elfogadtak egy közös nyilatkozatot, amelyben nemzetközi normákat fogalmaztak a szólásszabadság kulcsfontosságú területeivel kapcsolatban. 2011-ben a közös nyilatkozat az interneten megvalósuló szólásszabadság okán fogalmazódott meg *Joint Declaration on Freedom of Expression and the Internet* címmel, amely kimondta, hogy a kormányok vagy a kereskedelmi szolgáltatók által bevezetett tartalomszűrő rendszerek, amelyek nem a végfelhasználó irányítása alatt állnak, az előzetes cenzúra egy formáját valósítják meg, és nem igazolhatók a véleménynyilvánítás szabadságának korlátozásaként. Azaz a különmegbízottak szerint a szólásszabadság az internetre is vonatkozik, és hogy a szólásszabadság korlátozása az interneten csak nemzetközileg elfogadott előírásoknak megfelelően és bizonyos érdekek védelmében lehetséges. Felismerték tehát, hogy – hiába egyre kifinomultabbak az erre a célra létrehozott technológiák – a tartalomszűrés alapvetően veszélyt jelenthet a véleménynyilvánítás szabadságára, hiszen egyfajta előzetes cenzúraként, a tartalom előzetes korlátozásaként működik, függetlenül a használt eszköz „pontosságától" *(Llansó 2020)*.

A mesterséges intelligencia mint a biztonság és a kockázat megvalósítója

Egy 2021-ben publikált kutatásban *(Röttger et al. 2021)* az Oxfordi Egyetem és az Alan Turing Intézet kutatói az MI tévedhetőségét vizsgálták a gyűlöletbeszéd detektálásában. A legjobb tartalomszűrő MI-k különféle, a természetes nyelvben a gyűlöletbeszéd árnyalatainak felismerésére irányuló tesztelése során kiderült, hogy a gépek még mindig egyértelmű nehézségekkel küzdenek akkor, amikor a sértő és nem sértő tartalmak között kell dönteniük. Ez talán nem is annyira meglepő, viszont az ilyen és ehhez hasonló vizsgálatok segíthetnek abban, hogy felismerjük: hol és hogyan hibázik az MI – ez pedig elengedhetetlen a további fejlesztésekhez.

Mindazonáltal ennek fényében – összegezve az eddigiekben már hosszabban vagy rövidebben érintett problématerületeket – érdemes áttekinteni azt, hogy melyek azok a legfőbb kihívások, amelyekkel az online platformok tartalmának moderálása során az MI-rendszerek továbbra is szembesülnek.

- *Az előzetes moderáció mint cenzúra:* Az előzetes tartalommoderáció legnagyobb problémája, hogy a tartalom szűrése esetén általa egyfajta cenzúra valósul meg, ami veszélyt jelent a véleménynyilvánítás szabadságára nézve az egyének szintjén, a platformok vonatkozásában pedig hatással lehet a diskurzusok alakulására.
- *A kontextus és a nyelvi árnyalatok problémája:* A szöveg tartalmának dekódolásához elengedhetetlenül szükséges a szavak jelentése mellett a közlő szándékának, a nyelvi árnyalatoknak, vagy például a kulturális kontextusnak az ismerete. Az MI-k ennek felismerésében még kihívásokkal küzdenek, különösen pedig bizonyos tartalomtípusok, például az audiovizuális tartalmak esetében jelent nagy nehézséget a moderálás. De az MI-k nem teljesítenek jól a sértő tartalmak felismerésében például (a szándékos vagy nem szándékos) elgépelések, a GIF-ek, mémek esetében sem. Ez különösen akkor jelent problémát, ha emiatt legális tartalmakat, például jogsértésekről szóló posztokat, tudósításokat szűr ki az MI, gyűlöletkeltőnek gondolva őket.
- *Elfogult MI:* Az MI a tartalommoderálás során olyan elveket követ, amilyenekre tanítjuk. Amennyiben az MI a tanulási folyamat során elfogult, diszkriminatív tartalmakat magában foglaló adathalmazon tanulta meg a tartalomszűrés irányelveit, akkor „élesben" is ezeket az elveket fogja követni, ez pedig értelemszerűen olyan tartalmak moderálásához és meghagyásához fog vezetni, ami újratermeli és erősíti a platformokon megvalósuló diskurzus problémáit.
- *Átláthatatlan döntések és a transzparencia hiánya:* Fontos lenne, hogy egyértelmű legyen az, hogy az online platformok a tartalommoderálás során milyen elveket alkalmaznak és hogyan hozzák meg a döntéseket adott tartalmak szűrésekor. Ezzel szemben viszont inkább az látható, hogy a legnagyobb vállalatok tartalomszűrése egyfajta „black box"-ként működik, azaz szinte semmilyen információ vagy tudás nem szivárog ki az MI-k ilyen jellegű működéséről a platformokon. Ennek leginkább az az oka, hogy a techóriások óriási erőforrásokat invesztálnak a fejlesztésekbe, és nem szeretnék úgymond kiadni az általuk elért eredményeket; viszont hosszú távon alapvető problémákat vet fel az, ha ismeretlen elvek és módszerek alapján történik az online platformok tartalmának moderálása. Ez akkor is probléma lehet, ha a felhasználók kifogással szeretnének élni egy, a platform által károsnak ítélt tartalom moderálásával szemben. Ha nem ismertek a moderálás alapját képező elvek, akkor nagyon nehéz „fellebbezni" ellenük.

- *Nyelvi változatosság (hiánya):* A globális nyelvrendszerben az olyan, százmilliók, vagy éppen milliárdok által beszélt nyelvek esetében, mint az angol, a spanyol vagy a mandarin, a problémás tartalmak felismerésében az MI-k már jól teljesítenek. A vállalatok és a fejlesztők rendkívüli kapacitásokat fektetnek ezeknek a sokak által beszélt nyelveken íródott tartalmaknak a minél pontosabb moderálására, hiszen ezek hatalmas piacot jelentenek a számukra. Ugyanakkor a kevesek által beszélt nyelvek esetében ebből adódóan a szűrés sokkal pontatlanabb lehet.
- *Nem rendeltetésszerű használat:* Probléma lehet, ha az alkalmazás túlterjeszkedik saját határain és már nem csak és kizárólag azokra a célokra használják, amelyekre megalkották: például az online platformok tartalomszűrésére kifejlesztett MI-t a felhasználók szélesebb körű megfigyelésére, magánbeszélgetéseik monitorozására is lehet használni.

Konklúzió

Az online platformok diskurzusának moderációjához kapcsolódó legfontosabb fogalmak áttekintését, valamint a tartalommoderációt végző MI-k jelentette biztonsági kockázatok és ezen fenyegetések gyakorlati megjelenésének bemutatását követően időszerű válaszolni a bevezetőben feltett kérdésre, hogy milyen biztonsági aggályokat jelent az, ha az online platformok diskurzusát egyre nagyobb mértékben az MI ellenőrzi.

Az online diskurzusok MI általi moderálásának vonatkozásában meghatározhatunk egyrészt belülről kifelé, másrészt pedig kívülről befelé irányuló biztonsági aggályokat és konkrét fenyegetéseket.

Az első esetben, a belülről kifelé irányuló fenyegetések kapcsán olyan tényezőkről kell beszélnünk, amelyek az MI gyakorlati működése, így pedig a diskurzust befolyásoló szerepe okán jelentkeznek, és amelyek így az online platformokon megvalósuló diskurzusokon túli világot, például a társadalmat, a politikát vagy éppen a gazdaságot veszélyeztethetik. Olyan „beépített hibákról" van itt szó, amelyek az MI tartalomszűrő működéséből adódóan vannak hatással a diskurzusokra, ezáltal a diskurzusban részt vevőkre. Ilyen tényezők például az MI-k által az online platformok diskurzusaiban végzett előzetes moderáció, ami megakadályozza, hogy bizonyos – nem sértő, nem káros és legális – tartalmak, hozzászólások megjelenjenek. De szintén ebben a keretben jelentkezhet problémaként az, hogy az MI még mindig nehézségekkel küzd a szöveg kontextusának vagy a nyelvi árnyalatoknak az értelmezésében. Az MI tévedései, azaz a moderálást nem szükségeltető tartalmak kiszűrése, vagy az egyébiránt moderálandó tartalmak „átengedése" a résztvevőkre gyakorolt káros hatásokon túl megváltoztathatja a diskurzus értelmezési keretét és kontextusát is, valamint teret engedhet a sértő, káros vagy illegális megnyilvánulásoknak, ami egyértelmű biztonsági kockázatokat jelenthet.

A második esetben, a kívülről befelé irányuló fenyegetések vonatkozásában az MI működését kísérő kockázatok alapvetően nem az algoritmus beépített hibájából, azaz működésének egyediségéből adódnak, hanem az MI tulajdonosainak, fejlesztőinek, irányítóinak, az MI mögött álló embereknek a döntéseiből. Ha az MI-t használó platform úgy dönt, hogy elfogult MI-t hoz létre és így moderálja az online diskurzust, akkor a kárt alapvetően nem az MI, hanem az emberi tényező okozza. Szintén ugyanez a helyzet akkor, ha az MI-t nem rendeltetésszerűen, például tartalommoderációra használják, hanem a felhasználók magánbeszélgetéseinek megfigyelésére. Ezek a döntések ugyanolyan biztonsági fenyegetéseket hordoznak magukban, mint az első esetben, hiszen ugyanúgy az MI működéséből adódóan jelentenek problémát; de a felelősség itt a gép mögött álló emberé.

Mindebből az látható, hogy az MI-alapú döntéshozatali rendszerek, például az online platformok diskurzusának moderációjánál az elszámoltathatósághoz nem elegendő a platformok lelkiismeretére és önkorlátozására alapozni. Mindenképpen szükséges az MI-t a tartalommoderálásban használó platformokat magasabb szintű, külső és objektív ellenőrzésnek alávetni, ehhez kapcsolódóan pedig egy olyan jogi-értelmezési keretet megalkotni, ami meghatározza az online platformokon megvalósuló, az MI által moderált diskurzusok legfontosabb irányelveit, és lehetővé teszi azok kikényszerítését is.

Irodalomjegyzék

Belényesi P. (2015) A digitális piacok időszerű versenyjogi vonatkozásai. Róma, John Cabot University

Buzan, B., Waever, O. & de Wilde, J. (1998) Security: A new framework for analysis. Lynne Rienner Publishers

Castells, M. (1996) The Information Age: Economy, Society and Culture. Vol. 1. The Rise of the Network Society. Oxford, Blackwell

Corich, S., Kinshuk, H. L. & Lynn, M. (2004) Assessing discussion forum participation: In search of quality. International Journal of Instructional Technology & Distance Learning, Vol. 1. No. 12. pp. 3–12.

Dahlberg, L. (2007) Rethinking the fragmentation of the cyberpublic: from consensus to contestation. New Media & Society, Vol. 9. No. 5. pp. 827–847.

Európai Bizottság (2016) A Bizottság közleménye az Európai Parlamentnek, a Tanácsnak, az Európai Gazdasági és Szociális Bizottságnak és a Régiók Bizottságának. Online platformok és a digitális egységes piac: Lehetőség és kihívás Európa számára. https://eur-lex.europa.eu/legal-content/HU/TXT/PDF/?uri=CELEX:52016DC0288&from=HU [Letöltve: 2022. 03. 15.]

Gimmler, A. (2001) Deliberative democracy, the public sphere and the internet. Philosophy & Social Criticism, Vol. 27. No. 4. pp. 21–39.

Jullien, B. & Sand-Zantman, W. (2021) The economics of platforms: A theory guide for competition policy. Information Economics and Policy, Vol. 54. 100880. https://doi.org/10.1016/j.infoecopol.2020.100880

Koltay A. (2019) A social media platformok jogi státusa a szólásszabadság nézőpontjából. In Medias Res, Vol. 8. No. 1. pp. 1–56.

LaRue, F., Mijatović, D., Botero-Marino, C. & Tlakula, F. P. (2011) Joint Declaration on Freedom of Expression and the Internet. https://www.osce.org/files/f/documents/e/9/78309.pdf [Letöltve: 2022. 03. 15.]

Llansó, E. J. (2020) No amount of "AI" in content moderation will solve filtering's prior-restraint problem. Big Data & Society, Vol. 7. No. 1. 2053951720920686. https://doi.org/10.1177%2F2053951720920686

Llansó, E., van Hoboken, J., Leerssen, P. & Harambam, J. (2020) Artificial intelligence, content moderation, and freedom of expression. One in a series: A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. https://lirias.kuleuven.be/retrieve/594053 [Letöltve: 2022. 03. 15.]

Machlup, F. (1962) The Production and Distribution of Knowledge in the United States. Princeton, New Jersey, Princeton University Press

Masuda, Y. (1980) The Information Society as Post-industrial Society. Tokyo, Institute for the Information Society

Nahmias, Y. & Perel, M. (2021) The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. Harvard. Journal on Legislation, Vol. 58. No. 1. pp. 145–194.

OECD (2019) An Introduction to Online Platforms and Their Role in the Digital Transformation. Paris, OECD Publishing. https://doi.org/10.1787/53e5f593-en

Papacharissi, Z. (2002) The virtual sphere: The internet as a public sphere. New Media & Society, Vol. 4. No. 1. pp. 9–27.

Poster, M. (1997) Cyberdemocracy: The Internet and the Public Sphere. In: Holmes, D. (ed.): Virtual Politics: Identity and Community in Cyberspace. London, SAGE Publications Ltd., pp. 212–228.

Roberts, S. T. (2017) Content Moderation. In: Schintler L. & McNeely C. (eds): Encyclopedia of Big Data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_44-1

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H. & Pierrehumbert, J. B. (2021) HateCheck: Functional tests for hate speech detection models. In: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021) https://doi.org/10.48550/arXiv.2012.15606

von Solms, R. & van Niekerk, J. (2013) From information security to cyber security. Computers & Security, Vol. 38. pp. 97–102. https://doi.org/10.1016/j.cose.2013.04.004

Toffler, A. (1980) The Third Wave. New York, William Morrow & Company, Inc.

Wolfers, A. (1952) "National security" as an ambiguous symbol. Political Science Quarterly, Vol. 67. No. 4. pp. 481–502.

Ződi Zs. (2018) Platformok. Robotok és a jog. Budapest, Gondolat Kiadó