A mesterséges intelligencia felé fordulás a kormányzásban
Online kommunikáció

Gollatz, Kirsten; Beer, Felix; Katzenbach, Christian

Publikációs verzió / Published Version Sonstiges / egyéb

KIRSTEN GOLLATZ, FELIX BEER, CHRISTIAN KATZENBACH

A mesterséges intelligencia felé fordulás az online kommunikáció irányításában

Workshop jelentés

HIIG WORKSHOP JELENTÉS2018-09

ABSZTRAKT

Jelenleg intenzív vita folyik a mesterséges intelligencia (AI) kutatásának technológiai fejlődéséről és annak különböző társadalmi területeken és összefüggésekben való alkalmazásáról. Ebben az összefüggésben a média és a kommunikáció az egyik legkiemelkedőbb és legvitatottabb terület. Botok, hangalapú asszisztensek, automatizált (ál)hírgenerálás, tartalommoderálás és -szűrés - mindezek példák arra, hogy a mesterséges intelligencia és a gépi tanulás hogyan alakítja át a digitális kommunikáció dinamikáját és rendjét.

Márciusban20 az 2018Alexander von Humboldt Intézet az Internetért és a Társadalomért az Access Now civil szervezettel közösen egynapos szakértői workshopot rendezett "Az online kommunikáció irányítása a mesterséges intelligencia felé fordul" címmel. Berlinben a tudományos élet, a politika, a civil társadalom és az üzleti élet nemzetközi szakértői gyűltek össze, hogy megvitassák a mesterséges intelligencia technológiákkal, a gépi tanulási rendszerekkel, a tartalom moderálásában való alkalmazásuk mértékével, valamint az internetes társadalmi kommunikáció irányítására szolgáló mesterséges intelligencia-rendszerek helyzetének és jövőbeli hatásának megértéséhez szükséges különböző megközelítésekkel kapcsolatos összetett társadalmi-technikai kérdéseket és problémákat.

Ez a műhelybeszélgetésről szóló jelentés összefoglalja és dokumentálja a szerzőknek a megbeszélésekből levont főbb következtetéseit. A viták, a felvetett észrevételek és kérdések, valamint a szakértők válaszai szintén beépültek a jelentésbe. A jelentést szétosztották a munkaértekezlet résztvevői között. A jelentés célja, hogy hozzájáruljon a mesterséges intelligenciáról és a kommunikáció irányításáról szóló diskurzus aktuális szempontjaihoz.

KULCSSZAVAK

Tartalmak moderálása, gépi tanulás, platformirányítás, társadalom a hurokban, mesterséges intelligencia, közösségi média

KÖZLEMÉNYEK

A szerzők mindegyike a berlini Alexander von Humboldt Internet és Társadalom Intézetben dolgozik vagy dolgozott korábban.

A szakértői workshopot és e jelentés közzétételét a Volkswagen Alapítvány (Hannover, Németország) támogatta.

TARTALOM

1 BEVEZETÉS

A "mesterséges intelligencia" és a "gépi tanulás" rendszerei átalakítják és átszervezik a társadalom különböző területeit. A digitális platformokon a tartalom moderálása és a kommunikáció irányítása az automatizált döntéshozatali rendszerek egyik kiemelkedő, de egyre inkább vitatott alkalmazási területévé vált.

Az olyan nagy technológiai vállalatok, mint a Facebook, a Twitter, a YouTube vagy a Google alakítják a világ nagy részén a kommunikációs ökoszisztémát. Az okostelefonok és a táblagépek, a keresőmotorok és a közösségi média nagyjából felváltották a hagyományos médiát, mint az információ elsődleges kapuit. Miközben ez új lehetőségeket teremtett az emberek számára, hogy világszerte különböző módon kapcsolódjanak egymáshoz, a felhasználók számára lehetőséget kínál arra is, hogy kifogásolható tartalmakat töltsenek fel, beleértve a gyermekbántalmazásról és az indokolatlan erőszakról készült képeket, valamint a zavaró, gyűlöletkeltő üzeneteket.

Különösen a félretájékoztatás és a gyűlöletbeszéd gyorsan növekvő politikai hatása miatt egyre többen követelik, hogy az online platformok akadályozzák meg és távolítsák el a problémás tartalmakat. A kormányok világszerte - többek között Franciaország, Vietnam, Oroszország, Szingapúr és Venezuela - szabályozási politikát kezdeményeztek az általuk jogellenesnek ítélt online beszéd korlátozására. A német "Netzwerkdurchsetzungsgesetz" (hálózati végrehajtási törvény) például az egyik ilyen kísérlet a platformok nemzeti jogérvényesítésének javítására azáltal, hogy a platformok üzemeltetőit kötelezi a német jog szerint jogellenesnek minősülő tartalmak gyors felderítésére és eltávolítására. Közösen elfogadott beszédnormák és koherens szabályozási keretek hiányában ezeket a kormányzati szabályozási politikai kezdeményezéseket a határokon átnyúló tartalomszabályozás transznacionális kihívásainak összefüggésében kell vizsgálni (Gollatz, Riedl & Pohlmann, 2018).

E politikai nyomás és a platformokra naponta feltöltött hatalmas mennyiségű tartalom miatt a vállalatok a tartalom moderálásának nagyobb mértékű automatizálása felé fordultak. A gépi tanulási technológiákat rutinszerűen úgy mutatják be, mint a gyűlöletbeszéd, a félretájékoztatás és a szerzői jogok megsértésének felderítésére és szűrésére szolgáló mindenre kiterjedő megoldást. Az algoritmikus döntéshozatal ilyen mértékű kiterjesztése azonban számos problémával jár. Az ilyen mesterséges intelligenciával működő rendszerek átláthatatlan megvalósítása, homályos definíciói és az elszámoltathatóság hiánya olyan problémákat okozhat, mint a túlzott blokkolás vagy az elfogult döntéshozatal. Szakértők és aktivisták arra figyelmeztetnek, hogy a mesterséges intelligencia által vezérelt megoldások elhamarkodott bevezetése káros hatással lehet a szólásszabadságra és az információhoz való egyenlő hozzáférésre az interneten.

2 A NAP PROGRAMJA

Márciusban az 20,2018,Alexander von Humboldt Intézet az Internetért és a Társadalomért és az Access Now egynapos transzdiszciplináris workshopot szervezett "A mesterséges intelligencia felé fordulás az online kommunikáció irányításában" címmel. Berlinben a tudományos élet, a politika, a civil társadalom és az üzleti élet nemzetközi szakértői gyűltek össze, hogy megvitassák a technológiai fejlődést, a mesterséges intelligencia alkalmazásának mértékét, valamint a mesterséges intelligencia-rendszerek helyzetének és jövőbeli hatásának megértéséhez szükséges megközelítések körét az internetes kommunikáció irányításában.

Már20,ciusi műhelyprogram 2018

Idő	Ülés	Tematikus hatókör
9:00 am	Üdvözlés és bemutatkozás	
9:45	Kick-off nyilatkozatok	- Malavika Jayaram: *Napalm, árnyalat és nem hot dogok* - a gyűlölködés elleni küzdelem kiegyensúlyozásának gyakorlatáról tartalom és védelem a szólásszabadság Ázsiában - Nick Feamster: A platformok manipulálásáról és szűréséről: *AI and the Future of Free Expression Online* - on manipulation and filtering of platforms
10:30 1. **ülés:**	Tartalom felderítése és osztályozása	- A kifogásolható tartalom felderítése és annak technikai megvalósítása - A változás mozgatórugói és a mesterséges intelligencia korlátai
12:00 pm2. **ülés**	: Emberek és gépek - munkamegosztás, gyakorlatok	- Emberek ÉS gépek bevonásával végzett tartalommoderálási gyakorlatok - Az emberi hibák mérséklése, de egyben bizonyíték arra is, hogy több emberi döntéshozatalra van szükség.
14:00 3. **ülés:**	Politikai és irányítási eszközök	- A mesterséges intelligencia és a platformok irányítása - Viták és szabályozási kezdeményezések uniós szinten - Konkrét alkalmazások és szabályozási beavatkozások
15:30 Távoli	beavatkozás Tarleton Gillespie által	- A tartalom moderálása, mint a platformok nélkülözhetetlen alapanyaga - A mesterséges intelligencia használata a platformok irányításának demokratizálására
15:45 4. **szekció:**	AI és a társadalom - a hurokban	- A társadalom bevonása a vitába, a mesterséges intelligenciával kapcsolatos felfogása és elvárásai - Etika az ML-rendszerek társadalomtudatos tervezésében
17:00 Távoli	beavatkozás David Kaye által	- Milyen mértékben keretezik az automatizált szabályok azt, hogy az emberek mit beszélhetnek és mit fogadhatnak, vagy mit gondolnak a globális szólásszabadság normáiról?
17:30 Befejezés és a jövő útja		- A felelősség nélküli hatalom paradoxona

AZ EMBERI KOMMUNIKÁCIÓ HASONLÓSÁGAI ÉS KÜLÖNBSÉGEI A GÉPI FORDÍTÁSHOZ	A tartalom megfelelő/nem megfelelő meghatározása	- Műszakilag kivitelezett fordítások

* Néhány szakértő távolról vett részt.

Négy ülésen a résztvevők különböző témákat vizsgáltak a mesterséges intelligencia online tartalmak moderálásával kapcsolatos következményeiről.

A résztvevők hátterének, tudásának és szakértelmének sokfélesége nagy előnye volt ennek a műhelynek. Ez tükrözte a mesterséges intelligencia és az online kommunikáció irányítása körül kialakuló nyilvános és szakértői viták összetettségét és a nézőpontok sokféleségét. A műhely különösen arra törekedett, hogy túllépjen a diszciplináris határokon, és nem akadémikusokat is bevont az üzleti élet, a politika és a civil társadalom képviselői közül.[1]

Azáltal, hogy a műhely különböző érdekelt felek széles körét hozta össze, a résztvevők interdiszciplináris nézőpontot kaptak, hogy megosszák egymással szakértelmüket és megvitassák a témával kapcsolatos nézeteiket. Az alábbiakban tematikusan bemutatjuk a műhelybeszélgetések releváns meglátásait és tanulságait.

[1] A részt vevő szakértők listája a mellékletben található.

3 JELENTÉS A TEMATIKUS ÜLÉSEKRŐL

1. ülés: A tartalom felismerése és osztályozása

Terjedelem: Az első ülés arra a kérdésre összpontosított, hogy miként figyelhetjük meg és írhatjuk le az online platformokon található problémás tartalmak felderítésére és osztályozására szolgáló, mesterséges intelligencia alapú megoldások felé történő jelenlegi fordulást. A megbeszélés e fejlődés mögött meghúzódó indokokat boncolgatta, és feltárta a technológiák képességeit és korlátait, valamint a nagyobb társadalmi következményeket.

Közreműködik: Frank (Google), Emma Llansó (Center for Democracy & Technology), Fabrizio Augusto Poltronieri (De Montfort University), Betty van Aken (Beuth University) és Zeerak Waseem (University of Sheffield).

Moderátor: Christian Katzenbach (HIIG)

A workshop kiindulópontja egy vita volt arról, hogy milyen tényezők állnak az online platformok növekvő figyelmének hátterében a mesterséges intelligencia rendszerek iránt a tartalom moderálásában. A résztvevők egyetértettek abban, hogy e fejlődés egyik kulcstényezője minden bizonnyal az utóbbi időben egyre növekvő nyilvános nyomás a problémás tartalmakkal szembeni gyors fellépés érdekében. A félretájékoztatás és a gyűlöletbeszéd észlelt elterjedtsége a közelmúlt politikai eseményei, például a Brexit, Donald Trump megválasztása és általánosabban a jobboldali populizmus térnyerése olyan széles körű sürgősségérzet a nem megfelelő bűnüldözés és a szabályozás a digitális nyilvánosságban. Az új jogalkotási kezdeményezések nyomást gyakorolnak az online platformokra, hogy eltérjenek korábbi, meglehetősen semleges álláspontjuktól, és proaktívan lépjenek fel a problémás tartalmakkal szemben.

AZ EMBERI TARTALOMELLENŐRÖKKEL SZEMBEN AZ AUTOMATIZÁLT RENDSZEREK ELŐNYE ÉS VONZEREJE A SKÁLÁZHATÓSÁGUKBAN REJLIK.

Ez azt jelenti, hogy a platformoknak óriási mennyiségű a feltöltött tartalom mennyisége napról napra. Az emberi tartalomellenőrökkel szemben az automatizált rendszerek előnye és vonzereje egyértelműen a skálázhatóságuk. Az ilyen rendszerek azt ígérik, hogy sokkal egyszerűbbé, gyorsabbá és olcsóbbá teszik a tartalommoderálási folyamatot, mintha emberi munkaerőt alkalmaznának.

Bár az automatizált szűrést jelenleg elsősorban az emberi tartalomellenőrök munkájának kiegészítésére használják, az iparág nagy elvárásokat támaszt azzal kapcsolatban, hogy az automatizált döntéshozatal a belátható jövőben képes lesz-e az emberi moderátorok árnyalt ítélőképességét megismételni. A mesterséges intelligencia-alkalmazások kutatásába és fejlesztésébe történő nagy vállalati és állami beruházások ellenére egyes szakértők ezeket az elvárásokat irreálisnak és túlzottan optimistának tartják. Szerintük az

automatizált tartalommoderációs rendszerekre való törekvést a mesterséges intelligencia iránti lelkesedés szélesebb légkörében kell szemlélni. Ezzel a technológiai megoldásokon alapuló elképzeléssel szemben a workshop résztvevői rámutattak azokra a veszélyekre és hátrányokra, amelyeket az automatizált online kommunikáció szűrése hozhat magával. Az ezzel kapcsolatban felvetett aggályok között szerepelt a túl széles körű cenzúra,

a szólás- és egyesülési jogok megsértése, valamint a kisebbségekkel és a nem angolul beszélőkkel szembeni elfogult döntéshozatal.

Ezt követően az ülés fókusza a jelenlegi tartalommoderációs gyakorlatok képességeinek és korlátainak vizsgálatára helyeződött át (Duarte, Llansó & Loop, 2018). Az automatizált tartalomszűrés nem újdonság. Az évek során számos eszközt vetettek be a tartalom elemzésére és szűrésére, többek között a spam-felismerésre vagy a hash-illesztésre szolgáló eszközöket. Ezek az eszközök bizonyos élesen meghatározott kritériumok alapján azonosítják a nem kívánt tartalmakat, amelyek korábban megfigyelt kulcsszavakból, mintákból vagy metaadatokból származnak.

Az automatizált közösségi média moderációs eszközök hatékonysága azonban nagymértékben függ attól, hogy képesek-e pontosan elemezni és osztályozni a tartalmakat a kontextusukban. A szöveg jelentésének elemzésére való képesség rendkívül fontos a kétértelmű esetekben történő fontos megkülönböztetéshez, azaz a gyűlöletbeszéd és az irónia megkülönböztetéséhez. E feladathoz az iparág mostanában egyre inkább a gépi tanuláshoz fordul, hogy programjaikat úgy képezzék ki, hogy azok még inkább érzékenyek legyenek a kontextusra.

A résztvevők ezután megvitatták e jelenlegi megközelítés lehetőségeit és korlátait. Az automatizált tartalommoderációs rendszerek sikerét általában a pontossági arányok alapján értékelik, amelyek azt mutatják, hogy a rendszer megítélése átlagosan mennyire felel meg az emberi döntéshozatalnak. A magas pontossági arányok eléréséhez az algoritmikus képzésnek egy világosan meghatározott adattípusra kell összpontosítania. Ez azt jelenti, hogy az automatizált osztályozó és észlelő rendszereket általában kifejezetten ezeknek az eseteknek az értékelésére képzik ki, és ezért nem vihetők át más területekre. Minél többértelműbbé és kontextusfüggőbbé válnak azonban az osztályozási kritériumok, annál nehezebbé válik az algoritmusok pontos képzése. Egyrészt a mesterséges intelligencia megoldások hatékony eszközt jelentenek az olyan egyértelműen körülhatárolt esetek szűrésére, mint a gyermekpornográfia. Másrészt még az embereknek is nehézséget okoz bizonyos esetekben következetes ítéleteket hozni - például amikor egyértelmű különbséget kell tenni a politikai aktivizmus és az erőszakra való felhívás között -, és az automatizált rendszerek e tekintetben messze elmaradnak az emberektől.

AZ AUTOMATIZÁLÁS VALÓSZÍNŰLEG A REAKTÍV MODERÁLÁSRÓL A PROAKTÍV MODERÁLÁS FELÉ VALÓ ELMOZDULÁSHOZ VEZET, AMI AZ ELSZÁMOLTATHATÓSÁGOT, AZ ÁTLÁTHATÓSÁGOT ÉS A NYILVÁNOSSÁG RÉSZVÉTELÉT LÉTFONTOSSÁGÚVÁ TESZI.

A mesterséges intelligencia rendszerekből (legalábbis egyelőre) hiányzik az emberek nyelvi érzékenysége és szemantikai megértése, ami ehhez a nehéz feladathoz szükséges. Több résztvevő kifejtette, hogy a mai szűrési technikák többsége a tartalom bizonyos kulcsszavak alapján történő megjelölésére korlátozódik. A jelentés
A nyelv rendkívül érzékeny a kontextusra, és folyamatosan változik; egy szó jelentése gyökeresen megváltozhat, ha idővel különböző helyeken használják. Egy olyan tartalommoderációs rendszer, amely az osztályozást pusztán bizonyos kulcsszavakra alapozza, nem képes elérni ezt a komplexitási szintet, és fennáll annak a veszélye, hogy a kontextus hiányában váratlanul hamis pozitív és negatív eredményeket produkál.

A szerkezeti túlzárás elkerülése érdekében az emberi közreműködés következésképpen továbbra is

fényeges része a tartalom moderálásának, legalábbis a nagyon érzékeny kontextusú esetekben, hogy elkerülhető legyen a szerkezeti túlzárás. Ezen túlmenően aggályok merültek fel azzal kapcsolatban, hogy az automatizálás a képzési adatokban rejlő társadalmi torzítások és hibák miatt nem vezethet-e olyan csoportok további marginalizálásához, amelyek már most is diszkriminációval szembesülnek. Az online platformoknak ezeket a következményeket is figyelembe kell venniük mesterséges intelligencia rendszereik tervezési folyamatában.

Erre építve a vita azt a kérdést vetette fel, hogy milyen körülmények között tekinthető hasznosnak az automatizálás. Ez a kérdés nemcsak az ellenőrizendő tartalom típusától függ. Az automatizált rendszerek alkalmazása a tartalommoderálás folyamatának különböző szakaszai tekintetében is eltérő. A moderálás előtti, a moderálás utáni, a reaktív moderálás vagy az elosztott moderálás csak néhány az ebben az összefüggésben megvitatott fogalmak közül. Az automatizálás minden formája egyéni esélyekkel és kockázatokkal jár. Nagy volt az egyetértés abban, hogy az automatizálás valószínűleg a reaktív (azaz a felhasználó által jelzett) moderálásról a proaktív moderálásra (azaz az összes feltöltött tartalom elemzésére) való áttéréshez vezetne. A résztvevők kifejtették, hogy ez a forgatókönyv komoly aggályokat vet fel, és súlyosbítja az elszámoltathatóság és az átláthatóság hiányának meglévő problémáit. Erre válaszul a munkaértekezlet megvitatta a tartalommoderáció általános kialakításának és alkalmazásának demokratizálására irányuló lehetőségeket, biztosítva a folyamat érthetőségét és a nyilvánosság megfelelő részvételét.

2. ülés: Emberek és gépek - munkamegosztás és gyakorlatok

Terjedelem: A második ülésen azt vizsgáltuk, hogyan fog kinézni az emberek és a gépek közötti munkamegosztás a tartalom moderálásában a jövőben. Ezen az ülésen a résztvevők megvitatták, hogy a mesterséges intelligencia milyen lehetőségeket kínál az emberi munka helyettesítésére vagy segítésére a tartalom moderálási folyamatban.

Közreműködik: Roberts (UCLA), Jeremy Rollison (Microsoft), Mirko Vossen (die medienanstalten) és Jillian C. York (EFF).

Moderátor: Gollatz Kirsten (HIIG)

Mindennaposak azok az állítások, hogy a mesterséges intelligencia rendszerek elavulttá teszik az embert. Ezt a fekete-fehér érvelést számos szakértő megkérdőjelezi, akik kifejtik, hogy az automatizálás inkább átalakítja, mint helyettesíti az emberi munkát. Ugyanez igaz az online tartalom moderálására is. Kétségtelen, hogy az online platformok tartalmi ellenőrzésének és a sértő anyagok eltávolításának módja a változás küszöbén áll. Ellentétben azzal a vélekedéssel, hogy a mesterséges intelligencia teljesen kiszorítja majd az emberi felülvizsgálatot, résztvevőink között erős konszenzus alakult ki abban, hogy a hatékony moderálásnak belátható időn belül egy hibrid modellre kell támaszkodnia: míg egyes feladatok jól automatizálhatók, például a közösségi irányelveket egyértelműen sértő mondatok azonosítása vagy a gyanús esetek előzetes kiválasztása nagy mennyiségű adatból, szinte valós időben, mások továbbra is emberi ítélőképességet igényelnek majd - azaz a szürke területek eldöntéséhez a kontextuális tudás felhasználását. Röviden, az emberek és a gépek szinergikus kapcsolatra fognak támaszkodni. Széles körben elterjedt az a nézet, hogy a mesterséges intelligenciának csupán segítő technológiának kell lennie a skálázhatóság növelése, valamint az emberi hatékonyság és eredményesség javítása érdekében a tartalom megítélése során.

AZ AI RENDSZEREK A TARTALOM MODERÁLÁSÁBAN NEM FOGJÁK AZ EMBERT FELESLEGESSÉ TENNI, HANEM INKÁBB ÁTALAKÍTJÁK A TARTALOM MODERÁTOR MUNKÁJÁT.

A gépek és az emberek közötti megfelelő felosztás meghatározása tehát kihívást jelentő feladat lesz. Jelenleg a legtöbb nagy online platform azt állítja, hogy csak automatizáltan jelöli a tartalmakat, és a végső eltávolításról szóló döntést az emberi ellenőrökre bízza.[2] A tartalom moderálása azonban még mindig a következőktől függ alacsony képzettségű, többnyire indiai és Fülöp-szigeteki alvállalkozók által foglalkoztatott munkaerő, akiknek a bére jóval az átlagos Szilícium-völgyi technológiai dolgozók bére alatt van. Ezek a moderátorok naphosszat hatalmas mennyiségű tartalom áttekintésével töltik napjaikat, hogy eldöntsék, hogy azt el kell-e távolítani vagy sem, gyakran kétértelmű és átláthatatlan alkalmassági kritériumokat alkalmazva (Arsh & Etcovitch, 2018).

Megbeszélésünk felhívta a figyelmet arra, hogy egyre több bizonyíték utal arra, hogy a tartalmi moderáció a

[2] Lásd például: YouTube Official Blog. (2017. december 4.). A platformunkkal való visszaélések elleni munkánk kiterjesztése. Letöltve a https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html oldalról.

jelenlegi formájában jelentős pszichológiai kockázatoknak teszi ki a munkavállalókat (Gillespie, 2018).[3] Sokuknak elviselhetetlen számszerű kvótáknak kell megfelelniük, mivel olyan zavaró tartalmakat szűrnek, mint a lefejezés, az öngyilkosság vagy a pornográf videók. Oknyomozó újságírók hangsúlyozzák, hogy a moderátorok körében gyakran PTSD-szerű tünetek és más mentális egészségügyi problémák jelentkeznek ennek következtében. A jelenlévők kiemelték, hogy a meglévő rendszerek pontos értékelését nehezíti, hogy az online platformok csak nagyon kevés információt adnak a témáról. A vállalatok gyakran szándékosan átláthatatlanok, ellenállnak minden olyan kísérletnek, amelyet harmadik felek tesznek a gyakorlatuk ellenőrzésére, és titoktartási megállapodásokkal akadályozzák meg, hogy a munkavállalók beszéljenek a munkakörülményeikről.[4]

Számos megfigyelő problematizálta ezt a modellt, de csak kevesen javasoltak alternatív megoldásokat. A mesterséges intelligencia azonban lehetőséget nyithat az emberek és a gépek közötti új, szinergikusabb munkamegosztásra. A félautomata moderációs modellek nagyszámú alacsony képzettséget igénylő pozíciót helyettesíthetnek, és az emberi bírálóknak új szerep juthat, amelyet a mesterséges intelligencia egészít ki. Az AI-eszközök következő generációja a jelenleginél többféle jellemző alapján lesz képes azonosítani és értékelni a tartalmat, beleértve a tartalom forrását és kontextusát is. Ennek alapján egy relatív kockázati pontszámot lehet kiszámítani annak meghatározására, hogy valamit azonnal, felülvizsgálat után vagy egyáltalán nem kell-e közzétenni.

A gépi tanulással az eredmények felhasználhatók az algoritmusok autonóm optimalizálására, hogy folyamatosan javítsák pontosságukat. Ez a folyamat nagy mennyiségű tartalmat távolíthat el a nyomozók sorából, és lehetővé teheti a tartalommoderátorok számára, hogy az összetett szürke területeken a döntéshozatalra koncentráljanak. Speciális szakértelemmel, empátiával és kontextuális ismeretekkel járulhatnak hozzá a
megítélni ezeket a konkrét tartalomtípusokat. Ez azt jelentené, hogy valószínűleg azt látjuk majd, hogy a mai tartalommoderátorok helyét a pénzügyi nyomozókéhoz hasonló speciális képzéssel rendelkező tartalomnyomozók veszik át.[5] Ezek a tartalommoderátorok képzést kaphatnak a nyelvi, regionális, piaci, szabályozási és jogi sajátosságokra vonatkozóan, hogy jól tájékozottan tudjanak döntések a szürke zónás tartalmakkal kapcsolatban. Ez az átalakulás segíthet a gyorsan növekvő online platformoknak abban, hogy megfizethető költségek mellett, a kockázatok minimalizálása mellett, és az emberi ellenőrök karrierlehetőségeit jelentősen javítva, a tartalom moderálását is skálázni tudják.

AZ AUTOMATIKUS MODERÁCIÓS ESZKÖZÖKET A JÖVŐBEN TOVÁBB KELL FEJLESZTENI, HOGY A MODERÁTOROK KEVÉSBÉ LEGYENEK KITÉVE A ZAVARÓ TARTALMAKNAK.

Ennek ellenére számos hibrid modell még mindig jelentős kihívásokkal és megoldatlan kérdésekkel küzd. Míg a tartalommoderáció hatékonyságának és skálázhatóságának növelése elérhetőnek tűnik, a már meglévő problémák, mint például az átláthatatlanság vagy az elszámoltathatóság hiánya, továbbra is nagyrészt megoldatlanok, sőt tovább súlyosbodhatnak.

[3] Lásd még: Chen, A. (2017. január 28.), The Human Toll of Protecting the Internet from the worst of Humanity, The New Yorker. Letölthető: https://www.newyorker.com/tech/elements/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity; Buni, C. & Chemaly, S. (201613,. április), The Secret Rules of the Internet, The Verge. Retrieved from

https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech

[4] Powers, B. (2017. szeptember 9.), The Human Cost of Monitoring the Internet, The Rolling Stone. Retrieved from https://www.rollingstone.com/culture/features/the-human-cost-of-monitoring-the-internet-w496279

[5] Lásd még: Accenture: Tartalom-moderálás: A jövő bionikus. Letöltve: https://www.accenture.com/cz-en/_acnmedia/PDF-47/Accenture-Webscale-New-Content-Moderation-POV.pdf.

részleges automatizálással.[6] Számos aggály merült fel például azzal kapcsolatban, hogy az automatizálás tovább növelheti a moderálási folyamat átláthatatlanságát, és proaktívan hozzájárulhat a tartalom túlszabályozásához. Bár a továbbfejlesztett automatizált előválogatási eszközök csökkenthetik a zavaró tartalmaknak való általános kitettséget, a moderátoroknak valószínűleg továbbra is hatalmas mennyiségű ilyen anyaggal kell majd foglalkozniuk. Ezért a vállalatoknak fokozniuk kell erőfeszítéseiket, hogy egészséges munkakörnyezetet teremtsenek moderátoraik számára.

[6] Lásd még: Irányelvek a tartalom-moderálás jövőjéhez: Négy tudós és szószóló beszélgetése, 2018. február 1. Retrieved from https://atm-ucla2017.net/

3. ülésszak: Szakpolitikai és irányítási eszközök

Terjedelem: A harmadik ülésszak az online kommunikáció szabályozásának politikai és kormányzási eszközeivel foglalkozott a mesterséges intelligencia felé fordulással összefüggésben. A résztvevők először a szabályozási status quo-t, valamint a fokozott szabályozásra való jelenlegi törekvés indoklását és következményeit vitatták meg. A vita során áttekintették, hogy mi működött és mi nem működött eddig, majd a jobb szabályozási modellek lehetőségeit és kihívásait vizsgálták.

Közreműködik: Prabhat Agarwal (EU Bizottság), Eimear Farrell (Amnesty International), Amélie Heldt (Bredow Intézet), Michael Latzer (Zürichi Egyetem), Ramak Molavi (iRights), Erin Saltman (Facebook, távolról), Florent Thouvenin (Zürichi Egyetem) és Joris van Hoboken (Vrije Universiteit Brüsszel).

Moderátor: Hidvégi Fanny (Access Now)

Eddig a digitális platformok kevés ellenőrzésnek és csökkentett kormányzati beavatkozásnak örvendtek, így nagy mozgásteret kaptak a tartalommérséklési intézkedések végrehajtása tekintetében. Ez a régóta fennálló status quo azon a bináris feltevésen alapul, hogy a platformok vagy hírkiadóként működnek, akik teljes felelősséget viselnek a tartalmukért, vagy semleges közvetítőként, akiknek nincs jogi felelősségük (Tushnet, 2008). Általában az utóbbi modellt tekintették a jobb megközelítésnek, mivel úgy vélték, hogy ez erősíti az ágazati innovációt és a véleménynyilvánítás szabadságát a digitális szférában.

A problémás tartalmak miatt egyre több kritika éri az online platformokat és a politikai döntéshozókat egyaránt egyre nagyobb nyomás nehezedik a cselekvésre, ami új, ágazatközi konszenzushoz vezetett: a platformok tehetnek és kell is tenniük.

A CSELEKVÉSRE IRÁNYULÓ NYOMÁS, DE EGYÚTTAL EGY ÚJ, ÁGAZATKÖZI KONSZENZUS IS KIALAKULT: A PLATFORMOK TÖBBET TEHETNEK ÉS TÖBBET IS KELL TENNIÜK TARTALMAIK SZABÁLYOZÁSÁÉRT, AZ MI-RENDSZEREK HASZNÁLATÁVAL IS.

jobban szabályozzák a tartalmukat. A növekvő figyelem egyidejűleg a mesterséges intelligencia megoldások felé, ez a nyomás vezérelte a tartalom moderálási folyamatok jelenlegi algoritmikus fordulatát.

A tartalommoderációnak kényes egyensúlyt kell teremtenie a biztonságos online környezet megteremtése és a szólásszabadság biztosítása között. Egyesek úgy vélik, hogy a kormányoknak lépéseket kellene tenniük szolgáltatásaik szabályozása érdekében, és a problémás tartalmak ellenőrizetlen terjedése ellen energikusan küzdeni. Mások azzal érvelnek, hogy a platformoknak jobb, ha maguk szabályozzák a tartalmat, hogy elkerüljék az üzleti innovációra és a szólásszabadságra gyakorolt káros hatásokat. Milyen irányítási modell valószínűsíthető és előnyös az online kommunikáció szabályozására a jövőben?

Az elmúlt években a közösségi médiaplatformok egyre inkább önkéntes magatartási kódexek mellett kötelezték el magukat a problémás tartalmak elleni küzdelem érdekében, és fokozták az ilyen tartalmak proaktív megelőzésére és eltávolítására irányuló erőfeszítéseiket. A

A Twitter például decemberben új irányelveket 2017,vezetett be a zaklatás és a gyűlöletbeszéd megakadályozására, a [7]YouTube további emberi tartalommoderátorokat alkalmaz, és bővíti a megjelölési algoritmusokat, [8]és a Facebook is azt tervezi, hogy idénre20,000 növeli a tartalommoderátorok számát, Mark Zuckerberg vezérigazgató pedig kijelentette, hogy a Facebookon történő visszaélések orvoslása még személyes célja is a következő évre.Bár 2018.[9]ez a fejlődés széles körben elismerést vált ki, a kritikusok azt gyanítják, hogy ezek az erőfeszítések nem lesznek elegendőek egy egészséges és biztonságos online kommunikációs ökoszisztéma létrehozásához.

Emellett a kormányok egyre inkább előírják és ösztönzik a tartalom szigorúbb moderálását. A kemény szabályozással az a probléma, hogy könnyen a tartalom túlszabályozásához vezethet, és így drasztikusan korlátozhatja a szólásszabadságot a digitális szférában. A német NetzDG értelmében például a közösségi hálózatokat akár 50 millió eurós (kb. millió60 USD) bírsággal is büntethetik, ha az azonosított illegális gyűlöletbeszédet nem távolítják el órákon24 belül. Ez arra késztetheti az online platformokat, hogy a büntetés elkerülése érdekében túlreagáljanak, és felgyorsítsák az átláthatatlan és pontatlan automatizált tartalommoderációs rendszerek bevezetését. A platformok és a tartalom közötti kapcsolatokban való kormányzati szerepvállalás ellenére ezek a megközelítések még mindig nagymértékben a platformok önszabályozási mechanizmusaira támaszkodnak a problémás tartalmak azonosításához és eltávolításához szükséges mesterséges intelligencia-rendszerek kifejlesztésében.

A munkaértekezlet résztvevői arra a következtetésre jutottak, hogy a megfelelő irányítási modellek a köz- és a magánszféra beavatkozásainak kiegyensúlyozott keverékén alapulnak. A résztvevők általában egyetértettek abban, hogy a kormányzati beavatkozás csak olyan mértékben kívánatos, amennyire az önszabályozás nem képes kezelni a problémás tartalmak mérséklését. Ezért annyi állami szabályozásra van szükség, amennyire szükséges, de a lehető legkevesebbre. Természetesen sok vita van arról, hogy ez valójában mit is jelent. Hogyan szabályozhatjuk a tartalomáramlást anélkül, hogy az online platformok növekedését és innovációját akadályoznánk? Hogyan tükrözhetik a kormányzási modellek az automatizált tartalommoderáció átláthatóságával, legitimitásával és elszámoltathatóságával kapcsolatos közérdekű aggályokat? A többi érdekelt féllel szoros párbeszédet folytatva a politikai döntéshozóknak folyamatosan fel kell tenniük a következő kérdéseket maguk is ilyen kérdéseket tesznek fel az állami szabályozás és az önszabályozó rendszerek erősségeinek és hiányosságainak értékelésére, és ezáltal meghatározzák szerepüket az automatizált tartalommoderáció folyamatában.

A KÜLÖNBÖZŐ ORSZÁGOK ÉS KULTÚRÁK BESZÉDAKTUSAINAK KONTEXTUSFÜGGŐ KÜLÖNBSÉGEI RENDKÍVÜL BONYOLULTTÁ VÁLNAK, HA A PROBLÉMÁS TARTALMAK AUTOMATIKUS FELISMERÉSÉRE ALKALMAZZÁK ŐKET.

Ezen túlmenően a különböző szereplők közötti összetett kölcsönhatásról is szó esett az irányításban.
online kommunikáció. Sok más területtel párhuzamosan a digitális szférában a formális hatalom a központi államoktól a nemzetek feletti intézmények, a multinacionális technológiai óriások és a globalizált digitális civil társadalom felé oszlott szét. E különböző érdekelt felek érdekeinek összeegyeztetése nagyfokú összetettséget eredményez. Ha tehát az online kommunikáció hatékony és globális szabályozási rendszerének kialakítása a cél, milyen területi és intézményi szinteken lehetséges a döntéshozatal?

[7] https://blog.twitter.com/official/en_us/topics/company/2017/safetypoliciesdec2017.html

[8] https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html

[9] https://www.facebook.com/zuck/posts/10104380170714571?pnref=story

Ezt tovább bonyolítja az a tény, hogy a különböző országokban eltérő törvények és kulturális korlátozások vonatkoznak a beszédre. Például az "erőszakra való felbujtás" vagy a "gyűlöletbeszéd" másra utal Németországban - ahol a náci propaganda törvénytelen -, mint Spanyolországban - ahol a király megsértése törvénytelen. Ezek a kontextusfüggő különbségek rendkívül bonyolulttá válnak, ha a problémás tartalmak automatikus felismerésére alkalmazzuk őket. Már ezek az egészen egyszerű különbségek is azt mutatják, hogy lehetetlen egyetlen, univerzálisan használható tartalomosztályozót kiképezni. Ebben az esetben lényegében minden egyes joghatóságnak más-más tartalomosztályozóra lenne szüksége. A

A TARTALOMMODERÁCIÓ KÖZÖS KERETÉT NEM LEHET RÖGZÍTENI, HANEM FOLYAMATOSAN TÁRGYALNI KELL AZ ÖSSZES ÉRINTETT FÉLLEL.

a tartalom moderálásának közös keretét ezért nem lehet rögzített, de folyamatosan tárgyalni kell minden érintett féllel. érdekeltek. Ez jelentős kihívás elé állítja a politikai döntéshozókat: a nemzetközi irányítási modellnek eléggé specifikusnak kell lennie ahhoz, hogy szabályozási jogkört gyakoroljon, ugyanakkor eléggé adaptívnak is ahhoz, hogy figyelembe vegye a kontextustól függő árnyalatokat.

Összefoglalva, a résztvevők egyetértettek abban, hogy az online kommunikáció irányításának egy több érdekelt felet érintő és többszintű folyamatnak kell lennie, amelynek célja globális iránymutatások kidolgozása és a legjobb gyakorlatok cseréje az automatizált tartalom moderálásával kapcsolatos kihívások kezelése érdekében. Mivel ez a folyamat még mindig a fejlődés korai szakaszában van, résztvevőink rámutattak számos olyan akadályra, amelyek gátolják a hatékony irányítás előrehaladását:[10] Először is, hiányoznak az intézményi eljárások és platformok, amelyek megkönnyítenék a témáról folytatott ágazatközi beszélgetést. Másodszor, az online platformok jelenleg a tartalom moderálásának minden aspektusát ellenőrzik, és nagyon titokzatosak az irányítási modelljeikkel kapcsolatban. Különösen a releváns adatok és információk feletti monopóliumuk nehezíti a platformok által jelenleg alkalmazott eszközök ellenőrzését. Harmadszor, megbízható adatokhoz való hozzáférés nélkül nagyon nehéz közös kutatásokat folytatni ebben a témában. Végül, mivel nagyon kevés a tudás és az erőforrások megosztása, különösen a kisebb vállalkozások számára nehéz a semmiből kifejleszteni és betanítani saját algoritmikus moderációs rendszereiket.

[10] További információért lásd: Stern Center for Human Rights and Business (2017. november). Káros tartalom: The Role of Internet Platform Companies in Fighting Terrorist Incitement and Politically Motivated Disinformation, Fehér könyv. Letöltve a http://www.stern.nyu.edu/experience-stern/faculty-research/harmful-content-role-internet-platform-companies-fighting-terrorist-uszítás-és-politikusan

4. ülés: AI és társadalom a hurokban: Társadalmi következmények

Terjedelem: A negyedik ülés az automatizált tartalommoderációs folyamatok társadalmi következményeire terelte a vitát. Míg az előző vita a különböző irányítási modellek képességeire és korlátaira összpontosított, ez az ülés felvetette azt a kérdést, hogy miként lehet az algoritmikus hatóságot átlátható, demokratikus és elszámoltatható módon kialakítani.

*Közreműködik: (*Berkman Klein Center), Lisa Gutermuth (Ranking Digital Rights), Aphra Kerr (Maynooth University), Tilo Mentler (University of Lübeck), Kevin Morin (Institut National de la Recherche Scientifique), Jörg Pohle (HIIG) & Matthias Spielkamp (Algorithm Watch).

Moderátor: Christian Katzenbach (HIIG)

Egyrészt az online platformokra egyre nagyobb nyomás nehezedik, hogy fokozzák a tartalom belső szabályozását. Másrészt az emberi jogok védelmezői és aktivistái aggodalmuknak adnak hangot amiatt, hogy az automatizált folyamatok megkönnyítik a tartalmak túlszabályozását, és hibás döntésekhez vezetnek. Az online platformok a problémás megnyilvánulásokra adott válaszként egyre gyakrabban alkalmaznak automatizált cenzúrát a tartalom törlése, blokkolása és szűrése formájában. Ez a megközelítés azzal fenyeget, hogy sérti az egyének véleménynyilvánítási szabadságát, és aránytalanul nagy hatással lesz azokra a csoportokra, akiket a társadalomban már most is diszkrimináció ér, vagyis azokra a csoportokra, amelyek a közösségi médiát használják fel hangjuk felerősítésére, társulások létrehozására és a változás érdekében történő szerveződésre.

SÜRGŐSEN SZÜKSÉG VAN ARRA, HOGY NE CSAK AZ AUTOMATIZÁLT TARTALOMMODERÁCIÓS FOLYAMATOK KÁROS KÖVETKEZMÉNYEIT VITASSUK MEG, HANEM OLYAN AI TERVEKET IS VIZSGÁLJUNK, AMELYEK A TÁRSADALMAT IS BEVONJÁK A FOLYAMATBA.

A különböző társadalmi érdekek holisztikus kezelése érdekében ezeknek az algoritmikus rendszereknek be kell tartaniuk az átláthatóság és elszámoltathatóság, a vitatás és a részvétel demokratikus normáit, valamint fellebbezéseket és jogorvoslatokat kell tartalmazniuk. Ezért sürgősen szükség van nemcsak az automatizált tartalommoderációs folyamatok káros következményeinek megvitatására, hanem olyan mesterséges intelligenciatervek feltárására is, amelyek a társadalmat is bevonják a folyamatba (Rahwan, 2018; Link et al., 2016).

A "problémás tartalom" meghatározása vagy sem?

A "társadalom a hurokban" koncepcionális gondolata felvetette azt a kérdést, hogy a társadalomnak milyen szerepet kell és lehet játszania az online tartalom moderálási folyamatokban. Résztvevőink azzal kezdték, hogy megvitatták, vajon az online platformoknak egyoldalúan kellene-e meghatározniuk, hogy milyen tartalom elfogadható. Eddig a vállalatok szolgáltatásaik referenciakeretét önállóan, a szolgáltatási

felületekre vonatkozó szabályzatainkon keresztül határozták meg.

Szükség van az átláthatóságra, de miben és kinek?

Jelenleg az online platformok kevés ellenőrzés mellett működnek, és gyakran még úgy is döntenek, hogy gyakorlatukat

átláthatatlan a külső megfigyelők számára (O'Neil, 2016). A kormányok és a független kutatók felé való átláthatóság azonban elengedhetetlen ahhoz, hogy a társadalom megértse a platformok tartalommoderálási gyakorlatának következményeit. Az állampolgároknak tudniuk kell, hogyan működnek ezek a platformok, hogyan alakítják a felhasználói élményeket, és mit csinálnak a vállalatok a tartalmaikkal. Emiatt számos szakértő hozzáférést kérnek az adataikhoz és rendszereikhez. Bár ez a kérdés több okból is rendkívül bonyolult, szükséges annak biztosítása, hogy a társadalom hosszú távon profitáljon ezekből a technológiákból. Résztvevőink azt javasolták, hogy nyissunk új beszélgetési csatornákat, amelyek középpontjában az áll, hogy hogyan lehet olyan kereteket és mechanizmusokat létrehozni, amelyek biztosítják az ilyen, több érdekelt felet érintő ellenőrzést.

ÚJ BESZÉLGETÉSI CSATORNÁKAT KELL NYITNUNK, AMELYEK KÖZÉPPONTJÁBAN AZ ÁLL, HOGY HOGYAN LEHET OLYAN KERETEKET ÉS MECHANIZMUSOKAT LÉTREHOZNI, AMELYEK BIZTOSÍTJÁK A TÖBB ÉRDEKELT FÉL ÁLTAL VÉGZETT ELLENŐRZÉST.

Emberi jogi hatásvizsgálat és a felelősségek kijelölése az automatizált döntéshozatalban

Tekintettel arra, hogy a vállalatok folyamatosan új termékeket vezetnek be, frissítik politikáikat, és új joghatóságokba terjeszkednek, az emberi jogi hatásvizsgálatokat folyamatosan el kell végezni, és nem lehet egyszeri esemény (ENSZ Emberi Jogi Tanács, 2018). Az emberi jogi hatásvizsgálatoknak ki kell terjedniük az összes olyan emberi jogra, amelyre a vállalatok politikái hatással lehetnek, a véleménynyilvánítás és a magánélet szabadságán túlmenően, hogy többek között a gazdasági, szociális és kulturális jogokra, az erőszakmentességhez való jogra és a közéletben való részvételhez való jogra is kiterjedjenek. Ezen túlmenően azt is mérlegelniük kell, hogy politikáik hogyan erősíthetik, és nem pedig alááshatják a tisztességes eljárást.

Jogorvoslati mechanizmusok és megfelelő eljárás

Az irányítási modelleknek nemcsak az online platformok teljesítményének értékelésére kell alkalmasnak lenniük, hanem lehetővé kell tenniük a személyre szabott korrekciós intézkedések elfogadását is.

E modell megvalósítása bizonyos gyakorlati kihívásokat és problémákat vethet fel. Úgy véljük azonban, hogy ezeket tovább kell vitatni és vizsgálni. A mai digitális társadalmakban közösen kell foglalkoznunk azokkal a felmerülő kérdésekkel, amelyek alapvetően érintik a véleménynyilvánítás szabadságát és tágabb értelemben az emberi jogokat.

4 AZ ÚT ELŐTT

A berlini workshopon folytatott vita feltérképezte és feltárta a problémateret, annak összetettségi szintjeit és szempontjait. A jelentést szétosztottuk a résztvevők között, és közzétettük a honlapunkon. A projekt weboldalán kívül a Twitteren a #Turn2AI[11] hashtag alatt rendszeresen tweetelünk a problémákról és interakcióba lépünk egy egyre növekvő közösséggel.

KÖVESSE ÉS VEGYEN RÉSZT A VITÁBAN! Használja a #Turn2AI

Reméljük, hogy a jövőben is fenn tudjuk tartani a terület tudományos és nem tudományos szakértői hálózatát. A workshop fórumként szolgált a tudományos tudományágak közötti, valamint a különböző hátterű és érdeklődésű akadémikusok, a civil társadalom és a gyakorlati szakemberek közötti gyümölcsöző eszmecseréhez. A műhely tagjai például már rendeztek egy panelbeszélgetést a 2018. május 17-én Torontóban[12] megrendezésre kerülő RightsConon "This Panel May Contain Sensitive Content" címmel: Automated Filtering and the Future of Free Expression Online" címmel.

A *Big Data & Society* című rangos folyóiratban különszámot is kiadunk a szakértői workshop témája és tartalma alapján. A különszám olyan interdiszciplináris tudományos cikkeket tartalmaz majd, amelyek a mesterséges intelligencia szerepével foglalkoznak az online kommunikáció szabályozásában. Ide tartoznak - de nem kizárólagosan - azok a hozzászólások, amelyek a következő kérdésekre irányítják a figyelmet:

- Milyen tényezők és szereplők mozgatják ezt a változást az automatizálás és a mesterséges intelligencia felé a tartalom moderálása és szabályozása terén a közösségi médiaplatformokon?
- Milyen technikai lehetőségek és korlátok vannak a mesterséges intelligencia fejlesztésével és alkalmazásával kapcsolatban a kommunikációs irányításban?
- Milyen társadalmi és jogi elvárásokat támasztanak ezzel a technológiával szemben? Befolyásolják-e ezek az elvárások a szoftverfejlesztést? Ha igen, hogyan?
- Hogyan lehet a jövőbeni mesterséges intelligencia rendszereket fejleszteni és képezni? Valóban lehetséges-e optimalizálni őket a közjó érdekében?

Miközben belevágunk ebbe a kritikus és előremutató munkaterületbe, reméljük, hogy fenn tudjuk tartani ezt a kialakulóban lévő szakértői hálózatot, és a jövőben kutatási és érdekérvényesítő hálózatainkon belül hozzá kívánunk járulni a mesterséges intelligenciáról és a kommunikáció irányításáról szóló vitához.

[11] https://twitter.com/hashtag/Turn2AI

[12] A RightsCon egy csúcstalálkozó-sorozat, amely az emberi jogok és a technológia témájában hívja össze a globális közösséget.

Házigazdája az Access Now. A RightsCon Toronto májusban került megrendezésre a16 kanadai 18,2018Torontóban.

5 HIVATKOZÁSOK

Arsht, A. & Etcovitch, D. (20182,. március). The Human Cost of Online Content Moderation, Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Letölthető a következő címen: https://jolt.law.harvard.edu/digest/the-human-cost-of- online-content-moderation.

Duarte, N., Llansó, E. & Loup, A. (2018): Vegyes üzenetek? The Limits of Automated Social Media Content Analysis, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:106-106. Retrieved from http://proceedings.mlr.press/v81/duarte18a.html.

Gillespie, T. (2018). *Az internet letéteményesei : Platformok, tartalommoderálás és a közösségi médiát alakító rejtett döntések.* New Haven, London: Yale University Press.

Gollatz, K., Riedl, M. J. & Pohlmann, J. (2018. augusztus 9.). Az online gyűlöletbeszéd eltávolítása számokban. HIIG Science Blog, Alexander von Humboldt Institute for Internet and Society, Berlin, Németország. Cross-posted at Media Policy Project Blog, London School of Economics, London, Egyesült Királyság. Letölthető a következő oldalakról: https://www.hiig.de/en/removals-of- online-hate-speech-numbers/ és http://blogs.lse.ac.uk/mediapolicyproject/2018/08/16/removals-of-online-hate- speech-in-numbers/.

ENSZ Emberi Jogi Tanács (2018). David Kaye, a véleménynyilvánítás és a véleménynyilvánítás szabadságának előmozdításával és védelmével foglalkozó különmegbízott jelentése a felhasználók által generált online tartalmak szabályozásáról. Emberi Jogi Tanács harmincnyolcadik ülésszak június 618-július Retrieved2018. from http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35.

Link, D., Hellingrath, B. & Ling, J. (2016). A Human-is-the-Loop Approach for Semi-Automated Content Moderation, Long Paper - Social Media Studies Proceedings of the ISCRAM Conference 2016- Rio de Janeiro, Brazil, May Retrieved2016. from https://pdfs.semanticscholar.org/2223/f7245f7c310db2c4a24e6e4603c85936d460.pdf.

O'Neil, C. (2016). *Weapons of Math Destruction. Hogyan növelik a nagy adatok az egyenlőtlenséget és fenyegetik a demokráciát.* New York: Crown Books.

Rahwan, I. (2018). Társadalom a hurokban. Az algoritmikus társadalmi szerződés programozása, Rahwan, I. Ethics Inf Technol (2018) 205.: https://doi.org/10.1007/s10676-017-9430-8.

Tushnet, R. (2008). Hatalom felelősség nélkül: Közvetítők és az első módosítás, Geo76. Wash. L. Rev. 101. Visszakeresve a http://scholarship.law.georgetown.edu/fwps_papers/76 oldalról.

MELLÉKLET: A RÉSZTVEVŐK LISTÁJA

(ábécésorrendben)

Prabhat Agarwal Európai Bizottság, Belgium

Amar Ashar Berkman Klein Center for Internet & Society at Harvard University, Egyesült Államok

Johannes Baldauf Tanácsadó, Németország

Renata Barreto University of California Berkeley, Egyesült Államok

Eimear Farrell Amnesty International, Németország

Nick Feamster Princeton University, Egyesült

Államok **Sabine Frank** Google, Németország

Tarleton Gillespie Microsoft Research New England, Egyesült Államok (távolról)

Kirsten Gollatz Alexander von Humboldt Institute for Internet & Society,

Németország **Lisa Gutermuth** Ranking Digital Rights at New America, Egyesült

Államok

Amélie Heldt Hans-Bredow-Institut ; Alexander von Humboldt Intézet az Internet és Társadalomért, Németország

Fanny Hidvégi Access Now, Belgium

Jeanette Hofmann Berlini Társadalomtudományi Központ, Németország

Malavika Jayaram Digital Asia Hub, Hongkong

Christian Katzenbach Alexander von Humboldt Intézet az Internet és Társadalomért, Németország

David Kaye ENSZ különmegbízott a véleménynyilvánítás és a véleménynyilvánítás szabadságának előmozdításával és védelmével kapcsolatban (távolról)

Aphra Kerr Maynooth Egyetem, Írország

Ulrike Klinger Freie Universität Berlin ; Weizenbaum Institute for the Networked Society, Németország

Michael Latzer Zürichi Egyetem, Svájc

Emma Llansó Center for Democracy & Technology, Egyesült Államok

Tilo Mentler Lübecki Egyetem, Németország

Ramak Molavi iRights.Law, Németország

Kevin Morin Institut National de la Recherche Scientifique, Kanada

Iva Nenadic Európai Egyetemi Intézet, Olaszország

Jörg Pohle Alexander von Humboldt Institute for Internet & Society,

Németország **Fabrizio Augusto Poltronieri** De Montfort University, Egyesült

Királyság **Sarah T. Roberts** University of California Los Angeles, Egyesült

Államok (távolról) **Jeremy Rollison** Microsoft, Belgium

Erin Saltman Facebook, Egyesült Királyság (távolról)

Björn Scheuermann Humboldt-Universität zu Berlin ; Alexander von Humboldt Institute for Internet & Society, Németország

Matthias Spielkamp Algorithm Watch, Németország

Florent Thouvenin Zürichi Egyetem, Svájc

Betty van Aken Beuth University of Applied Sciences, Németország

Joris van Hoboken Vrije Universiteit Brüsszel, Belgium

Mirko Vossen die medienanstalten, Németország

Zeerak Waseem University of Sheffield, Egyesült Királyság

Jillian C. York Electronic Frontier Foundation ; Center for Internet and Human Rights, Németország