Az óriásplatformok tartalomtörlési, fiókfelfüggesztési, fióktörlési és shadow banning gyakorlata a magyar felhasználókkal szemben

Bevezetés

Az itt bemutatott kutatás célja az óriásplatformok legkomolyabb, a felhasználókat a legmélyebben érintő tartalom- és fiókkorlátozási intézkedéseinek vizsgálata, elsősorban a magyar felhasználók szemszögéből.

Az írás két nagyobb részre bomlik.

Az első részben (1., 2., és 3. pontok) átfogó képet adok az óriásplatformok vonatkozó gyakorlatáról, elsősorban a saját honlapjaikon és szabályzataikban, valamint a nyilvános transzparenciajelentéseikben közzétett adatok alapján. Vizsgálódásaim két nagy platformra terjedtek ki, a Facebookra és a Youtube-ra, elsősorban azért, mert ezek azok, amelyeknél egyáltalán ezek az adatok hozzáférhetők. Ebben az első részben tisztázok néhány fogalmat is.

A második részben (4. pont) a magyar felhasználók tapasztalatait írom le, az NKE információs társadalom kutatóintézetének longitudinális reprezentatív, kérdőíves kutatásai, valamint három magyar médium (nagy elérésű óriásplatformon jelen levő account) tapasztalatai alapján, amelyeket mélyinterjúk során szereztem be.

Kutatásom eredeti célkitűzése az volt, hogy megszólaltom az óriásplatformok képviselőit is, de ezt nem tudtam teljesíteni. Sajnos a tisztségviselők annyira nehezen elérhetők (sokszor azt sem lehet tudni, hogy ki felelős az adott területért), hogy minden ilyen irányú kísérletem kudarcba fulladt.[1]

1. Az óriásplatformok tartalom- és fiókkorlátozási gyakorlatáról általában – a fogalmak tisztázása

1.1. Átláthatatlan, kusza szabályozás

Mielőtt a részletekre rátérnék, két általános megjegyzést kell tenni a tiltásokkal és a korlátozásokkal kapcsolatban. Mindkét ténynek nagy jelentősége van a téma szempontjából.

Az első, hogy az ezekre a szankcionáló jellegű intézkedésekre vonatkozó szabályok rendkívül átláthatatlanok, és ez minden óriásplatformra igaz. A moderálással és a tiltásokkal kapcsolatos

[1] Sem a Meta, sem a Google nem közöl semmilyen speciális (pl. a policy) területhez tartozó elérhetőségeket (emailcímeket vagy telefonszámokat) a nyilvános felületein. A felhasználói panaszokat általában erre létrehozott űrlapokon keresztül lehet továbbítani.

szabályokat ugyanis sehol sem találjuk meg egyetlen, átlátható, összefoglaló dokumentumban (egy „normában"), hanem ezek csak egyes részletkérdéseket tartalmazó, egymáshoz linkelt oldalak tucatjain férhetőek hozzá. Ez azért rendkívül problematikus, mert az oldalakon kattintgatva a felhasználó sohasem lehet biztos benne, hogy valóban elolvasott minden információt, a „végére ért" a tilalmaknak és az előírásoknak, és ismer minden szabályt.

Ennek a problémának a része az is, hogy a különböző természetű szabályok tartalmilag is tökéletes kuszaságban jelennek meg ezeken az egyébként is átláthatatlan szerkezetű oldalakon. Megkülönböztethetetlenek azok a szabályok, amelyeket a felhasználási feltételek tartalmaznak és a belső „közösségi" szabályok, az anyagi és az eljárási szabályok, a különböző súlyosságú normaszegések, de legfőképp az, hogy mi az, ami jogellenes, és mi az, ami csak a belső szabályokat sérti. Igen gyakran ráadásul ezek a szabályok, tilalmak redundánsan (ugyanaz az előírás több oldalon, többféle szövegkörnyezetben) jelennek meg. A következő két alpontban részletezem az ezzel kapcsolatos konkrét problémákat az általam vizsgált két óriásplatformon.

1.1.1. Facebook

A Facebook ÁSZF-e a „Felhasználási feltételek" oldalon található. Itt a törzsszövegben tucatnyi alkalommal történik egyéb, külső szabályzatokra hivatkozás, amelyekről nem eldönthető, hogy részét képezik-e a felhasználási feltételeknek, vagy sem. A dokumentum végén két pontot találunk, a „4. Egyéb" és az „5. Egyéb feltételek és szabályzatok, amelyek vonatkozhatnak rád" pontokat. A 4. pontban a szöveg három olyan szabályzatot említ, amelyben kiegészítő feltételek is találhatók, és amelyeket el kell fogadni, ha az adott „terméket" használjuk. Ezután említ három szabályzatot „Például" bevezetéssel. Hogy azonban teljes legyen a bizonytalanság, az 5. pont pár sorral lejjebb egy listában 19 (!) szabályzatot sorol fel, amelyek egy része nagyon specifikus célcsoportokra (pl. a „platformhasználati szabályzat" az API-kon keresztül fejlesztő programozókra) vonatkozik, mások azonban (pl. a „Panaszkezelési folyamat") nyilvánvalóan minden felhasználóra. Nem mellékesen ráadásul ez utóbbit nem is szabályzatnak hívják.

Nem jobb a helyzet a Facebook (magyar nyelvű), „közösségi alapelveket" közlő oldala esetén sem. Az oldal összesen 28 (!) külső hivatkozást tartalmaz, amelyek a különböző jogsértő, vagy csak „sértő" tartalmakat, illetve néhány, a tartalomgondozással kapcsolatos eljárási szabályt (pl. a fiókok integritására vonatkozó, vagy a „Felhasználói kérések" című) részletezik. A „jogsértő" és az egyszerűen csak „bántó", „nem biztonságos" tartalmakat az oldal összekeverve említi, jóllehet jogi szempontból ezek teljesen eltérő megítélés alá esnek. Rendkívül zavaró, hogy a tartalmi és az eljárási szabályok sincsenek elválasztva egymástól, nem lehet tudni, hogy melyik link mutat egy tilalomra, és melyik magyaráz például egy jogorvoslatot vagy egy gyakorlatot. Emellett a különböző alcímek alá nehezen követhető logika mentén sorolták be az egyes tilalmakat tartalmazó kategóriákat is. Például a „Biztonság" alcím alatt szerepel az egyik legsúlyosabban tilalmazott tartalomtípus (a gyermekek szexuális kizsákmányolása), míg az „Erőszakot megjelenítő tartalmak" a „Kifogásolható tartalmak" csoport alá vannak besorolva. És egyik sem az egyébként logikus „Erőszak és bűncselekmények" alcím alatt található, ahová egyébként tartoznának, mert ez alatt az alcím alatt pl. olyan magatartások szerepelnek, mint a „Csalás, átverés és félrevezető gyakorlat", amely bizonyos esetekben egyáltalán nem jogellenes, és szinte biztosan nem erőszakos. Végül érdemes azt is megjegyezni, hogy az aloldalak némelyikén váratlanul angol nyelvű szövegek bukkannak fel, és van, ahol a magyar szöveggel összekeveredve.

1.1.2. Youtube

A Youtube szabályozása, a Metahoz képest talán még kuszább.

Már az sem teljesen tiszta, hogy a Google számára melyik szabályzat jelenti a kiindulópontot, melyik lenne az „alap-szabályzat". Ha mondjuk a „Youtube szabályzatok" keresést beütjük a Google-ban, akkor ez egy magyar nyelvű gyűjtőoldalra navigál, amelyen a „Közösségi irányelvek" áttekintése szerepel, kéttucatnyi további linkkel. Ha a „Súgó központ" oldalra navigálunk, itt pedig összesen 37 szabályzatot találunk, mint amelyek a Közösségi irányelvek hátterét biztosítják.

A „lényeg" azonban ennek ellenére nem itt található, hanem a Youtube főoldalról (bal oldalt alul, 12 másik link társaságában) a „Feltételek" linkre kattintva, amely előhozza a Youtube általános szerződési feltételeit. Ebben a dokumentumban találunk utalást a felfüggesztésre és a kitiltásra is, de a részleteket illetően az oldal továbbirányít bennünket a „Közösségi irányelvek" oldalra, amely azonban nagyjából csak egy gépelt oldalnyi hosszúságú, és szintén 27 (!) linket tartalmaz a szövegbe beszúrva és egy külön panelen öt továbbit, amelyek olyan oldalakra mutatnak, mint a „YouTube irányelvei" (amely visszavisz az összes szabályzatot tartalmazó oldalra), vagy a „bejelentés és jogérvényesítés". A linkek nagy része a főoldalon található szabályzatokra mutat, de értelemszerűen nem tartalmazza mindegyiket.

Összefoglalóan: az óriásplatformok tartalommoderálási és szolgáltatás-korlátozási szabályai – amellett, hogy valószínűleg tartalmilag is komplexek – formailag is szinte reménytelenül kusza rendszerben vannak közzétéve. Az össze-vissza, olykor körbeforgóan linkelt, és illogikusan tagolt oldalak alapján az átlagfelhasználó – a teljesen egyszerű eseteket leszámítva – bizonyosan nem fogja tudni egyértelműen megállapítani, hogy mi fog tartalomkorlátozás alá esni és mi nem. A mélyinterjúk során egyértelműen kiderült, hogy még a professzionális médiavállalkozások sem az írott információk alapján tájékozódnak, hanem a saját empirikus tapasztalataikra támaszkodnak.

1.2. A tömeges szankcionálást algoritmusok végzik

A másik fontos előzetes megjegyzés, hogy bár némileg antropomorfizálóan azt mondjuk, hogy a „platformoknak" van tiltási, stb. „gyakorlata", valójában az esetek elsöprő többségében (és egyre növekvő arányban) ezt a „tevékenységet" (mind a moderálást, mind a rangsorolást) algoritmusok, és ezen belül is a leggyakrabban mesterséges intelligenciák (gépi tanuláson alapuló algoritmusok) végzik. A platformok ma már sokféle célra használnak mesterséges intelligenciákat, már attól a fázistól kezdve, hogy valaki fiókot akar nyitni, és különböző algoritmusok ellenőrzik, hogy nem ún. „fake accountról" van-e szó, a gépi fordítást végzőkön keresztül egészen az arcfelismerő algoritmusokig (amelyeknek a Meta 2021-ben leállította a használatát), de a legfontosabb két algoritmuscsoport a moderáló és a rangsoroló algoritmusok.

A moderáló algoritmuscsoport a platformra feltöltött tartalmak *előszűrését* (a közösségi médiafelületeken: *moderálását*) végzi, míg a rangsoroló algoritmusok a tartalom „szerkesztésére", egy adott felhasználó számára történő *sorrendbe rakására* szolgálnak. Az előbbi a „banning" míg a második a „shadow banning" szempontjából játszik szerepet. A két algoritmuscsoport működése sok ponton összefügg ugyan, de egy jellegzetességükben radikálisan különböznek: az egyik a feltöltés pillanatában fut, és arról dönt, hogy az adott tartalom megfelel-e a közösségi szabályoknak (és egyáltalán megjelenhet-e), míg a másik arról, hogy egy már „átengedett" tartalom hány felhasználónál (milyen gyakran) és mennyire hangsúlyosan (a hírfolyamban milyen helyen) fog megjelenni.

Az algoritmusok pontos működése, forráskódja üzleti titok. Működésükről egyrészt csak azt lehet megtudni, amit maguk a platformok elárulnak, másrészt ami a felhasználók használat során szerzett tapasztalataiból leszűrhető. A legnagyobb problémát az jelenti, hogy ezeket az algoritmusokat, miközben mind a szólásszabadság, mind a monetizálás szempontjából mondhatni életbevágóan

fontosak, a platformok kényük-kedvük szerint teljesen átláthatatlanul változtatgatják és állítgatják. Később, az esettanulmányokból lehet majd látni, hogy azok a nagyfelhasználók, akiknek komoly érdekük fűződik ahhoz, hogy ne kapjanak kitiltást vagy felfüggesztést, kétségbeesetten próbálnak folyamatosan megfelelni ezeknek a nem transzparensen, hanem önkényesen és kiszámíthatatlanul változtatgatott algoritmusoknak, de ez ennek ellenére nem mindig sikerül nekik.

1.3. Fogalmi tisztázás

1.3.1. A banning: tartalom- és fióktörlés és korlátozás a felhasználó tudtával

A banning (tiltás) az internetes platformoknak a felhasználói tartalmak, interakciók terjedésére, láthatóságára, vagy használatára, illetve a felhasználói fiókok használatára vonatkozó olyan korlátozások összefoglaló elnevezése, amelyekről az érintett felhasználót értesítik.

A tartalomkorlátozás (törlés és korlátozás) célpontja szerint sújthatja a felhasználó által létrehozott tartalmat, vagy a más által létrehozott tartalom megosztását, vagy hozzá interakció végrehajtását, illetve magát a fiókot. A tartalomkorlátozás terjedelme szerint lehet teljes (törlés), vagy részleges (pl. időben korlátozott) (felfüggesztés).[2]

1.3.2. A shadow banning

Az elérhetőségében (terjesztésében) történő korlátozás a legtöbb esetben nem explicit, arról a felhasználót nem tájékoztatják. A shadow banningnek („árnyéktiltás") nevezett módszer a felhasználó által létrehozott tartalmak korlátozása oly módon, hogy ezt *nem közlik* a felhasználóval.

A shadow banning elég nehezen definiálható fogalom, és sokféle tartalom-terjedési korlátozást jelent, amelyek közül a legtöbb nem nyilvános, sőt gyakran a platformok el sem ismerik. Ugyanakkor van elismert, nyilvánosságra hozott verziója is, amelyet a Facebook „A problémás tartalmak terjesztésének csökkentése" címszó alatt ismertet is, ebből szerezhetük valamiféle benyomást a shadow banning folyamatáról.

A vonatkozó magyarázat szerint, „ha egy facebookos tartalom nem sérti ugyan a Facebook közösségi alapelveit, de mégis problémás lehet, vagy egyéb módon alacsony színvonalú, a Meta csökkentheti a terjesztését – a felhasználói beállításokkal összhangban."

---

[2] Mint azt említettem, ez a tanulmány nem szól a monetizáció korlátozásáról, mint szankcióról. A monetizációval kapcsolatos korlátozások ugyanakkor igen érdekesek. Itt ugyanis arról van szó, hogy vagy egy speciális tartalomtípus (a hirdetés), vagy a nem hirdetési tartalom egy speciális felhasználási módja, (hirdetéselhelyezésre történő felhasználás), a platform a „normál" közösségi szabályok *feletti, azoknál jóval szigorúbb* mércét alkalmaz. A hirdetésekre vonatkozó szabályokat a Youtube-on például egy összesen több mint 30 aloldalból álló oldalhalmazban találhatjuk meg, amely a hirdetések tartalmi elemein kívül, (pl. tiltott termékek, gyermekeknek szóló hirdetések, a szerzői jogilag védett anyagokra, vagy a márkavédjegyekre vonatkozó szabályok) bizonyos eljárási szabályokat is tartalmaz, (pl. a hirdetésekkel való visszaélés, rossz formátum, stb.) Külön szabályzata van a hirdetések engedélyeztetésének, amelyet nagyon valószínű, hogy szintén mesteréges intelligencia végez.
A hirdetés vagy megkapja az „Eligible" (alkalmas) státuszt, vagy feltételekkel kapja meg.
Az ezzel kapcsolatos szabályzat

Ezen az oldalon példakánt három tartalomtípust említ, de utal arra, hogy ennél jóval többféle ok van, amely „terjesztés-csökkentéshez” vezethet:

- Alacsony színvonalú – például kattintásvadász és aktivitásvadász – tartalmak.
- Olyan webhelyekre mutató hivatkozások, amelyekben a tartalmat elfedik a hirdetések, illetve amelyek lassan töltődnek be vagy hibásak.
- Alacsony színvonalú hozzászólások, amelyeket emberek ismétlődően odamásolnak több helyre.

Ez értelemszerűen azt is jelenti, hogy a „magas minőségű” „eredeti” tartalom (bármit is jelentsen ez) előnyt élvez. A kapcsolódó oldalon egy tartalom-típus, a hír-jellegű tartalmak rangsorolásával kapcsolatos szempontokat találjuk. Első lépcsőben a rendszer eldönti, hogy egyáltalán hírről van-e szó. Ezt öt szempont alapján teszi.

- A tartalom aktuális eseményekről, friss információkról vagy egy folyamatban lévő vizsgálatról tájékoztat, vagy szerkesztői közlésnek, illetve véleménynyilvánításnak minősül?
- A tartalom közvetlenül egy megnevezett szerzőnek, újságírónak vagy tartalomkészítőnek tulajdonítható?
- A tartalom megnevezi tényállításainak forrását?
- A tartalomnak átlátható a szerkesztői háttere (azaz szerepel a szerző vagy a közreműködők teljes neve, illetve a munkatársak elérhetősége)?
- Rendelkezik a tartalom dátummal és/vagy időbélyeggel?

A rendszer második lépcsőben a hír minőségét vizsgálja meg öt aspektusban: eredetiség, hitelesség, informativitás, pontosság és átláthatóság. Hogy ezek a szempontok pontosan hogyan érvényesülnek, milyen „jeleket” használ a Facebook ezeknek a szempontoknak az operacionalizálására, arról nincsen információnk. Nagyon valószínű, hogy egy folyamatosan változó szempontrendszerről van szó. Annak ellenére, hogy a Meta több helyen is igyekszik a hírrangsorolás szempontjait elmagyarázni (a Meta Journalism Project keretén belül, amelyben igyekszik általánosabb és részletesebb leírást is adni az általa megfelelő minőségűnek tartott tartalomról, sőt külön magyar oldal is részletezi a hírtartalmak minőségi követelményeit,) az algoritmusok tényleges viselkedése a gépi tanulás, és a folyamatos változtatások miatt nehezen számítható ki. Ugyanakkor az, hogy egy paraméter „jelként” történő felhasználását a rangsorolásban megszüntetik, vagy a jel erősségét, ennek végső hatását a rangsorolásra lecsökkentik vagy megnövelik, valószínűleg mindennapos eset, és a megkérdezett magyar médiumok tapasztalata is ezt támasztja alá.

Külön kategóriát képez – mert egy igen kényes határesetet jelöl - a dezinformáció („hamis tájékoztatás”) kategóriája. A tényellenőrzéssel azonosított, félretájékoztatást tartalmazó posztokat a Meta terjedésükben bevallottan korlátozza, annak ellenére, hogy elismeri, hogy itt nincsen szó jogellenességről, sőt még a közösségi alapelvekbe történő ütközésről sem. A félretájékoztatás definíciójáról lentebb szólok, az erre vonatkozó statisztikákat pedig a 3. pontban ismertetem.

2. A tartalom- és fiókkorlátozások anyagi és eljárási szabályai

2.1. Facebook

2.1.1. Anyagi szabályok: milyen tartalmakat moderál a Facebook?

Ahogy fentebb azt már bemutattam, a Facebook a Felhasználási feltételekben utal arra, hogy az ún. közösségi alapelvek (community standards) elnevezésű dokumentumban teszi közzé, hogy milyen típusú tartalmak számítanak tiltottnak, vagy diszpreferáltnak a felületén, azaz milyen tartalmakat nem fog egyáltalán kiengedni a nyilvános térbe, és melyeket fog korlátozni. Ahogy azt már szintén fentebb jeleztem, a dokumentum egészében nem olvasható el, azt szövevényesen egymáshoz kapcsolt weboldalak tucatjaiból lehet megismerni. A moderálóalgoritmusok működésének alapját tehát ez a szabályrendszer adja oly módon, hogy a Meta ezt igyekszik különféle mesterséges intelligenciákkal, szabályalapú szoftverekkel és emberi erőfeszítéssel kikényszeríteni.

Ami biztos, hogy a közösségi alapelvek alapvetően kétféle csoportba sorolják a tartalomkorlátozás alá eső tartalmakat: a „nem engedélyezett" és a „korlátozásokkal, vagy feltételekkel terjeszthető" tartalmak csoportjaiba. Erről azt hihetnénk, hogy a „jogellenes", és a „nem jogellenes, de káros" megkülönböztetést fedik, de *nem erről van szó*. A nem engedélyezett tartalmak közt egy sor olyan tartalmat is találunk, amely nem jogellenes.

A tiltott kategóriák tehát a következők: „amelyek erőszakra vagy bűncselekményekre buzdítanak (például veszélyes szervezeteket népszerűsítő, támogató vagy dicsérő tartalmak), veszélyeztetik az emberek biztonságát (például megfélemlítés és zaklatás, öngyilkosság és önsértés, valamint gyermekek vagy felnőttek szexuális kizsákmányolása), kifogásolhatók (például gyűlöletbeszéd), nem hitelesek (például kéretlen tartalmak, félretájékoztatás vagy hamis profilok), vagy sértik valaki más szellemi tulajdonát." – írja a vonatkozó oldal.

Ami igazán érdekes, az a „félretájékoztatás" kezelése. Ez a kategória az „integritás és hitelesség" csoportban található a közösségi alapelvek oldalon, és bár javarészt a következménye a tartalom terjedésének korlátozása (shadow banning), bizonyos fajtáit (amely „feltehetően közvetlenül növelné a fenyegető fizikai bántalom kockázatát", illetve „ha az vélhetően közvetlen szerepet játszik politikai folyamatok működésébe való beavatkozásban") nem korlátozza a Facebook, hanem azonnal eltávolítja.

2.1.2. Eljárási szabályok és folyamat

2.1.2.1. Észlelési fázis MI-vel

Az észlelés elvileg kétféleképp lehetséges, a mindenre kiterjedő általános monitorozással és a bejelentéssel. A bejelentés jelentősége az idők folyamán fokozatosan csökkent, ma már korántsem olyan fontos, mint korábban volt, és már a bejelentések sem kerülnek automatikusan emberi moderátor elé, ezeket is először a mesterséges intelligencia vizsgálja át.

A moderálás teljes folyamatát nem ismerjük, de többféle leírást is találunk róla a Facebookon. A legalapvetőbb (teljesen semmitmondó) a help-ben található leírás arról, hogy „hogyan használja a Facebook a mesterséges intelligenciát a tartalmak moderálására?"

Részletesebb információt a Facebook „transzparenciaközpontjában" találunk. Az automatikus észlelés a Meta saját bevallása szerint a szabályzatellenes tartalmak több mint 90%-t még a bejelentés előtt,

(azaz a feltöltés után közvetlenül) megtalálja. Amit a cég erről elárul, azok elég általános ismeretek. Így pl. megtudhatjuk a vonatkozó oldalon, hogy az általános MI-építés után egy külön fejlesztői csapat dolgozik a specifikus, egy adott jogsértés-típus (pl. gyűlöletbeszéd) felismeréséért felelős MI kidolgozásán. A modellek multimodálisak, egyes modellek a képeket elemzik a specifikus jogsértés-kategória mentén (pl. meztelenséget keresve), mások a szöveget. Az ismétlődő szabálysértések (ugyanazon tartalom újbóli feltűnése) esetén a technológia igen hatásos, mert a „vírusszerűen terjedő félretájékoztatási kampányoknál, mémeknél, valamint más, rendkívül gyorsan terjedő tartalmaknál" automatikusan milliószámra korlátozza az egyszer már szabályellenesnek címkézett adott tartalmakat. Ez a magyarázata egyébként annak is, hogy a transzparenciajelentésekben miért találunk néha teljesen megmagyarázhatatlan ugrálásokat a számokban.[3]

Az évek során alkalmazott fokozatos fejlesztésekkel a Facebook elérte, hogy a szabályzatellenes tartalmak elsöprően nagy része (90% feletti része, bár ez az egyes kategóriáknál változatos) már a

[3] Jelen tanulmány kereteit szétfeszítené, hogy az automatikus moderálás technológiai hátteréről részletesen beszéljek, de itt lábjegyzetben röviden összefoglalom, amit erről tudni lehet. A szabályellenes tartalmakat felismerő MI tanítása megerősítő tanítás formájában zajlik, és sohasem áll le, folyamatos: „Idővel a szabályaink is változnak, hogy kövessék a termékünk, a társadalmi normák és a nyelv változásait." – írja a vonatkozó oldal.
Mivel a technológiának többféle formátumú tartalmat kell tudnia értelmezni mindenféle nyelveken, a Meta szemmel láthatóan hatalmas erőforrásokat mozgósít ezen a területen. Ebből egy példát emelnék ki, a gyűlöletbeszéd felismerését végző fejlesztésüket. A Facebook abból indul ki, hogy bár a gyűlöletbeszédet egy ember könnyen fel tudja ismerni, de, mint írják, ennek a *gépi* felismerése mégis nagyon nehéz. Az első nehézség a kontextusból fakad. A vonatkozó oldal egy olyan képet hoz magyarázatként, ahol csak egy sírkő és egy felirat látható: „az etnikumod minden tagjának itt kellene lennie". A képen tehát sem offenzív képi tartalom, sem offenzív szöveges elemek nincsenek, mégis az egészet lehet gyűlöletbeszédként értelmezni.
A dolgot még tovább bonyolítja, hogy mivel az emberek tudnak a korlátozásokról, igyekeznek kicselezni a rendszert (pl. épp a fentebbi trükkös kép-szöveg kombinációval). Végül az utolsó, lényegében legnagyobb probléma a rengeteg nyelvi és kulturális különbözőséggel van. Vannak kisebb nyelvek, ahol nem is áll rendelkezésre elegendő tanítóadat. (Ezeknek a nyelveknek az esetében egy speciális szoftver a más nyelveken rendelkezésre álló adatokat fordítja le az adott nyelvre – ami persze sokszor igen félrevezető lehet.)
A fejlesztéseket bemutató oldalon a mesterséges intelligenciával kapcsolatos fejlesztések illusztrálására a Facebook több fejlesztési projektet sorol fel, amelyeket az elmúlt években a szabályzatellenes tartalmak szűrésére dolgoztak ki. Ezek közül hármat emelek ki: az egyik tartalmi, míg a másik kettő technológiai jellegű projekt, de mindhárom jól illusztrálja, hogy milyen gigantikus erőforrásokat mozgósít a Facebook a tartalommoderáló MI-k fejlesztésekor.
A Hateful Memes Challenge (Gyűlölködő mémek kihívás) kiindulópontja az volt, hogy a gyűlöletbeszéd megértéséhez gyakran a teljes kontextus megértésére szükség van. A teljes kontextus azonban a potenciálisan figyelembe veendő tartalom méretét és variabilitását is megnövelte. A Facebook összegyűjtött egy tízezres nagyságrendű valós mém-adatbázist, amelyet azután a Gettytől licenszelt képek segítségével emberek elkezdtek variálni (pl. egy állat képét a mémben kicserélni egy másik ugyanolyan, de másképp ábrázolt állat képére). A gép a variált, mutált képeket is megkapta, így sokkal érzékenyebb lett a mémek felismerésében. Ez eredményezte a 97%-os felismerési pontosságot a gyűlöletbeszéd kategóriájában.
A másik projekt a transformer technológia, a generatív MI-k „lelkének" továbbfejlesztését célozta. A Linformer tartalomelemző technológia előnye, hogy a hagyományos transzformerekkel szemben, amelyek hosszabb szövegeket és képeket a hosszúsággal négyzetesen növekvő gépi erőforrások segítségével tudtak csak elemezni, a hosszúsággal csak egyenesen arányosan növekszik a számítási igénye. (A „lin" a linearitást jelöli). Tulajdonképpen a linformer tette lehetővé, hogy a Facebook képes legyen a posztokat valós időben (azaz nagyon gyorsan) elemezni.
Végül a Reinforcement Integrity Optimizer (RIO) modell a tanítást optimalizálja azzal, hogy visszacsatolja az ajánlórendszer által adott rangsorolási döntést a tanítási fázisba, azaz azokkal az adatokkal tanítja vissza az előrejelző rendszert, amelyeket az ajánlórendszer rangsorolt amely így finomhangolja magát).

feltöltés előtt kiszűrődik. Bizonyos jogsértés-típusok esetén (pl. a gyűlöletbeszéd esetén) ez akár 97% környékén is lehet.

Itt mindenképpen meg kell említeni ismét a félretájékoztatás, mint kategória ellentmondásosságát. Egyfelől a vonatkozó oldal elismeri, hogy a félretájékoztatás esetében „nincs lehetőség arra, hogy átfogó listát állítsunk össze a tiltott dolgokról", és a tényellenőrzés elsődleges módszereként a külső (emberi) tényellenőrzőket említi. Másfelől azonban a tényellenőrzést részletesen magyarázó oldalon azt olvashatjuk, hogy „technológiánk számos országban képes azonosítani a feltehetően félretájékoztatásnak minősülő bejegyzéseket különböző jelek alapján – például az alapján, hogy az emberek hogyan reagálnak rá, és hogy a tartalom milyen gyorsan terjed."

2.1.2.2. Az emberi ellenőrzés fázisa

A cég maga is elismeri, hogy az egyértelmű eseteken kívül, (amely a bejegyzések elsöprő többsége), vannak kevésbé könnyen felismerhető esetek, amelyek azonban még mindig a napi milliós nagyságrendben mozognak. Ahhoz, hogy ezt kezelni lehessen, még ebben a halmazban is priorizálni kell, hogy mely tartalmak kerüljenek előbb emberi ellenőrzés alá. Az emberi ellenőrzésre való felkínálás szempontjai: a súlyosság, a vírusszerű terjedés és a szabálysértés valószínűsége. Az emberi ellenőrök ma már szinte kizárólag azokat a tartalmakat ítélik meg, amelyekről a mesterséges intelligencia végképp nem tud dönteni. Ez azért rendkívül érdekes, mert korábban a bejelentés dominálta a korlátozások nagy részét, és a bejelentéseket nagyobbrészt emberek intézték.

Az ellenőrök munkájában a főszabály az, hogy a Facebook és az Instagram közösségi irányelveit tartják szem előtt. Nagyon valószínű, hogy továbbra is egy részletes kézikönyvből dolgoznak, amelynek a tartalmát továbbra sem ismerjük részleteiben. Az emberi ellenőrök által hozott döntések azonnal tanítóadatként visszakerülnek a vonatkozó moderáló MI-hez.

2.1.2.3. Szankcionálás

Eltávolítás, csökkentés, tájékoztatás (Remove, reduce or inform): ez a háromféle intézkedéstípus („ban") létezik a Facebook szerint (2016 óta).

Az eltávolítás valójában egy gyűjtőfogalom, és többféle súlyosabbnak tekinthető intézkedést és szankciót foglal magában.

A legelső típus valóban az adott tartalmi egység tényleges és teljes láthatatlanná tétele. A láthatatlanná tételt természetesen a mesterséges intelligencia-rendszerek végzik. Ilyenkor értesítik a tartalom közzétevőjét („We removed your post"), és a „lehetőségekhez mérten" jelzik, hogy milyen szabályokat nem tartott be az illető, mégpedig a „tágabb kontextus" (azaz azon szabálycsoport, amelyet megsértett) felidézésével és ennek magyarázatával. A törléskor lehetőséget biztosítanak az ellenőrzésre (amelyet valószínűleg még mindig mesterséges intelligencia bírál el).

A második szankció, (amely az eltávolítással együtt automatikusan bekövetkezik) a hibapontok felírása az adott fiókhoz. „A hibapont alkalmazása függ a tartalom súlyosságától, a megosztás kontextusától, valamint a közzététel idejétől is." – írja az útmutató, de itt sincsen egzakt számítási módszer, ezt is nagy valószínűséggel egy gépi tanuláson alapuló szoftver végzi. Ez annál is aggasztóbb, mert a pontszámoknak nagy jelentősége van, sávos, egyre súlyosbodó következményekkel járnak a gyűlő pontok. Ennek viszonylag részletes leírását is megtalálhatjuk. Eszerint egy hibapontra csak figyelmeztetés jár, kettő-hat közt bizonyos funkciók kerülnek letiltásra, míg tíz hibapont felett 30 napra korlátozzák a fiókot.

A harmadik szankciótípus a fiókkorlátozás, ezt ismétlődő, súlyos szabálysértések esetén alkalmazzák, az időtartama változhat, pár naptól 30 napig.

A negyedik, és legsúlyosabb szankció a fióktiltás. *A gyermekpornográfia közzététele azonnali fióktiltást eredményez*, de az ún. veszélyes személyek és szervezetek (pl. terrorszervezetek, ha lelepleződnek) is ilyet kapnak. Ezeket a kiemelt eseteket leszámítva a fióktiltást többször ismétlődő súlyos szabálysértések esetén alkalmazzák, bár ennek szempontjai szintén nem transzparensek.

A fióktiltáshoz hasonló az oldal és csoporttiltás (ez esetben a fő profilt nem tiltják le, csak a profil által gondozott csoportot vagy oldalt). Ez a közösségi szabályok megsértésén felül bekövetkezhet akkor is, ha a csoport neve, leírása, vagy borítóképe sértő.

Minden szankcionáló döntés esetén lehetséges a panasz, a felülvizsgálat, amely szintén MI alapú, de ennek részleteit ebben a szakértői jelentésben nem ismertetem. A tiltásokról a fiók tulajdonosa mindig kap értesítést („inform"), de a magyarázat általában csak a fő okot tartalmazza, pl. „gyűlöletbeszéd", így a szabályok és a konkrét poszt vagy aktivitás közötti kapcsolat részletes magyarázata mindig elmarad. Ez a mesterséges intelligencia jellegzetességeiből fakad: az MI statisztikai mintafelismerés alapján dönt, így nem képes a konkrét „tényállást" „felállítani", és a szabályok szövegét úgy „értelmezni", hogy az hasonlítson egy ember magyarázatára. (Példánkban, képtelen arra, hogy elmagyarázza, hogy mi konkrétan az adott posztban a gyűlöletbeszéd, melyik védett kisebbséget sérti, melyik kifejezés, miért sértő ez rájuk nézve, stb.)

2.2. Youtube

2.2.1. Anyagi szabályok

A Youtube az általános szerződési feltételeiben az alábbi módon szabályozza a tartalomeltávolítást:

> „A Tartalom YouTube általi eltávolítása
>
> Arra az esetre, ha észszerű okunk van azt feltételezni, hogy az Ön bármely Tartalma (1) sérti a jelen Szerződést, vagy (2) kárt okozhat a YouTube-nak, a felhasználóinknak vagy harmadik feleknek, fenntartjuk a jogot az ilyen Tartalom egészének vagy egy részének eltávolítására. Ilyen esetekben értesítjük a döntésünk okáról, kivéve, ha okkal feltételezhetjük, hogy ez (a) sértené a törvényeket vagy valamely végrehajtó hatóság utasítását, vagy egyéb módon a YouTube vagy Társult vállalkozásaink jogi felelősségre vonását eredményezné; (b) akadályozná egy nyomozás menetét vagy a Szolgáltatás integritását vagy működését; vagy (c) kárt okozna bármely felhasználónak, egyéb harmadik félnek, a YouTube-nak vagy Társult vállalkozásainknak. A Súgónk Problémamegoldás oldalán további információt talál a bejelentésekről és a szabályok betartatásáról, ezen belül a fellebbezés módjáról."

Ami a fentebbi feltételekben figyelemreméltó, hogy nemcsak a kifejezetten jogellenes, sőt még csak nem is csak a szabályzatellenes tartalmakat távolíthatja el a Youtube, hanem lényegében bármit, ami „kárt okozhat" a Youtube-nak. Az indokolási kötelezettség szintén elég lazán van megfogalmazva, ugyanis ugyanezzel a szóhasználattal, „amennyiben az kárt okozna a Youtube-nak vagy a társult vállalkozásainak", a platformnak nem kötelessége közölni az indokokat.

A Fiók megszüntetést és felfüggesztést a Youtube hasonlóan szabályozza, azzal a különbséggel, hogy itt a szabálysértésnek és a károkozásnak „jelentős mértékűnek" kell lennie. Az indokolási kötelezettsége alól a Youtube itt is könnyedén kibújhat, ha akar.

> „A YouTube által kezdeményezett megszüntetés és felfüggesztés
>
> A YouTube fenntartja a jogot, hogy felfüggessze vagy megszüntesse az Ön Google-fiókját, illetve a Szolgáltatás egy részéhez vagy egészéhez való hozzáférését, ha: (a) Ön jelentős mértékben vagy ismételten megszegi a jelen Szerződést; (b) erre jogszabályi követelménynek vagy bírósági végzésnek való megfelelés érdekében van szükség; vagy (c) észszerű okunk van azt feltételezni, hogy olyan

tevékenység történt, amely kárt okozhat a felhasználóknak, egyéb harmadik feleknek, a YouTube-nak vagy a Társult vállalkozásainknak, illetve ezek felelősségre vonását eredményezheti.

Megszüntetésről vagy felfüggesztésről szóló értesítés

Ilyen esetekben értesítjük a YouTube általi megszüntetés vagy felfüggesztés okáról, kivéve, ha okkal feltételezhetjük, hogy ez (a) törvényszegésnek vagy valamely végrehajtó hatóság utasítása megszegésének minősülne; (b) akadályozná egy nyomozás menetét; (c) akadályozná a Szolgáltatás integritását, működését vagy biztonságát; vagy (d) kárt okozna bármely felhasználónak, egyéb harmadik félnek, a YouTube-nak vagy Társult vállalkozásainknak."

A Youtube hasonlóan a Facebook-hoz alapvetően szintén a közösségi irányelvekben rögzíti azokat a szabályokat, amelyek megsértése korlátozásokhoz vagy szankciókhoz vezethet. Itt rögtön érdemes két megjegyzést tenni.

Rendkívül problematikus, hogy hasonlóan a Facebook szabályaihoz, itt sem lehet egymástól a jogellenes (illegal) és a sértő („lawful-but-awful"[4]) tartalomtípusokat elválasztani. Az oldalon az „Erőszakos vagy veszélyes tartalom" csoporton belül található, mondjuk a bombakészítés bemutatását tiltó szabály, ugyanúgy, mint mondjuk a dezinformációra („téves információ") vonatkozó tilalom. Első látásra is nyilvánvaló, hogy ezek nem egy súlycsoportba tartozó problémák.

A másik, hogy annak ellenére, hogy összességében csaknem 100 aloldalról, és több száz oldalnyi, egészen részletes információról beszélünk, az oldal tetején egy rövid mondat utal arra, hogy koránt sincsen szó a teljes normarendszerről: a Youtube tulajdonképpen ezeken felül és ezeken kívül akármikor kiszabhat szankciókat. „Ez a lista nem teljes." - írják a „Káros vagy veszélyes tartalmakra vonatkozó irányelv" oldal „Példák káros vagy veszélyes tartalomra" fejezetének elején.

A tiltásokat a Youtube öt, logikailag elég nehezen értelmezhető kategóriába sorolja. Az első kategória a *Spam és megtévesztő módszerek* címet viseli, és tulajdonképpen a technikai jellegű visszaéléseket foglalja össze, a spamtől a mutatóhamisításig, amely a sok szempontból kulcsfontosságú nézettség-számláló manipulálását, mesterséges pörgetését tilalmazza. A második kategóriába, egy nem pontosan érthető szempont miatt lényegében a *pornográfia, a gyermekpornográfia és gyermekekkel kapcsolatos erőszak, a vulgáris beszéd és az önsértés* különböző formái kerültek. A harmadik kategória az *erőszak különböző formáinak* bemutatását tilalmazza. Ez alá a kategória alá szintén egészen különböző tartalomtípusok kerültek, az extrém módon veszélyes kihívásoktól és a gyilkosságok élőben történő közvetítésétől a gyűlöletbeszédig. A negyedik a „*Szabályozott áruk*" címet viseli, és lényegében a drog és a fegyverkereskedelemre irányuló tartalmakat tilalmazza, végül az ötödik a *dezinformációs* tartalmakat.

2.2.2. Eljárási szabályok és folyamat

A Youtube elvileg minden jogsértés-típus esetén ugyanazt a szankciórendszert alkalmazza, amely a feltöltési lehetőség felfüggesztése – figyelmeztetés – tanfolyam elvégzése – fiókmegszüntetés kombinációját takarja. A döntések ellen a Youtube minden esetben fellebbezési lehetőséget biztosít.

Ez az egyöntetűség – akárcsak a Facebook esetében – két helyen biztosan csorbul. Egy helyen találunk arra utalást, hogy léteznek olyan komoly visszaélések, hogy azok azonnali csatornamegszüntetéshez vezetnek, tehát ezek esetén nem a sztenderd eljárást alkalmazza a Youtube. A magyarázó oldalon

[4] Eric Goldman; Jess Miers, "Online Account Terminations/Content Removals and the Benefits of Internet Services Enforcing Their House Rules," Journal of Free Speech Law 1, no. 1 (2021): 191-226

három ilyen magatartást említenek, az ún. ragadozó viselkedést, a spamet és a pornográfiát. A pornográfiával kapcsolatban nincsenek magyarázó oldalak, de az ún. „ragadozó viselkedésre", és a spamre igen. Előbbi lényegében a gyermekekkel kapcsolatos szexuális vagy erőszakos tartalmak közzétételét jelenti. Mivel a vonatkozó oldalon az általános szankciókat találjuk, nem egyértelmű, hogy ezen a tartalomtípuson *belül* milyen jellegű tartalom lehet az, amely azonnal csatornabezárást eredményez. A spam és a megtévesztő módszerek és átverések esetén a Youtube jelzi, hogy a bevételszerzés felfüggesztése és a csatorna megszüntetése az elsődleges szankció. Mivel a bevételt szerző csatornákra és a bevételszerzésre használt tartalmakra a Youtube egy teljesen külön, önálló szabályozási univerzumot alakított ki, ennek a részleteire nem térek itt ki. A lényeg tehát az, hogy a Youtube egy relatíve kiszámítható eljárási és szankciórendszerrel dolgozik, kivéve persze az említett három tartalomtípust.

Ha nem erről a három esetkörről van szó, a Youtube első alkalommal ún. figyelemfelhívást küld (a felhasználó emailcímére, vagy más megjelölt helyre). A figyelmeztetés a következő információkat tartalmazza:

- Az eltávolított tartalom
- A megsértett irányelvek (pl. zaklatás vagy erőszak)
- Az irányelvsértés hatása a csatornára
- A további lehetőségek

A további lehetőségek közt egy tanfolyam elvégzésének lehetősége áll fenn: ennek az az előnye, hogy a csatornára nyilvánosan kitett figyelmeztetés a tanfolyam elvégzése után 90 nappal törlődik. Ha a csatorna az első „figyelemfelhívást" követő 90 napon belül másodszor is megszegi a szabályokat, akkor „figyelmeztetést" („strike") kap, amely már viszonylag komoly szankciókkal, funkciókorlátozásokkal is együtt jár. Ezek a funkciókorlátozások a következők:

- Videó vagy élő közvetítés feltöltése
- Ütemezett élő közvetítés indítása
- Videó nyilvánossá tételének ütemezése
- Premier létrehozása
- Előzetes hozzáadása közelgő premierhez vagy élő közvetítéshez
- Egyéni indexkép és közösségi bejegyzés létrehozása
- Lejátszási lista létrehozása, szerkesztése és együttműködők lejátszási listához való hozzáadása
- Lejátszási lista hozzáadása a videóoldalhoz és eltávolítása onnan a Mentés gomb használatával

Ha a 90 napos periódus alatt még egy figyelmeztetést kap a csatorna, akkor 2 hétig nem lehet a fentebbi funkciókhoz hozzáférni. A harmadik figyelmeztetés után a csatornát végleg eltávolítják.

Ha bármilyen döntés születik, a felhasználó előtt két út áll: vagy fellebbez, vagy elvégez egy tanfolyamot. Mindkettőre van lehetőség akár egyidőben is. Ha a csatornát megszüntették, további fellebbezésre van lehetőség. Nem világos, hogy a fellebbezés emberhez kerül-e, illetve mikor kerülhet oda. A panaszok nagyságrendjét látva nagyon valószínű, hogy első körben mindenképpen gépek bírálják el a panaszt, de hogy milyen esetekben kerülhet végül emberi ügyintéző elé, az teljesen

homályos, illetve elég valószínű, hogy például az ügyfél erőszakossága, vagy fenyegetőzése is szerepet játszik.[5]

3. A banning és a shadow banning a számok tükrében – az európai összkép

3.1. A tartalomkorlátozásokkal kapcsolatos európai statisztikák

Az óriásplatformok moderálási gyakorlatáról a Digitális Szolgáltatásokról szóló rendelet[6] (a továbbiakban: DSA) hatályba lépésig szinte semmit nem lehetett tudni.

A DSA alapján az óriásplatformok elkezdték publikálni a moderációs (tartalom- és fiókkorlátozási) döntéseiket. Ezekről a döntésekről alapvetően három forrásból tájékozódhatunk.

1. *Európai transzparencia-adatbázis*: a DSA alapján a platformok kötelesek tájékoztatni a felhasználókat a moderációs döntéseik okairól, és erről adatokat is kell, hogy szolgáltassanak egy adatbázisba. A platformok által szolgáltatott adatokat egy, a Bizottság által karbantartott oldalon találhatjuk, amely egyrészt (elvileg) a platformok által szolgáltatott adatokat nyers formátumban tartalmazza, (CSV fájlok formájában), másrészt különböző módon lekérdezhetővé teszi, és vizualizálja azokat.[7]
2. *Transzparenciajelentések*: a két, általam vizsgált platform (a két utolsó jelentés, a Facebooké itt , a Google-é itt) közzéteszi (a két fentebbi link mindkét nagy platform utolsó közzétett negyedéves jelentésére mutat) egy dokumentum formájában a DSA-ban előírt transzparencia-jelentéseket is. [8]
3. Végül mindkét nagy platformnak van saját szempontok szerint lekérdezhető, *és az egész világra vonatkozó* transzparencia-adatbázisa is. A Facebook például egy hatalmas excel file-ba gyűjti világszinten az adatokat. Ennek a legfőbb számait a *Függelékben* közlöm.

---

[5] Stöckert Gábor: Kitiltás magyarázat nélkül, visszaszerezhetetlen profilok – kalandjaink a techcégek ügyfélszolgálati útvesztőjében. *Telex.hu*, 2024. 03. 18. https://telex.hu/techtud/2024/03/18/ugyfelszolgalat-big-tech-kitiltas-facebook-paypal-dsa

[6] 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) a továbbiakban DSA

[7] A Youtube statisztikáit tartalmazó fájlok például 37 adatmezőt tartalmaznak, (a döntés alapjául szolgáló tényektől a döntés jellegén keresztül a jogalapokig). Sem az oldalon, sem a fájlokban nincsen utalás arra, hogy a felhasználó milyen országból származik, így tagállami szintű statisztikák nem készíthetők. Egy felirat utal arra, hogy van egy „territorial scope" elnevezésű mező is, de ez elég valószínűen a tartalomkorlátozás területi hatályát jelöli, és nem a felhasználó nemzetiségét. A nyers fájlok, - legalábbis a Youtube esetében – átlagosan 20-30, (de olykor több mint 100) MB méretű fájlt jelentenek *napi szinten*. Bár próbáltam a fájlokat letölteni és feldolgozni, (a Youtube esetében) ez a legnagyobb igyekezetemre sem sikerült: a tömörített fájlok fejléce letöltődött, de a bennük található adatok nem. A vizualizáció (Dashboard https://transparency.dsa.ec.europa.eu/dashboard) egyelőre szintén nehezen használható, mert egyrészt nem lehet országra szűkíteni, másrészt nem lehet az adatokat a „dashboard-ról" kiexportálni.

[8] Itt is egy sor nagyon bosszantó problémával szembesülünk. A jelentések számadatai nem exportálhatók, de még problémásabb, hogy más és más időszakokat ölelnek fel. A Google jelen sorok írásakor rendelkezésre álló legutolsó jelentése a 2024. március 1. és június 30. közti időszakot öleli fel. (4 hónap) A Facebook esetében ez az utolsó hozzáférhető jelentés a 2024. április és a 2024. szeptember közötti időszakot (6 hónap) jelenti.

Az itt következő néhány oldalon a jobb összehasonlíthatóság és átláthatóság kedvéért a DSA által előírt transzparenciajelentésekben található táblázatokat (2. pont) közlöm, ezek közül is mindkét nagy platformszolgáltatás esetén a moderált bejegyzésekkel kapcsolatos adatokat.

3.2. Facebook adatok a DSA transzparenciajelentés alapján

A DSA alapján közzétett jelentés 10 fejezetre tagolódik. Ezek közül a 3. ('Tartalommoderáció a cég saját kezdeményezése alapján'), a 4. ('A Meta belső panaszkezelési mechanizmusa alapján kapott panaszok'), és az 5. ('A tartalommoderáció automatizált eszközei') pontok alatt található adatok lehetnek számunkra érdekesek.

Itt két táblázatot közlök, a saját hatáskörben történő *eltávolításokról*, és a saját hatáskörben *fékezett terjesztéséről* (demoted) szólót. Ezek a táblázatok két ok miatt sem feleltehetők meg a Függelékben később közölt táblázatoknak. Egyrészt azok világszintű adatokat tartalmaznak, másrészt az időszak is más: a DSA transzparenciajelentések a 2024 április és szeptember közötti időszakot ölelik fel, míg a világszintű jelentések naptári negyedévesek. Még egy furcsaságot érdemes megjegyezni, az első táblázat (moderálási döntések kategóriánként) végösszege 35 millióval magasabb, mint a táblázatban szereplők összege – erről a 35 milliónyi moderációs döntésről, amely az „egyéb" soron szerepel, semmit nem tudunk.

Jogsértés jelleg	Eltávolított tartalmak száma	Arány a teljesen belül	Automatikusan eltávolított tartalmak	Automatikus %
Felnőtt meztelenség és szexuális tevékenység	2 246 208	4,6%	2 125 014	94,6%
Bullying és zaklatás	915 187	1,9%	508 435	55,6%
Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás és	168 641	0,3%	26 476	15,7%
Gyermekek veszélyeztetése: szexuális kizsákmányolás	312 717	0,6%	287 710	92,0%
Veszélyes szervezetek: gyűlölködő szervezetek	191 002	0,4%	148 544	77,8%
Veszélyes szervezetek: terrorizmus	292 393	0,6%	269 003	92,0%
Gyűlöletbeszéd	1 151 054	2,3%	949 336	82,5%
Korlátozott áruk: drogok	83 109	0,2%	21 524	25,9%
Korlátozott áruk: fegyverek	170 542	0,3%	153 205	89,8%
Spam	7 382 238	15,0%	7 250 783	98,2%
Öngyilkosság és önsértés	129 153	0,3%	114 297	88,5%
Erőszakos és erőszakot ábrázoló képi tartalom	565 616	1,1%	425 712	75,3%
Erőszak és uszítás	67 295	0,1%	57 762	85,8%
Más, a táblázatban nem jelzett okok miatti eltávolítás	35 686 213	72,3%	34 421 105	96,5%
Összes (beleértve más szabálysértéseket is)	49 361 368	100,0%	46 758 906	94,7%

1. táblázat A Facebook által 2024 áprilisa és 2024 szeptembere közt eltávolított tartalmak az Európai Unió területén (darab)

Ami a táblázatból azonnal kiolvasható, az a moderációs döntések hatalmas mennyisége, és ezen belül a gépi moderáció elsöprő aránya. A második megfigyelés, amit tehetünk (és ez igaz világszinten is), hogy a legtöbb eltávolítás a spamek (kéretlen üzenetek, posztok) miatt van, bár a Függelékben található világméretű adatokban a spam és a hamis fiókok aránya jóval nagyobb az itt közöltnél. A harmadik megfigyelés, amit fentebb is említek, hogy az „egyéb" okok miatti eltávolítások aránya hatalmas. Ez elég valószínűen nagyobbrészt a „hamis fiókokat" jelenti. Érdekes, hogy a gépi detektálás aránya a pornográfiánál a legnagyobb, és meglepően kicsi a gyermekek veszélyeztetésének enyhébb eseteinél, vagy a drogoknál. A Facebook kommunikációjától némileg eltérő módon a gyűlöletbeszéd gépi felismerési aránya csak 82,5%-os. Végül megjegyzem, hogy a Függelékben található világszintű adatok sok szempontból más arányokat és képet mutatnak, mint a fentebbi (csak Európára vonatkozó) táblázat, de ennek okai – az eltérő módszertan és csoportosítások mellett, amelyek csak részben magyarázzák a különbségeket – nem világosak.

Ennél is jóval érdekesebb a „demoted" azaz a tartalom terjedésének korlátozásáról szóló táblázat.

Organikus korlátozott terjesztésű tartalmak	Korlátozás mennyisége	A tartalom aránya a teljesen belül	Automatikusan korlátozott	Automatikus%
Felnőtt meztelenség és szexuális tevékenység	1 088 521	4,0%	1 088 518	100,0%
Bullying és zaklatás	311 158	1,2%	311 157	100,0%
Ellenőrzött dezinformáció	18 989 014	70,3%	18 977 541	99,9%
Gyűlöletbeszéd	205 577	0,8%	205 576	100,0%
Korlátozott áruk: drogok	384 783	1,4%	384 783	100,0%
Öngyilkosság és önsértés	43 955	0,2%	43 955	100,0%
Erőszak és uszítás	398 226	1,5%	398 226	100,0%
Erőszakos és erőszakot ábrázoló képi tartalom	5 600 461	20,7%	5 600 449	100,0%
Összesen	27 021 695	100,0%	27 010 205	100,0%

2. táblázat A Facebook által 2024 áprilisa és 2024 szeptembere közt korlátozott tartalmak az Európai Unió területén (darab)

Ebből ugyanis világossá válik, hogy a korlátozás (shadow banning) legfőbb célpontja a dezinformáció (több mint 70%). A dezinformációról viszont azt érdemes tudni, hogy nem egy egyértelmű 0/1 jelenség, hanem rengeteg átmeneti kategóriája van. A dezinformáció legtöbbször nem kontrafaktuális közléseket takar, hanem csak egy eltérő narratívába helyezést, vagy egyszerűen csak egy túlzó, érzelmektől fűtött véleménynyilvánítást.

3.3. Youtube adatok a DSA transzparenciajelentés alapján

Ahogy fentebb említettem a Youtube részben eltérő időszakra vonatkozóan közöl számokat, és a transzparencia jelentésének szerkezete is más jellegű. Először is, mivel a Youtube-nak nincsen önálló transzparenciajelentése, a Google jelentéséből lehet a Youtube-ra vonatkozó sorokat kiolvasni. Másrészt ez a jelentés a március 1.-június 30. közötti időszakra vonatkozik. Az első táblázat az időszakon belül a korlátozott videók száma.

	Korlátozott videók	Automatizált eszközökkel korlátozottak	Automatikus %
Felhasználói videók	24 848 592	24 582 688	98,9%

3. táblázat Illegális, vagy szabályzatba ütköző eltávolított tartalmak (videók) a Youtube-on az Európai Unióban 2024 március 1 – június 30 közt

Témánk szempontjából lényeges, hogy a Youtube is korlátozza a tartalmak elérhetőségét, ő is alkalmaz shadow banninget.

	Láthatóság korlátozása	**Monetizáció korlátozása**	**Account szintű korlátozások**
Felhasználói videók	29 760 176	328 698	265 805
Hirdetők videói	50 162 201	N/A	29 468

4. táblázat A láthatóság korlátozása a Youtube-on az Európai Unióban 2024 március 1 – június 30 közt

Mivel ez a jelentés nem tartalmazza a típusonkénti bontást, ugyanerre az időszakra a 3. pont alatt szereplő, EU Bizottság által működtetett adatbázist kérdeztem le. A lekérdezés végösszegének nagyjából összhangban kellett volna lennie a fentebbi transzparencia jelentésben található számmal, hiszen ugyanazt az időszakot választottam a lekérdezés során (2024. március 1. június 30.), de meglepetésemre ez nem így volt, és részben a korlátozások csoportosítása is más, mivel itt két

kategóriát találunk, a „visibility” és a „service provision” kategóriákat, amely nagyjából a tartalomkorlátozás és a fióktiltás kategóriáinak felel meg. Összességében a két kategória 46 240 061 beavatkozást takar a második negyedévben, amelyből 291 164 darab a fióktiltás. Ha a transzparenciajelentésben található két számot (tartalom tiltása és láthatóság-korlátozás) összeszámítjuk, ennél jóval nagyobb számot (54 millió körüli számot) kapunk. Az eltérésre a magyarázatot egyelőre nem találom, de az mindenképp tanulság, hogy a különböző adatokat összhangba kellene hozni egymással. (Azt fel sem merem tételezni, hogy a Google nem valós adatokat szolgáltat.)

A tartalomkorlátozások és tiltások okok szerinti megoszlása az adatbázisban az alábbi:

	Darab videó	% aránya az egészen belül
A platform szabályzatába ütköző	29 514 639	63,8%
Átverések, csalások	13 141 904	28,4%
Szellemi alkotások tulajdonjoga	2 300 341	5,0%
Nem biztonságos és illegális termékek	598 115	1,3%
Kiskorúak védelme	488 078	1,1%
Pornográfia és szexuális tartalom	163 123	0,4%
Adatvédelem és privacy	22 184	0,0%
Illegális és veszélyes beszéd	11 666	0,0%

5. táblázat Illegális, vagy szabályzatba ütköző eltávolított tartalmak (videók) a Youtube-on az Európai Unióban kategóriák szerint

Ami igazán meglepő, hogy a platform saját szabályzatába ütköző, (tehát nem illegális) tiltások teszik ki a teljes tiltás-mennyiség csaknem kétharmadát. Ezen belül nem tudjuk, hogy mennyi a gyűlöletbeszéd (illetve a „veszélyes beszéd” kategóriája tartalmazza-e), és semmilyen információnk nincsen a dezinformációnak minősített tartalmak miatti korlátozásokról. Pedig ezek a kategóriák azok, amelyek a leginkább érintik a szólásszabadságot, és ahogy az a mélyinterjúk során kiderült, az egyik médiumnak a Youtube-ról történő teljes kitiltása az egyik alkalommal részben éppen a dezinformáció miatt történt.

4. A magyar helyzet

4.1. Általános kép három felmérés tapasztalatai alapján

A nagy platformok nem közölnek országonként lebontott számokat a moderálási döntéseikről. (A transzparenciajelentésekben is csak a hatósági megkeresések láthatók országonkénti bontásban.) Némi indikációt a számokkal kapcsolatban – és csak a Facebook kitiltási gyakorlatának egyes aspektusait illetően – az NKE Információs Társadalom Kutatóintézete által már harmadik alkalommal lefolytatott reprezentatív kérdőíves felmérésből lehet szerezni.

A kutatásról és a lentebbi táblázatokról nagy vonalakban azt érdemes tudni, hogy azok egy, a legfőbb demográfiai jellemzőkre nézve (korra, nemre, településtípusra és végzettségre) reprezentatív (~1000 fős) mintán készültek. A kitiltással kapcsolatban a kérdőív négy kérdést tartalmaz, ezek, és az ezekre adott válaszok láthatók az alábbiakban táblázatos és diagramm formában. A táblázatok alsó sorában szerepel, hogy az 1000 fős mintán belül hányan válaszoltak az adott kérdésre, illetve hány főre értelmezhető egyáltalán a kérdés. (Így pl. az első táblázatban az 1000 főből 2020-ban valamivel több, mint 700 fő volt Facebook felhasználó, így összesen 704 fő válaszát tartalmazza az első oszlop, hiszen

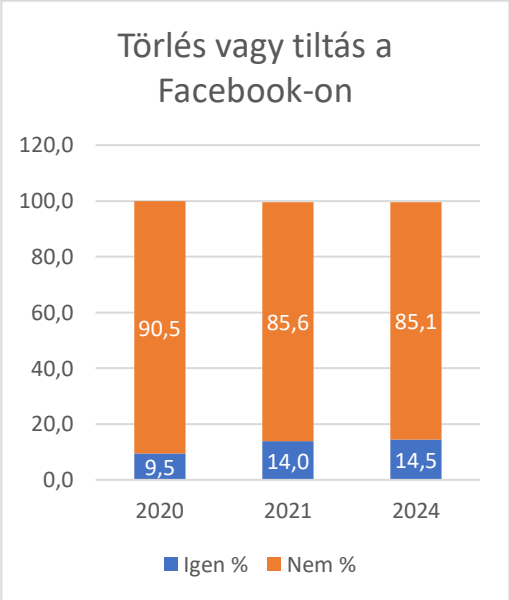
a kitiltással kapcsolatban csak ennek a populációnak a válaszai értelmezhetők. Ugyanígy, a második táblázatban már csak a törlést, vagy letiltást elszenvedett – 2020-ban 9,5 %, – 65 fő válaszainak megoszlását mutatja a második táblázat).

Teljes adatsor (mindhárom kutatásban) csak a négy alapkérdés esetén áll rendelkezésre, (1. törölték-e a bejegyzést, 2. kapott-e tájékoztatást, 3. kérvényezte-e a döntés megváltoztatását és 4. mi lett a kérvényezés eredménye) a 2024-es kutatásban ugyanakkor három további, a téma mélyebb megértését, a bannelés típusára, (tartalom vagy fiók) és gyakoriságára rákérdező kérdést is tartalmaz: 1. hány alkalommal vette észre, hogy a Facebook törölte a hozzászólását, 2. előfordult-e hogy a Facebook megszüntette a személyes profilját, 3. hány alkalommal függesztette fel a hozzáférést nem személyes profilhoz. Ezek a kérdések tehát csak a 2024-es kutatásban állnak rendelkezésre, de a releváns részeknél beszúrom ezeket is.

Nézzük tehát először, hogy mennyire tekinthető gyakorinak a bannelés Magyarországon

6. táblázat és diagramm Előfordult Önnel, hogy egy bejegyzését törölte a Facebook, vagy egy időre letiltották a szolgáltatás használatáról?

		2020	2021	2024
Igen	%	9,5	14,0	14,5
Nem	%	90,5	85,6	85,1
Nincs válasz	%	0,0	0,4	0,4
Elemszám	darab	704	702	725



Miközben a Facebook felhasználók száma nem növekszik szignifikánsan az évek során, a negatív döntést (bannelést) elszenvedettek száma 2020 és 2021 közt enyhe emelkedést mutat. Ez sok mindennek lehet betudható, de a legvalószínűbb az, amit maga a Facebook is elismer, hogy a rendszerei fejlesztése nyomán azok egyre hatékonyabban szűrik ki a jog- és szabályzatellenes aktivitásokat. 2021 és 2024 közt ugyanakkor ebben már nincsen szignifikáns emelkedés.

Összességében jelenleg a magyar Facebook felhasználók kb. 15%-a szenvedett már el valamilyen moderációs döntést. A döntés többféle lehet (az egyszerű poszttörléstől az ideiglenes profiltiltáson keresztül a profilfelfüggesztésig), és ez a táblázat ömlesztve mutatja ezeket a döntéseket.

A 2024-es felmérés ugyanakkor már tartalmaz arra vonatkozó adatokat is, hogy a 2023-as év végi adatfelvételt megelőző 3 évben a moderációs döntést elszenvedett felhasználók mekkora hányada kapott, és milyen gyakran tartalomtörlési üzenetet és fiókfelfüggesztést.

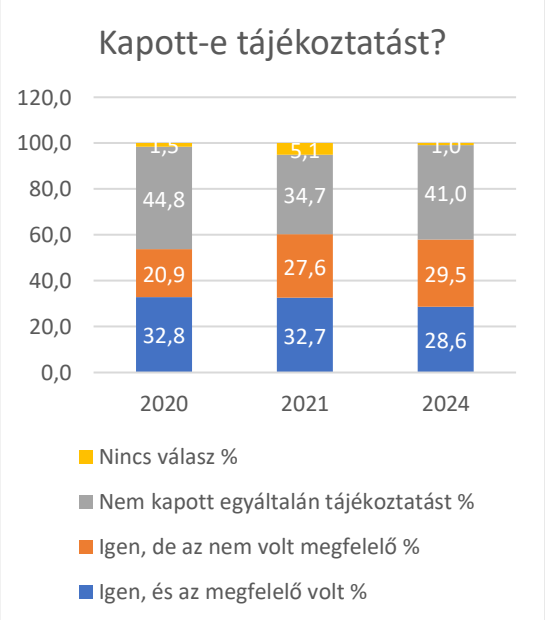
7. táblázat Kb. hány alkalommal vette észre az elmúlt 3 év során, hogy törölte hozzászólását, vagy felfüggesztette a fiókját a Facebook algoritmusa vagy dolgozói?
Kb. hány alkalommal vette észre az elmúlt 3 év során, hogy a Facebook felfüggesztette az Ön személyes profilját?

	% tartalomtörlés	Összes törlés %	Összes fiókfelfüggesztés %
Egyszer sem	85,9		
1 alkalommal	3,2	12,5	3,3
2-5 alkalommal	6,7		
6-15 alkalommal	1,3		
15-50 alkalommal	1,2		
Több mint 50 alkalommal	0,1		
Nem tudja	1,6		
Total	100,0		

Ebből látható, hogy a Facebook-felhasználók 12,5%-a a 2023 év végi adatfelvételt megelőző három évben szembesült valamilyen tartalomtörlési döntéssel[9], -és 3,3%-uknak legalább egyszer felfüggesztették a fiókját is. Ez nem tűnik nagy számnak, de – mivel reprezentatív felmérésről van szó – a 7 millió Facebookfelhasználóból kiindulva ez több, mint 200 000 felhasználót érint.

8. táblázat és diagramm Kapott tájékoztatást a törlés és/vagy letiltás okáról?

		2020	2021	2024
Igen, és az megfelelő volt	%	32,8	32,7	28,6
Igen, de az nem volt megfelelő	%	20,9	27,6	29,5
Nem kapott egyáltalán tájékoztatást	%	44,8	34,7	41,0
Nincs válasz	%	1,5	5,1	1,0
Elemszám	darab	67	98	105

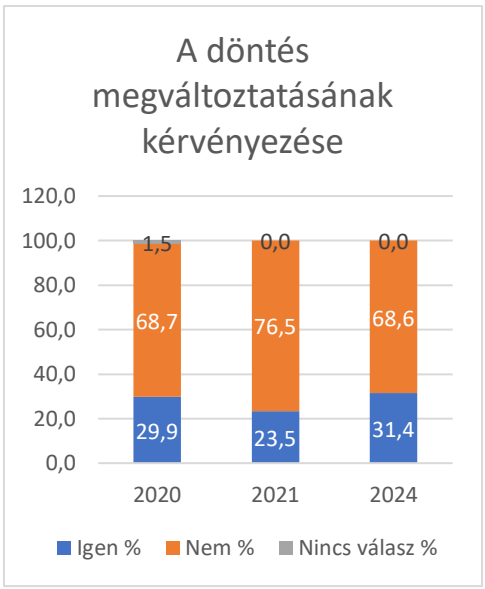


[9] A 11. táblázatban található 12,5% 2%-kal kevesebb, mint a 10. táblázatban található 14,5%. A kérdéseket a kérdéssor két különböző pontján tették fel, a különbség ebből adódik.

A második kapcsolódó táblázat azt tartalmazza, hogy akik moderációs döntést szenvedtek el (tehát az a 9,5 – 14,5%), kaptak-e tájékoztatást a döntés okáról. A válaszadók nagyjából 40%-a semmilyen tájékoztatást nem kap a moderációs döntésekről, akik pedig kapnak, azoknak nagyjából a fele elégedetlen a magyarázattal. Mivel a moderációs döntést gépek hozzák, a magyarázat a legtöbb esetben általános, és csak azt tartalmazza, hogy a szabálysértés melyik általános, nagy kategóriába esik bele. Így tulajdonképpen azt mondhatjuk, hogy az a csoda, hogy a kitiltottak fele egyáltalán megelégszik ezzel az általános magyarázattal, és csak a felük elégedetlen.

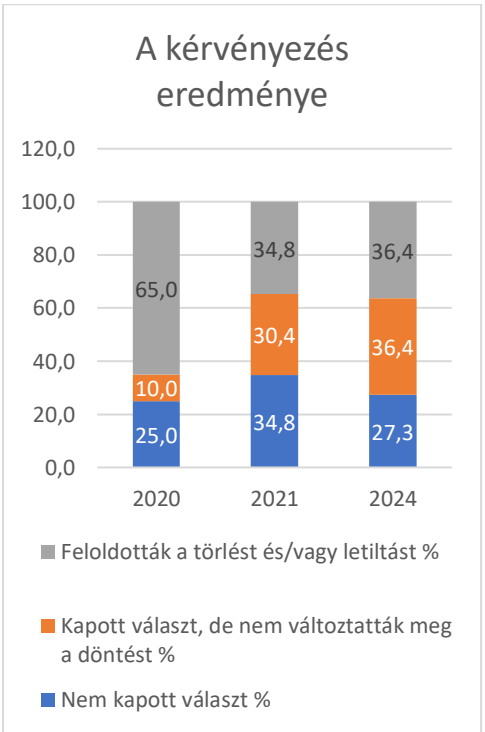
9. táblázat és diagramm Kérvényezte Ön, hogy változtassák meg a törlésről és/vagy letiltásról szóló döntést?

		2020	2021	2024
Igen	%	29,9	23,5	31,4
Nem	%	68,7	76,5	68,6
Nincs válasz	%	1,5	0,0	0,0
Elemszám	darab	67	98	105



10. táblázat és diagramm Mi lett a kérvényezés eredménye?

		2020	2021	2024
Nem kapott választ	%	25,0	34,8	27,3
Kapott választ, de nem változtatták meg a döntést	%	10,0	30,4	36,4
Feloldották a törlést és/vagy letiltást	%	65,0	34,8	36,4
Elemszám	darab	20	23	33



A valamilyen módon korlátozottak nagyjából 30%-a kéri a döntés felülvizsgálatát, amelyből a 2020-as kutatásban az elsöprő többség (65%) nyilatkozta, hogy a panasz után feloldották a döntést, azonban ez a 2021-es és a 2024-es adatfelvétel során 34-35%-ra esett vissza. A jelentős változás magyarázatát nem tudjuk, gyaníthatóan köze lehet ahhoz, hogy a Facebook hatékonysági (profit) szempontok miatt minden döntést mesterséges intelligenciáknak szeretne kiszervezni.

4.2. Esettanulmányok

A tanulmány írásához lefolytattam három mélyinterjút viszonylag nagy elérésű magyar médiumokkal.[10] (Egy esetben a vezető írásban válaszolt.) Ebből a három válaszból nagyon eltérő kép rajzolódott ki a médiumok platformokon szerzett moderációs és fióktiltási szubjektív élményeivel kapcsolatban. Míg két médium (két kormánykritikus, inkább baloldali beállítottságú médium) szubjekítve nem éli meg tragikusként a platformok bannelési és árnyéktiltási gyakorlatát, bár néhány jelenséget ők is kényelmetlennek vagy furcsának találnak, a megkérdezett jobboldali irányultságú médium szinte apokaliptikusként értékelte a helyzetet. Mivel ennyire eltérő vélemények vannak, az alábbiakban a két baloldali médium válaszait összevontan, míg a jobboldali véleményportál válaszait külön közlöm.

4.2.1. Akik szerint csak kicsi a baj...

A kérdéseket három csoportra osztottam: a közösségi médiában való jelenléttel kapcsolatos adatok, a banninggel kapcsolatos adatok és a shadow banninggel kapcsolatos benyomások. A válaszadókat nem kértem a konkrét adatok közlésére (ez túl hosszú lett volna, és nem is biztos, hogy mindenki képes rá), hanem azt kértem, hogy nagyjából becsüljék meg, ha számokról van szó.

Ami a közösségi médiajelenlétet illeti, mindegyik médium kulcsfontosságúnak gondolta. A Partizánnak a fő disztribúciós platformja a Youtube, ahol 500 000 feliratkozott követője van, de aktív a többi médiában is, a Facebookon, az Instagramon, a TikTokon, és podcastok formájában az rss.com-on és a Spotifyon is. Ezeken a platformokon nagyjából 120 - 150 000 követővel rendelkezik. A 24.hu az rss-t és a Spotify-t leszámítva szintén minden közösségi médiumon jelen van. A Facebookon több, mint 1 millió követővel rendelkezik, a többi közösségi médiumon jellemző feliratkozó-követőszáma a 30 000 – 170 000-es sávban van.

Arra a kérdésre, hogy milyen gyakran fordul elő valamilyen tiltás vagy figyelmeztetés, mindkét médium a havi 1-2 alkalmat említette. Egyöntetűen azt mondták, hogy nem jelent komoly problémát számukra, és folyamatosan tanulnak, egyre jobban tudnak alkalmazkodni. Indokolást szinte soha nem kapnak, az esetek egy részében csak az általános kategóriára utal a platform. Ezek közül a 24.hu a szubjektív benyomás szerint a szerzői jogi problémákat találta a leggyakoribbnak (látszik valamilyen logo, kép vagy hallatszik valamilyen zene), a második leggyakoribb ok az erőszak és a vélt vagy valós pornográf tartalom. A pornográf tartalmakkal kapcsolatban két extrém esetet is megemlített a válaszadó, az egyik alkalommal a német médiapartnerük filmjének a promócióját korlátozták, egy másik alkalommal pedig egy szexuális segédeszközöket gyártó üzemben készült vágóképek miatt, egészen hosszú időre. A csatorna képviselője azt hangsúlyozta, hogy nagyon magas az érzékenységi küszöb mind a Facebook, mind a Youtube esetén, és főképp az első néhány percre kell odafigyelni a videókban. . Amikor az ismert „Elkúrtuk” filmről volt szó az egyik tudósításukban, a Youtube korlátozta a szó miatt az elérhetőséget.

Ezek az esetek mutatnak rá a gépi moderáció gyengeségeire. Ismert és elég gyakori jelenség ugyanis, hogy a médium valamilyen negatív társadalmi jelenségről tudósít, amelyet azután a platform éppen az

[10] A médiumok és az adatközlők a következők: Partizán (Naszályi Natália), 24.hu (Szigeti Péter) és Pesti Srácok (Huth Gergely)

adott jelenség „propagandájának” minősít, mert a mesterséges intelligenciák nem tudják értelmezni a tágabb kontextust. Ilyen, a tágabb kontextus megértésének a hiánya miatt tiltások többször megjárták az Oversight Boardot is, [11] de ilyen tiltás áldozata lett a Partizán csatornája is 2021-ban, amikor a 64 Vármegye mozgalomról tudósított.[12]

Azt mind a két médium kiemelte, hogy nagyon ritkán kapnak érdemi magyarázatot a tiltásokra, de a panaszkezelési mechanizmusokat eltérően értékelték. A csatorna a Facebook panaszkezelési mechanizmusát „egészen jónak” nevezte, és a supportjukkal is jó kapcsolatuk van, szemben a TikTok-kal, amelyről még nem kaptak soha semmilyen érdemi választ, és amely egyszer 30 napra korlátozta a fiókelérésüket, azonban nem tudtak vele érdemben kommunikálni. A portál azonban sommásan úgy látja, hogy nemcsak magyarázatot nem kapnak, de a panaszkezelési mechanizmus sem működik. A hirdetési csatornákon keresztül esetleg el lehet érni emberi ügyintézőket, de már fordult elő velük, hogy csak ügyvédi felszólításra volt hajlandó a portál reagálni.

Mindkét médium sajátos taktikákat alakított ki arra vonatkozóan, hogy meg tudjanak felelni a platformok igényeinek. A Partizán például olyankor is feltölti a videóit előzetes kontrollálásra a hirdetésekre vonatkozó ellenőrző felületre, ha közvetlenül nem áll szándékában monetizálni azt, mert tudja, hogy ha itt átmegy, akkor biztosan nem lesz vele gond. A 24.hu inkább tartalmilag szelektál – például gyermekpornográfiával összefüggő tartalmakat semmilyen, még bűnügyi tudósítás, stb. tehát neutrális vagy elítélő körítéssel sem tölt fel platformokra, ismerve a „kontextust nem értő” MI-k problematikáját.

A shadow banneléssel kapcsolatban mindkét médiumnak az volt a véleménye, hogy a jelenség létezik, de nagyon nehéz benne szabályszerűséget felfedezni, azt kiismerni, hogy az elérések miért ugrálnak.

4.2.2. Aki szerint nagy a baj...

A Pesti Srácok küzdelme a platformokkal évekre megy vissza. Ők is nagyon rá vannak utalva a platformokra, mert direkt kattintás csak a törzsolvasóiktól érkezik. A forgalmuk érezhető része, jelenleg kb. 10%-a érkezik a platformokon keresztül.

A csatorna nevében nyilatkozó szerint ők rendszeresen és folyamatosan szenvednek el tiltásokat és korlátozásokat, többször korlátozták a tartalmaik monetizálását is, többször letiltották őket ideiglenesen (pl. a 2022-es választások előtt), valamint a Youtube örökre törölte őket a felületéről gyermekpornográfiára hivatkozva. Az eset még évekkel ezelőtt történt, egy „privátra” állított rendőrségi videó miatt. Mivel a Youtube jelenlegi szabályai szerint (kérdés, hogy akkor is ez volt-e a szabály), a privátra állított (azaz csak a felhasználó által látható) videóknak is meg kell felelniük a közösségi rányelveknek, és a gyermekpornográfiáért azonnal tiltás jár, a Youtube valószínűleg a szabályainak megfelelően járt el, más kérdés, hogy ezek a szabályok ésszerűek-e. Az sem világos, hogy ez a szabály mindig része volt-e a Youtube szabályainak, a korábbi szövegverziók nem férhetők hozzá. Ráadásul, ahogy ezt több helyen jeleztem, a Youtube szabályai igen nagy teret adnak az önkényes kitiltásoknak a gumiszabályok miatt.

A válaszadó szerint nemcsak az jelenti a problémát, hogy az elvárások folyamatosan változnak és nagyon nehezen kiszámíthatók, hanem két további probléma is van a kitiltásokkal. Az egyik, hogy úgy tűnik, a korlátozások jobban sújtják a jobboldali véleményeket és csatornákat, mint a baloldaliakat, ahol

[11] Facebook Oversight Board 2022-005-FB-UA döntése: „A tálibok említése híradásban”, és Facebook Oversight Board 2022-004-FB-UA döntése: „Kolumbiai rendőrségi karikatúra”.
[12] Kasza János: Cenzúrát emleget és tiltakozik a Partizán, 2021. 12. 20. *Media1.hu*

szinte az összes szabályzatellenes tevékenységfajta (és ezek közül a leggyakoribb gyűlöletbeszéd és dezinformáció) sokkal jobban tolerált. A másik - a nyilatkozó szerint talán még nagyobb – probléma pedig az, hogy a platform egy csatorna bezárásakor és korlátozásakor nincsen tekintettel a követőbázisra. A Youtube csatornájuk bezárásakor annak százezret meghaladó követőbázisa volt. 2023-ban a Facebook csatornájukat törölték le, amelynek szintén 140 000 követője volt. A Facebook visszaengedte a csatornát, szemben a Youtube-bal, amely a médium által később nyitott fiókokat is letörölte utólag, arra hivatkozva, hogy azokat meg sem lehetett volna nyitni.

A válaszadó szerint minden tiltás és korlátozás alkalmával megpróbáltak kapcsolatba lépni az adott platformmal, de ez érdemben egyszer sem sikerült, és a korlátozásokat sem oldották fel egyetlen esetben sem.

Végül a válaszadó szerint a shadow banning igenis mutat szabályszerűséget. Teljesen egyértelmű, hogy a korlátozások azért sújtják elsősorban őket, és más jobboldali médiumokat, mert jobboldali beállítottságúak, és néhány társadalmilag érzékeny témát a mainstream liberális narratívától eltérően tárgyalnak. Ez az oka annak is, hogy szubjektív megítélésük szerint „(a) konzervatív oldal elérései azonos intenzitás és tartalom mellett ötödére-tizedére estek vissza az elmúlt 2 évben a Facebookon, de a demonetizálás, karanténba zárás a YouTube-on is rendszeres az egyéni csatornán publikáló kollégáknál.”

Összefoglalás, következtetések és ajánlások

A tanulmány az első részében a két óriásplatform (Facebook és Youtube) moderálási szabályairól az alábbi megállapításokat teszi.

1. Az óriásplatformok moderálási gyakorlatát szabályozó előírások egy hétköznapi átlagfelhasználó számára teljességgel átláthatatlanok. Már azt is nehéz kideríteni, hogy a tucatnyi szabályzat hogyan kapcsolódik egymáshoz, de ezt tovább nehezíti a szabályzatok megjelenítési módja, amely jellemzően egy-egy téma köré egymásba linkelt weboldalak tucatjait jelenti. Ez lényegében azt eredményezi, hogy egy átlagfelhasználó szinte képtelen átlátni, hogy milyen játékszabályok szerint tiltja le a platform, vagy korlátozza az általa közzétett tartalmak láthatóságát.

2. A szabályzatok folyamatosan változnak, de ezekről a változtatásokról a felhasználókat nem értesítik, legtöbbször csak az őket érő szankciók nyomán szereznek róluk tudomást. Erről lentebb a tanulmány részletesen szól. Ez azokat a felhasználókat, akik rendszeresen posztolnak, (mert pl. médiavállalkozások) folyamatos nyomás alatt tartja és folyamatos kísérletezésre, óvatosságra kényszeríti.

3. A szabályzatokat az esetek elsöprő többségében mesterséges intelligenciák tartatják be, sőt még a panaszkezelést is ezek végzik első körben. Ezek a mestereséges intelligenciák képtelenek valódi, érdemi magyarázatot nyújtani a moderálás (tartalomeltávolítás) és a fiókkorlátozás és törlés eseteiben. Legtöbbször csak egyszerűen közlik, hogy a korlátozás melyik nagy kategória alá esik. Működésükről lényegében semmit nem lehet tudni.

4. A panaszkezelési mechanizmusok nyomán a tartalom és fiókvisszaállítások, a „sikeres fellebbezések” ma már inkább kivételnek számítanak, mint főszabálynak. A panaszkezelést ugyanúgy mesterséges intelligenciák végzik, mint a moderációt, így nehezen várható el, hogy felülbírálják a másik MI döntését. Annak szempontjai, hogy mi kerül emberi ügyintéző elé, lényegében szintén teljesen átláthatatlanok.

A magyar felhasználókkal kapcsolatban az írás az alábbi következtetésekre jut. A moderálás problémája tömegeket érint Magyarországon, hiszen a felhasználók 10-12%-a említette (ez nagyjából 500 000 embert jelent), hogy már érte törlés vagy tiltás a Facebookon. Ennek a részhalmaznak ráadásul a felét

többször is érintette ilyen intézkedés, és nagyjából a negyedüknek a fiókját is felfüggesztették, amely még mindig százezres nagyságrend. Az érintett felhasználók nagy része, 35-44%-a úgy érzi, nem kapott megfelelő magyarázatot a moderációs döntésre, illetve a fióktörlésre, és 20-30%-uk, ha kapott is, elégedetlen volt a magyarázattal. Harmaduk kérte az intézkedés feloldását, de – a 2021-es és a 2024-es adatfelvételnél – csak a tizedüknek oldották fel a korlátozást. Ez az arány még jóval nagyobb (nagyjából a felfüggesztettek ötöde) volt 2019-ben. Ez a változás, akárcsak sok más, sokszor megmagyarázhatatlan „ugrálás" az adatokban gyakran annak köszönhető, hogy ezeket a döntéseket egyre inkább nem emberek, hanem mesterséges intelligenciák hozzák.

A megkérdezett magyar médiumok szintén elégedetlenek, bár eltérő mértékben a platformok moderálási gyakorlatával: főként a nehezen kiismerhető és változó szabályokat, a konkrét magyarázat hiányát és az emberi ügyintéző elérhetetlenségét említették problémaként.

A tanulmányban leírtak alapján az alábbi ajánlásokat fogalmazhatjuk meg:

Mivel az óriásplatformok moderációs döntései, a platformok „büntetőjoga" nagyon komolyan érinti a felhasználók jogait, az óriásplatformok gyakorlata néhány ponton fejlesztésre szorul.

1. Az írott szabályzatoknak egységes, áttekinthető és a teljes szöveget tartalmazó verzióit publikálni és széles körben megismerhetővé kellene tenni.
2. A szabályok változtatását folyamatosan vezetni kellene, ezeknek nyomonkövethetőknek kellene lenniük.
3. Egységes terminológiát és tilalom csoportokat („bűncselekmény-katalógust") kellene kidolgozni, egységes definíciókkal.
4. A transzparenciajelentéseknek egységesebb szerkezetet kellene adni, és az egyes időszakokat is egységesen kellene riportolni. Ezt a problémát a Bizottság is felismerte és jelen sorok írásakor tette közzé a jelentések egységes mintasablonjait.[13] Ezen felül az adatoknak könnyen exportálhatónak kellene lenniük, hogy azok kutatási célokra könnyebben legyenek használhatók.
5. A döntésekhez érdemi magyarázatokat kellene adni minden esetben, és nem csak egy általános kategóriára történő utalással kellene az indokolási kötelezettséget letudni.
6. Jóval több esetben kellene biztosítani az emberi ügyintézőhöz fordulás jogát. Ehhez a platformoknak nagyságrendekkel nagyobb számú munkatársat kellene foglalkoztatniuk.
7. A szólásszabadságot érintő kategóriáknál különösen gondosan kellene eljárni az indokolás és a felfüggesztés terén is. Az örökre történő kitiltás lehetőségét meg kéne szüntetni.

[13] A BIZOTTSÁG (EU) 2024/2835 VÉGREHAJTÁSI RENDELETE (2024. november 4.) a közvetítő szolgáltatók és az online platformot üzemeltető szolgáltatók (EU) 2022/2065 európai parlamenti és tanácsi rendelet szerinti átláthatósági jelentési kötelezettségeire vonatkozó mintadokumentumok meghatározásáról (Hatálybalépés: 2024. november 25.)

Függelék - Adatok és következtetések a Facebook transzparencia-adatbázisából

Az adatbázisról: kategóriák, kontextus, módszertan

A transzparenciajelentések csak nagyon nagy vonalakban tartalmaznak információkat, a Bizottság transzparencia Dashboardjáról pedig, ahogy azt fentebb jeleztem, nagyon nehéz aggregált információkat szerezni. Ugyanakkor a Meta a „Közösségi elvek végrehajtásáról szóló jelentése" (Community Standards Enforcement Report) egyszerűsített formában aggregáltan tartalmazza a moderálással kapcsolatos adatokat, méghozzá világszinten. Ebből lehet véleményem szerint a legpontosabb képet kapni a Meta moderálási gyakorlatáról.

A Meta tizenegy esetkört jelöl meg, amely miatt valamilyen módon korlátozhatja a tartalmat, illetve letilthatja a fiókot. (1) Felnőtt meztelenség és szexuális tevékenység (2) Bullying és zaklatás (3) Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás (4) Gyermekek veszélyeztetése: szexuális kizsákmányolás (5) Veszélyes szervezetek: terrorizmus és veszélyes szervezetek: szervezett gyűlölet (6) Hamis fiókok (7) Gyűlöletbeszéd (8) Korlátozott áruk: drogok és korlátozott áruk: fegyverek (9) Spam (10) Öngyilkosság és önsértés (11) Erőszak és uszítás és erőszakos és erőszakot ábrázoló képi tartalom.

Mivel a Facebook-on naponta kb. 1 milliárd posztot tesznek közzé, természetesen a moderálást szinte teljes egészében algoritmusok végzik.

Mielőtt a statisztikák lényegi elemeit ismertetném, érdemes néhány szót szólni a mérési módszertanról. A jelentés központi kategóriája a „content actioned" („intézkedéssel érintett tartalom"). Ez a magyarázat szerint (https://transparency.meta.com/hu-hu/policies/improving/content-actioned-metric/) az adott tartalom blokkolását és figyelmeztető üzenet küldését jelenti „bizonyos célcsoportoknak" (sic!) esetleg a tartalmat közzétevő fiókok felfüggesztésével együtt. („Taking action could include removing a piece of content from Facebook or Instagram, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts.") Sajnos arról nincsen statisztika, hogy mely esetekben történt tartalomtiltás és mely esetekben fióktiltás is (és mennyi időre), illetve arról sincsen adat, hogy ezek közül mennyi esetet jelentettek a hatóságoknak.

Az adatok tartalmaznak még három információt, mégpedig azt, hogy az intézkedéssel érintett tartalmak közül mennyit fellebbeztek meg, mennyit állítottak vissza a fellebbezés nyomán, és mennyit állítottak vissza fellebbezés nélkül.

Az adatok 2017 negyedik negyedévétől 2024 második negyedévéig állnak rendelkezésre mindkét Meta platformon. Mivel jelen tanulmánynak nem célja a banning és a shadow banning gyakorlatának elemzése *általában*, a lenti táblázatokban és magyarázatokban én egyrészt csak a Facebook-os adatokat közlöm, másrészt csak a Meta 2023 – 24-es éveinek hat negyedévére vonatkozó statisztikákat. Ezekből az mindenképpen jól kirajzolódik, hogy mely jog-, vagy szabályzatsértés típusok a leggyakoribbak, milyen arányokról beszélünk, milyen arányban fellebbezik meg ezeket és milyen arányban állítja helyre a tartalmat (vagy a fiókot) a Facebook.

„Intézkedéssel érintett tartalmak" (és fiókok) a Facebookon

Kezdjük két olyan adatsorral, amely az összes intézkedéssel érintett tartalmat tartalmazza.

Kategória	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
1. Felnőtt meztelenség és szexuális tevékenység	38 600	51 200	35 100	44 100	39 400	32 200	240 600
2. Bullying és zaklatás	**6 900**	**7 900**	**8 300**	**7 700**	**7 900**	**7 800**	**46 500**
3. Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás és	1 900	1 700	1 800	1 900	8	1	7 309
4. Gyermekek veszélyeztetése: szexuális kizsákmányolás	8 900	7 200	16 900	16 200	14 400	9 700	73 300
5. Veszélyes szervezetek: szervezett gyűlölet	9	1 100	8	10	5	7	1 139
6. Veszélyes szervezetek: terrorizmus	14 500	13 600	8 200	13 900	8 400	7 500	66 100
7. Hamis fiókok	426 000	675 900	827 400	691 400	630 900	1 200 000	4 451 600
8. Gyűlöletbeszéd	**10 700**	**18 000**	**9 600**	**7 400**	**7 400**	**7 200**	**60 300**
9. Korlátozott áruk: drogok	3 100	2 600	1 800	1 800	1 700	1 700	12 700
10. Korlátozott áruk: fegyverek	1 300	9	8	2 300	3 300	1 900	8 816
11. Spam	1 600 000	1 100 000	413 300	964 200	436 000	322 300	4 835 800
12. Öngyilkosság és önsértés	3 100	6 400	5 300	7 100	7 100	6 600	35 600
13. Erőszak és uszítás	**12 400**	**10 600**	**8 600**	**8 200**	**8 700**	**7 400**	**55 900**
14. Erőszakos és erőszakot ábrázoló képi tartalom	13 600	13 800	9 000	14 600	10 600	14 900	76 500
Összesen	2 141 009	1 910 009	1 345 315	1 780 810	1 175 813	1 619 208	9 972 164

F1. Táblázat: intézkedéssel érintett tartalmak a Facebook-on 2023 – 2024 években
(x1000 intézkedés)

Kategória	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
1. Felnőtt meztelenség és szexuális tevékenység	1,8%	2,7%	2,6%	2,5%	3,4%	2,0%	2,4%
2. Bullying és zaklatás	0,3%	0,4%	0,6%	0,4%	0,7%	0,5%	0,5%
3. Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás és	0,1%	0,1%	0,1%	0,1%	0,0%	0,0%	0,1%
4. Gyermekek veszélyeztetése: szexuális kizsákmányolás	0,4%	0,4%	1,3%	0,9%	1,2%	0,6%	0,7%
5. Veszélyes szervezetek: szervezett gyűlölet	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%
6. Veszélyes szervezetek: terrorizmus	0,7%	0,7%	0,6%	0,8%	0,7%	0,5%	0,7%
7. Hamis fiókok	19,9%	35,4%	61,5%	38,8%	53,7%	74,1%	44,6%
8. Gyűlöletbeszéd	0,5%	0,9%	0,7%	0,4%	0,6%	0,4%	0,6%
9. Korlátozott áruk: drogok	0,1%	0,1%	0,1%	0,1%	0,1%	0,1%	0,1%
10. Korlátozott áruk: fegyverek	0,1%	0,0%	0,0%	0,1%	0,3%	0,1%	0,1%
11. Spam	74,7%	57,6%	30,7%	54,1%	37,1%	19,9%	48,5%
12. Öngyilkosság és önsértés	0,1%	0,3%	0,4%	0,4%	0,6%	0,4%	0,4%
13. Erőszak és uszítás	0,6%	0,6%	0,6%	0,5%	0,7%	0,5%	0,6%
14. Erőszakos és erőszakot ábrázoló képi tartalom	0,6%	0,7%	0,7%	0,8%	0,9%	0,9%	0,8%
Összesen	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**

F2. táblázat: az F1. táblázat százalékos arányokban kifejezve

A tartalomkorlátozásokkal kapcsolatos intézkedések a táblázatokból kitűnően negyedévente nagyjából 1 és a 2 milliárd intézkedés közé esnek. A Meta magyarázata szerint a hatalmas hullámzást a számokban nem a saját rendszereinek a működése, hanem külső tényezők okozzák. Ha megfigyeljük a számokat, az intézkedések több mint 93%-a két kategória: a hamis fiókok törlése és a spam (amely jelentheti a spam törlését és a spammelő fiókok törlését egyarán, de gyaníthatóan inkább az előbbi). A fennmaradó 12 kategóriából egyedül a felnőtt meztelenség emelkedik ki, a maga 32-51 millió tartalomkorlátozási intézkedésével, amely a teljes intézkedési szám 1,8 – 3,4 %-át teszi ki.

Ha ezeket a számokat a Facebook napi kb. egymilliárdos (így évi szinten a sokszázmilliárdos nagyságredű) posztjaihoz hasonlítjuk, amelyből az európai felhasználók nagyjából 12,5%-ot (408 millió / 3,27 milliárd[14]) tesznek ki, akkor nagyjából 50 milliárd európai posztot kapunk egy évben, havonta nagyjából 4 milliárdot. Ha pedig a fenti számokból csak a valóban szólásszabadsághoz kapcsolódó kategóriákat vesszük (azaz épp a két leggyakoribb tartalomintézkedési típust, a hamis fiókok törlését és a spamet kivesszük ebből), akkor valójában egy évi kb. 400 milliós moderált tartalomegység áll szemben az évi kb. 50 milliárdos összes európai poszttal, amely 0,8%, azaz 8 ezreléknyi „intézkedést" jelent.

Ezzel nem azt szeretném mondani, hogy a banning és a shadow banning nem probléma, hiszen évente 400 millió „intézkedés" mégiscsak történik Európában, azaz tartalom, vagy fióktiltás, és ezek közt természetesen ott vannak azok a fiókok is, amelyek hatalmas követőbázissal rendelkeznek. Egy több százezer követővel rendelkező fiók letiltása éppen ugyanúgy 1 darab „content actioned", mint egy spamként feltöltött tartalom, vagy egy „fake"-nek bélyegzett fiók letiltása a több millióból.

Fellebbezések és visszaállított tartalmak és fiókok

Rendkívül tanulságosak azok a statisztikák is, amelyek *a megfellebbezett és a visszaállított* tartalmakat és fiókokat mutatják. Előbb a megfellebbezett tartalmakat tartalmazó táblázatot mutatom be, ezt immár nem 1000 darabban, hanem az eredeti darabszámot feltüntetve, mert itt vannak olyan kategóriák, amelyek 1000-nél kisebb számot tartalmaznak (F3. táblázat), majd ezt az összes intézkedéssel érintett tartalomra vetítve (F4. táblázat). Ezután a „content restored with appeal" és a „content restored without appeal" táblázatokat, azaz, hogy az intézkedéssel érintett tartalmak közül mennyi volt az, amelyet voltaképpen helytelenül távolított el a Facebook.

Ezekben a táblázatokban, és különösen az F3. számúban érdekes jelenségekre bukkanhatunk. Míg bizonyos területeken az emberek belenyugszanak a döntésbe, három olyan tartalomkorlátozás típus van, amely esetén lényegében minden ötödik döntést (14 és 20% közötti értékek) megfellebbezik. Mindhárom szorosan kapcsolódik a szólásszabadság kérdéséhez, és egyúttal azt jelzi, hogy az emberek számára ezek a korlátozások komoly problémát jelentenek, bántja őket, ha a Facebook korlátozza a szólásukat. Ez a három kategória a 'bullying és zaklatás', a 'gyűlöletbeszéd' és az 'erőszak és uszítás' kategóriái. Mindháromra nagyjából a 7 millió és a 12 millió eltávolított tartalom-egység a jellemző negyedévente, amely eltávolításból nagyjából 2 millió körüli intézkedést megfellebbeznek.

[14] A legfrissebb adatok szerint a Meta-nak Európában 408 millió aktív felhasználója van, (https://www.statista.com/statistics/745400/facebook-europe-mau-by-quarter/) amely a világszintű 3,27 milliárd felhasználóhoz képest annak 12,48%-a.

	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
1. Felnőtt meztelenség és szexuális tevékenység	2 400 000	2 800 000	2 700 000	2 300 000	2 500 000	2 500 000	15 200 000
2. Bullying és zaklatás	**1 400 000**	**1 500 000**	**1 500 000**	**1 300 000**	**1 300 000**	**1 200 000**	**8 200 000**
3. Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás	918	944	1 122	1 351	781	835	5 951
4. Gyermekek veszélyeztetése: szexuális kizsákmányolás	1 045	1 468	2 666	1 000 000	3 809	410	1 009 398
5. Veszélyes szervezetek: szervezett gyűlölet	132	1 636	1 209	1 245	687	1 185	6 094
6. Veszélyes szervezetek: terrorizmus	5 435	662	6 325	1 200 000	6 484	5 242	1 224 148
8. Gyűlöletbeszéd	**2 000 000**	**2 500 000**	**1 800 000**	**1 300 000**	**1 200 000**	**1 200 000**	**10 000 000**
9. Korlátozott áruk: drogok	3 171	2 853	1 776	1 715	1 494	1 821	12 830
10. Korlátozott áruk: fegyverek	1 106	756	68	2 643	5 066	2 711	12 350
11. Spam	2 600 000	6 500 000	6 200 000	3 800 000	2 300 000	6 600 000	28 000 000
12. Öngyilkosság és önsértés	1 023	1 059	955	801	1 399	1 872	7 109
13. Erőszak és uszítás	**2 400 000**	**1 800 000**	**1 600 000**	**1 400 000**	**1 600 000**	**1 300 000**	**10 100 000**
14. Erőszakos és erőszakot ábrázoló képi tartalom	351	387	529	1 095	1 134	1 172	4 668
Összesen	**10 813 181**	**15 109 765**	**13 814 650**	**12 308 850**	**8 920 854**	**12 815 248**	**73 782 548**

F3. táblázat Fellebbezések száma

	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
1. Felnőtt meztelenség és szexuális tevékenység	6%	5%	8%	5%	6%	8%	6%
2. Bullying és zaklatás	**20%**	**19%**	**18%**	**17%**	**16%**	**15%**	**18%**
3. Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás és	0%	0%	0%	0%	10%	91%	0%
4. Gyermekek veszélyeztetése: szexuális kizsákmányolás	0%	0%	0%	6%	0%	0%	1%
5. Veszélyes szervezetek: szervezett gyűlölet	1%	0%	16%	13%	13%	16%	1%
6. Veszélyes szervezetek: terrorizmus	0%	0%	0%	9%	0%	0%	2%
8. Gyűlöletbeszéd	**19%**	**14%**	**19%**	**18%**	**16%**	**17%**	**17%**
9. Korlátozott áruk: drogok	0%	0%	0%	0%	0%	0%	0%
10. Korlátozott áruk: fegyverek	0%	9%	1%	0%	0%	0%	0%
11. Spam	0%	1%	2%	0%	1%	2%	1%
12. Öngyilkosság és önsértés	0%	0%	0%	0%	0%	0%	0%
13. Erőszak és uszítás	**19%**	**17%**	**19%**	**17%**	**18%**	**18%**	**18%**
14. Erőszakos és erőszakot ábrázoló képi tartalom	0%	0%	0%	0%	0%	0%	0%
Összesen	**1%**	**1%**	**1%**	**1%**	**1%**	**1%**	**1%**

F4. táblázat Fellebbezések százalékos aránya az összes intézkedéssel érintett tartalomhoz képest a Facebookon Európában 2023-2024 években

Mindennek fényében érdemes egy pillantást vetni arra, hogy mit mutatnak a számok a Facebook helyreállítási gyakorlatát illetően. Mennyi tartalmat állít vissza a Facebook abból, amit korlátoz, és ebből mennyi a saját belátásból visszaállított és mennyi az, amit fellebbezés miatt állítanak vissza. (F4. és F5. táblázatok)

Egyrészt 2023 II. negyedévében valami furcsa történt a gyűlöletbeszédnek minősített tartalmakkal. Ebben az időszakban a korábbi és későbbi negyedévek átlagos tiltási számát kb. 8 millió (valószínűleg téves) tiltással meghaladó tiltás történt, amelyet a Facebook azonnal korrigált is, hiszen a szokásos, saját hatáskörben visszavont pár tucatnyi tiltás helyett ezeknek a visszavonásoknak a száma 6 millióra ugrott fel. Nehéz utólag rekonstruálni, hogy mi történhetett, mindenesetre elég aggasztó, hogy ilyen óriási mennyiségű tartalmat korlátoztak tévesen.

Másrészt a számokból egy elég egyoldalú kép bontakozik ki, amelyet röviden úgy lehetne jellemezni, hogy a Facebook által (mint említettem, valójában egy MI rendszer által) hozott döntést a Facebook kivételesen, és nagyon ritkán változtat meg. A számok alapján elmondható, hogy nagyjából minden ezredik fellebbezés eredményes (F6. és F7. táblázatok).

	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
1. Felnőtt meztelenség és szexuális tevékenység	8 197	1 000 000	7 133	4 908	5 166	5 035	1 030 439
2. Bullying és zaklatás	2 101	2 604	2 808	2 064	239	2 123	11 939
3. Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás	123	205	203	361	112	93	1 097
4. Gyermekek veszélyeztetése: szexuális kizsákmányolás	208	387	878	3 175	1 239	90	5 977
5. Veszélyes szervezetek: szervezett gyűlölet	323	483	248	94	91	26	1 265
6. Veszélyes szervezetek: terrorizmus	869	1 536	933	744	104	846	5 032
8. Gyűlöletbeszéd	2 302	9 233	3 062	1 337	1 447	1 537	18 918
9. Korlátozott áruk: drogok	719	585	244	171	143	285	2 147
10. Korlátozott áruk: fegyverek	176	162	133	705	2 055	104	3 335
11. Spam	4 048	6 816	1 862	8 895	3 809	2 100 000	2 125 430
12. Öngyilkosság és önsértés	377	383	217	148	194	237	1 556
13. Erőszak és uszítás	3 069	2 061	1 807	1 727	2 869	2 038	13 571
14. Erőszakos és erőszakot ábrázoló képi tartalom	64	66	69	103	88	12	402
Összesen	**22 576**	**1 024 521**	**19 597**	**24 432**	**17 556**	**2 112 426**	**3 221 108**

F5. táblázat Fellebbezés után visszaállított tartalom (fiók) („content restored with appeal”)

	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
1. Felnőtt meztelenség és szexuális tevékenység	2 837	1 700 000	344	493	223	311	1 704 208
2. Bullying és zaklatás	202	92	95	84	103	69	645
3. Gyermekek veszélyeztetése: meztelenség és fizikai bántalmazás és	54	14	18	2 795	500	11	3 392
4. Gyermekek veszélyeztetése: szexuális kizsákmányolás	172	413	1 166	1 200 000	734	523	1 203 008
5. Veszélyes szervezetek: szervezett gyűlölet	168	3 071	76	2	17	28	3 362
6. Veszélyes szervezetek: terrorizmus	3 325	7 983	57	204	49	49	11 667
8. Gyűlöletbeszéd	63	6 000 000	69	42	34	29	6 000 237
9. Korlátozott áruk: drogok	104	143	48	3	28	117	443
10. Korlátozott áruk: fegyverek	11	108	86	217	68	231	721
11. Spam	102 200 000	128 800 000	16 300 000	35 900 000	26 000 000	32 800 000	342 000 000
12. Öngyilkosság és önsértés	143	32	19	108	11	13	326
13. Erőszak és uszítás	78	46	39	32	41	211	447
14. Erőszakos és erőszakot ábrázoló képi tartalom	24	14	13	14	800	16	881
Összesen	**102 207 181**	**136 511 916**	**16 302 030**	**37 103 994**	**26 002 608**	**32 801 608**	**350 929 337**

F6. táblázat Fellebbezés nélkül visszaállított tartalom („content restored without appeal”)

		2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
2. Bullying és zaklatás -	érintett tartalom	6 900 000	7 900 000	8 300 000	7 700 000	7 900 000	7 800 000	46 500 000
	megfellebbezett	1 400 000	1 500 000	1 500 000	1 300 000	1 300 000	1 200 000	8 200 000
	helyreállított saját hatáskörben	202	92	95	84	103	69	645
	helyreállított fellebbezés után	2 101	2 604	2 808	2 064	239	2 123	11 939
8. Gyűlöletbeszéd	érintett tartalom	10 700 000	18 000 000	9 600 000	7 400 000	7 400 000	7 200 000	60 300 000
	megfellebbezett	2 000 000	2 500 000	1 800 000	1 300 000	1 200 000	1 200 000	10 000 000
	helyreállított saját hatáskörben	63	6 000 000	69	42	34	29	6 000 237
	helyreállított fellebbezés után	2 302	9 233	3 062	1 337	1 447	1 537	18 918
13. Erőszak és uszítás	érintett tartalom	12 400 000	10 600 000	8 600 000	8 200 000	8 700 000	7 400 000	55 900 000
	megfellebbezett	2 000 000	2 500 000	1 800 000	1 300 000	1 200 000	1 200 000	10 000 000
	helyreállított saját hatáskörben	78	46	39	32	41	211	447
	helyreállított fellebbezés után	3 069	2 061	1 807	1 727	2 869	2 038	13 571

F7. táblázat A három szólásszabadságot is érintő tartalomtípus korlátozása, az ezek elleni fellebbezések száma, és a helyreállított tartalmak száma

	2023Q1	2023Q2	2023Q3	2023Q4	2024Q1	2024Q2	Összesen
2. Bullying és zaklatás -	**0,15%**	**0,17%**	**0,19%**	**0,16%**	**0,02%**	**0,18%**	**0,15%**
8. Gyűlöletbeszéd	**0,12%**	**0,37%**	**0,17%**	**0,10%**	**0,12%**	**0,13%**	**0,19%**
13. Erőszak és uszítás	**0,15%**	**0,08%**	**0,10%**	**0,13%**	**0,24%**	**0,17%**	**0,14%**

F8. táblázat Sikeres fellebbezések aránya a szólásszabadságot is érintő kategóriáknál