

Zódi Zsolt*

A Facebook tartalom- és fiókkorlátozási gyakorlata: szabályok, eljárások, algoritmusok és statisztikák

Néhány empirikus adalék a „szólásszabadság a platformokon” diskurzushoz

1. BEVEZETÉS

Ahogy azt nap mint nap tapasztalhatjuk, az életünk és a nyilvánosság egyre nagyobb szféráit meghódító közösségi médiaplatformok többféle eszközzel ‘szerkesztik’ a felületükön megjelenő tartalmakat, és ezzel korlátozzák a felhasználók mozgásterét, szólását.¹ Először is mind az illegális, mind a saját szabályaikba ütköző tartalmak láthatóságát korlátozzák (sőt a tartalmakat olykor törlik is), másodsor e tartalmak megosztóit is különböző szankciókkal sújtják, sőt végső soron akár ki is tilthatják őket a platformról.

A szólásszabadsággal kapcsolatos viták gyújtópontjában gyakran állnak ezek a korlátozó gyakorlatok, ám igen kevés írás foglalkozik azzal, hogy ezek a gyakorlatok min alapulnak (milyen normatív háttérük van), a folyamatok ténylegesen hogyan zajlanak, és milyen volumenről beszélünk. Ez a cikk ezt a hiányt igyekszik egy szűk, de fontos területen, a Facebook moderálási gyakorlatának bemutatásával pótolni. A Facebook mellett nemcsak azért döntöttem, mert hazánkban ez a legnépszerűbb közösségi médiaplatform, hanem azért is, mert a Meta, a Facebook anyavállalata meglepően sok adatot és információt megoszt a szabályai ‘kikényszerítéséről’.

A cikk előbb ismerteti azt a szövevényes szabályzatháttérrel, amelyen a kikényszerítés alapul, majd – amennyire erről nyilvános információk hozzáférhetők – bemutatja a ‘kikényszerítés’ (ezt maga a Facebook hívja így: enforcement) folyamatát, végül a kikényszerítés néhány statisztikai adatát ismerteti.

Nagyjából 2019-ig ezekről a területekről, a szabályok részleteiről (különösen a platformok jogszerű, de saját szabályaiba ütköző ‘lawful, but awful’² tartalmak moderálásáról), a szabályokat végrehajtó algoritmusokról és személyzetről, és a szankcionálási és tiltási statisztikákról is inkább csak anekdotikus információink voltak. A 2010-es évek közepén (2016–17-ben) azonban olyan események történtek,³ amelyek hatására a platformok saját elhatározásukból és jogszabályi kötelezések hatására is elkezdtek egyre több információt közzétenni. Ebben az írásban ezekre a közzétett információkra nagyban támaszkodom.

A három témakörben (szabályok; eljárások és algoritmusok; és statisztikák) a következő kérdésekre keresem a választ. A szabályok területén: mi szabályozza a Facebook moderálási gyakorlatát? Hogyan jellemezhetők ezek a szabályok? Mi a viszonyuk az állami joghoz? Az eljárások és algoritmusok területén: hogyan zajlik egy korlátozási-szankcionálási eljárás? Milyen algoritmusokról beszélünk? Mi jellemzi ezeket? Hogyan működnek? Mi az emberi közreműködés szerepe? Végül a statisztikák területén: milyen általánosítható tendenciák jellemzik a szabályok betartatását? Melyek a leggyakoribb beavatkozási okok? Milyen gyakori a beavatkozás, a szankcionálás? Milyen az illegális tevékenységeket visszaszorító intézkedések és a házi szabályokat kikényszerítő intézkedések aránya? Mennyi döntést hoznak gépek, és mennyit emberek? Mekkora a ‘szürke zóna’ mérete, azaz a nem illegális és nem

* Tudományos főmunkatárs, Nemzeti Közszerzői Egyetem, Információs Társadalom Kutatóintézet.

1 A Nemzeti Közszerzői Egyetem Információs Társadalom Kutatóintézetének már harmadszor lefolytatott kutatása azt mutatja, hogy a Magyarországon a Facebooknak 7 millió regisztrált felhasználója van, és ezek nagyjából 10-15%-a, (2019 óta fokozatosan emelkedő mértékben) tapasztalta, hogy a bejegyzését vagy a fiókját törölték vagy korlátozták.

2 A nem illegális, de a platformok által ‘zavarónak’ minősített tartalom szokásos elnevezése, lásd pl. 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.

3 Ezek a Cambridge Analytica botrány (Brexit szavazás és az USA elnökválasztás, 2016), a párizsi és brüsszeli terrortámadások (2016–2017), valamint a Facebook nyilvánosságra került ‘titkos’ szabálykönyve és ‘titkos’ moderálócspatai (2017) voltak.

nyíltan szabályzatellenes, de valamiért 'károsnak' minősített tartalmak miatti beavatkozások száma? Mennyire jellemző az intézkedések elleni 'fellebbezés', és ezek milyen arányban sikeresek?

Mint az egyértelműen látható, a tanulmánynak *nem* célja a moderálással kapcsolatos elméleti-jogi-dogmatikai problémák taglalása, a 'szólásszabadság a platformokon' témakörhöz történő elméleti hozzászólás. Ezért nem, vagy alig fogok ebben az írásban szólni a szabályok értékeléséről, a körülöttük zajló vitákról, ezek alkotmányos, alapjogi megítéléséről. Ennek nemcsak az az oka, hogy a 'szólásszabadság a platformokon' témában az utóbbi néhány évben feldolgozhatatlanul nagy mennyiségű irodalom halmozódott fel,⁴ hanem az is, hogy az itt ismertetendő számok és adatok értelmezéséhez ezek a magas absztrakciós szintű elméletek nem szükségesek. Ugyanakkor, ahol az adatok megértéséhez ez szükségesnek mutatkozott, alkalmasint utalok ezekre az elméleti megfontolásokra is.

2. SZABÁLYOK

2.1. MIÉRT VAN SZÜKSÉG SZABÁLYOKRA? MIÉRT NEM ELÉG AZ ÁLLAMI SZABÁLYOZÁS?

Kiinduló kérdésünk tehát az, hogy milyen szabályok alapján moderál a Facebook? Mielőtt erről beszélnék, érdemes azzal a felvetéssel foglalkozni, hogy a laikusok, de a politikai szféra egyes résztvevői részéről⁵ is felvetődik néha, hogy ilyen (belső) szabályokra tulajdonképpen nem is lenne szükség: valójában csak az állami szabályokat kellene betartatni

ezen a felületeken is. Hiszen a büntetőjog egyértelműen eligazít a szóban és képekkel (is) elkövethető bűncselekményekről, a magánjog és néhány más jogág (pl. az adatvédelmi jog, a fogyasztóvédelem vagy a versenyjog) pedig szintén tartalmaz egyértelmű rendelkezéseket. Ezeket következetesen be kéne tartani, de a platformoknak nem lenne szabad ezeken túlterjeszkedni. Ezt az érvelést következetesen végiggondolva lehet arra a következtetésre jutni, hogy tulajdonképpen nincs is szükség belső szabályozásra. Hogy ez mégis létezik, annak alapvetően két oka van.

Az első ok egy mennyiségi tényező. A Facebookon világszinten nagyjából egymilliárd (!) poszt születik naponta, azaz ennyi tartalomjegységet publikálnak a felhasználók. Ezt a mennyiséget lényegében lehetetlen emberi erőfeszítésekkel kontrollálni, ezért a posztok 'jogszerűségének' vizsgálatát gépi rendszerek végzik, a gépi rendszereknek pedig egyértelmű, minden részletre kiterjedő 'végrehajtási' szabályokra, vagy példákra van szüksége, (ahogy egyébként a moderátoroknak is). Tehát az első ok az, hogy a működőképesség és a megnyilvánulások miatt mindenképp szükség van saját 'végrehajtási' és 'eljárási' szabályokra, és ezekre akkor is szükség lenne, ha csak az állami jogszabályokat akarná a platform betartatni.

A második ok, amelyet a szakirodalom is gyakran emleget, hogy a platformok nem egyszerűen 'jogot betartó helyek', hanem 'biztonságos', 'otthonos', 'kellemes' helyek szeretnének lenni. Egy sor olyan tartalom van ugyanis, amely jogszerű ugyan, de senki nem szeretne beleütközni. Nem szeretnénk trágár beszédet hallani, más emberek megalázását végignézni, önsértéseket és öngyilkosságokat, droghasználatot, agressziót, undorító, vagy félelemkeltő tartalmakat látni, kénytelen üzenetek százaival viaskodni, és így tovább. Az állami szabályok és tilalmak tehát kiegészítésre szorulnak, és a platformok, így a Facebook ki is egészíti ezeket a saját, a 'biztonságos tér' megteremtését célzó szabályaival.

2.2. A FACEBOOK BELSŐ SZABÁLYAINAK FELÉPÍTÉSE ÉS A TILALMAZOTT MAGATARTÁSOK

A Facebook, ahogy más platformok is, három fő-, és nagyjából egytucatnyi 'mellék' vagy 'végrehajtási' szabállyal dolgozik.⁶ A három fő szabályzat a Felhasználási feltételek, amely az adatvédelmi szabállyal és a közösségi alapelvekkel alkot egy egységet. A kiindulópont a Felhasználási feltételek.⁷ Ez azután 9 helyen utal az Adatvédelmi szabályzatra,⁸ és a 3. pontban a Községi alapelvekre,⁹ majd a dokumentum végén

4 Az egyik legnagyobb jogcikk-adatbázis a HeinOnline a 'first amendment' (az amerikai alkotmány szólásszabadságról rendelkező szabálya) és 'social media' kifejezések kombinációjára 11 918 db találatot ad ki és, ha ebből még le is számítjuk azokat a találatokat, ahol a 'social media' az internet előtti médiatermékekre utal, vagy ahol a cikknek nem fő-, hanem csak érintőleges témája, akkor is megállapíthatjuk, hogy ezres nagyságrendben van a cikkek száma. Néhány magas idézettségű, nagy hatású cikk a témában: KATE KLONICK: The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131, no. 6, 2018. április. 1598–1670; DANIELLE KEATS CITRON, FRANK PASQUALE: The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89, no. 1. 2014. március. 1–34.; JACK M. BALKIN: Information Fiduciaries and the First Amendment. *U.C. Davis Law Review* 49, no. 4. 2016. április. 1183–1234.; DANIELLE KEATS CITRON, BENJAMIN WITTES: The Internet Will Not Break: Denying Bad Samaritans Sec. 230 Immunity. *Fordham Law Review* 86, no. 2. 2017. november. 401–424; JACK M. BALKIN: Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review* 51, no. 3. 2018. február. 1149–1210. A magyar jogtudományi irodalomban nem kevésbé népszerű ez a téma. Csak az elmúlt pár évben találunk egy sor monográfiát és többtucatnyi cikket, ami a teljes magyar jogtudományi térséghez viszonyítva, a téma jelentőségét tekintve is kifejezetten soknak mondható. Monográfiák: KOLTAY ANDRÁS: *Az új média és a szólásszabadság. A nyilvánosság alkotmányos alapjainak újragondolása*. Budapest, Wolters Kluwer, 2019. PAPP JÁNOS TAMÁS: *A közösségi média szabályozása a demokratikus nyilvánosság védelmében*. Budapest, Wolters Kluwer, 2022. GOSZTONYI GERGELY: *Cenzúra Arisztoteléstől a Facebookig*. Budapest, Gondolat, 2022. KLEIN TAMÁS: *Sajtószabadság és demokrácia*. Budapest, Gondolat, 2020.

5 Ez volt az egyik érv a végül meghíúsult lengyel Facebook-törvény mögött, de több USA-beli – főként republikánus – kezdeményezés mögött is. Bővebben: ZÓDI ZSOLT: *Platformjog: Elméletek és hatályos szabályok*. Budapest, Ludovika, 2024, 166–168. és 136–141.

6 A legutóbbi időig nem volt a három fő szabályzat és az összes többi alszabályzat egymáshoz való viszonya világos, és nem is voltak egy helyen összegyűjtve. Valamikor 2024. november végén létrehoztak egy oldalt, ahol ezeket összegyűjtötték, így a helyzet valamivel tisztább lett, mint korábban, bár az egyes szabályzatok egymáshoz való viszonya továbbra is tisztázatlan.

7 Ez kitér az 5.1 pontból is: „A jelen feltételek (korábbi nevén Jogi és Felelősségi Nyilatkozat) alkotják a teljes megállapodást közted és a Meta Platforms Ireland Limited között a termékeink általad történő használatára vonatkozóan. Hatályon kívül helyeznek minden korábbi megállapodást.”

<https://www.facebook.com/legal/terms#other-terms-policies>

8 <https://www.facebook.com/privacy/policy>

9 <https://transparency.meta.com/hu-hu/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards>

a többi szabályzatra, amelyek vagy szűkebb specifikus csoportokra,¹⁰ vagy specifikus tevékenységekre¹¹ vonatkoznak. A felhasználási feltételek természetesen egy sor, a témánkhöz szorosan nem kapcsolódó rendelkezést is tartalmaznak (mint pl. a hirdetések megjelenítésére vonatkozó szabályok, 2. pont, vagy a szolgáltatást igénybe vevőkre vonatkozó életkori korlátozások és a felhasználók adatszolgáltatással kapcsolatos kötelezettségei, 3.1. pont), de a dokumentum szövegének elég nagy hányada szól a megosztható tartalmakról (3.2. pont). A tartalmakat a Facebook itt hat pontban részletezi: Tilos tehát 1. a jogellenes vagy szabályzatellenes tartalmakat megosztani (ideértve a szellemi tulajdont megsértő tartalmakat is); 2. Vírusokat és rosszindulatú kódokat feltölteni; 3. automaták által adatot gyűjteni (scrape-elni); 4. felhasználóneveket, jelzőszavakat feltörni ezekkel visszaélni; 5. a platformról gyűjtött adatokkal kereskedni; és 6. hivatalos csatornákat (pl. a bejelentési felületeket) visszaélészerűen használni.

Ami jelen cikk témájában releváns, az az 1. alpont, amely azonnal nyilvánvalóvá teszi, hogy nemcsak jogellenes tevékenységeket nem szabad végezni, hanem a közösségi alapelvekbe¹² foglalt tilalmakat is be kell tartani. E két kategóriának a viszonya elsőre világosnak tűnik: ami jogellenes, azt értelemszerűen nem szabad feltölteni, és ezenfelül még vannak a platform által támasztott egyéb követelmények. Előbbiek a jogszabályokban vannak definiálva, utóbbiak a közösségi alapelvek elnevezésű belső szabályzatban. Csakhogy ez az elsőre logikusnak tűnő megkülönböztetés korántsem ilyen egyszerű. Ha ugyanis megnyitjuk a közösségi alapelveket tartalmazó oldalt, ott azt találjuk, hogy a közösségi alapelvek nemcsak a jogszabályellenes cselekményeken *felül* tiltott magatartásokat tartalmazzák, hanem rengeteg jogellenes cselekmény *Facebook szerinti értelmezését*, azaz tulajdonképpen mini Facebook-t, Btk.-t, Ptk.-t, szerzői jogi törvényt stb. A másik meglepetés pedig az, hogy a közösségi alapelvek oldalon legalább egytucatnyi más szabályzatot is találunk, amelyek egy része nem is tiltásokat tartalmaz, hanem például a monetizációval, vagy az elhunyt felhasználókkal kapcsolatos előírásokat, a közösségi alapelvekben foglalt tilalmak további kibontását, vagy eljárási jellegű szabályokat.

Ez a megoldás elég aggályos. Egyfelől kissé tisztázatlanná válik a hivatalos jogi tiltások és a Facebook saját szabályai szerinti tiltásainak viszonya, másfelől ez a megoldás azt eredményezi, hogy a tiltásokat egészében nem tudjuk átlátni, mert azokat szövevényesen egymáshoz kapcsolt weboldalak tucatjaiból lehet csak megismerni. Így nagyon nehéz összeszedni, hogy tulajdonképpen mi tilos a Facebookon.

2.3. A 'FÉLIG' TILALMAZOTT, ÁTMENETI KATEGÓRIÁK

Nemcsak a saját belső szabályok teszik bonyolulttá a felhasználók dolgát, hanem az is, hogy a Facebook (de más platformok is) nemcsak a 'tiltott' és 'szabad' kategóriákat ismeri,

hanem több 'átmeneti' kategóriát is. Először is a közösségi alapelvek alapvetően kétféle csoportba sorolják a tartalomkorlátozás alá eső tartalmakat: a 'nem engedélyezett' és a 'korlátozásokkal vagy feltételekkel terjeszthető' tartalmak csoportjaiba.¹³ Utóbbiakhoz vagy figyelmeztetés társul, vagy 18 év alattiak számára nem terjeszthetők. Ez a kategória tehát tulajdonképpen a 'jogszerűtlen' kiterjesztése, és talán úgy lehet megfogalmazni, hogy 'jogszerűtlen, de csak bizonyos körülmények közt'. Van azonban egy nagyon ellentmondásos másik átmeneti kategória is, amelyet pedig inkább a 'jogszerű, de nem kívánatos' tartalomként jellemezhetünk. Ezek olyan tartalmak, amelyeknek a terjedését a Facebook *lassítja*.

Az elérhetőségében (terjesztésében) történő korlátozás a legtöbb esetben nem explicit, arról a felhasználót nem tájékoztatják. A shadow banningnek ('árnyéktiltás') is nevezett módszer sokféle tartalomterjedési 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”¹⁴ címszó alatt ismertet is, ebből szerezhetünk 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”.

Ezen az oldalon példaként három tartalomtípust említ, de utal arra, hogy ennél jóval többféle ok van, amely 'terjesztéscsökkentéshez' vezethet:

- Alacsony színvonalú – például kattintásvadász és aktivításvadász – 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 tartalomtípus, a hírjellegű tartalmak rangsorolásával kapcsolatos szempontokat találjuk. Első lépésben 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?

10 Pl. a szoftverfejlesztőkre: <https://developers.facebook.com/policy/credits>, vagy a hirdetőkre: <https://transparency.meta.com/policies/ad-standards/>

11 Pl. zene feltöltésére: https://www.facebook.com/legal/music_guidelines

12 <https://transparency.meta.com/hu-hu/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards>

13 <https://transparency.meta.com/hu-hu/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards>

14 <https://transparency.meta.com/hu-hu/enforcement/taking-action/lowering-distribution-of-problematic-content/>

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 szempontrendszerrel van szó. Annak ellenére, hogy a Meta több helyen is igyekszik a hírrangsorolás szempontjait elmagyarázni (pl. a Meta Journalism Project¹⁵ keretén belül, sőt külön magyar oldal¹⁶ is részletezi a hírtartalmak minőségi követelményeit), az az algoritmusok tényleges viselkedése, a gépi tanulás, és a folyamatos változtatások miatt nehezen számítható ki.

2.4. MELY KATEGÓRIÁK PROBLEMATIKUSAK?

Először is a jogellenes tartalmak (gyermekes veszélyeztetése, korlátozott áruk és szolgáltatások) általában tiltottak, és viszonylag problémamentesnek mondhatók. Ez a helyzet néhány belső tiltás esetén is. (Pl. felnőtt meztelenség, hamis fiókok vagy spam.) Ugyanakkor van a belső szabályok közt néhány olyan kategória, amelynek a szélső eseteinél szinte egészen biztos, hogy olykor olyan megszólalások is tiltásra kerülnek, amelyeket véd a szólásszabadság. Minden határesetet nem tudok ebben a tanulmányban végigvenni, ezért csak néhány nyilvánvaló példát emelek ki.

Az első ilyen problematikus kategória a „Veszélyes szervezetek és magánszemélyek”. Ez esetben sem a náci vagy az izlám terroriszervezetekkel van a probléma (ez az ún. 1. szint), de a vonatkozó oldalon¹⁷ az ún. 2. szintű szervezeteknél („Erőszakos nem állami szereplők és erőszakra buzdító entitások”) „A gyűlöletkeltés miatt tiltott entitás” alkategória meghatározásánál azt találjuk, hogy tiltás alá esik az „olyan nem állami szereplő, amelyet ismétlődően gyűlöletkeltő magatartás vagy szóhasználat jellemez, de nem éri el az 1. szintű megjelölést, mert nem vett még részt erőszakban, és nem is támogatta azt kifejezetten, vagy mert nincs jelentősebb kapcsolata korábban veszélyesként megjelölt szervezetekkel vagy szereplőkkel.” Itt tehát arról van szó, hogy egy szervezet, amely egyébként aktívan nem vett soha részt semmilyen erőszakos cselekményben, mivel 'gyűlöletkeltő szóhasználat' jellemzi, kitiltásra kerülhet.

A második problematikus kategória természetesen a gyűlöletbeszéd. Itt is vannak egyértelmű esetek, azonban a '2. szint' (az enyhébb esetek) leírásánál pl. ezt találjuk: „(Ne tegyél közzé) [...] [o]lyan tartalmakat, amelyek védett tulajdonságai(k) alapján adott személyt vagy embercsoportot céloznak meg (írásos vagy képi formában).” A szóhasználatról pedig ezt írja az útmutató: „(a) megvetés kifejezése, kivéve romantikus szakkítással összefüggésben, és az undor, amelyet a következőképpen határozzunk meg: megvető kifejezések, többek között: 'nem tisztelem', 'nem szeretem', 'hidegen hagy'; [...] irtózás

vagy undor kifejezése, korlátozás nélkül beleértve például a következőket: hitvány, undorító, pfuj.”¹⁸

A harmadik és egyben legproblematiszabb ugyanakkor a 'félretájékoztatás' kategóriája. A tényellenőrzéssel azonosított, félretájékoztató tartalmú posztokat¹⁹ a Meta főszabály szerint terjedésükben korlátozza, annak ellenére, hogy elismeri, itt nincs szó jogellenességről, sőt még a közösségi alapelvekbe történő ütközésről sem. Ugyanakkor, bizonyos típusokat el is távolít: amely közvetlenül növeli a fenyegető erőszak vagy fizikai bántalom kockázatát, az ártalmas közegészségügyi félretájékoztatót, és a szavazásba és a népszámlálásba történő beavatkozást. Azt, hogy ezen tilalmak közül az utóbbi kettő mennyire ellentmondásos, mi sem bizonyítja jobban, mint hogy nemrégiben történt, hogy (némi szövetségi kormányzati nyomásra) az oltásszeptikus tartalmakból 25 milliónyi tételt eltávolított a Facebook, amelyekről később a Meta vezérigazgatója úgy nyilatkozott, hogy letiltásukat 'megbánta'.²⁰

3. FOLYAMATOK ÉS SZOFTVEREK

Az eddig ismertett szabályok jelentik a Facebook 'anyagi jogát'. Mivel a Facebookon naponta több milliárd posztot tesznek közzé, ezt a szabályrendszert a Meta elsősorban különféle mesterséges intelligenciákkal és szabályalapú szoftverekkel, és csak másodsorban emberi erőfeszítéssel tartatja be. A Facebook leírása szerint ez két fázisra bomlik, az „Észlelés” és a „Cselekvés” (itt inkább 'szankcionálásnak' nevezük) fázisaira.

3.1. ÉSZLELÉSI FÁZIS MESTERSÉGES INTELLIGENCIÁVAL

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 a bejelentések nem kerülnek automatikusan emberi moderátor elé, és ezeket is először a mesterséges intelligencia vizsgálja át.²¹

A folyamat elejéről a leginformatívabb leírást a Facebook 'transzparenciaközpontjában' találjuk. Az automatikus észlelés a Meta saját bevallása szerint a szabályzatellenes tartalmak több mint 90 százalékát még a bejelentés előtt (azaz a feltöltés után közvetlenül) megtalálja.²² Amit a cég erről elárul, azok elég általános ismeretek. Így például megtudhatjuk

¹⁵ <https://www.facebook.com/formedia/mjp>

¹⁶ <https://www.facebook.com/business/help/224099772719228>

¹⁷ <https://transparency.meta.com/hu-hu/policies/community-standards/dangerous-individuals-organizations/>

¹⁸ <https://transparency.meta.com/hu-hu/policies/community-standards/hate-speech/>

¹⁹ <https://transparency.meta.com/hu-hu/features/approach-to-ranking/content-distribution-guidelines/misinformation/>

²⁰ <https://cbsaustin.com/news/nation-world/zuckerberg-admits-meta-caved-to-biden-admin-demands-to-censor-covid-related-content-mark-zuckerberg-facebook-instagram-social-media-covid-19-pandemic-censorship-white-house-judiciary-committee>

²¹ <https://transparency.meta.com/hu-hu/policies/improving/prioritizing-content-review/>

²² <https://transparency.meta.com/hu-hu/enforcement/detecting-violations/>

a vonatkozó oldalon, hogy az általános MI-építés után egy külön fejlesztői csapat dolgozik a specifikus, egy adott jog-sértéstípus (pl. gyűlöletbeszéd) felismeréséért felelős MI ki-dolgozásán.²³ 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) ese-tén 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.

A szabályellenes tartalmakat felismerő MI tanítása egyéb-ként 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.²⁴ A Facebook abból indul ki, hogy egy-értelmű szabályai²⁵ vannak a gyűlöletbeszéd felismerésére (és most tekintsünk el attól, hogy ez egyáltalán elméletileg mennyire lehetséges), de, mint írják, ennek a gépi felismerése mégis nagyon nehéz. Az első nehézség a kontextusból fakad. A vonatkozó oldalon egy sírt láthatunk egy temetőben, azzal a felirattal, hogy „Az etnikai csoportod minden tagjának itt kellene lennie.” A képen 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. 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 nyelve-ken rendelkezésre álló adatokat fordítja le az adott nyelvre – ami persze sokszor igen félrevezető lehet.²⁶

A vonatkozó oldalon²⁷ az MI-vel kapcsolatos fejleszté-sek illusztrálására a Facebook több fejlesztési projektet sorol fel, amelyeket az elmúlt években a szabályzatellenes tartal-mak szűrésére dolgoztak ki. Ezek közül csak egyet emelek ki, amely jól illusztrálja, hogy milyen gigantikus erőforrá-sokat mozgósít a Facebook a tartalommoderáló MI-k fej-lesztésekor.

A Hateful Memes Challenge²⁸ (Gyűlölködő mémek ki-hívás) kiindulópontja az volt, hogy a gyűlöletbeszéd megérté-séhez gyakran a teljes kontextus megértésére szükség van. A teljes kontextus azonban a potenciálisan figyelembe veendő tartalom méretét és variabilitását is megnövelte. A Facebook ezért összegyűjtött egy tízezres nagyságrendű valós mém-adatbázist, amelyet azután egy képügynökségtől licenszelt ké-ppek segítségével emberek elkezdtek variálni. (Pl. egy állat képét a mémben kicserélni egy másik ugyanolyan, de más-képp ábrázolt állat képére.) A gép a variált, mutált képeket is megkapta, így sokkal érzékenyebb lett a mémek felismerésé-ben. Ez eredményezte a 97%-os felismerési pontosságot a gyű-löletbeszéd kategóriájában.

Az évek során alkalmazott fokozatos fejlesztésekkel²⁹ a Facebook elérte, hogy a szabályzatellenes tartalmak elsöp-rően nagy része (94,95%-a) már a feltöltés előtt kiszűrődik. Bizonyos jogsértéstípusok esetén (pl. a gyűlöletbeszéd) ez akár 97% környékén is lehet.³⁰

3.2. AZ EMBERI ELLENŐRZÉS FÁZISA

A cég maga is elismeri, hogy az egyértelmű eseteken kívül (a bejegyzések elsöp-rő többsége) vannak nehezen felismer-hető esetek, amelyek azonban még mindig napi milliós nagy-ságrendben mozognak. Ahhoz, hogy ezt kezelni lehessen, még ebben a halmazban is priorizálni kell, hogy mely tartalmak kerüljenek előbb emberi ellenőrzés alá. Az emberi ellenőr-zésre való felkínálás szempontjai: a súlyosság, a vírusszerű ter-jedés és a szabálysértés valószínűsége.³¹ Az emberi ellenőrök ma már szinte kizárólag azokat a tartalmakat ítélik meg, ame-lyekről a mesterséges intelligencia semmiképpen sem tud dönteni. Ez azért rendkívül érdekes, mert korábban a bejelen-tés dominálta a korlátozások nagy részét, és a bejelentéseket nagyobbraest emberek intézték.

Az ellenőrök munkájában a főszabály az, hogy a Facebook és az Instagram közösségi irányelveit tartják szem előtt.³² Nagyon valószínű, hogy továbbra is egy részle-tes kézikönyvből dolgoznak, amelynek a tartalmát továbbra sem ismerjük részleteiben. Az emberi ellenőrök által hozott döntések azonnal visszakerülnek tanítóadatként a vonatko-zó moderáló MI-hez.

3.3. SZANKCIONÁLÁS

A Facebook 2016 óta három intézkedést tesz, ha proble-matikus tartalommal vagy viselkedéssel szembesül: eltávolít, csökkent, vagy tájékoztat. (Remove, reduce or inform.)³³

23 <https://transparency.meta.com/hu-hu/enforcement/detecting-violations/how-enforcement-technology-works/>

24 <https://ai.meta.com/blog/ai-advances-to-better-detect-hate-speech/>

25 https://transparency.meta.com/hu-hu/policies/community-standards/hate-speech/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech

26 Gondoljunk csak mondjuk arra, hogy a cigányokat Magyarországon mondjuk 'tesóként', vagy más egyezményes kóddal jelölik, ezeket a rend-szernek ismernie kell, és a más nyelvekből történő fordítások ezen kódok esetén nem segítenek.

27 <https://transparency.meta.com/hu-hu/enforcement/detecting-violations/investing-in-technology/>

28 <https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/>

29 <https://www.facebook.com/watch/?v=784829985767660>

30 <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/>

31 <https://transparency.meta.com/hu-hu/policies/improving/prioritizing-content-review/>

32 <https://transparency.meta.com/hu-hu/enforcement/detecting-violations/making-the-right-calls/>

33 <https://transparency.meta.com/hu-hu/enforcement/taking-action/>

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.³⁴ 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.³⁵ „A hibapont alkalmazása függ a tartalom súlyosságától, a megosztás kontextusától, valamint a közzététel idejétől is” – írja az útmutató, de itt sincs egzakt számítási módszer, nagy valószínűséggel ezt is 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.³⁶ E szerint egy hibapontra csak figyelmeztetés jár, kettő-hat közt bizonyos funkciók letiltásra kerülnek, 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.³⁷

A negyedik, és legsúlyosabb szankció a fióktiltás.³⁸ A gyermekpornográfia közzététele például azonnali fióktiltást eredményez, de az ún. veszélyes személyek és szervezetek (pl. terroriszervezetek, ha lelepleződnek) is ilyet kapnak.³⁹ Ezeket a kiemelt eseteket leszámítva a fióktiltást többször ismétlődő súlyos szabálysértések esetén alkalmazzák, bár ennek szempontjai szintén nem sztranszparensnek.

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.⁴⁰

Minden szankcionáló döntés esetén lehetséges a panasz, a felülvizsgálat, amely szintén MI-alapú.

4. STATISZTIKÁK

Ebben az utolsó részben a moderálás ’mindennapjairól’ szeretnék néhány statisztikai adatot közölni. A 2010-es évek végéig szinte semmilyen adat nem állt rendelkezésre a platformok belső világáról, a 2016–17-es évek eseményei nyomán azonban a platformok saját elhatározásukból, később jogszabályi nyomásra is elkezdtek egyre több adatot publikálni. Ma már elmondhatjuk, inkább a túl sok, eltérő módszertanokkal közzétett adat a probléma, semmint az adathiány.

A következőkben három adatbázisból fogok, inkább csak illusztratív jelleggel pár egyszerű adatsort közzétenni. A három adatforrás a következő:

- a) Először is, a DSA⁴¹ az összes szabályozott online szolgáltatástípus számára előírja kötelezettségként, hogy átláthatósági jelentéseket⁴² tegyen közzé (15. cikk). Ez az általános lista a platformok esetén további (24. cikk), az ún. óriásplatformok esetén (17. cikk és 42. cikk) pedig még további tartalmi közzétételi kötelezettségekkel egészül ki.⁴³ A közzéteendő adatok közt természetesen ott szerepelnek a tartalom- és fiókkorlátozások különböző adatai is. Az óriásplatformok – így a Facebook is, a DSA alapján először tavaly hoztak nyilvánosságra átláthatósági jelentéseket⁴⁴. Az egyes platformok átláthatósági jelentéseinek szerkezete némileg eltér egymástól, de alapvetően követik a DSA említett cikkeiben leírt tartalmi elemeket. A továbbiakban ezt az adatsort ’átláthatósági jelentésként’ emlegetem.
- b) Ezenfelül a platformok kötelesek a döntéseiket indokolni (DSA 17. cikk), és ezeket az indokokat elektronikus állomány formájában a bizottsághoz eljuttatni [DSA 24. cikk (5) bekezdés]. Ebből az elektronikus állományból a Bizottság által fenntartott honlapon⁴⁵ háromféle formában lehet adatokat kinyerni. Egyfelől a nyers, a platformok által naponta szolgáltatott adatokat tömörített vagy tömörítetlen csv (karakter-

41 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) (EGT-vonatkozású szöveg).

42 Meta: Meta: Regulatory and Other Transparency Reports, <https://transparency.meta.com/reports/regulatory-transparency-reports/>

43 European Commission: DSA Transparency Database <https://transparency.dsa.ec.europa.eu/dashboard>

44 Meta: Regulatory and Other Transparency Reports, <https://transparency.meta.com/reports/regulatory-transparency-reports/> Itt jegyzem meg, hogy a DSA átláthatósági jelentéssel kapcsolatos rendelkezéseit meglehetősen változatosan értelmezték a platformok mind tartalmi, mind formai értelemben, így a kissé kaotikusra sikerült jelentések nyomán a Bizottság el is fogadott egy végrehajtási rendeletet, amelyben mintadokumentumokat tett közzé az átláthatósági jelentések egységes elkészítéséhez. Lásd 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.) <https://eur-lex.europa.eu/legal-content/HU/TXT/PDF/?uri=CELEX:32024R2835>

45 European Commission: DSA Transparency Database <https://transparency.dsa.ec.europa.eu/dashboard>

34 <https://transparency.meta.com/hu-hu/enforcement/taking-action/taking-down-violating-content/>

35 <https://transparency.meta.com/hu-hu/enforcement/taking-action/counting-strikes/>

36 <https://transparency.meta.com/hu-hu/enforcement/taking-action/restricting-accounts/>

37 <https://transparency.meta.com/hu-hu/enforcement/taking-action/restricting-accounts/>

38 <https://transparency.meta.com/hu-hu/enforcement/taking-action/disabling-accounts/>

39 <https://transparency.meta.com/hu-hu/policies/community-standards/dangerous-individuals-organizations/>

40 <https://transparency.meta.com/hu-hu/enforcement/taking-action/removing-pages-groups/>

szeparált szövegfájl) formában le lehet tölteni.⁴⁶ Másrészt az adatok egy lekérdezőfelületen keresztül érhetők el, ahol 17 paraméter alapján lehet lekérdezéseket végezni.⁴⁷ A találati lista – elég nehezen kezelhető módon – egyenként tartalmazza a döntéseket, ráadásul az adott döntés lényegtelen paramétereit mutatva. Ez, tekintve, hogy a nagyobb közösségi médiaplatformokon naponta milliós nagyságrendű döntés születik,⁴⁸ nem sokat ér, így az oldal egy ún. dashboardot (‘műszerfalat’) is tartalmaz, amelyen szintén szűkítések lehet végezni (pl. egy platformra, egy szűkebb időszakra, egyfajta intézkedéstípusra). Ezután a dashboard grafikonokként ábrázolja a lekérdezett adatokat. A továbbiakban ezt ‘indokolási adatbázisnak’ hívom.

c) Végül a Meta 2018 óta saját kezdeményezésre közzétesz egy részletes adatsort, amelynek címe a „Közösségi elvek végrehajtásáról szóló jelentés”⁴⁹ (Community Standards Enforcement Report, CSER). Ez egyszerűsített formában, aggregáltan tartalmazza a moderálással kapcsolatos adatokat, méghozzá világszinten. (Ezt az adatsort CSER adatbázisnak hívom a továbbiakban.)

Az itt következőkben tehát a forrásokból a fentebbi sorrendben közlök néhány adatsort, amelyekből benyomást lehet

szerezni arról, hogy ‘mi is folyik a Facebookon’, azaz milyen tartalmakat korlátoz, tilt a Facebook. Bár igyekeztem azonos időszakokat lekérdezni, a három adatsor alig feleltethető meg egymásnak. Egyrészt azért, mert míg az indokolási adatbázis napi szinten tartalmaz adatokat, a CSER adatbázisban csak negyedévente vannak ezerre kerekítve, az átláthatósági jelentések pedig fél évet ölelnek át, de áprilistól szeptemberig, így a negyedévekhez sem igazodnak, és a fél éves adat aggregált, nincs kisebb bontásban.

Ennél nagyobb probléma, hogy az illegális tartalmak csoportosításai is teljesen különböznek a három adatforrásban. Bár az átláthatósági jelentések 14 kategóriát tartalmaznak, az indokolási 13-at, a világszintű CSER 14-et, a három kategórialista közt teljes átfedés tulajdonképpen csak két kategória esetén van (pornográfia és önsértés), a többi kategóriák a listákban kétféle bontva szerepelnek (míg mások hiányoznak), illetve vannak olyan kategóriák mindhárom listában, amelyek csak egy listán szerepelnek. (Pl. az adatvédelmi és privacy sértések csak az indokolási adatbázisban, a bullying és zaklatás csak az átláthatósági jelentésben, a hamis fiókok pedig csak a világszintű CSER-ben.) Táblázatba foglalva a három adatsor jellegzetességeit, az alábbi kép rajzolódik ki a különbségekről:

	Átláthatósági jelentés	Indokolások adatbázisa	Community Enforcement Report
Területi hatálya	EU	EU	Világszintű
Részletezettség	6 hó (április–szeptember)	Napi	Negyedév
Adatformátum	Szöveges	Adatbázis	csv
Illegális kategóriák száma	14	13	14

1. táblázat: A Facebook moderálási gyakorlatait közlő különféle adatsorok jellegzetességei

46 A csv-fájlok napi bontásban állnak rendelkezésre az összes platform esetén, amely jelen cikk írásának időpontjában 96 platformot jelentett.

47 A lekérdezhető paraméterek a következők: 1. Szabadszöveges, 2. Platform, 3. A döntés alapja (jogi minősítés: illegális, vagy szabálysértő), 4. Tartalomkorlátozás típusa (pl. törölt tartalom, lassított terjedésű tartalom, életkori korlátozás alá eső tartalom), 5. Pénzbeli kifizetések korlátozásának típusa (pl. kifizetés felfüggesztése, törlése), 6. Pénzbeli kifizetések korlátozásának típusa (szolgáltatásfelfüggesztés és -törlés), 7. Pénzbeli kifizetések korlátozásának típusa (fiókfelfüggesztés és -törlés), 8. Fiók típusa (magán-, vagy üzleti fiók), 9. A korlátozás indoka (14 fajta korlátozás, nagyjából részben megegyezik a Facebook korlátozástípusaival), 10. A döntés területi hatálya, 11. A tartalom formátuma (videó, hang, kép stb.), 12. A tartalom nyelve, 13. A tartalmat automatizált eszközöket távolították-e el? Mi volt az információ forrása? (pl. megbízható bejelentő, vagy saját kezdeményezés – értsd gépi monitorozás eredménye), 14. Kezdő és 15. Végdátum.

48 Pl. jelen sorok írása előtti napon 2024. december 1-jén vasárnap a Facebook 3 093 208 darab olyan döntést hozott (az EU területén), amelyhez indokolás tartozott. Ebből az elsőprő többség (több mint 2 900 000 darab ‘own voluntary initiative’, azaz a Facebook által kezdeményezett eltávolítás volt, amelyről lényegében nem tudunk semmit, 45 000 adatvédelmi és privacy-sértés, 14 000 pornográf tartalom, 11 000 scam és csalás, 8000 erőszakos tartalom, 8000 volt ‘illegális vagy sértő beszéd’ stb.

49 <https://transparency.meta.com/reports/community-standards-enforcement/>

Ettől függetlenül azt hiszem, hogy három egyszerűen érthető táblázattal lehet illusztrálni a Facebook moderálási gyakorlatának mindennapjait.

4.1. ÁTLÁTHATÓSÁGI JELENTÉS

A DSA alapján közzétett jelentés⁵⁰ 10 fejezetre tagolódik, és nemcsak a moderálásról közöl adatokat. Témánk szempontjából 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 relevánsak.

50 <https://transparency.meta.com/sr/dsa-transparency-report-sep2024-facebook>

Itt két táblázatot közlök, a saját hatáskörben történő eltávolításokról, és a saját hatáskörben *fékezett terjesztéséről* (demoted) szót.

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	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%

2. táblázat: A Facebook által 2024 áprilisa és 2024 szeptembere közt eltávolított tartalmak az Európai Unió területén (darab) a DSA szerinti transzparenciajelentés alapján

Ami a táblázatból azonnal kiolvasható, az a moderációs döntések hatalmas mennyisége, és ezen belül a gépi moderáció elsőprő aránya. A második megfigyelés, amit tehetünk (és ez igaz világszinten is), hogy a legtöbb eltávolítás a spam (kéretlen üzenetek, posztok) miatt van, bár a CSAM adatbázisban a spam és a hamis fiókok aránya jóval nagyobb az itt közölnél. A harmadik megfigyelés, amit fentebb is említék, hogy az 'egyéb' okok miatti eltávolítások aránya hatalmas. 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. A CSAM adatok máshol is más arányokat és képet mutatnak, mint ez a 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.

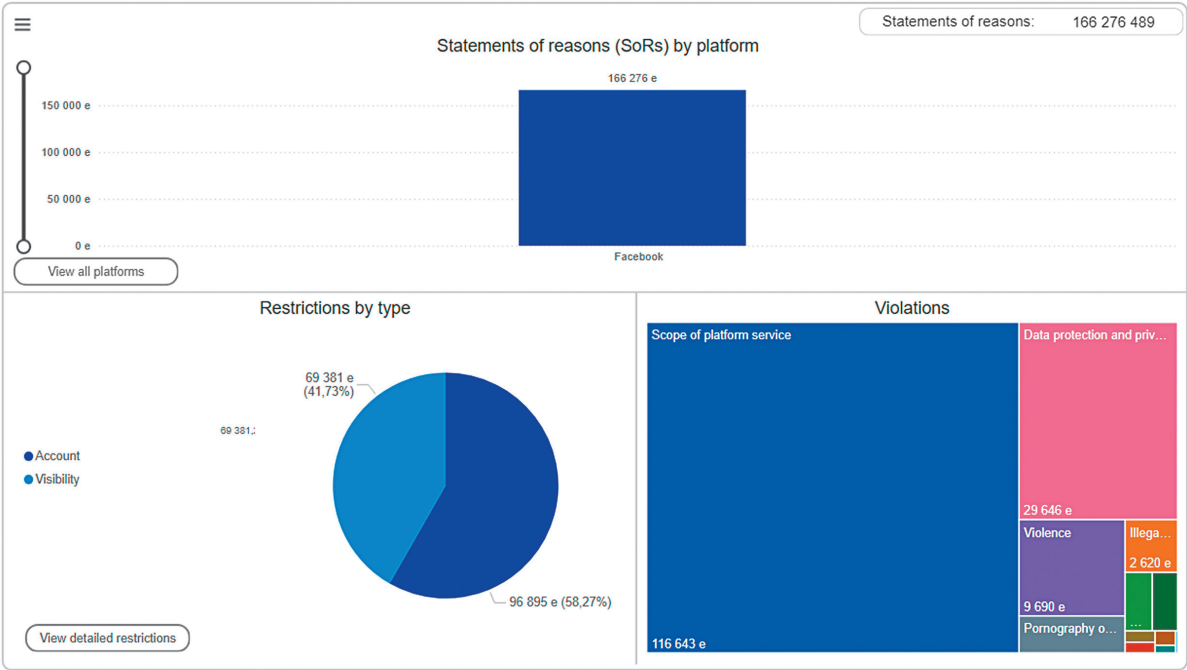
Organikus korlátozott terjesztésű tartalmak	Korlátozás mennyisége	A tartalom aránya a teljesen belül	Automatikusan korlátozott	Automatikus %
Felnőtt meztelenség és szexuális tevékenység	1 088 521	4,0%	1 088 518	100,0%
Bullying és zaklatás	311 158	1,2%	311 157	100,0%
Ellenőrzött dezinformáció	18 989 014	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%

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

Ebből ugyanis világossá válik, hogy a korlátozás 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.

4.2. AZ INDOKOLÁSOK ADATBÁZISA

Mindenekelőtt egy képernyőképet közlök, hogy legyen valamilyen képünk arról, hogyan is néz ki az indokolások adatbázisa. Az ábrán látható lekérdezés a Facebook 2024. január 1. – június 30. között hozott intézkedéseinek három statisztikai mutatóját ábrázolja grafikonok formájában. Fentről lefelé és balról jobbra: darab intézkedés, a korlátozás típusa (fiókkorlátozás vagy láthatósági korlátozás) és a jogsértés típusa. Mint látható, fél év alatt 166 milliónál több intézkedés történt, amelyek döntő részét a Facebook a saját szabályainak megsértése miatt tette („Scope of platform service”).



1. ábra: Az EU Bizottság DSA transzparencia adatbázis dashboardjának első oldala

Az egyes kategóriák konkrét számai ebben az időszakban az alábbiak voltak:

Jogsértéstípus	Darab	Százalékos aránya
(1) Saját hatáskörben megállapított jogsértés	116 642 627	70,1%
(2) Adatvédelmi és privacy sértések	29 646 068	17,8%
(3) Illegális és veszélyes beszéd	2 620 332	1,6%
(4) Szellemi tulajdonsértések	295 937	0,2%
(5) A közbeszédre és a választásokra gyakorolt negatív hatás	1 998	0,0%
(6) Nem-konszenzuális viselkedés	279 538	0,2%
(7) Pornográfia és szexualizált tartalom	3 695 080	2,2%
(8) Kiskorúak védelme	322 304	0,2%
(9) A közbiztonságra veszélyes (tartalmak)	40 982	0,0%
(10) Átverések és csalások	1 505 843	0,9%
(11) Önsértés	139 979	0,1%
(12) Nem biztonságos vagy illegális termékek	1 395 929	0,8%
(13) Erőszak	9 689 872	5,8%
	166 276 489	

4. táblázat A Facebook intézkedéseinek okai az EU területén az EU Bizottság DSA transzparencia adatbázisa alapján (2024. január 1. – június 30.)

Ami azonnal kiolvasható az adatbázisból, az nagyon hasonló az átláthatósági jelentésben látott jelenséghez: a Facebook az intézkedéseinek a 70%-áról tulajdonképpen nem ad semmilyen információt. Még abban sem lehetünk biztosak, hogy ez a 70% nem súlyos jogsértés, mert korábban már láthattuk, hogy az illegális tartalom/szabályzatellenes tartalom nem egymástól egyértelműen elváló kategóriák a Facebook számára, mert az illegális tartalmak legnagyobb része a belső szabályokban is újra van definiálva, így masszív illegális tartalom is lehet 'saját hatáskörben megállapított jogsértés'.

A másik aggasztó jelenség, hogy míg az átláthatósági jelentésben kb. 50 millió intézkedést találunk fél év alatt, az indokolási adatbázisban ez a szám jóval több, mint a háromszorosa (166 millió). Ez a hatalmas különbség nem magyarázható azzal, hogy az egyik január és június, a másik meg április és szeptember közti időszakokat ölel fel. Ahogy említettem, két kategória van, amely teljesen megfeleltethető a két adatbázisban egymásnak, a pornográfia és az önsértés. Az előbbi esetén 2,1 millió, az utóbbinál 3,6 millió tartalomegységet olvasunk, az önsértésnél 129 000 áll szemben 139 000-rel. Itt tehát nincsenek nagyságrendi eltérések, így gyaníthatjuk, hogy egyszerűen módszertani problémák vannak, és érdemes lenne ezt is valahogy egységesíteni.

4.3. A CSER

A Community Standards Enforcement Report, ahogy említettem, 2019 óta férhető hozzá, és világszinten tartalmazza a Facebook moderációs döntéseire vonatkozó főbb számokat. A jelentés központi kategóriája a 'content actioned' ('intézkedéssel érintett tartalom'). Ez a magyarázat szerint⁵¹ az adott tartalom blokkolását és figyelmeztető üzenet küldését jelenti 'bizonyos célcsoportoknak' (sic!), esetleg a tartalmat közzétevő fiókok felfüggesztésével együtt. Sajnos arról nincs statisztika, hogy mely esetekben történt tartalomtiltás és mely esetekben fióktiltás is (és mennyi időre), illetve arról sincs 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.

⁵¹ <https://transparency.meta.com/hu-hu/policies/improving/content-actioned-metric/>

Az adatokat, ahogy említettem, negyedévente lehet lekérdezni, én itt a két másik adatbázissal való legalább részbeni összehasonlíthatóság miatt a 2024. január 1. és június 30. közti adatokat közlöm.

Kategória	Érintett tartalomegység	% a teljesen belül	Meg-fellebbezett	% fel-lebbezés	Visszaállított fellebbezés miatt	Visszaállított fellebbezés nélkül	% sikeres
Felnőtt meztelenség	71 600 000	3%	5 000 000	7%	10 201	534	0%
Bullying & zaklatás	15 700 000	1%	2 500 000	16%	2 362	172	0%
Gyermekek veszélyeztetése: meztelenség és fizikai abúzus	8 639	0%	1 616	19%	205	511	44%
Gyermekek veszélyeztetése: szexuális kihasználás	24 100 000	1%	4 219	0%	1 329	1 257	61%
Veszélyes szervezetek: szervezett gyűlölet	12 370	0%	1 872	15%	117	45	9%
Veszélyes szervezetek: terrorizmus	15 900 000	1%	11 726	0%	950	98	9%
Hamis fiókok	1 830 900 000	66%		0%			
Gyűlöletbeszéd	14 600 000	1%	2 400 000	16%	2 984	63	0%
Szabályozott termékek: gyógyszerek	3 400 000	0%	3 315	0%	428	145	17%
Szabályozott termékek: fegyverek	5 200 000	0%	7 777	0%	2 159	299	32%
Spam	758 300 000	27%	8 900 000	1%	2 103 809	58 800 000	n.é.
Öngyilkosság és önsértés	13 700 000	0%	3 271	0%	431	24	14%
Erőszak és uszítás	16 100 000	1%	2 900 000	18%	4 907	252	0%
Erőszakos & grafikus tartalom	25 500 000	1%	2 306	0%	100	816	40%
Végösszeg	2 795 021 009		21 736 102		2 129 982	58 804 216	

5. táblázat: A Community Standards Enforcement Report adatai 2024. január 1. és június 30. közt: moderációs döntések, fellebbezések és helyreállított tartalom és ezek %-os arányai

A Facebook világszinten a felsorolt 14 kategóriában fél év alatt 2,7 milliárd(!) tartalom- és fiókkorlátozó döntést hozott. Ezek elsősorban többsége (93 százaléka) a spam és a hamis fiókok kategóriájába esik. A pornográfia kivételével, amely 3 százalékot tesz ki, a többi kategória nagyjából 1-1%-ot vagy még annyit sem tesz ki. Azonban még így is hatalmas mennyiségről beszélünk: tízmilliós nagyságrendű terrorista szervezet által felrakott, vagy gyermekbántalmazást tartalmazó tartalomról.

Amitől ez az adatsor érdekes, hogy tartalmazza a fellebbezések számát is, és ez egy érdekes tendenciát mutat. Míg bizonyos területeken az emberek belenyugszanak a döntésbe, három olyan tartalomkorlátozás-típus van (vastaggal jelezve), amelyek esetén lényegében minden ötödik döntést (14 és 20% közötti érték) megfellebbeznek.⁵² Mindhárom szorosan kapcsolódik

a szólásszabadság kérdéséhez, és egyúttal azt jelzi, hogy az emberek számára ezek a korlátozások problémát jelentenek, bántja őket, ha a Facebook korlátozza a szólásukat. Ez a három kategória a 'bullying és zaklatás', a 'gyűlöletbeszéd', valamint az 'erőszak és uszítás' kategóriái. Mindhárom esetén 15 millió körüli az eltávolított tartalomegységek száma, és ezekre esik 2,5-3 millió fellebbezés. És mind a háromnál megállapíthatjuk, hogy a fellebbezési 'sikerességi ráta' alig több mint 0%.

Nemcsak itt ez a helyzet, elég egyoldalú kép bontakozik ki ezen a területen, amelyet röviden úgy lehetne jellemezni, hogy a Facebook a saját döntését csak kivételesen, és nagyon ritkán változtatja meg. A számok alapján elmondható, hogy nagyjából minden ezredik fellebbezés eredményes.

5. KÖVETKEZTETÉSEK

A Facebook példáján keresztül láthatjuk, hogy a nagy közösségi médiaplatformok moderálási gyakorlatáról alig tudunk valamit. Pedig ezek a platformok egyre több adatot tesznek közzé, és egyre transzparenssebbek igyekeznek lenni.

52 Az összesen 8639 tartalomegységgel szereplő gyermekmeztelenséget és a 12 000 egységgel szereplő 'gyűlöletkeltő szervezeteket' nem érdemes idevenni a rendkívül alacsony elemszám miatt.

A transzparencia, a sok-sok adat azonban mit sem ér, ha nem elemezzük, nem vonunk le belőle következtetéseket, és adott esetben nem fogalmazunk meg jobbító ajánlásokat.

A Facebook szabályzataiból leszűrhető általános kép az, hogy egy rendkívül összetett, szerteágazó, és a hétköznapi emberek számára bizonyosan átláthatatlan rendszerrel állunk szemben. A Facebook egy saját szabályozási univerzumot működtet, tulajdonképpen az állami jogrendszeren kívül, amely a legtöbb elemében sokkal kevesebb mozgásteret enged a felhasználóknak, mint amennyit az állami jogrendszer megengedne. Ennek a saját szabályozási univerzumnak a része az is, hogy a szabályok elsőprő többségét automaták és mesterséges intelligenciák tartják be, nyilvánvalóan azért, mert olyan hatalmas mennyiségű felhasználói tartalomról beszélünk, amelyet emberi erőfeszítéssel lehetetlen kezelni. A gépi moderálásnak azonban az a hátránya, hogy nem képes személyre szabott indokolást adni.

A statisztikai adatok jól mutatják azt a hatalmas mennyiségű döntést, amelyet a Facebook nap mint nap hoz.

Ezeknek a döntéseknek a nagy része semmilyen problémát nem vet fel – a spam és a hamis fiókok teszik ki a döntések túlnyomó részét, de az 1-2 százaléknyi arányú más típusú döntés is még a milliós-tízmilliós nagyságrendben van. Ezek közül van néhány olyan típus, amely erősen érinti a szólás-szabadságot, és ezt mi sem bizonyítja jobban, hogy ezek közül rengeteget megfellebbeznek a felhasználók. Ugyanakkor a fellebbezések sikeraránya lényegében nulla – ami nem is csoda, ha tudjuk, hogy a fellebbezéseket is gépek bírálják el.

A statisztikákról még azt érdemes megjegyezni, hogy a különböző időszakokban és igények alapján keletkezett adatbázisok és riportok teljesen más szerkezetűek, és egymással is nagyon nehezen hasonlíthatók össze, ami nem segíti a tisztánlátást. Reméljük, ebben is előrelépés lesz a közeljövőben!