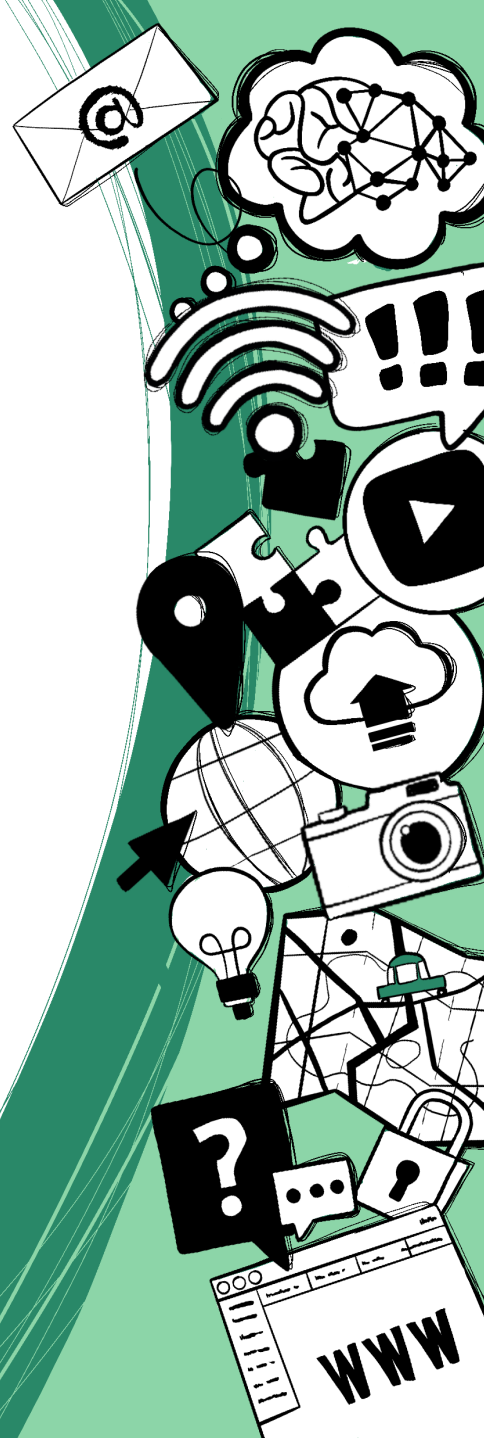
ENARIS

MI rendszerek torzítási hibái



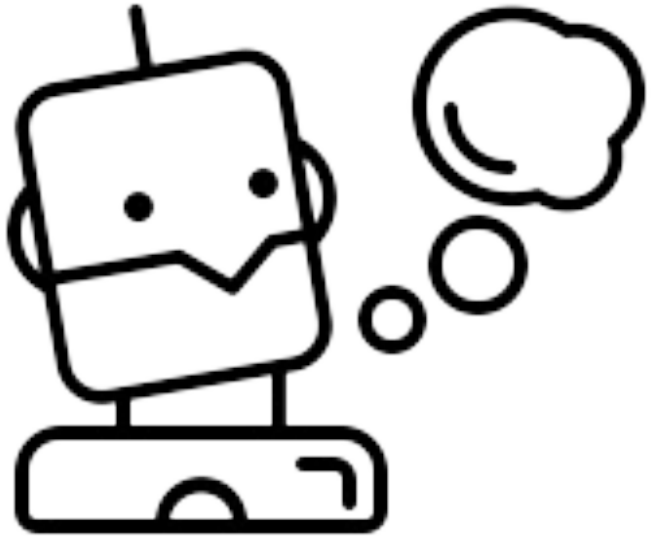
mobilis
interaktív kiállítási központ

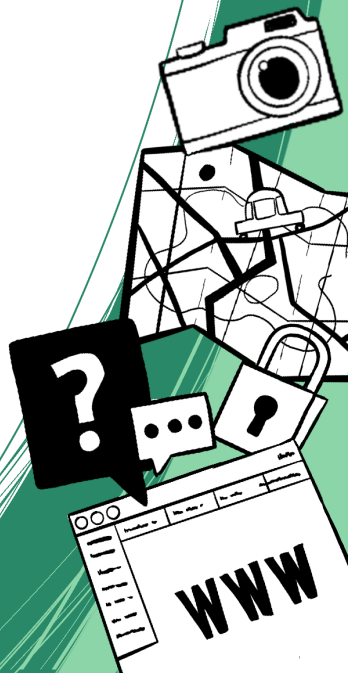
TU Graz

OESTERREICHISCHE COMPUTER GESELLSCHAFT
AUSTRIAN COMPUTER SOCIETY

Interreg
Austria–Hungary 2014–2020
European Union – European Regional Development Fund

Egyáltalán mik azok a torzítási hibák?



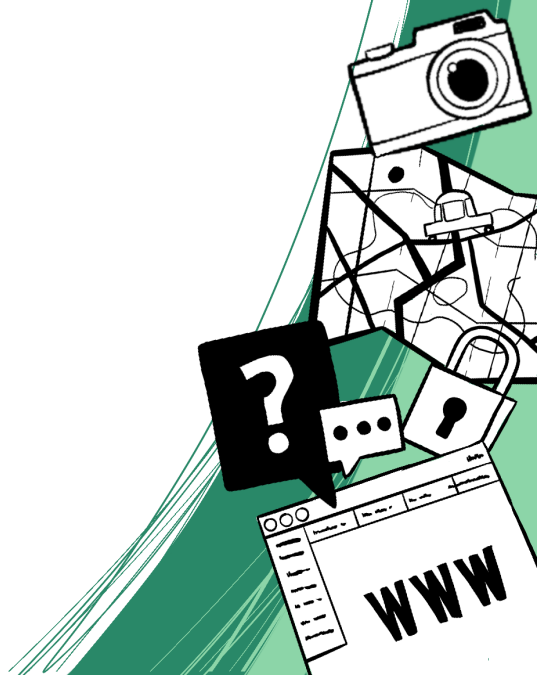


Egyáltalán mik azok a torzítási hibák?

- Torzítás vagy előítélet

  → a rendszerhez szükséges adatok gyűjtése során merülnek fel

  → komoly rendszerbeli problémákhoz és diszkrimincióhoz vezethet



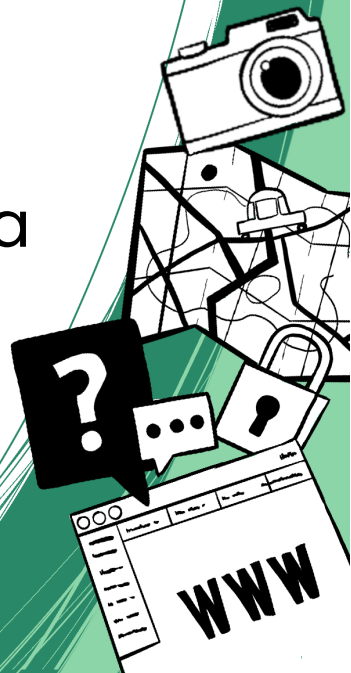
Egyáltalán mik azok a torzítási hibák?

Példa: **Tay Chatbot** (2016)

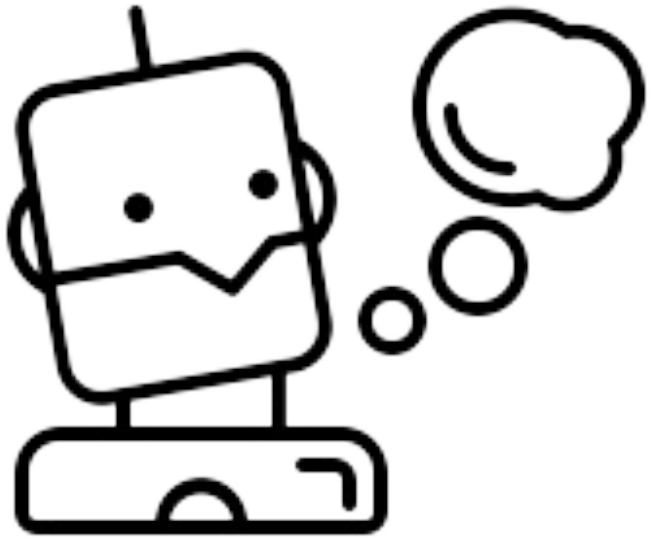
Twitter chatbot, amely ‘valódi felhasználók' tweetjeiből “tanult”

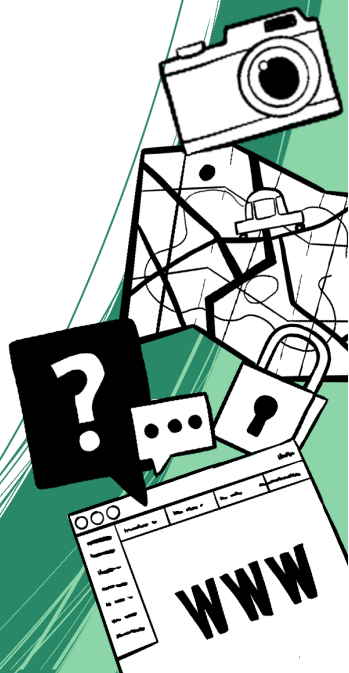
→ nagyon hamar elkezdett diszkriminatív üzeneteket posztolni

→ 16 óra elteltével a chatbotot visszakapcsolták offline módba



Milyen típusú torzítási hibákról beszélhetünk?

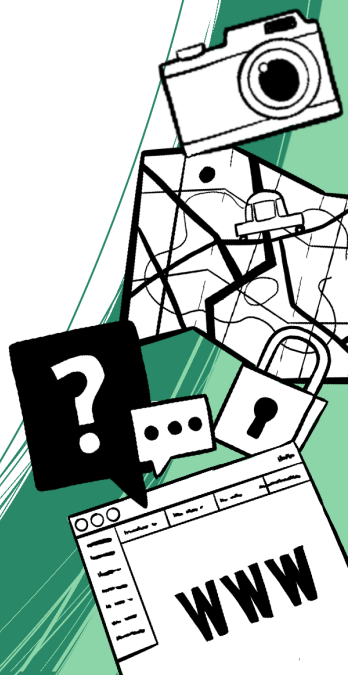




Milyen típusú torzítási hibákról beszélhetünk?

- **algoritmikus MI-torzítás** vagy “**adat torzítás**”: a betáplált adatok által okozott torzítási hiba (az adatok statisztikai jellegű torzítottsága)

- **társadalmi MI-torzítás**: a társadalom által „belénk programozott” normák; de a sztereotípiák is képezhetnek vakfoltokat vagy (társadalmi) előítéleteket

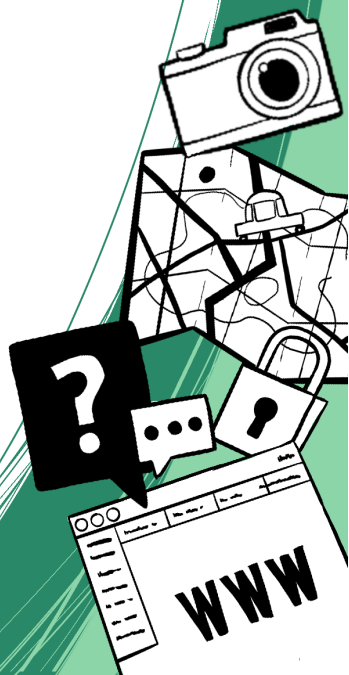


Milyen típusú torzítási hibákról beszélhetünk?

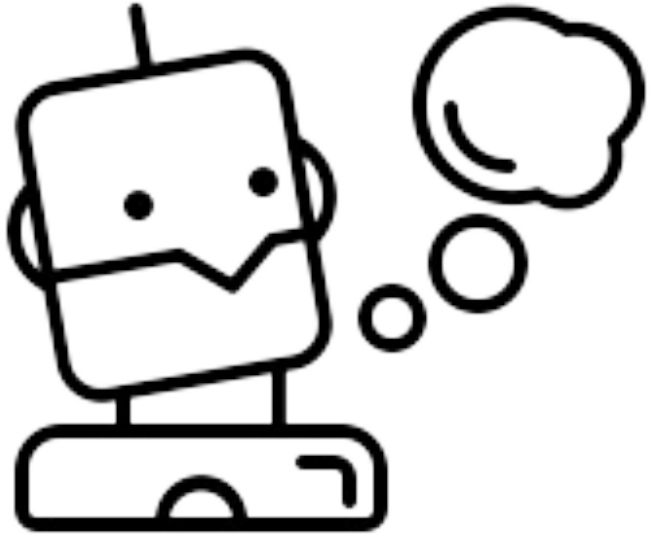
- **algoritmikus MI-torzítás** vagy "**adat torzítás**": a betáplált adatok által okozott torzítási hiba (az adatok statisztikai jellegű torzítottsága)

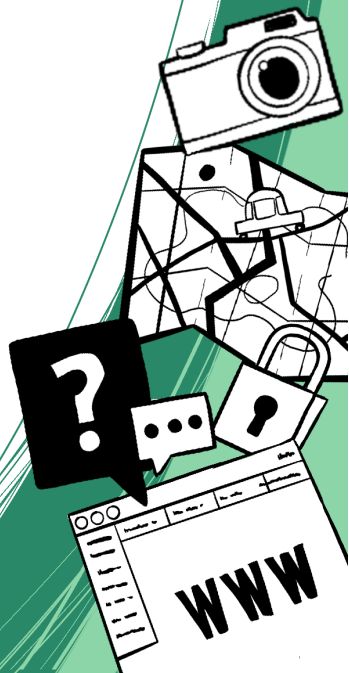
- **társadalmi MI-torzítás**: a társadalom által „belénk programozott" normák; de a sztereotípiák is képezhetnek vakfoltokat vagy (társadalmi) előítéleteket

→ A társadalmi előítéletek gyakran befolyásolnak/teremtenek algoritmikus torzítási hibákat!



Miért van szükségünk etikai szabályokra a mesterséges intelligenciához?

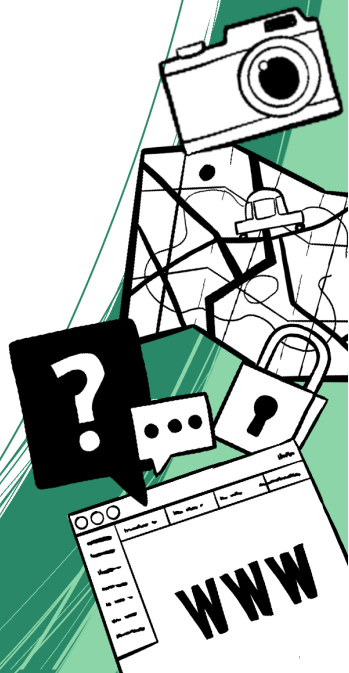




Miért van szükségünk etikai szabályokra a mesterséges intelligenciához?

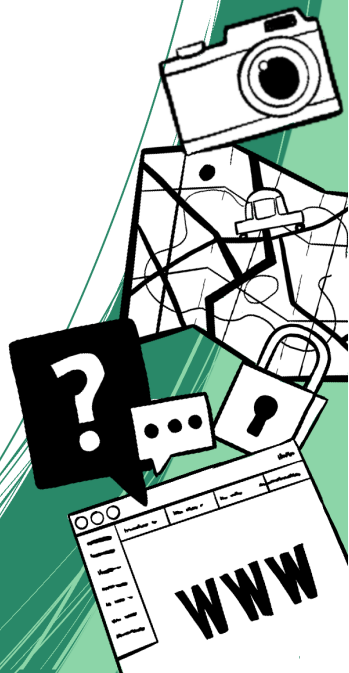
- A programozók (akaratlanul is) beleépíthetik saját előítéleteiket a programokba.

- Az etikai szabályok igyekeznek biztosítani, hogy senkit ne zárjanak ki és diszkrimináljanak.



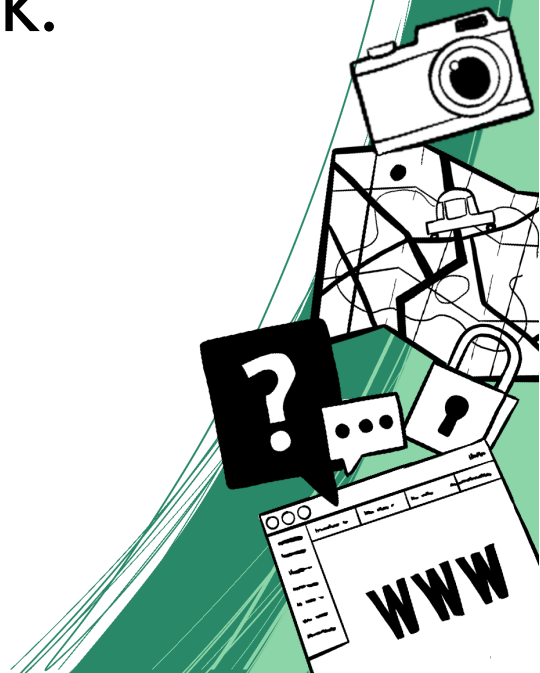
Etikai irányelvek egy megbízható MI-ért

- **Méltányosság**
- **Az emberi autonómia tiszteletben tartása**
- **Védelem az ártalmaktól**
- **Nyomonkövethetőség**



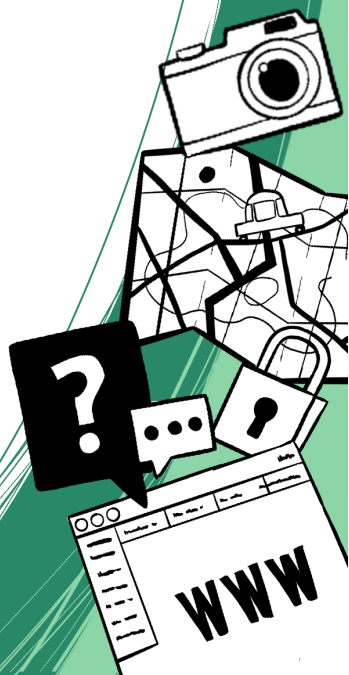
Méltányosság

- Véd a diszkrimináció ellen.
- Egyenlő lehetőségek
- Az embereket nem szabad megtéveszteni.
- Az MI-rendszereknek átláthatónak kell lenniük.



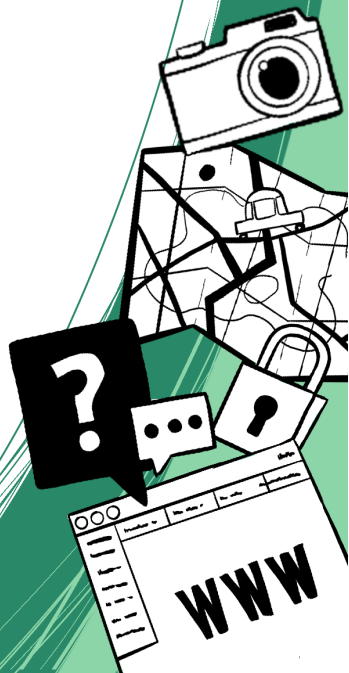
Az emberi autonómia tiszteletben tartása

- Önrendelkezés (tudok dönteni magamról)
- Alapvető jogok megélése
- Az MI-rendszereket úgy tervezték, hogy segítsék és támogassák az embereket.
- Az MI-rendszerek emberi felügyelete



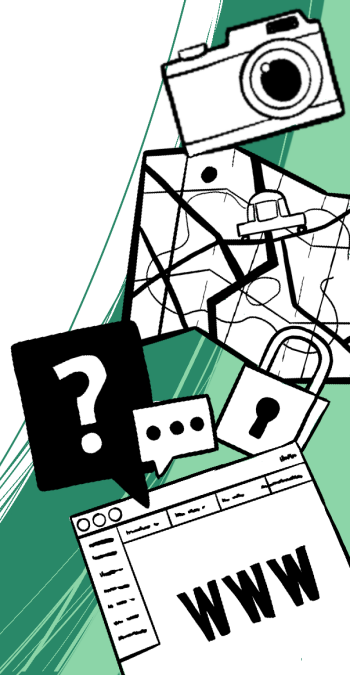
Védelem az ártalmaktól

- Az MI nem okozhat kárt, és nem is súlyosbíthatja a meglévő károkat (mentális és fizikai értelemben sem).
- A MI-nek technikai értelemben sziklaszilárdnak kell lennie.
- Kiszolgáltatott személyek (gyermekek, fogyatékkal élők ...) védelme
- Hatalom vagy információ egyenlőtlen elosztása (pl. kormány és állampolgárok)



Nyomonkövethetőség

- A folyamatoknak átláthatónak kell lenniük
- Óvakodjunk a "**fekete doboz algoritmusoktól**" (amelyeknél zavaros, hogy a rendszer hogyan jut el az adott végeredményhez).

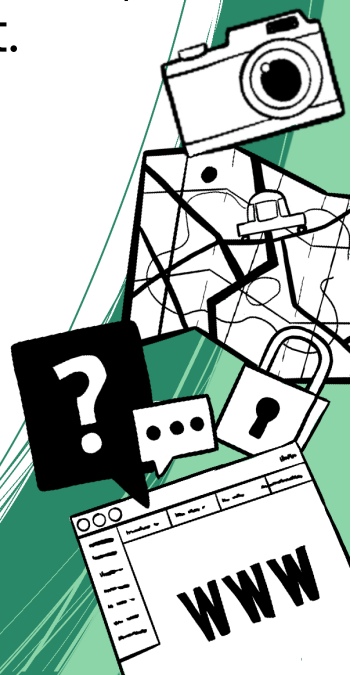


Vigyázat!

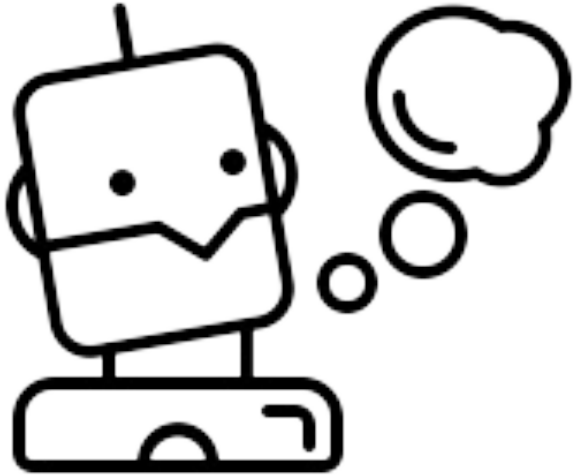
- Néha ez a négy elv nem fér meg egymás mellett!

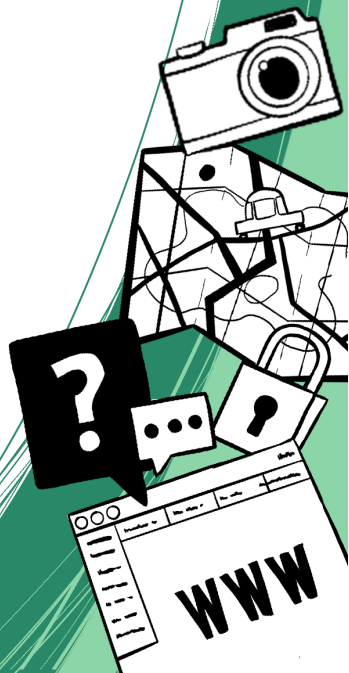
- Pl. "rendőrségi bűnmegelőzés"

  A speciális megfigyelési módszerek segíthetnek a bűnözés elleni küzdelemben, de egyúttal korlátozzák az egyéni szabadságot és az adatvédelmi jogokat.



Hogyan kerülhetők el a torzítási hibák?





Hogyan kerülhetők el a torzítási hibák?

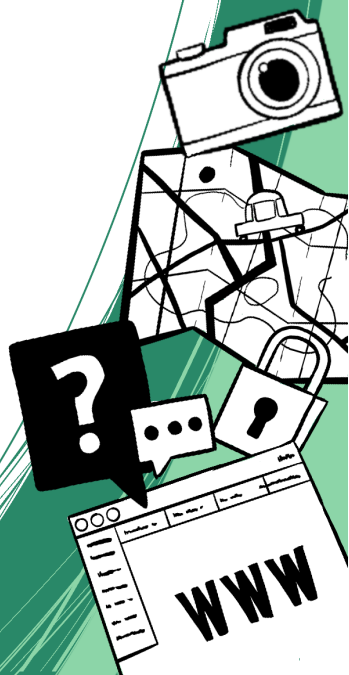
1. Önreflexió

Különböző szempontok figyelembevétele: Előfordul-e, hogy sztereotípiákban gondolkodom? Vannak-e előítéleteim másokkal szemben?

2. Aktív kommunikáció

Észrevettem valamit egy programban? Kirekesztve vagy diszkriminálva érzem magam?

→ foglalkozzunk vele közvetlenül



Hogyan kerülhetők el a torzítási hibák?

1. **Önreflexió**

Különböző szempontok figyelembevétele: Előfordul-e, hogy sztereotípiákban gondolkodom? Vannak-e előítéleteim másokkal szemben?

2. Aktív kommunikáció

Észrevettem valamit egy programban? Kirekesztve vagy diszkriminálva érzem magam?

→ foglalkozzunk vele közvetlenül

Más ötlet?