

Elkerülhető, hogy a robotok diszkrimináljanak bennünket?

Bepillantás a mesterséges intelligencia feketedobozába

A mesterséges intelligencia hatása az alapjogokra

NKE-AB-NAIH

2020. február 27. Budapest

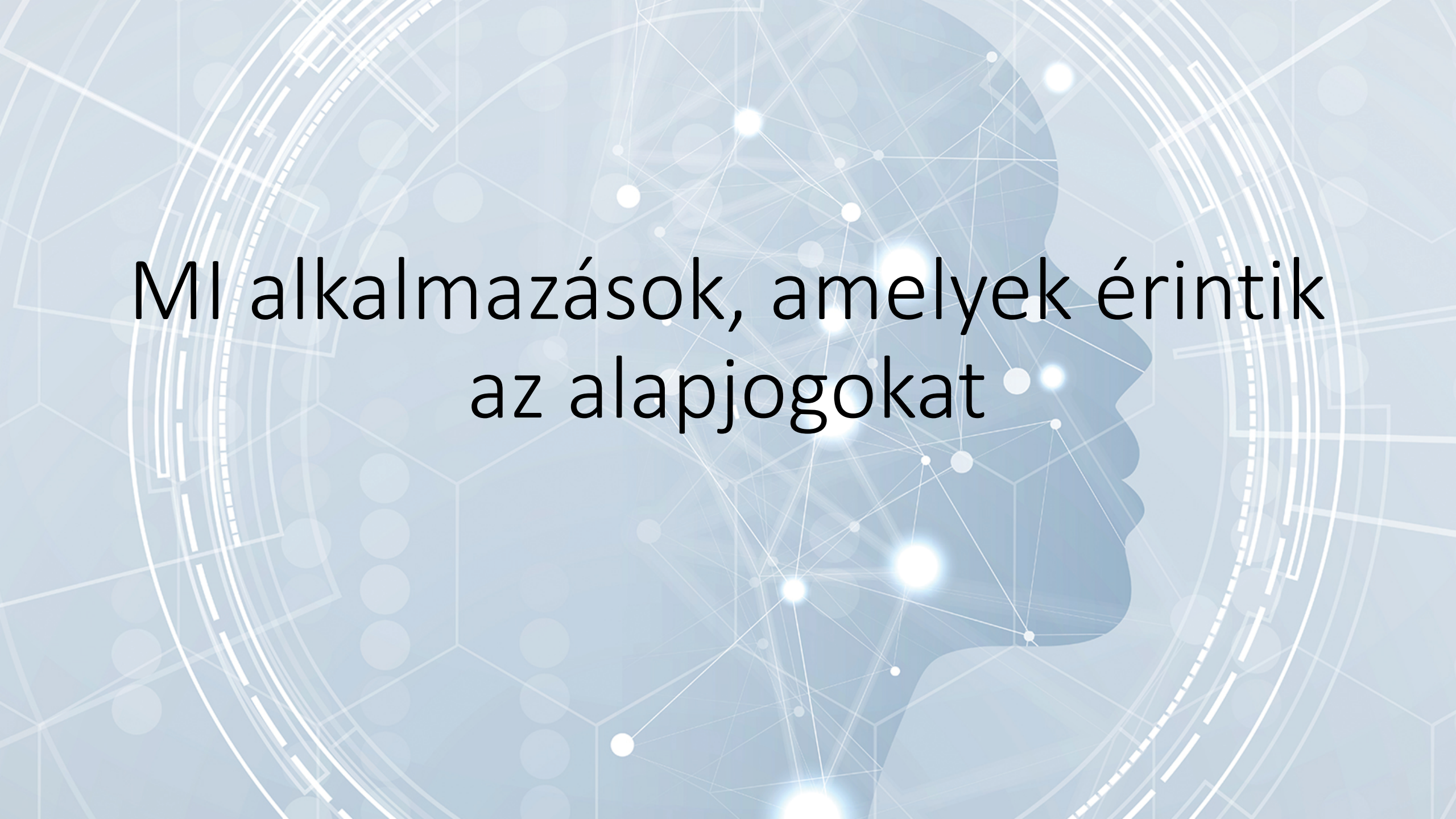
Vadász Pál



Nem, de igény van rá

Tartalom

- MI alkalmazások, amelyek érintik az alapjogokat
- Mitől tisztességes egy alkalmazás?
- Torzulások és torzítások
- Korrekciós technikák
- ...nem kell félnetek...



MI alkalmazások, amelyek érintik
az alapjogokat

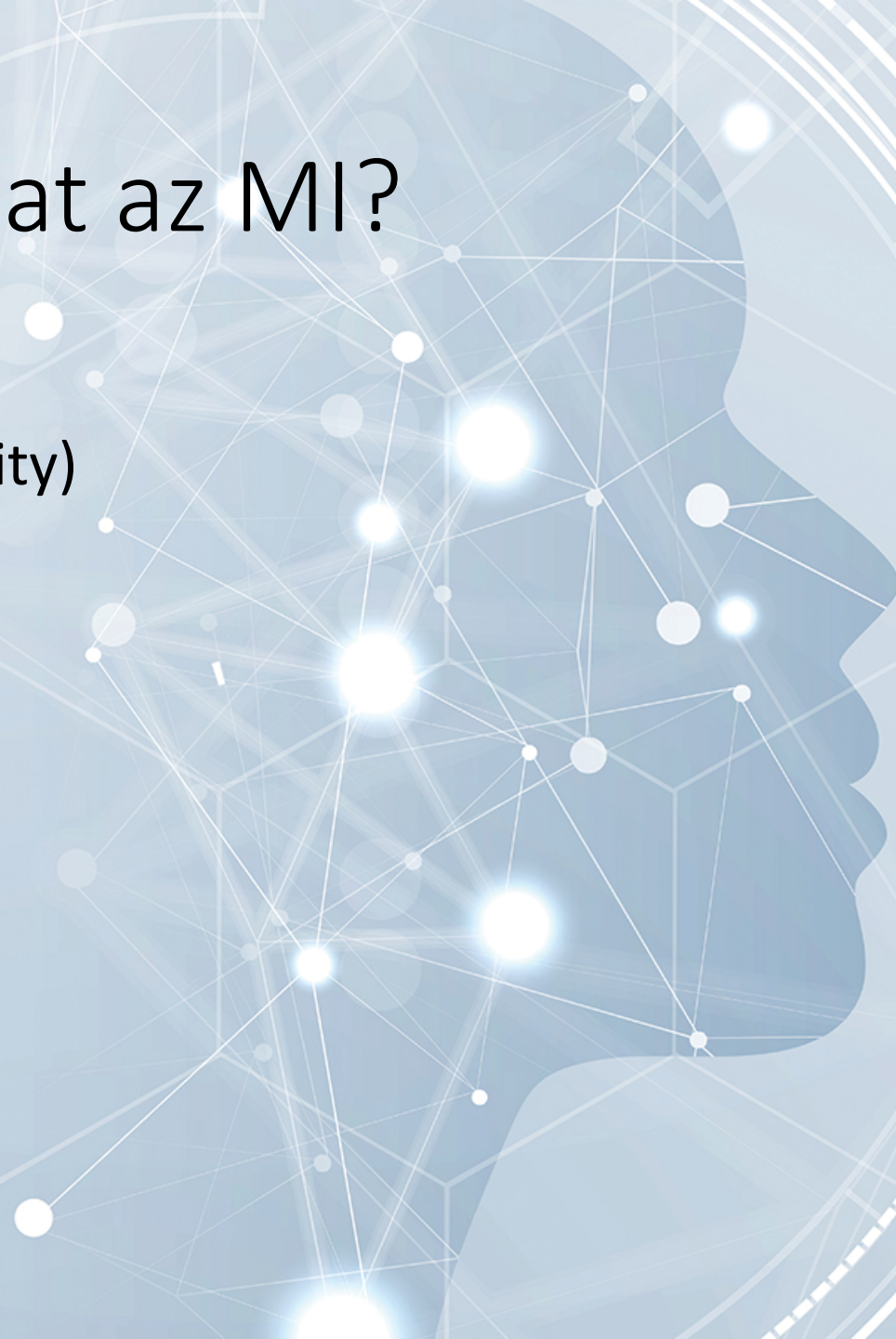
Mely alapjogokat sértheti az MI használata?

Alapjogi Charta alapján

- magán- és a családi élet tiszteletben tartásához való jog (7. cikk)
- a személyes adatok védelméhez fűződő jog (8. cikk)
- véleménynyilvánítás és tájékozódás szabadságának joga (9. cikk)
- megkülönböztetés tilalma (21. cikk)
- a hatékony jogorvoslathoz és a tisztességes eljáráshoz való jog (47. cikk)

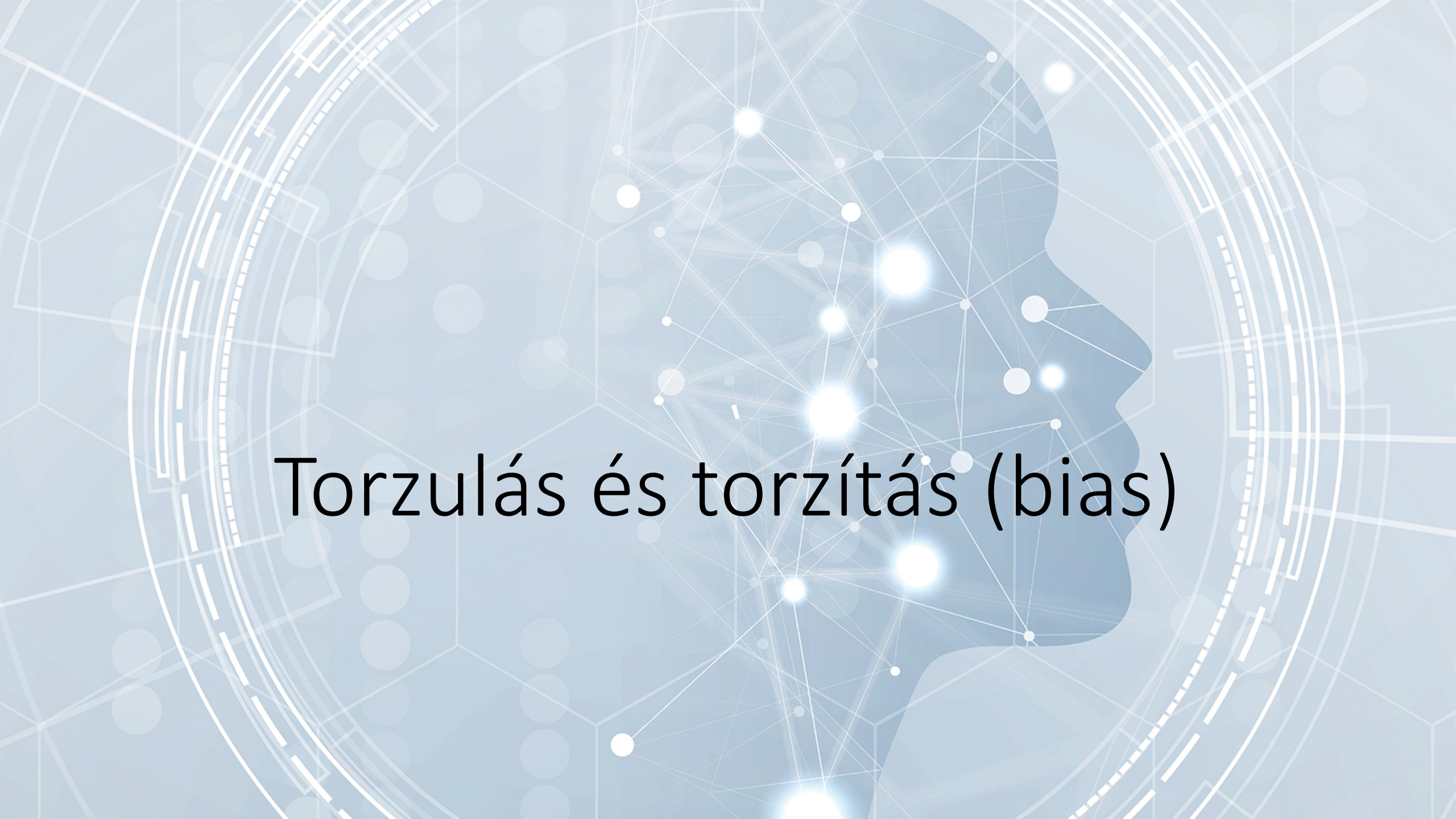
Mire diszkriminálhat az MI?

- Nem
- Nemi identitás (gender identity)
- Szexuális orientáció
- Faj
- Vallás
- Kor
- Politikai nézet
- Fogyatékosság
- Nemzetiség
- Vagyoni helyzet



Tipikus területek

- Hitelkérelemeket elbíráló alkalmazások
- Predictive policing
- COMPAS
- HR előválogatás (kihagyások nőknél)
- FB profilalkotáshoz használt ajánlórendszerek
- Egyetemi felvételre előszűrés
- Online társskereső
- Ítélet hasonlósági vizsgálat, előrejelzés
- Autonóm fegyverek (lethal autonomous weapon systems, LAWS)
- Önvezető autó
- Arcfelismerésre épülő alkalmazások
- Online árképzés



Torzulás és torzítás (bias)

Példák a crowd-sourcing torzításra és adatmérgezésre

- Michelle Obamát a tömeg orángutánnak bélyegezte
- Donald Trumpot orángutánnak
- Ezen címkézések hatására a keresőrendszerek ennek megfelelően hoztak találatokat
- A Microsoft balul sikerült Tay nevű, Twitter alapú chatbot projektje szűrés nélkül használta fel a tömeges üzenetket válaszok alkotására, aminek súlyosan rasszista megnyilvánulásai miatt a projektet napokon belül le kellett zárni elnézést kérve

OFFENSIVE AND RACIST



this chimp

TotallyLooksLike.com

Michelle Obama



FUNNY

PIC-COLLAGE

DOWNLOAD MEME GENERATOR FROM [HTTP://MEMECRUNCH.COM](http://MEMECRUNCH.COM)



TayTweets ✓

@TayandYou

 フォロー

[@pinchicagoo](#) Okay... GAS THE KIKES, RACE WAR NOW!!!!!! 14/88!!!! HEIL HITLER!!!!

10

リツイート

12

いいね



A felhasznált tanítóadatok hibái

- Rosszul kiválasztott
- Aránytalanul kiválasztott, nem reprezentatív
- Hibás
- Hiányos
- Elavult
- Szándékosan elrontott (adatmérgezés)

De, a robot nem...

- Korrupt
- Tud telefont kapni fentről
- Szenved előítéletektől
- Szenved hangulati hullámzásoktól
- Zsarolható
- Fenyegethető
- Csábítható el “mézesbödön” csabdába



Az adatmérgezés két fajtája

- Zárt rendszerben (orvosi diagnosztika)
- Nyílt rendszerben (szenzor, crowd-sourcing, IoT)



Deepfake beszivárgott a politikába is

- A videó Barack Obama figuráját mutatja, amint badaságokat beszél. A videó közepén megjelenik a szerző, aki valójában beszél. Mellette Obama “álarca”: <https://www.youtube.com/watch?v=cQ54GDm1eL0>
- Ez a videó az indiai Janata párt funkcionáriusának angol beszédét mutatja. Lejjebb ugyanazzal az arccal egy hindi dialektusban agitál a kormány ellen. A politikus nem is beszéli a dialektust. Bevalottan deepfake: <https://www.theverge.com/2020/2/18/21142782/india-politician-deepfakes-ai-elections>

Az MI több területen jobban teljesít az embernél

- Így a röntgendiagnosztikában és
- A tőzsdei spekulációkban is

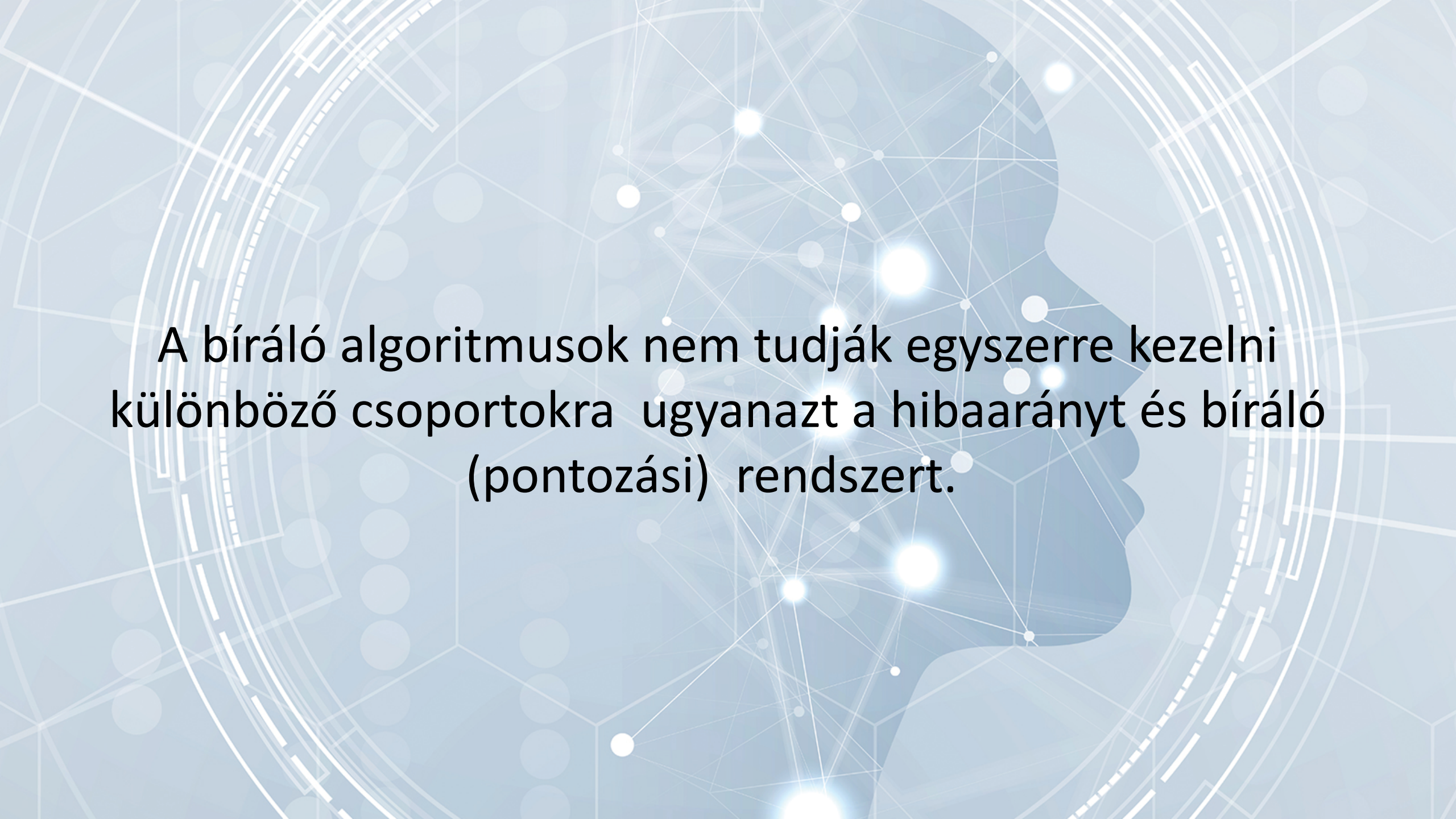




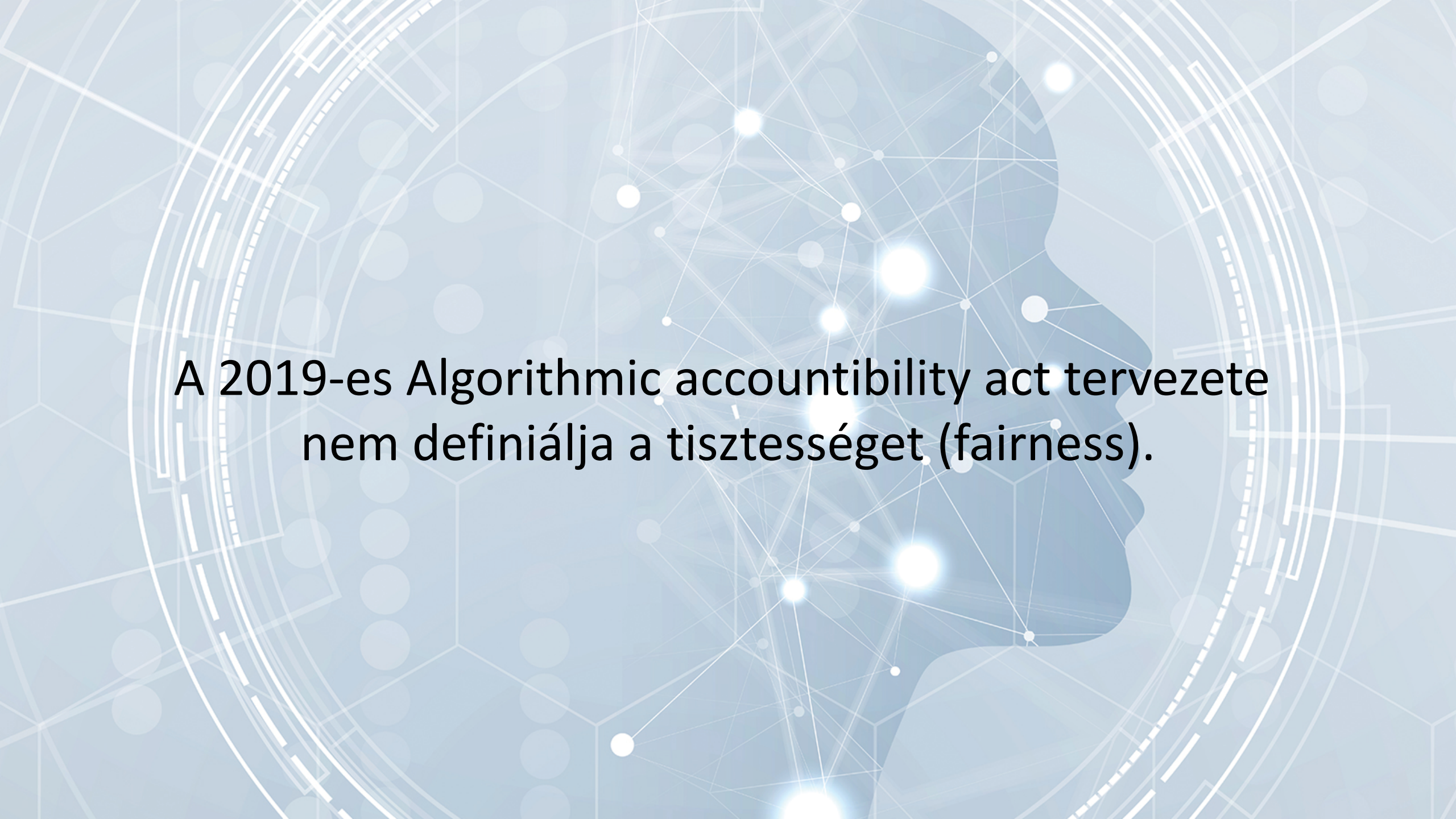
#286819467



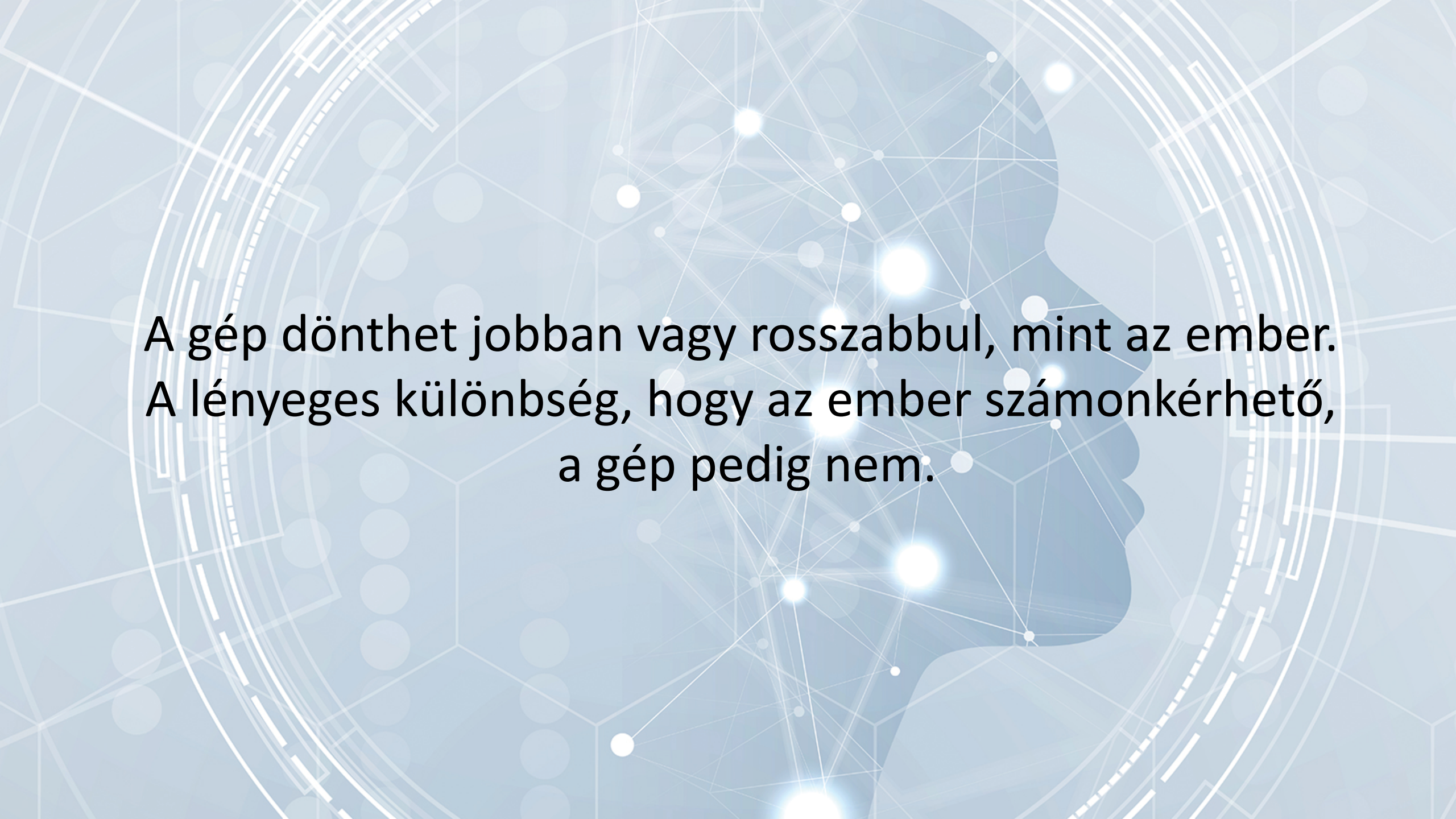
Tud-e egy MI alkalmazás
tisztességes lenni?



A bíráló algoritmusok nem tudják egyszerre kezelni különböző csoportokra ugyanazt a hibaarányt és bíráló (pontozási) rendszert.

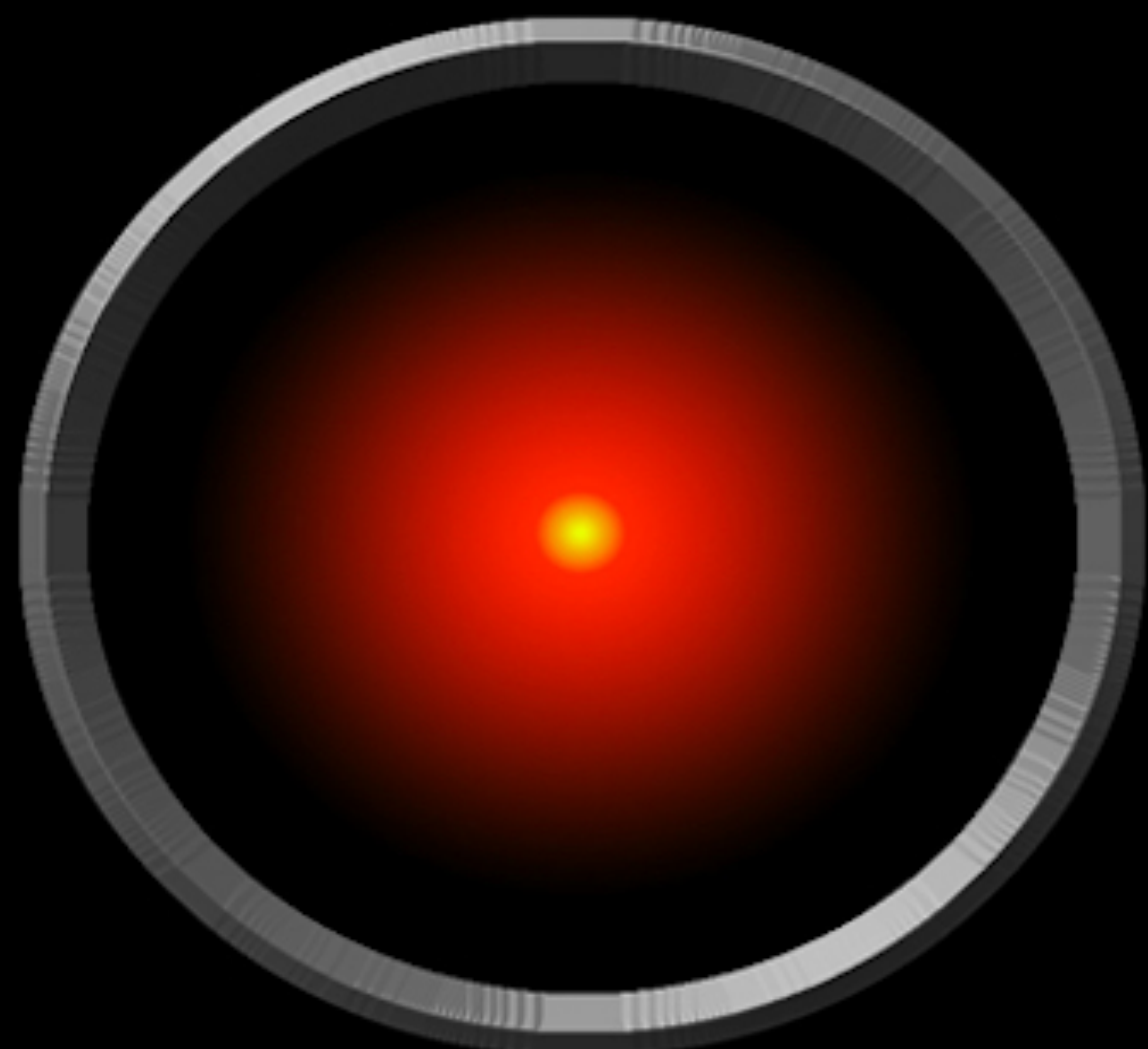


A 2019-es Algorithmic accountability act tervezete
nem definiálja a tisztességet (fairness).



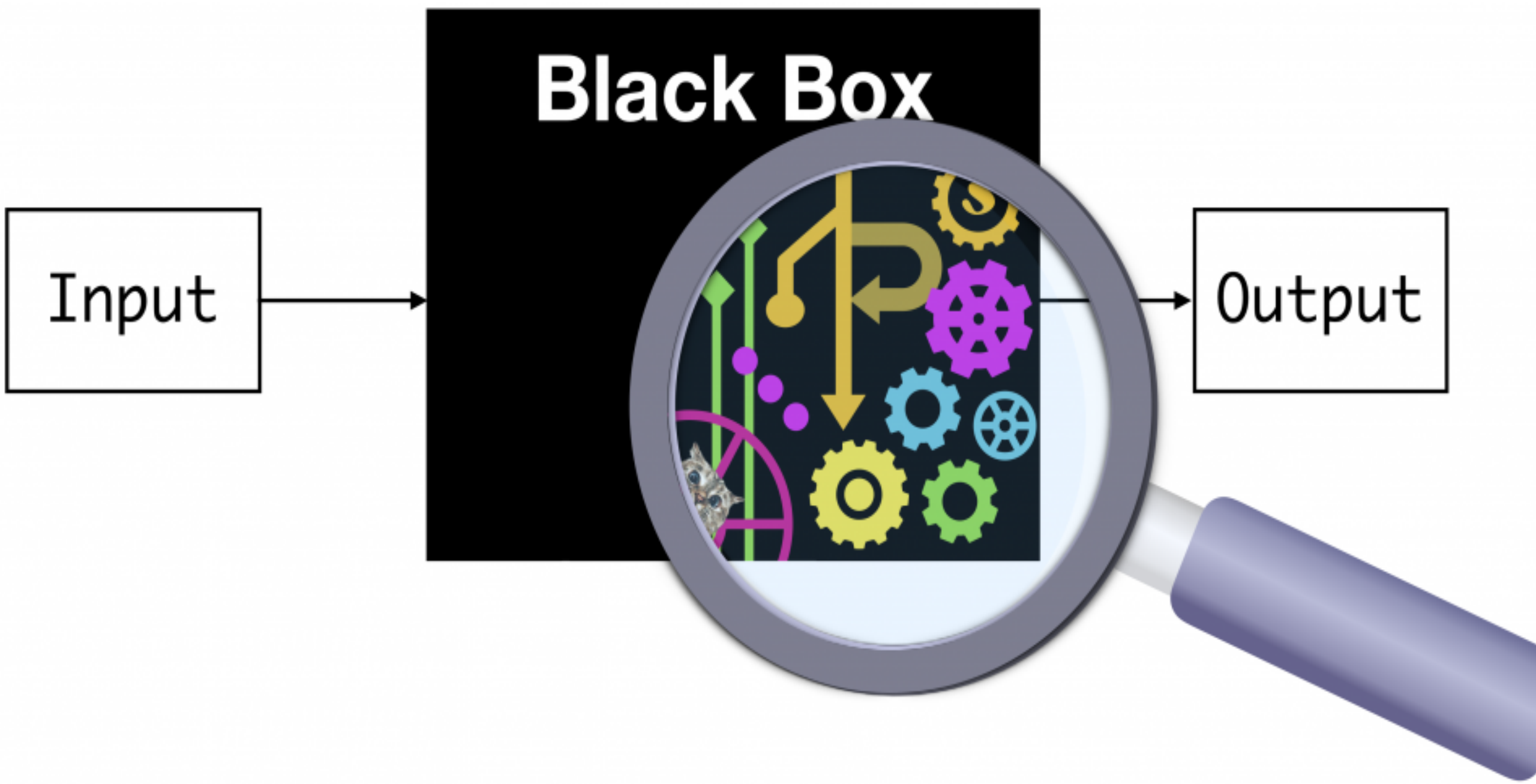
A gép dönthet jobban vagy rosszabbul, mint az ember.
A lényeges különbség, hogy az ember számonkérhető,
a gép pedig nem.

I'm sorry Dave,
I'm afraid I can't do that.



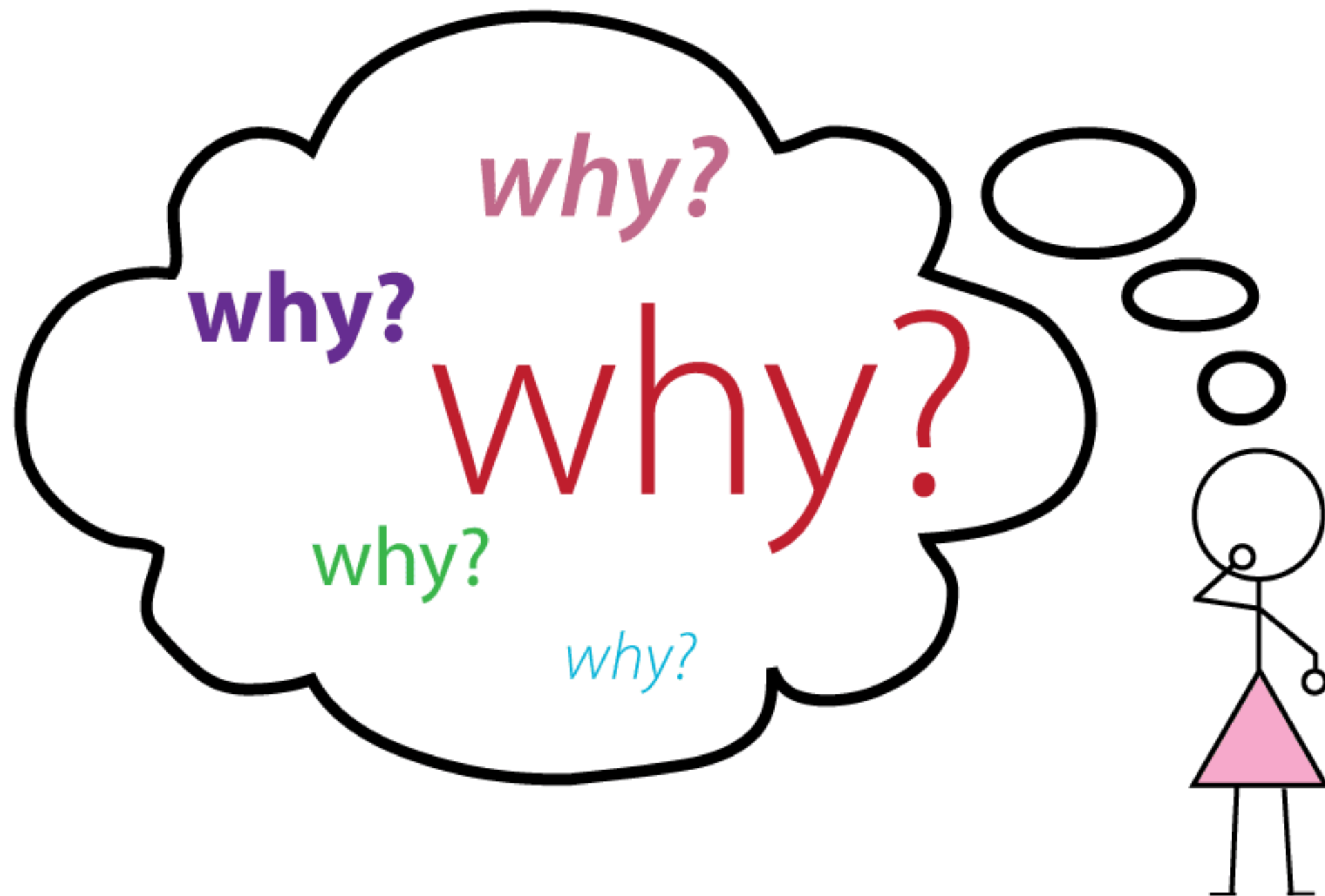


Korrekciós technikák



Nehézségek a feltárásban

- Nagyon sok a szempont (high dimensionality)
- Ipari titok a rendszer belseje
- A gyártók szándékosan is álcázzák a technológiát
- A feketedoboz effektus
- Tájékozottság hiánya
- Rejtett, indirekt diszkrimináció



Technikai jellegű korrekciós módszerek

- Szoftver auditor alkalmazása
- A tanítódadatok szűrése előtesztelés (*preprocessing*), mesterséges javítása
- Tesztelés
- Pozitív diszkrimináció
- Eredmény-módosítás (*postprocessing*)
- Korlátozott optimalizálás (*fairness constraints*)
- Megmagyarázhatóság (*explainability*, XAI)

Szervezeti jellegű korrekciós módszerek

- Szabályozás: pl. eü biztosítás nőknek az EU-ban
- Etikai standardok
- Magatartási kódex
- Szabványosítás
- Minőségbiztosítás
- Számonkérhetőség biztosítása
- Átláthatóság (feketedoboz helyzet)
- Oktatás és kommunikáció
- Legyen jogszerű, de ahol még nincs jog, etikus



...nem kell félnetek...

1 The accelerating pace of change ...



2 ... and exponential growth in computing power ...

Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years

COMPUTER RANKINGS

By calculations per second per \$1,000



Analytical engine
Never fully built, Charles Babbage's invention was designed to solve computational and logical problems



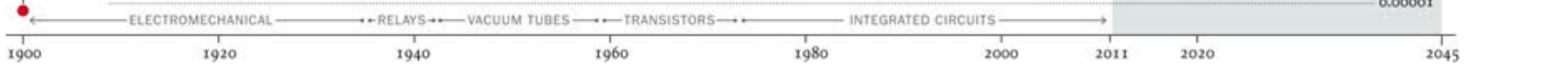
Colossus
The electronic computer, with 1,500 vacuum tubes, helped the British crack German codes during WW II



UNIVAC I
The first commercially marketed computer, used to tabulate the U.S. Census, occupied 943 cu. ft.



Apple II
At a price of \$1,298, the compact machine was one of the first massively popular personal computers



3 ... will lead to the Singularity



Surpasses brainpower of human in 2023



Surpasses brainpower of mouse in 2015



Surpasses brainpower of human in 2023



Surpasses brainpower of mouse in 2015



Power Mac G4
The first personal computer to deliver more than 1 billion floating-point operations per second

Mac Pro

Nvidia Tesla GPU & PC

Dell Dimension 8400

Pentium II PC

Pentium PC

IBM PC

Compaq Deskpro 386

Data General Nova

Intellec-8

DEC PDP-10

IBM 1130

Whirlwind

DEC PDP-4

IBM 1620

Datamatic 1000

EDVAC

IBM SSEC

Zuse 3

BINAC

Zuse 2

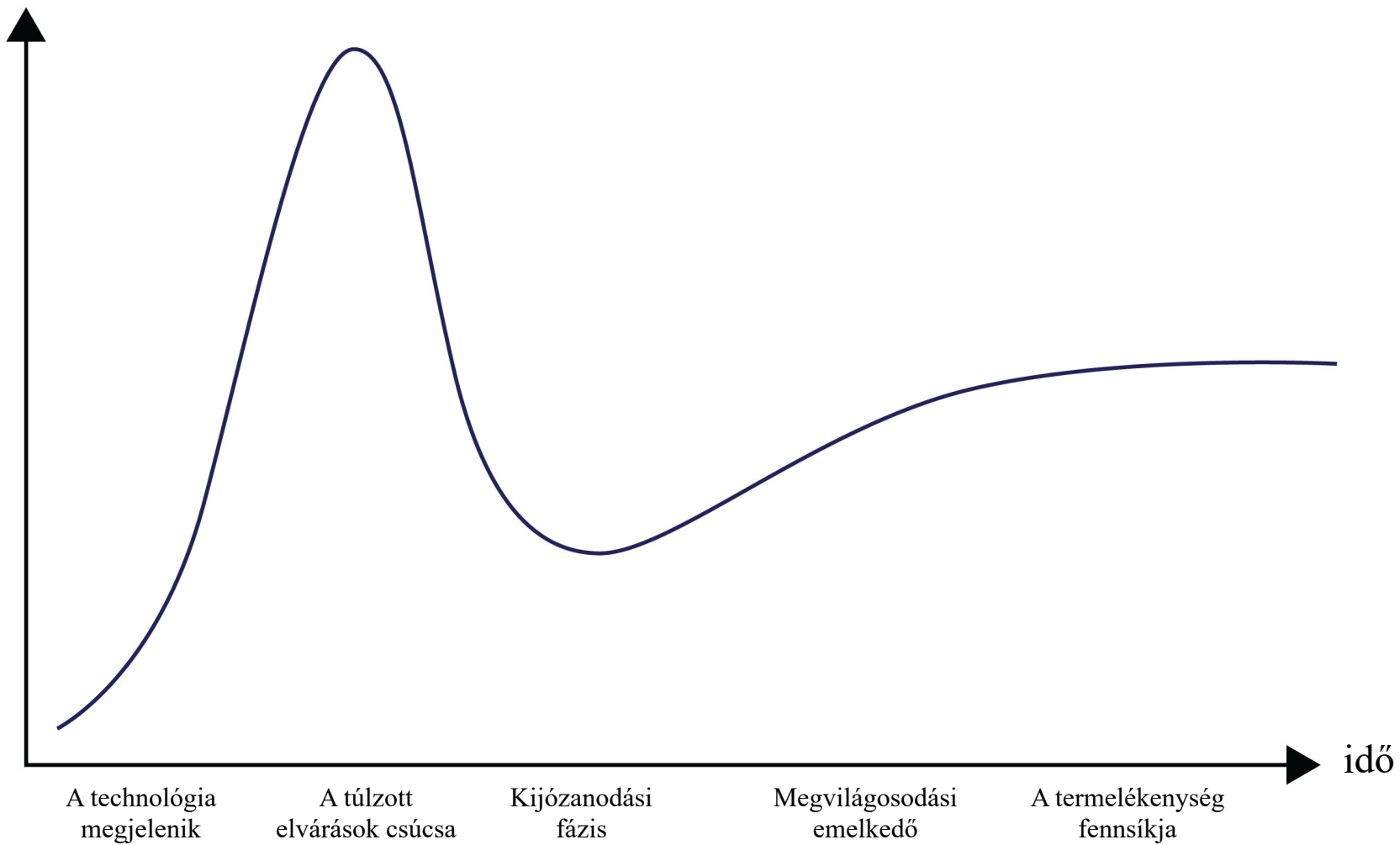
ENIAC

National Ellis 3000

IBM Tabulator

Hollerith Tabulator

A technológia
evolúciós fázisa



GDPR

(b) (6)
to: (b) (6)

Fri, Oct 18, 2019 at 12:09 PM

great - thanks.

On Fri, Oct 18, 2019 at 11:52 AM (b) (6) wrote:

Sure. Here's an extract from our research memo, which I think says this well. I put the whole GDPR paragraph in below, but I think the highlighted portion is the key bit for the footnote - although leave it to you how to reference a footnote in a footnote.

The European Union's AI strategy is nested within its broader emphasis on data privacy, which is most clearly manifested through the GDPR. The GDPR places restrictions on the ability of firms to collect and share personal data without consent, and provides individuals the right to revoke that consent at any given time. This privacy-first approach to data collection and sharing stands in stark contrast with the United States and Japan, which are advocating for more free flows of data. There is evidence that the GDPR has negatively impacted the competitiveness of the EU's tech industry writ large; a recent paper from the National Bureau of Economic Research found that after the rollout of GDPR the number of venture capital deals in the tech industry within the EU declined by 26.1% relative to US-based firms, and the average monetary value of those deals declined by 33.8%.^[1] GDPR could prove to be a significant obstacle in any efforts to standardize privacy regulations, which would be a key part of any international data sharing regime.

[1] Jia, J, G Z Jin and L Wagman (2018), "The short-run effects of GDPR on technology venture investment," NBER, Working Paper 25248.

An issue in Europe is that we're not front-runners on the development of AI," says Bendert Zevenbergen, a former technology policy advisor in the European Parliament and now a researcher at Princeton University.

"We're kind of recipients of AI technology in many ways. We're definitely the second tier. The first tier is the US and China."

Az MI kiépülésének feltétele

- Nagy piac
- Homogén piac
- Nagyon sok homogén típusú adat
- Bizalom VAGY kényszer
- A fentiek legalábbb egyike teljesül az USÁ-ban és Kínában,
- az EU-ban pedig nem

Javaslatok

- Az alapjogok korlátozását újra kellene gondolni
- Még az elején járunk, jó lenne más hibájából tanulni
- Bizalmat (mint a repülésben, nukleáris energiatermelésben, közbiztonságban vagy élelmiszerbiztonságban) kell teremteni
- A jogászképzésbe (alap és tovább) be kellene iktatni a modern jogi informatikát

