



MI és mi - bizalmi válság egy friss kapcsolatban

Hollósi Gergely, BME TMIT

Hollósi Gergely

BME VIK, TMIT

hollosi.gergely@vik.bme.hu



**BME
TMIT**

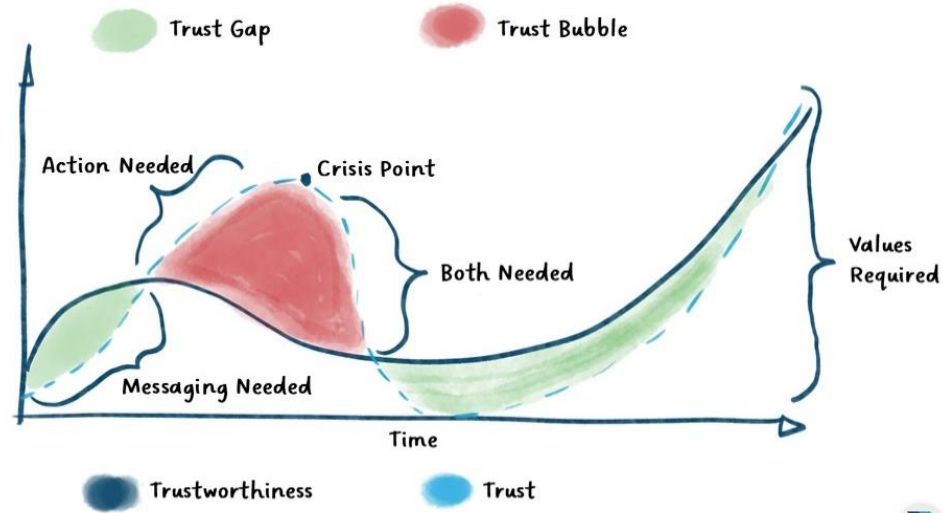
I. A randi - és a *rózsaszín kód*

Izgalom, kíváncsiság, csodálat.

II. A bizalmi válság

A Typical Trust Cycle:

The Trustworthiness Reversion Principle in Action



PROBABLY
APPROXIMATELY
CORRECT

Nature's Algorithms for Learning and
Prospering in a Complex World



LESLIE VALIANT

III. A párterapeuta - Miért várunk többet egy géptől, mint egymástól?

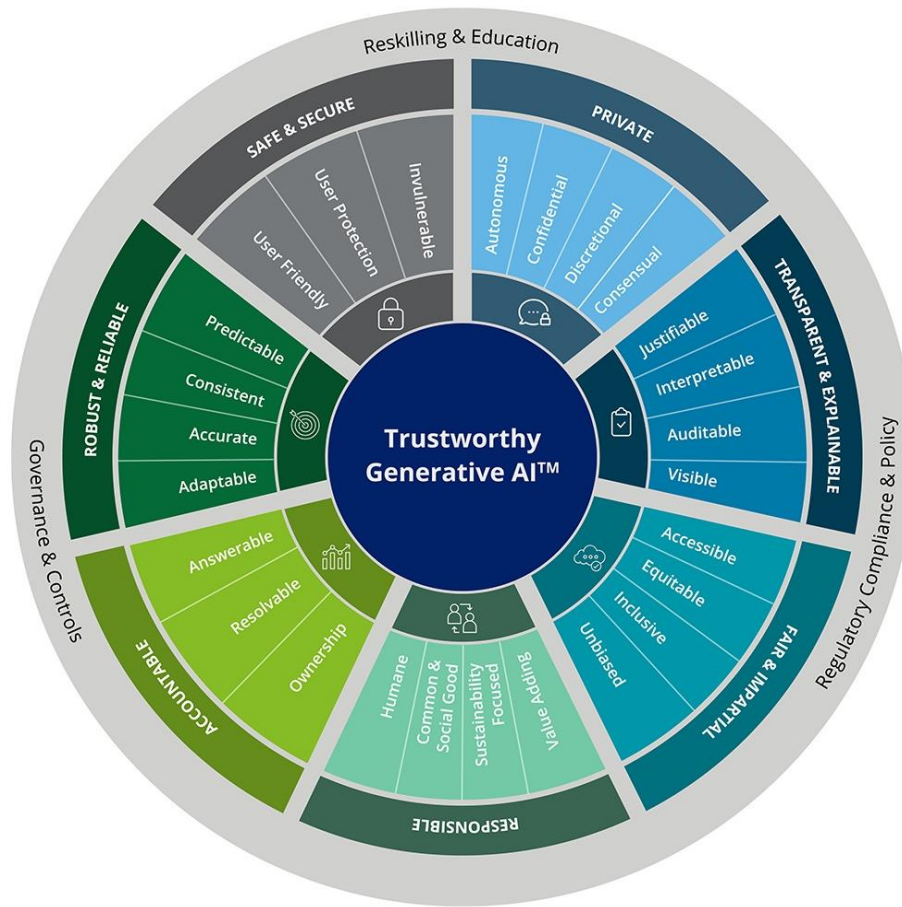
*"Much of everyday human decision making appears to [...] be based on a competent ability to predict from past observation without any good articulation of **how** the prediction is made or any claim of fundamental understanding of the phenomenon in question. The predictions **need not be perfect** or the best possible. They need merely to be **useful enough**."*

IV. Párkapcsolati gyakorlatok

- © Mit tegyünk a bizalomért?
 - © Mire van szükségünk?
 - © Mik a bizalmi *metrikák*?
-
- © **AI Trustworthiness**



Elszámoltatható
Robusztus és megbízható
Biztonságos
Bizalmas
Átlátható és magyarázható
Igazságos és pártatlan
Felelősségteljes
Elszámoltatható



Ártalmas lehet...

- emberekre
- szervezetekre
- ökoszisztémára

Forrás: Deloitte



A bizalom nem monolitikus –
a rendszer minden elemében
bízunk kell.

A bizalmi rétegek (trust stack)





- Platform sokszor alapvető jogi akadály
- Titkosítás, identitás menedzsment
- Modell verziókövetés, modell menedzsment
- Monitorozás, drift menedzsment, SLA
- Felhő – miben is különbözik?
- *"Gyenge lábakon nem lehet megbízhatóságot építeni."*



- Ágensek képességeinek megértése
- Tényszerűség: RAG, tudásgráfok, nyilvános források
- Eszköz korlátok: API whitelist, sandbox
- Be/kimenet korlátok: guardrails
- Ágens kooperáció (tanuljunk kritikus rendszerek üzemeltetéséből - emberek is hibáznak)
- *"Ágensek hatalommal bírnak - korlátozzuk és ellenőrizzük."*



- Human-in-the-loop
- Áttekintő dashboard-ok, élő metrikák (adatforrások, érvelési láncok, stb.)
- Visszacsatolás (tesztelői/felhasználói "jelek" visszacsatolása)
- **Oktatás és továbbképzés**

• *"A bizalom akkor nő, ha a mesterséges intelligencia működése átlátható és befolyásolható."*



- Kockázatkezelés, jogszabályok: NIST AI RMF, EU AI Act, ISO/IEC 42001.
- Auditok
- Bízunk egyáltalán a jogszabályokban?
- Felhasználói érzékenyítés, tudásbővítés (oktatás!)

• *"Minden kapcsolat korlátokra és közös szabályokra épül."*

V. A közös jövő

- ◎ "Lehet bízni az MI-ben?" → "Milyen feltételekkel és milyen keretek közt építjük a bizalmat?"
- ◎ Az MI **felelős és átlátható**, az ember pedig **tájékozott és tudatos**.
- ◎ Szakmai szempontból:
 - Megbízható szabályozási rendszerek
 - Explicit, domén specifikus korlátok beépítése (kritikus rendszerekben)
 - Monitorozás és megérthetőség
 - Szakmai képzések és oktatás

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles, and others with blue dots.

**Köszönöm a
figyelmet!**



BME
TMI

