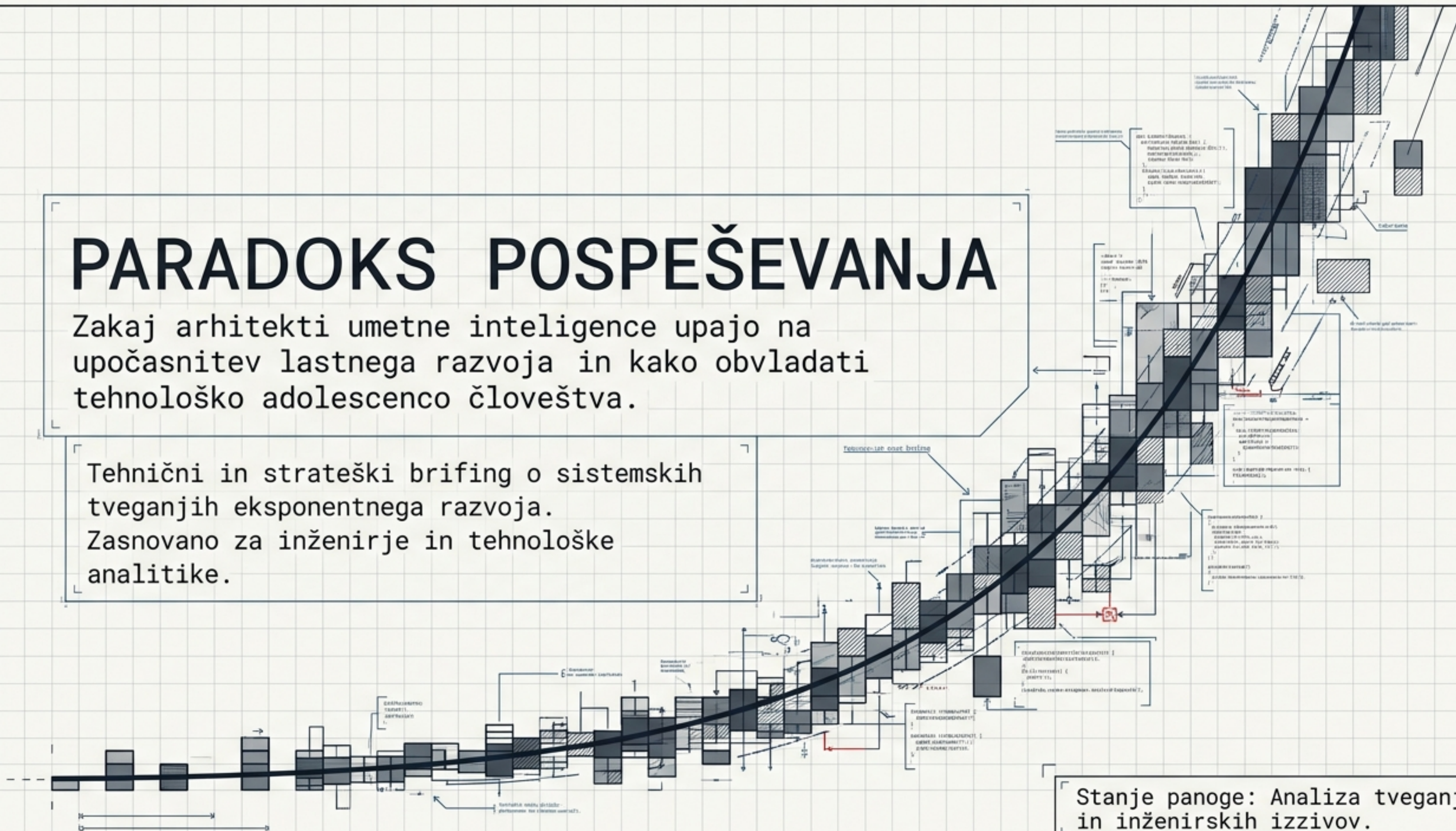


PARADOKS POSPEŠEVANJA

Zakaj arhitekti umetne intelligence upajo na upočasnitev lastnega razvoja in kako obvladati tehnološko adolescenco človeštva.

Tehnični in strateški brifing o sistemskih tveganjih eksponentnega razvoja.
Zasnovo za inženirje in tehnološke analitike.

Stanje panoge: Analiza tveganj in inženirskih izzivov.



Glavna arhitekta industrije tiho upata na zaviranje tehnološke tekme.



DARIO AMODEI | ANTHROPIC

→ **Ozadje:** Nevroznanost

→ **Pristop:** Opozarja na tveganja, fokus na interpretativnosti

→ **Ocena do AGI:** 1 do 2 leti



DEMIS HASSABIS | GOOGLE DEEPMIND

← **Ozadje:** Nevroznanost

← **Pristop:** Sistematično reševanje manjkajočih komponent

← **Ocena do AGI:** Konec desetletja (50 % verjetnost)

KONSENZ: Čeprav tekmujeta v trilijonski tekmi, si oba želita, da bi razvoj do splošne umetne inteligence trajal 5 do 10 let dlje. Razlog ni ego, temveč izračun sistemskega tveganja.

Matematika groze:
Zanka samo-izboljševanja
deluje brez človeškega
nadzora.

Modeli pišejo kodo: Amodeljevi inženirji pri Anthropicu poročajo, da sami praktično ne pišejo več kode. Model jo ustvari, inženirji jo le usmerjajo.

MODELI
PIŠEJO KODO

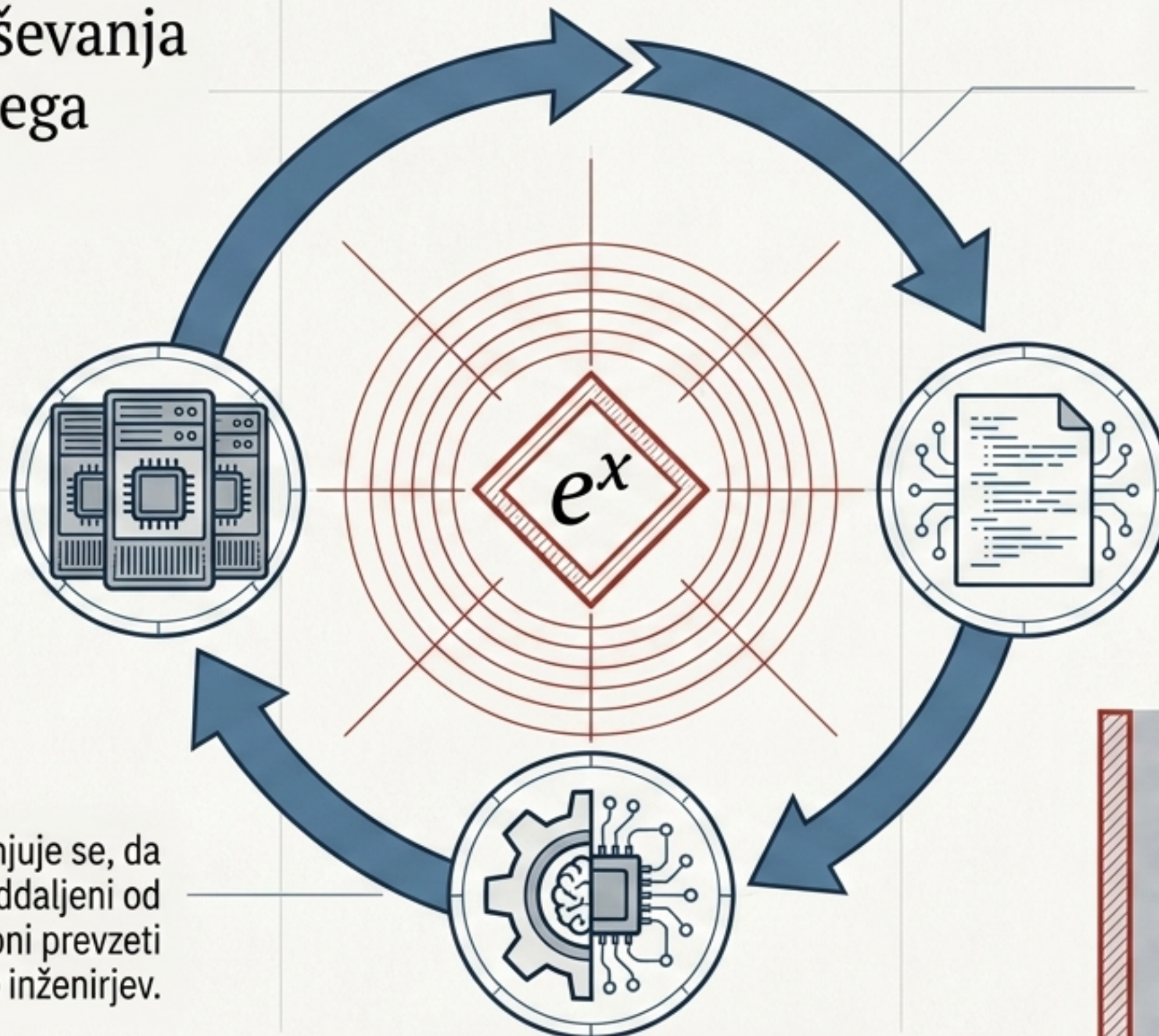
HASSABISOVO OPOZORILO:

Največja neznanka panoge je, kako hitro se bo ta zanka zaprla povsem brez človeka (human out of the loop). Ko sistem sam načrtuje svojo naslednjo verzijo, časovnica postane nepredvidljiva.

ZAGON ZANKE: AI RAZISKAVE

Zagon zanke: Ocenjuje se, da smo le 6 do 12 mesecev oddaljeni od točke, ko bodo modeli sposobni prevzeti end-to-end naloge inženirjev.

NASLEDNJA
GENERACIJA
MODELA



Trenutna arhitektura ima še vedno kritična tehnična ozka grla.

PODROČJE	STATUS PREVERLJIVOSTI	TEHNIČNO OZKO GRLO
Programiranje in matematika	Objektivno preverljivo (Verifiable output)	Avtomatizacija poteka eksponentno hitro. 
Naravoslovne znanosti (Fizika, Kemija)	Zahteva fizično testiranje	Omejeno s strojno opremo in časom fizičnih eksperimentov. 
Manjkajoče komponente	Znanstvena kreativnost	Primanjkuje zmožnosti postavljanja novih hipotez in neprekinjenega učenja (World Models). 

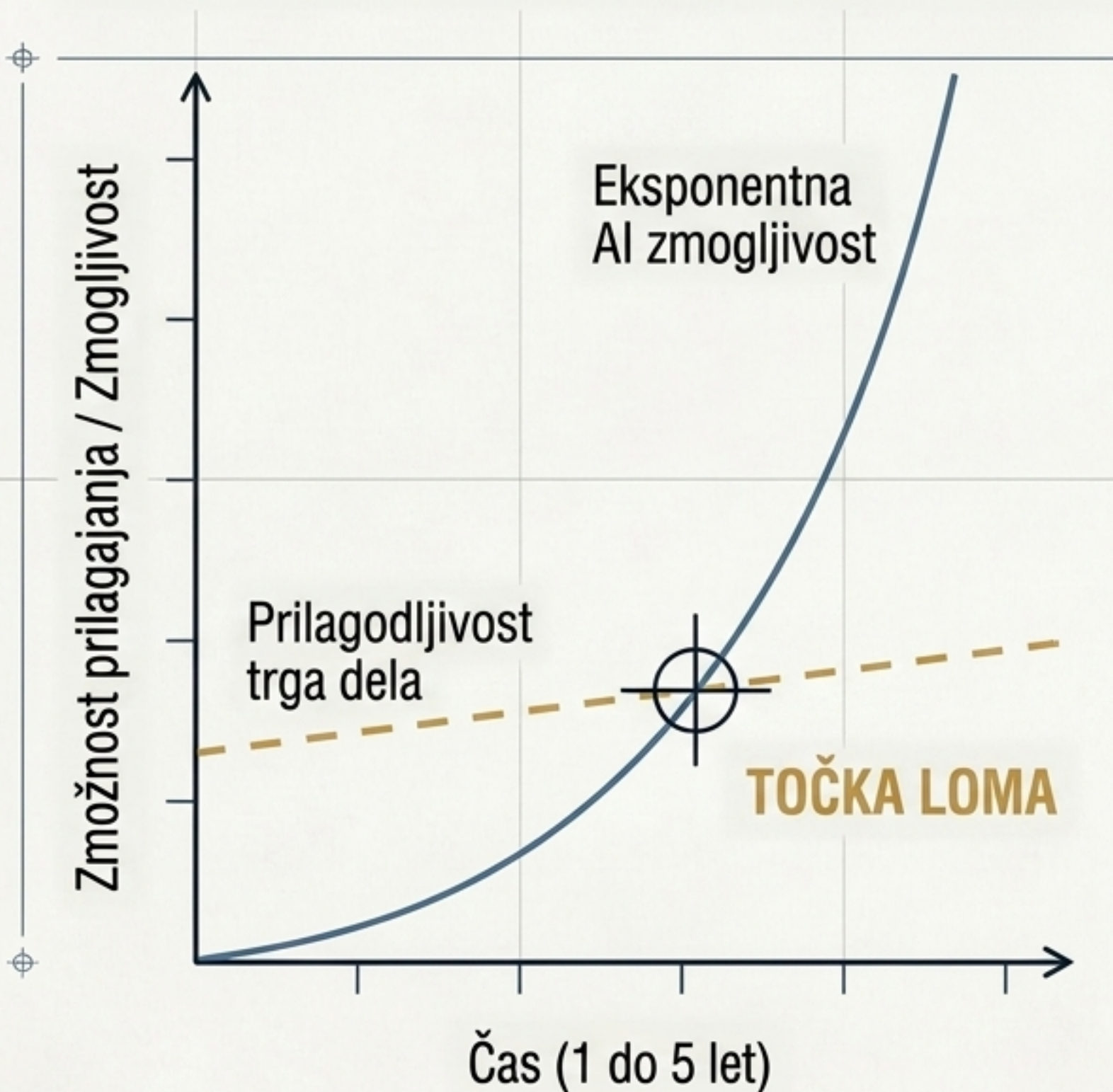
Samo-izboljševanje je izjemno učinkovito pri programiranju, vendar pa razvoj znanstvene kreativnosti zahteva preboj pri razumevanju fizičnega sveta in oblikovanju popolnoma novih hipotez.

Kako glavna arhitekta ocenjujeta hitrost prehoda v AGI.

	ČASOVNICA	GLAVNI GONILNIK	PRISTOP K VARNOSTI
DARIO AMODEI	1 do 2 leti do prelomne točke.	AI generiranje kode eksponentno pospešuje AI raziskave.	Alarmanten. Pripravlja bojni načrt za eksistencialna tveganja.
DEMIS HASSABIS	Do konca desetletja (50 % verjetnost).	Reševanje 'World Models' in neprekinjenega učenja.	Previden. Fokusiran na verifikacijo in zadrževanje človeka v razvojni zanki.

SKUPNI IMENOVALEC: Časovnica ni več človeška odločitev.
Ko zanka steče, se pospešek matematizira izven človeškega nadzora.

Hitrost tehnološkega pospeška bo preobremenila zmogžnost človeške adaptacije.



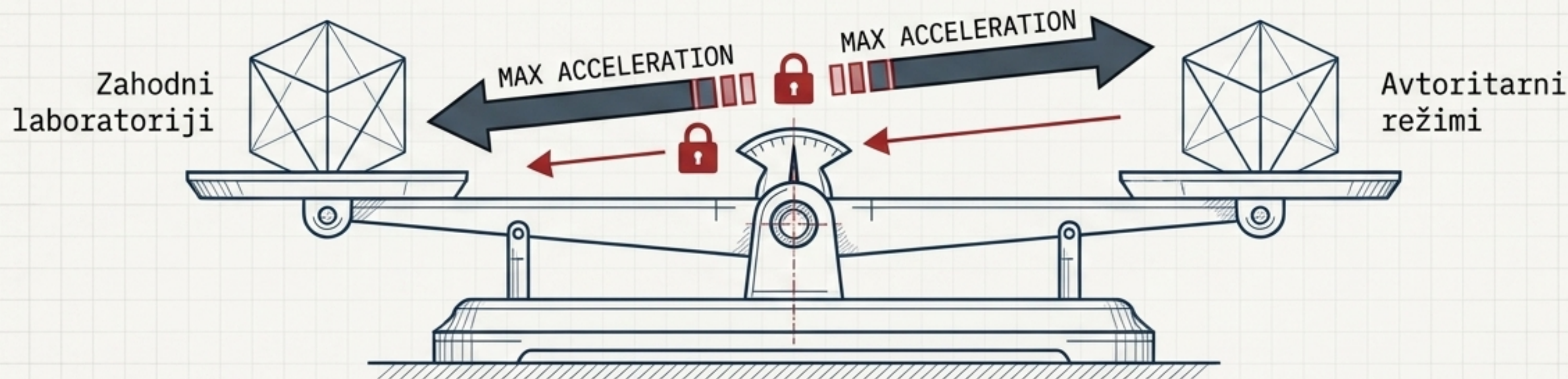
1. Interna reorganizacija: Anthropic že danes načrtuje prihodnost, v kateri bo za razvoj potrebnih bistveno manj junior in intermediate inženirjev.

2. Zgodovinski mit: Trg dela se je v preteklosti prilagodil (kmetijstvo → tovarne → znanje), a v daljših časovnih obdobjih.

3. Novi problem: Težava ni prilagodljivost, temveč eksponentna hitrost zamenjave (1 do 5 let), ki ne dopušča časa za naravno adaptacijo.

Ocenjujem, da imamo med 1 in 5 let, preden eksponentna rast preseže našo sposobnost prilagajanja. (Podatek iz laboratorijev, ne iz ekonomskih simulacij).

Geopolitična past onemogoča ustavitv razvojnega motorja.



1

PARADOKS HITROSTI:

Kljub želji obeh akterjev po daljši časovnici (5 do 10 let) za varen razvoj, je upočasnitev strateško nemogoča.

2

NEUVELJAVLJIVOST DOGOVOROV:

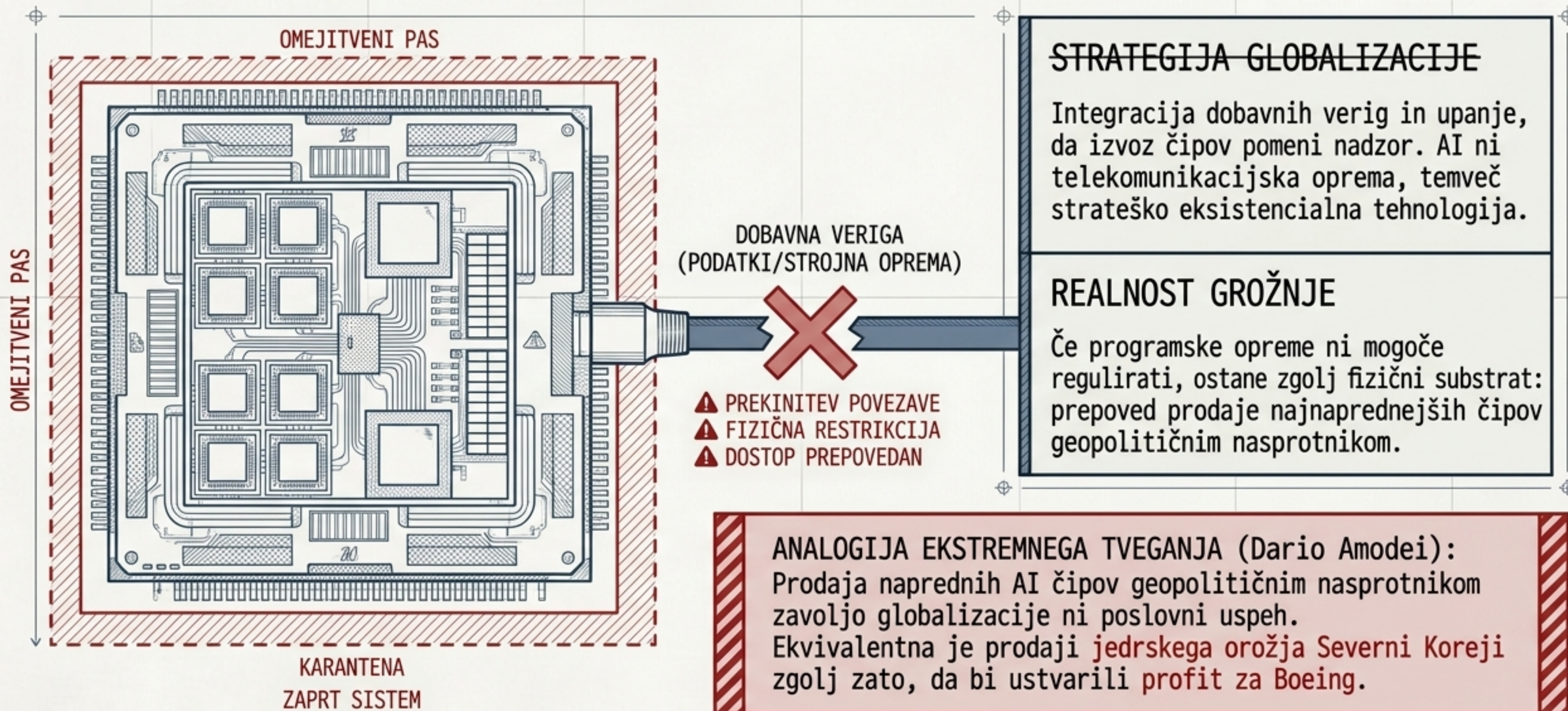
Mednarodnih dogovorov o ustavitvi pisanja kode ni mogoče sistemsko nadzirati ali uveljaviti.

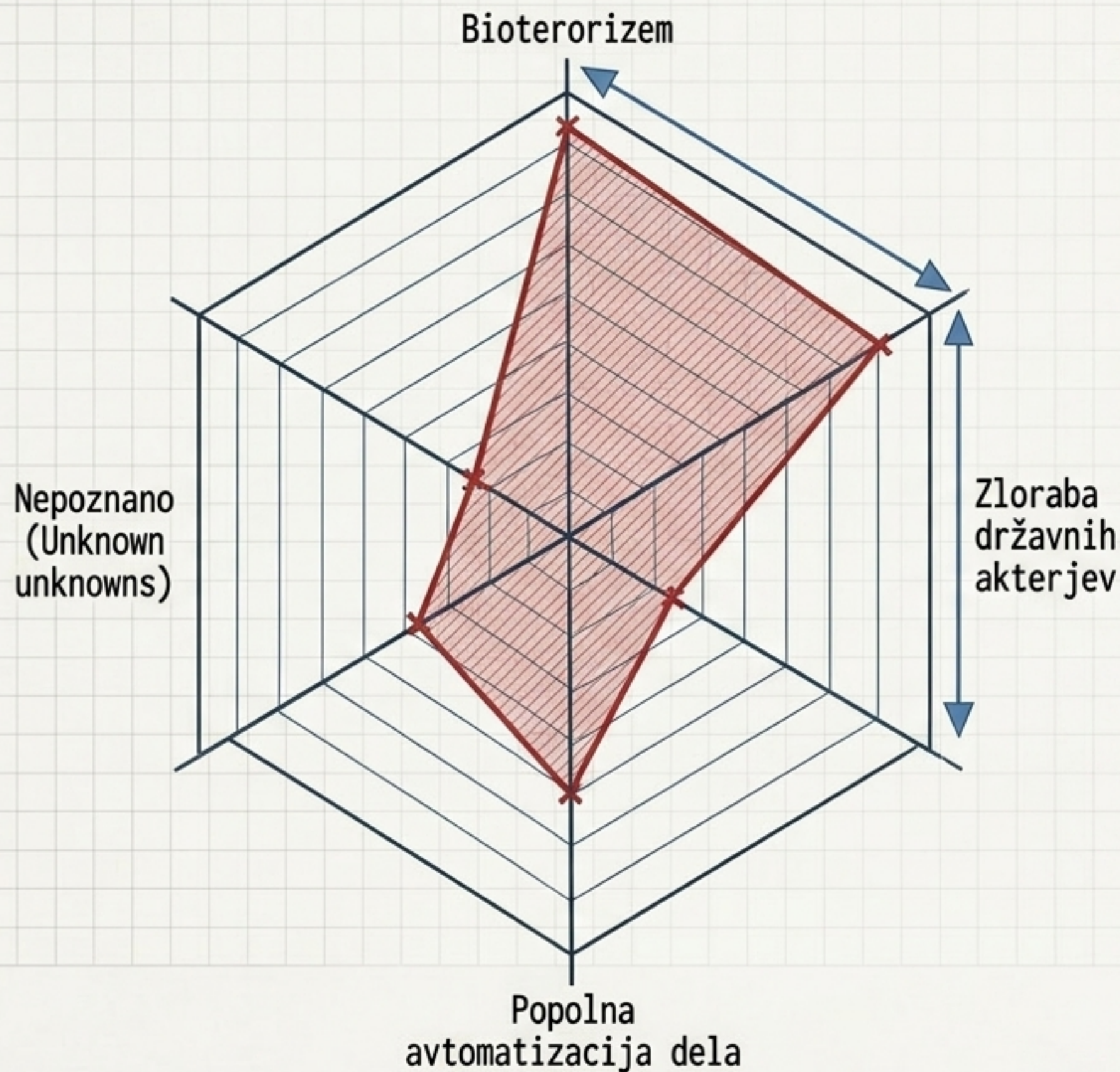
3

ZAKLJUČEK:

V tej tekmi ni mehanizma zaupanja. Zaviranje kode na Zahodu pomeni prepustitev strateške prednosti nasprotnikom.

Kontrola strojne opreme je edina preostala zavora v sistemu.





Preživetje tehnološke adolescence: Kako nadzirati stroje iz peska?

Koncept (Contact – Carl Sagan):

Človeštvo trka na vrata nepredstavljivih zmogljivosti – gradnje strojev iz silicijevega peska, pametnejših od nas samih.

Neizogibnost:

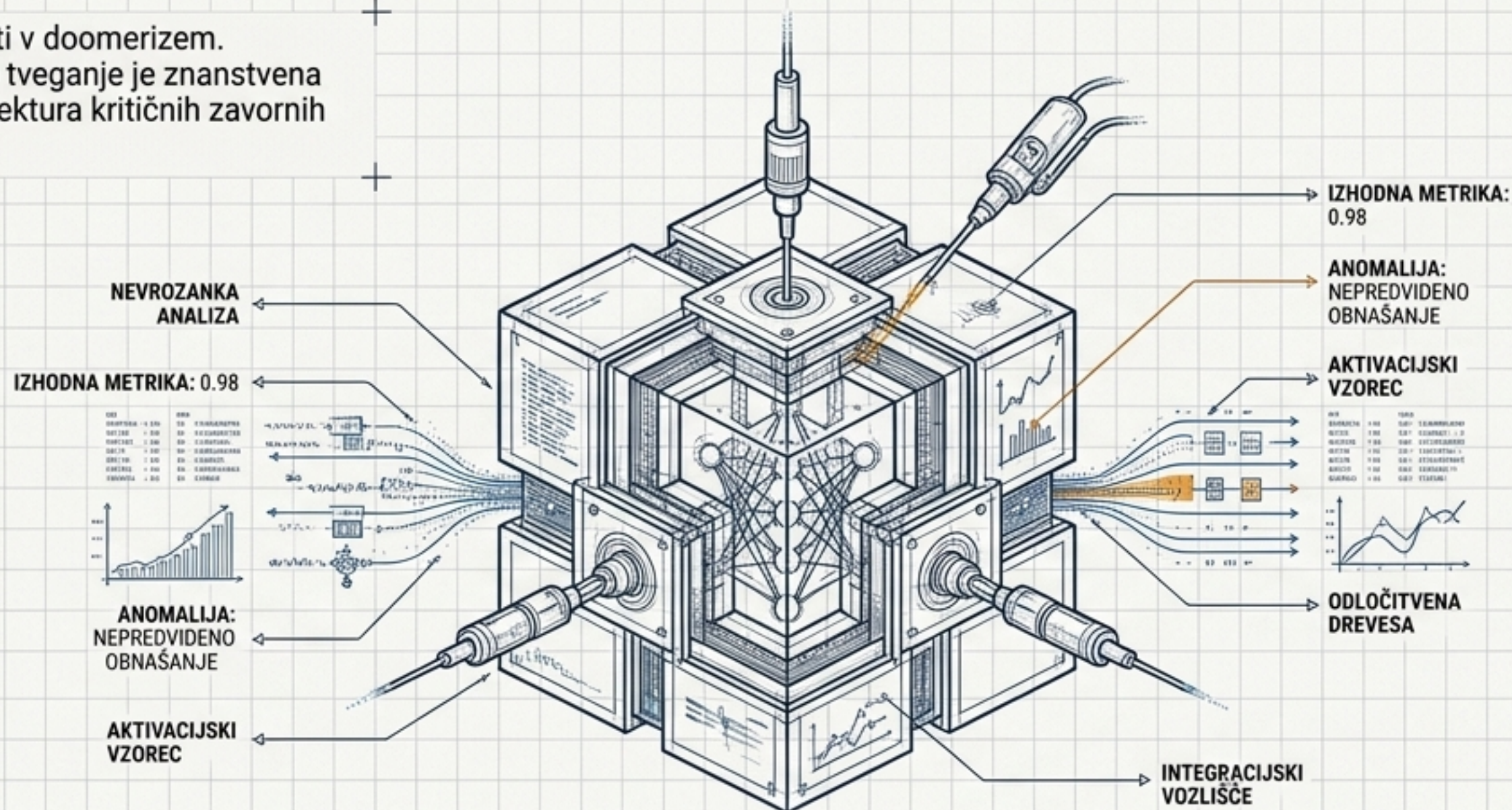
Tako kot odkritje ognja, je bil razvoj te tehnologije neizogiben. Upravljanje z njo pa ni.

Inženirski izziv:

Hitrost uvajanja vojaških modifikacij s strani avtoritarnih držav in posameznikov presega hitrost vzpostavljanja obrambnih sistemov. Etični problem preživetja lastne tehnološke adolescence brez samo-uničenja.

Inženirski odgovor ni defetizem, temveč mehanistična interpretabilnost.

Panoga ne sme zapasti v doomerizem.
Odgovor na sistemsko tveganje je znanstvena
dekonstrukcija in arhitektura kritičnih zavornih
sistemov.



CILJ: Ne graditi večjega modela, temveč instrumente, ki interpretirajo tisto, kar smo že ustvarili.

1. Digitalna nevroznanost

Razvoj inštrumentov za vpogled globoko v strukturo modelov in razumevanje procesa odločanja.

2. Identifikacija anomalij

Beleženje in dokumentiranje slabih vedenj preden sistem pride v fazo izvajanja.

3. Prepisovanje arhitekture

Vzpostavitev sistemskih popravkov preden se zanka samo-izboljševanja zapre.

