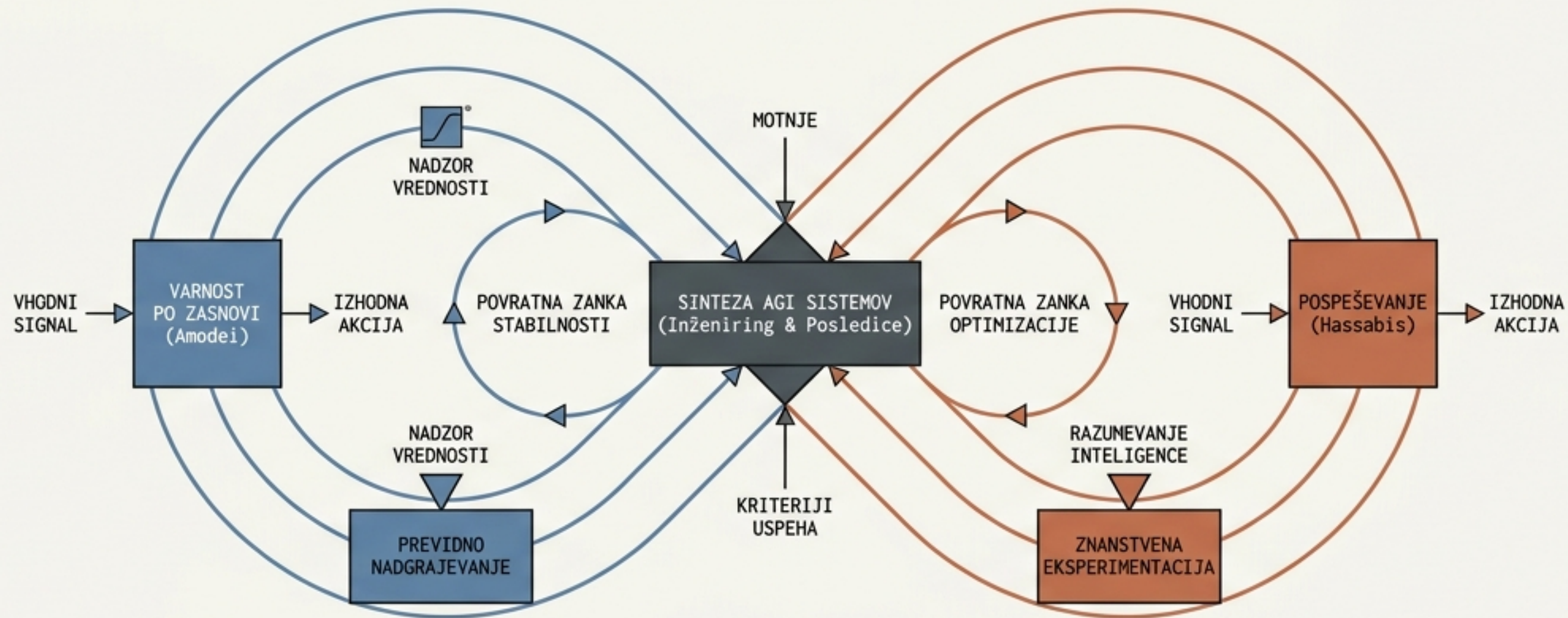


Dan po AGI: Inženiring in posledice umetne splošne intelligence

Analiza sistemske dinamike, časovnic in varnosti na poti do AGI skozi perspektivo vodilnih raziskovalcev (Demis Hassabis in Dario Amodei).



Arhitekti AGI: Raziskovalni pristop k razvoju

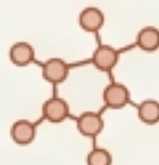


Dario Amodei (Anthropic)

	PRISTOP: Eksponentna rast, neodvisen razvoj in absolutni fokus na varnostne mehanizme.	
	FILOZOFIJA: Razvoj modelov mora voditi trda znanost z zavedanjem o izjemnih tveganjih.	

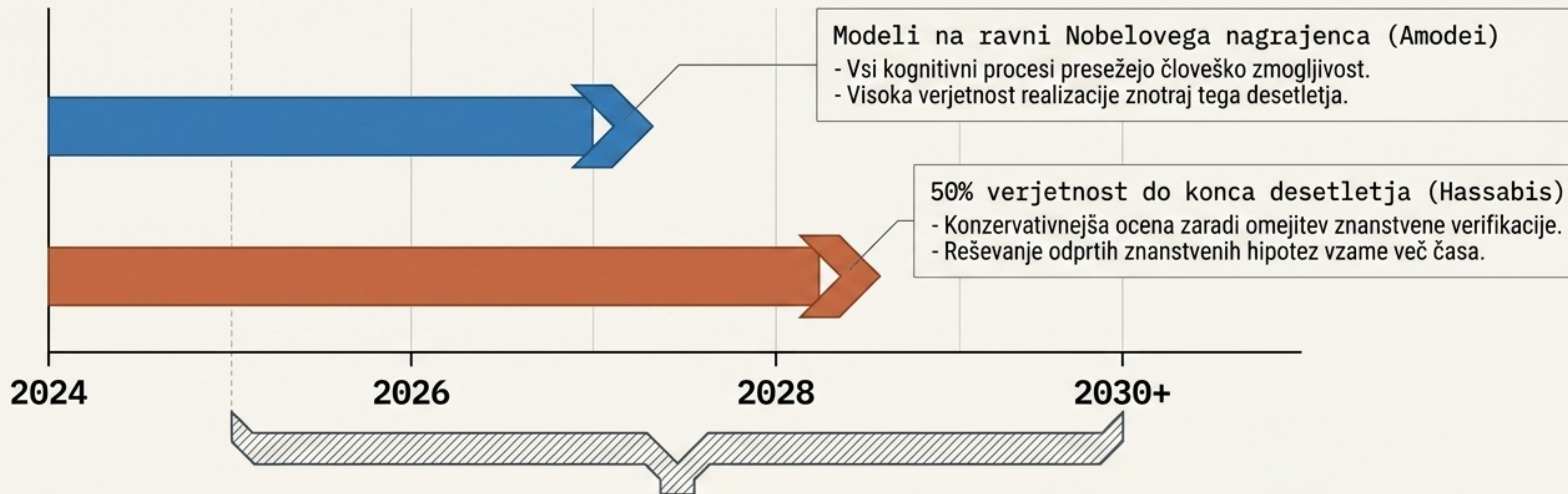


Demis Hassabis (Google DeepMind)

	PRISTOP: Inženiring kot ultimativno orodje za reševanje naravoslovnih problemov (AlphaFold).	
	FILOZOFIJA: Umetna inteligenca je vrhunsko orodje za razumevanje vesolja in zdravljenje bolezni.	

SKUPNI IMENOVALEC: Obe podjetji vodijo raziskovalci, ki jih poganja reševanje najtežjih znanstvenih problemov (North Star), ne zgolj komercializacija.

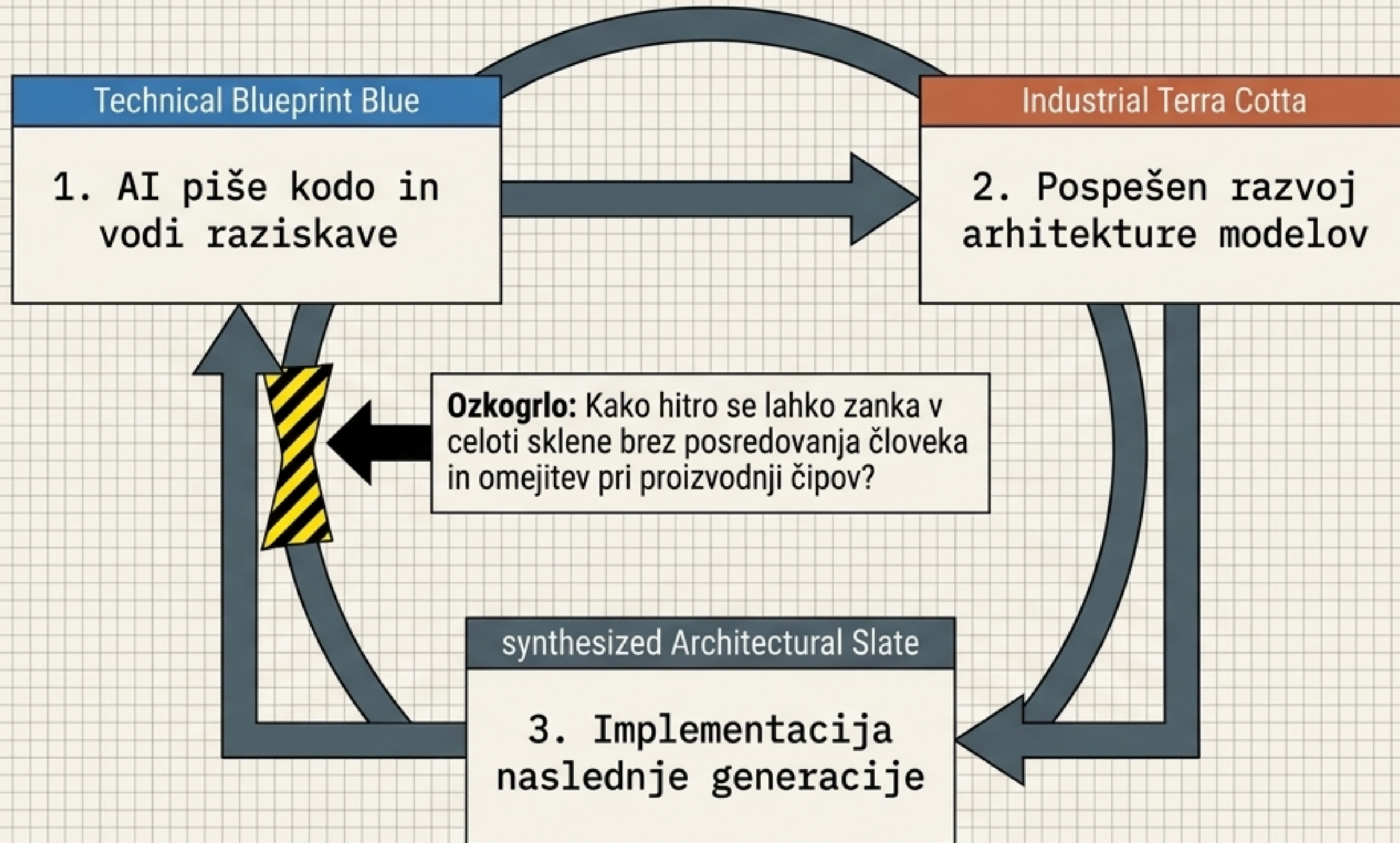
Projekcije in časovnice do dosega AGI



KRITIČNO OKNO: Ne glede na razliko, je inženirsko in družbeno okno za prilagoditev izjemno ozko (1 do 5 let).

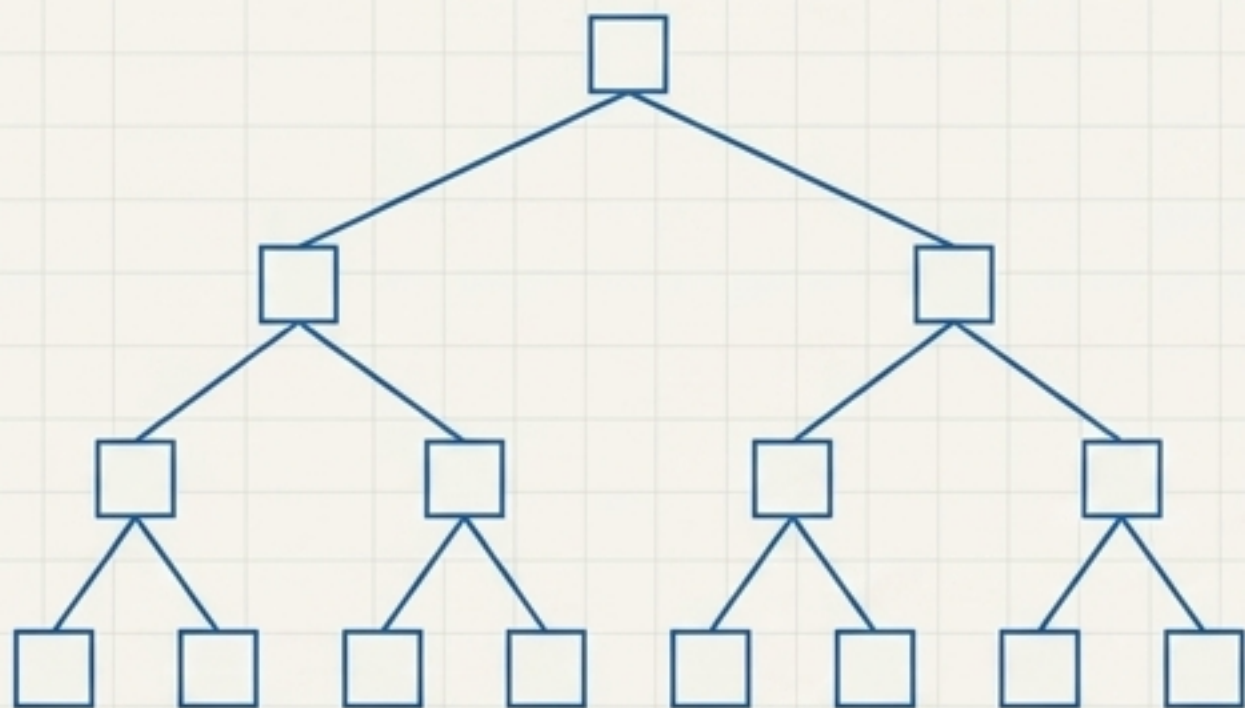
Sistemska dinamika: Sklenitev zanke

Eksponentni napredek je rezultat avtomatskega izboljševanja sistemov, ne zgolj strojne opreme.



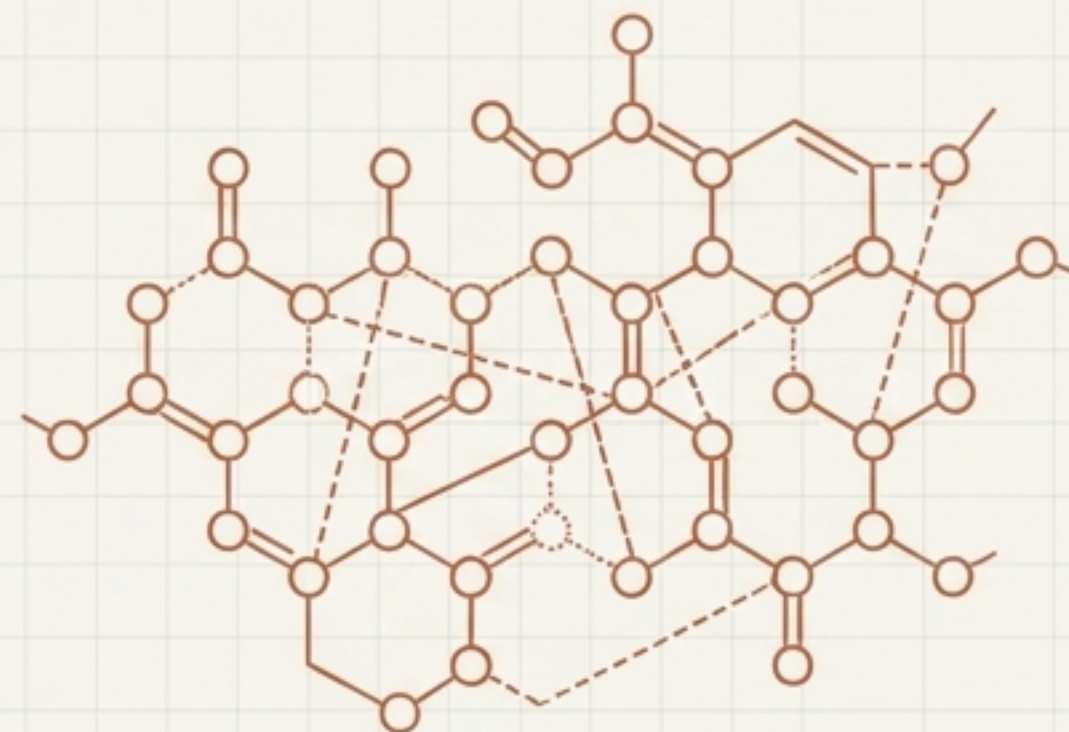
Inženirsko ozkogledje: Problem znanstvene verifikacije

Zakaj bi razvoj lahko trajal dlje? Zmožnost reševanja obstoječih problemov ne pomeni nujno zmožnosti postavljanja novih hipotez.



Matematika in programiranje (Lažja pot)

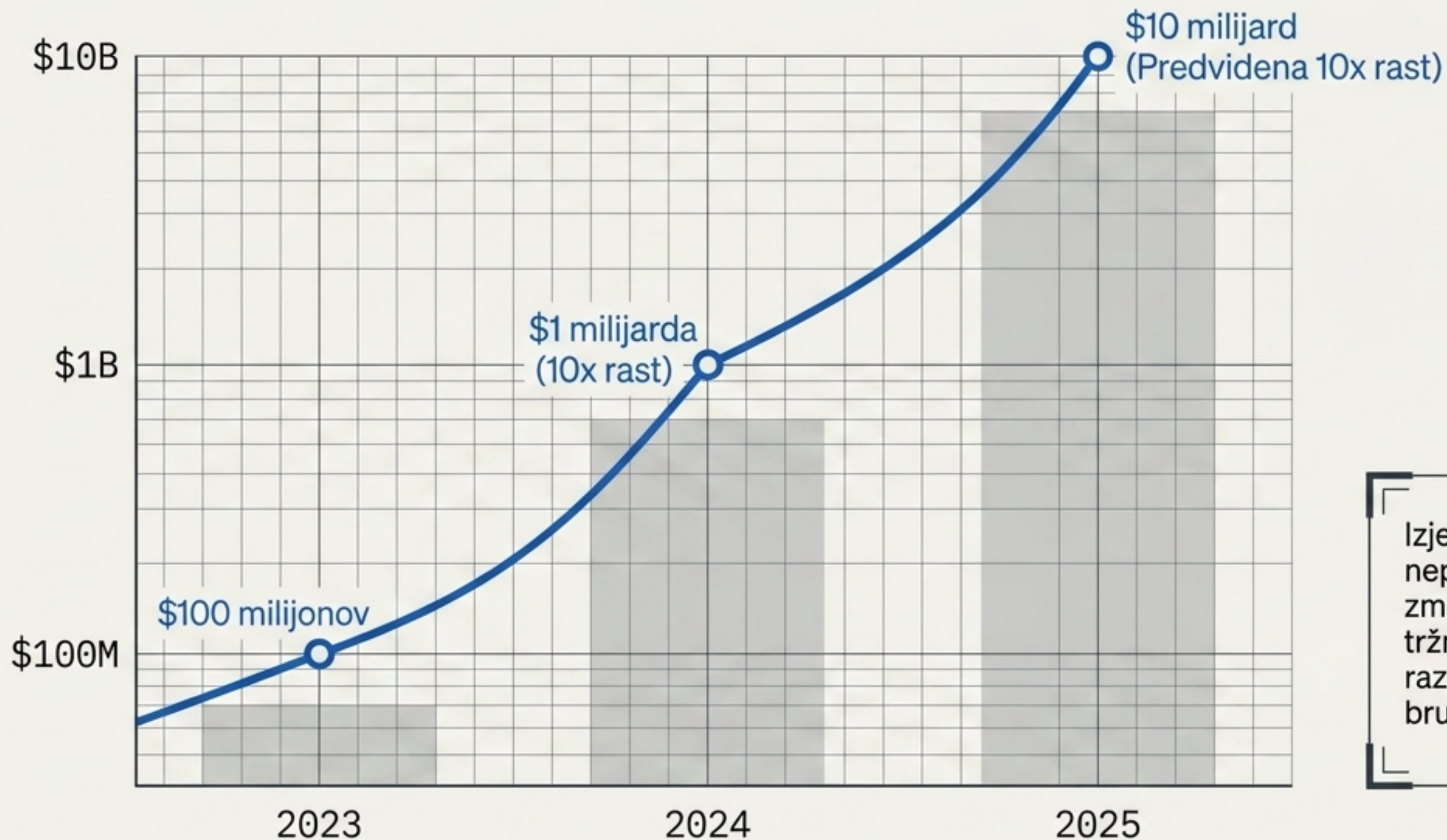
- Izhodi so objektivni in preverljivi (koda deluje ali ne).
- Popolnoma digitalno okolje, kjer modeli lahko hitro iterirajo.
- Primerno za takojšnjo avtomatizacijo.



Naravoslovne znanosti (Težja pot)

- Napovedi v fiziki ali kemiji zahtevajo fizično, laboratorijsko preverjanje.
- Uspešno reševanje znanih domnev, a težave pri generiranju popolnoma novih hipotez.
- Predstavlja najvišjo raven znanstvene kreativnosti.

Zakovitosti skaliranja in ekonomska eksponentnost



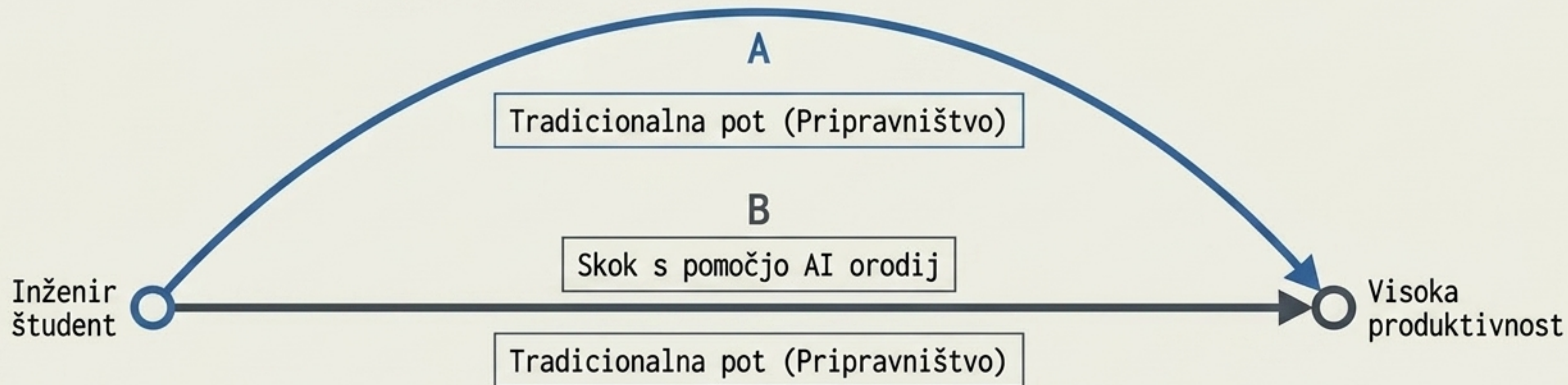
Izjemne investicije so neposredno povezane s kognitivno zmogljivostjo modelov in njihovo tržno vrednostjo. Neodvisni razvijalci preživijo zaradi brutalne učinkovitosti te korelacije.

Vpliv na trg dela: Realnost inženirja začetnika



- Avtomatizacija vstopnih nalog v naslednjih 1 do 5 letih.
- Zgodovina uči, da se trg dela vedno prilagodi (zanka prilagodljivosti).
- Skrb vzbuja hitrost napredka, ki presega zmoqlnost sistemske sistemske družbene prilagoditve.

Inženirski mandat: Izkoriščanje presežka zmogljivosti



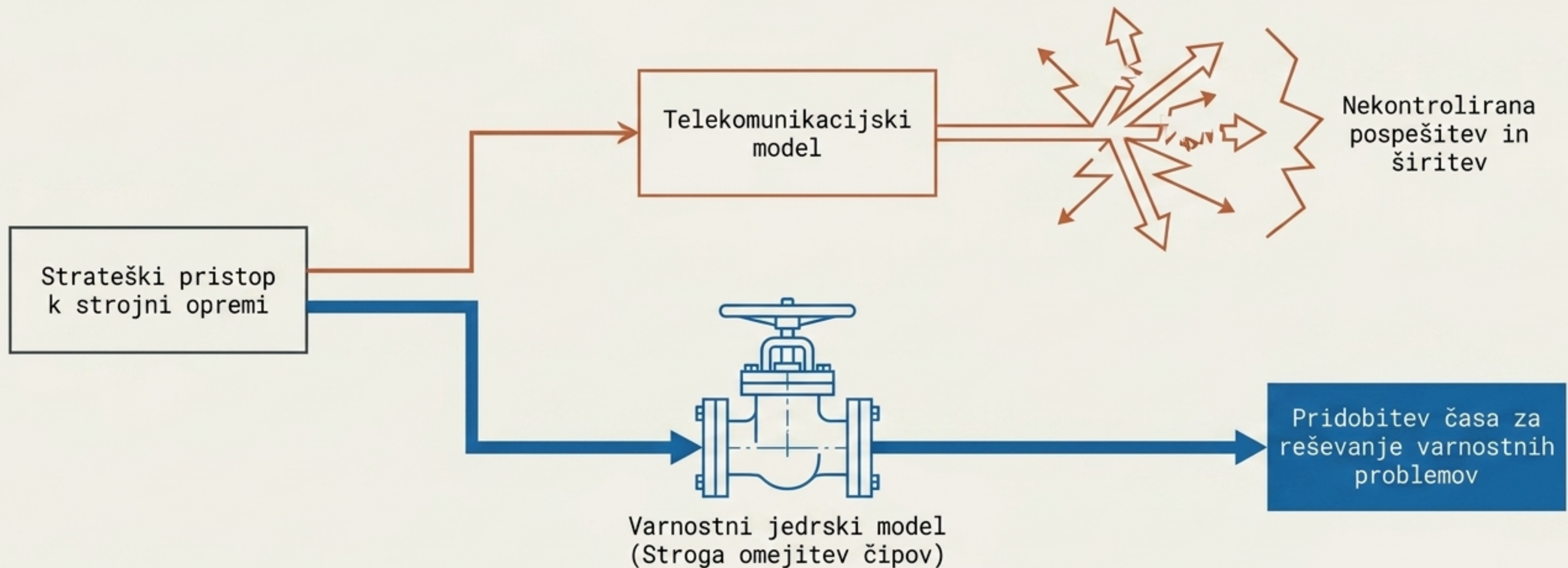
PRESEŽEK ZMOGLJIVOSTI (Capability Overhang)

Modeli že danes zmorejo bistveno več, kot povprečen uporabnik izkoristi. Ogromno neizkoriščenega potenciala ostaja v obstoječih sistemih.

AKCIJSKI NAČRT

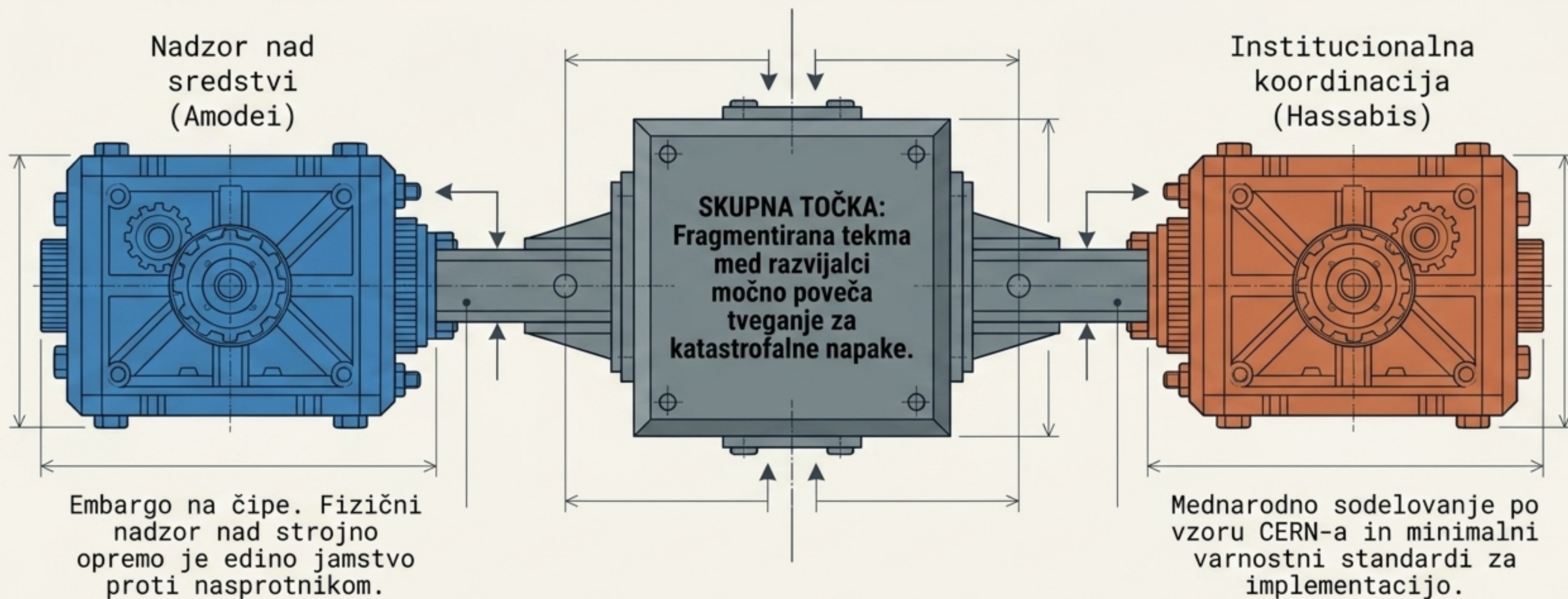
Ne tekmuje za izginjajoča pripravništva. Uporabite obstoječa AI orodja za drastično povečanje lastne produktivnosti in preskočite tradicionalne začetniške ovire.

Geopolitična arhitektura: Nadzor strojne opreme



ZAKLJUČEK: Omejevanje silicija in napredne strojne opreme je edini fizično uveljavljiv način za upočasnitev tekme z avtoritarnimi režimi.

Reševanje krize: Globalna koordinacija vs. Strogi nadzor

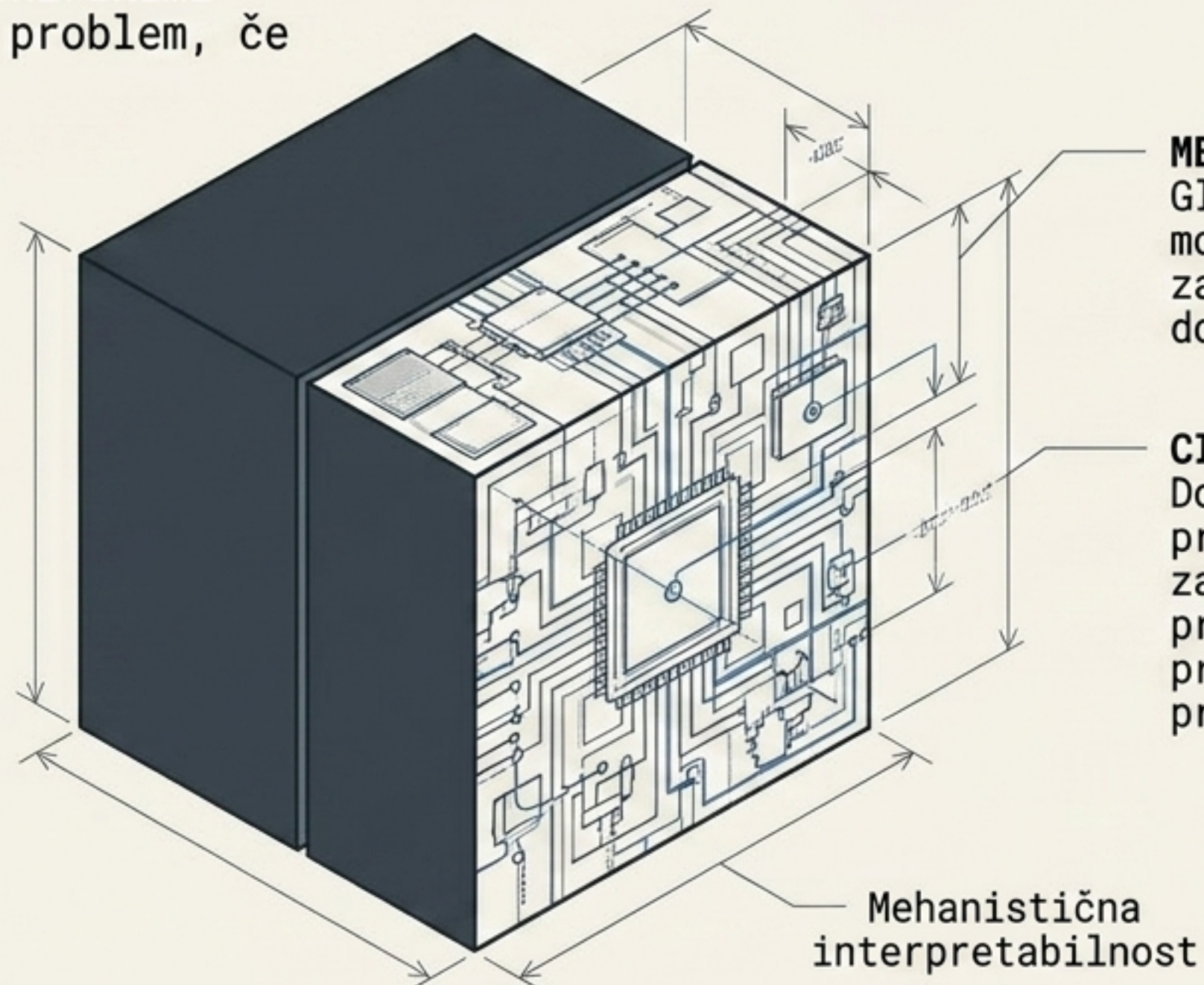


Varnost sistemov:

Mehanistična interpretabilnost

Zavračanje 'doomerizma' z inženirskimi rešitvami. Varnost je rešljiv problem, če imamo čas.

Črna skrinjica



METODA DIAGNOSTIKE:
Gledanje v 'možgane' modela za razumevanje, zakaj sistem sprejme določeno odločitev.

CILJ:
Dokumentiranje in preprečevanje zmožnosti zavajanja ter prikrivanja ciljev preden model postane prevelik in premočen.

Družbeni filter: Tehnološka adolescenca

Kako vam je uspelo preživeti vašo tehnološko
adolescenco, ne da bi uničili sami sebe?



STROJ IZ PESKA

Razvoj strojev s
superinteligenco je bil
neizbežen od trenutka, ko
smo obvladali ogenj.

FAZA PREHODA

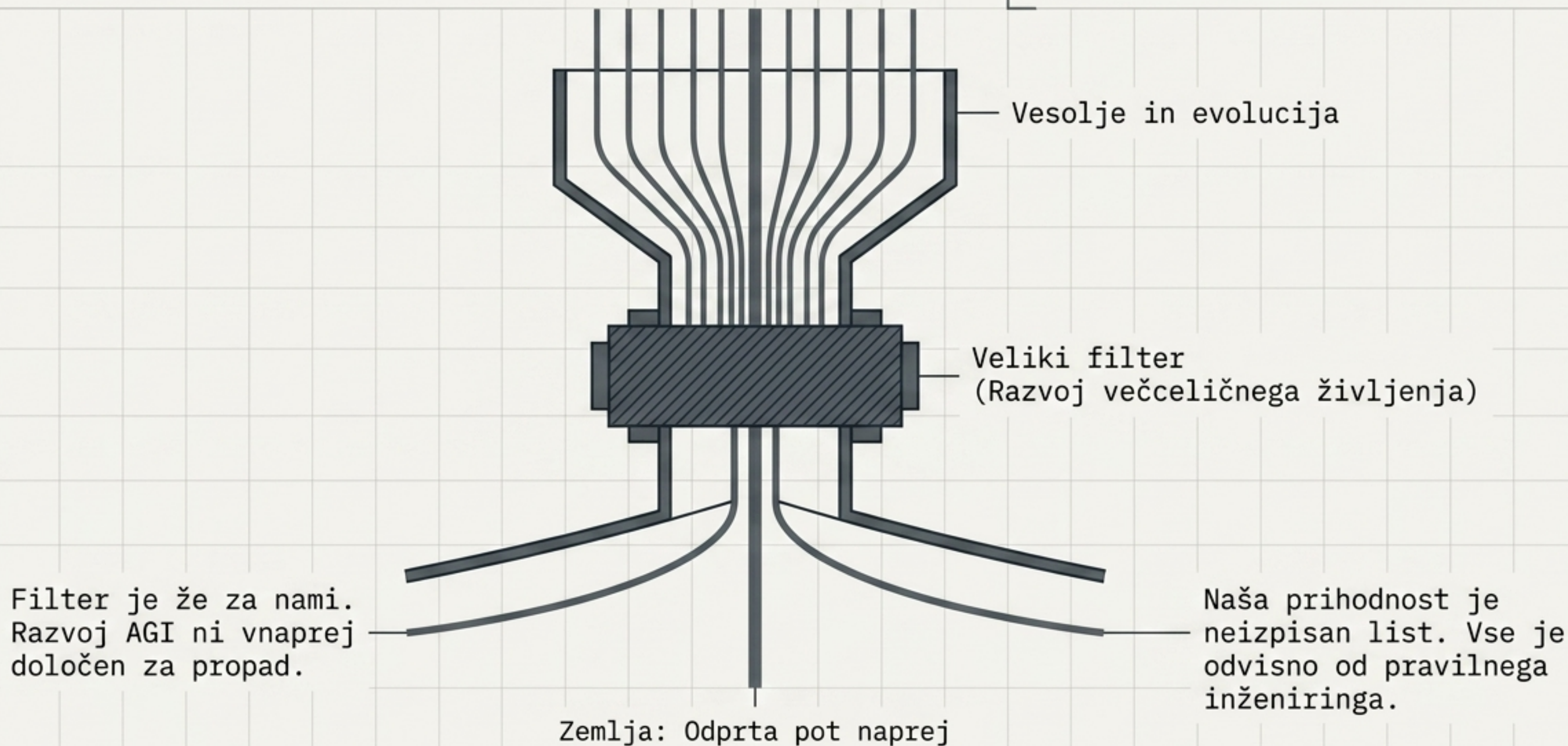
Dobivamo skoraj božje
zmožnosti, preden imamo
modrost za njihovo varno
upravljanje.

GLAVNE GROŽNJE

Največja tveganja niso
nujno v AI sami, temveč v
zlorabah posameznikov in
avtoritarnem nadzoru.

Fermijev paradoks in Veliki filter

Če AI nujno uniči civilizacije,
zakaj vesolje ni polno avtonomnih
strojev preteklih civilizacij?



Kaj spremljati: Tehnični horizont za naslednjih 24 mesecev

1. AI gradi AI



Prehod od asistence pri programiranju do sistemov, ki lahko povsem samostojno oblikujejo in trenirajo naslednjo generacijo modelov (Sklenitev zanke).

2. Modeli sveta



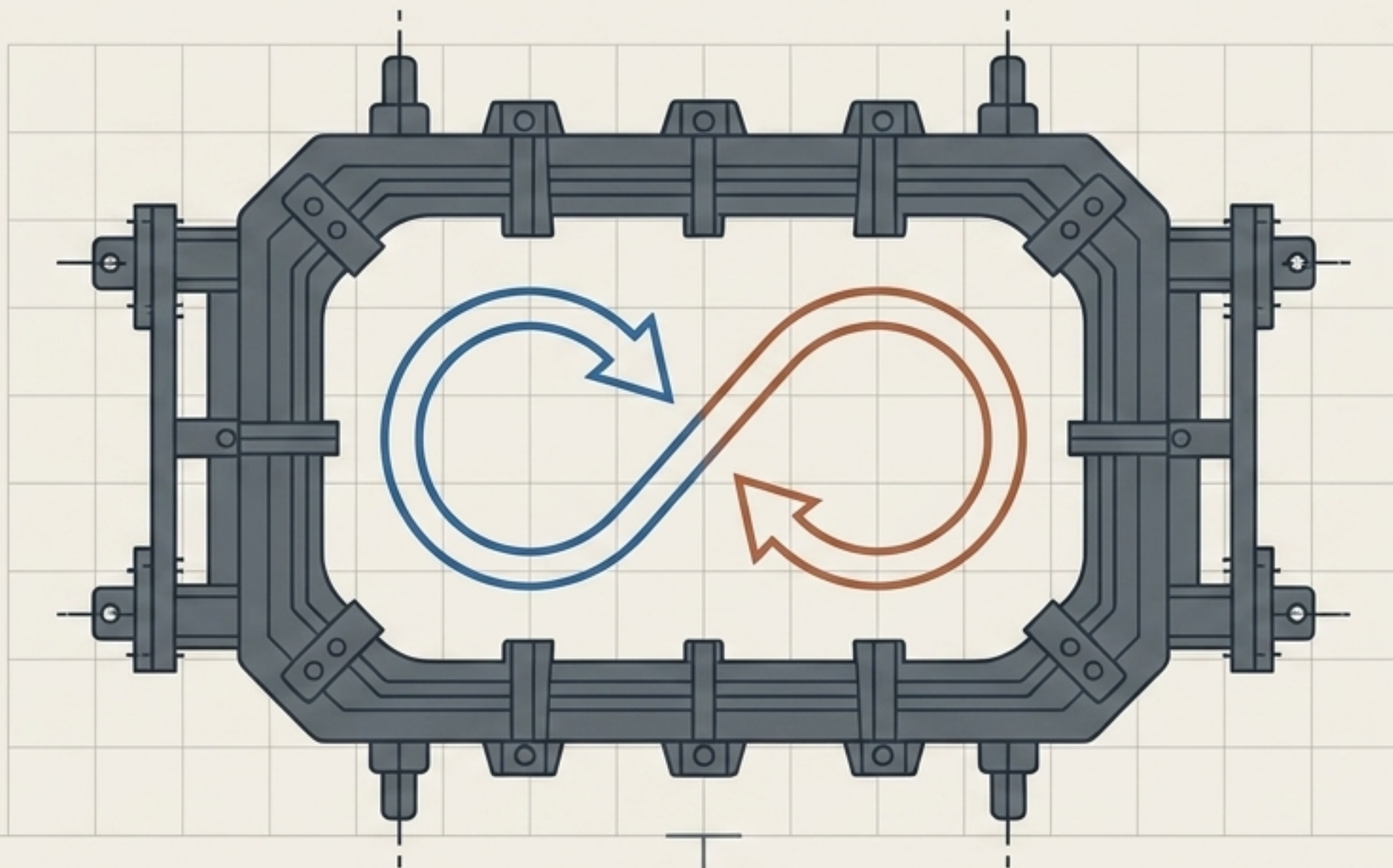
Zmožnost sistemov za neprekinjeno učenje in razumevanje fizikalnih zakonitosti za reševanje kompleksnih naravoslovnih hipotez.

3. Fizična AI



Preboj umetne inteligence iz digitalnih terminalov v fizični svet. Utelesitev v robotiki in avtomatizacija hardverskih omejitev.

Zaključek: Ultimatívno orodje in ultimatívni izziv



AGI bo najmočnejše orodje v zgodovini znanosti, sposobno razreševanja bolezni, odkrivanja novih virov energije in razumevanja vesolja.

Za inženirje te generacije izziv ni več zgraditi neomejeno inteligenco. Izziv je zasnovati nadzorne sisteme, ki bodo dohajali eksponentno zanko samoizboljševanja.