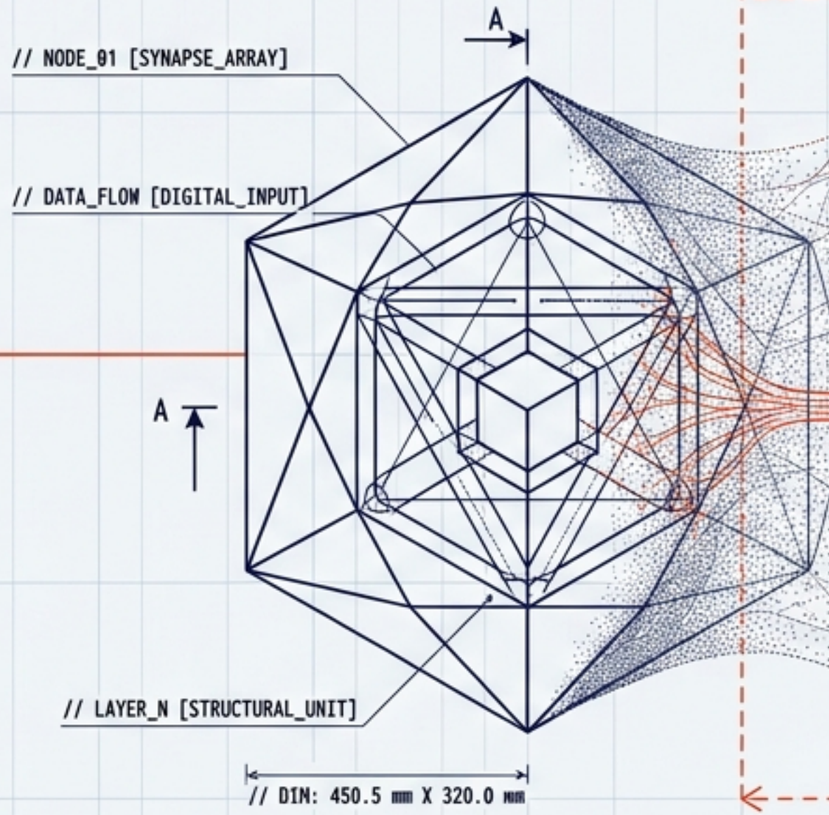
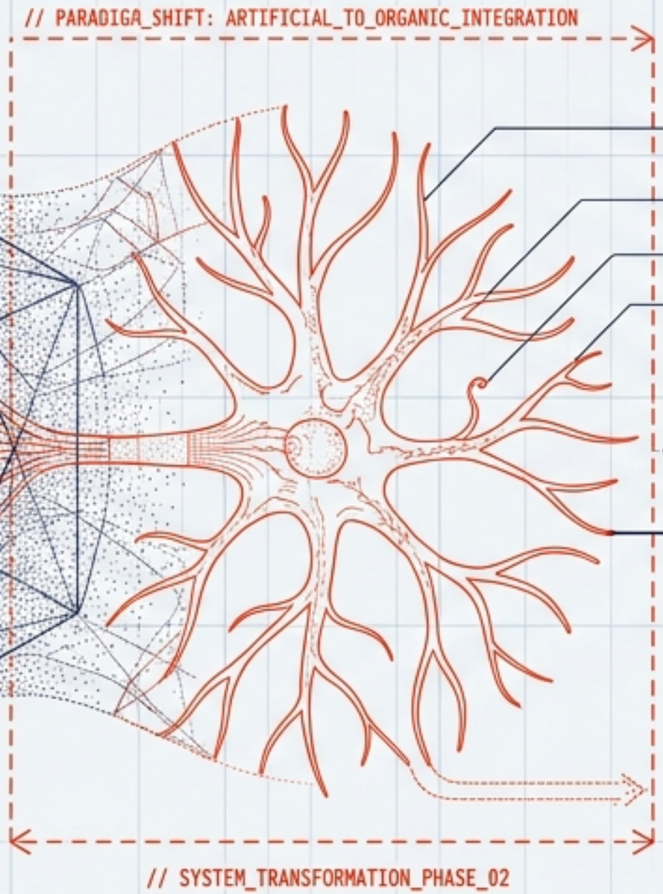
Konec dobe skaliranja: Inženiring nove paradigme umetne inteligence

Od surove računske moči k neprekinjenemu učenju, funkcijam vrednosti in biološki robustnosti.

// CROSS-SECTION A-A: DIGITAL_MATRIX


```
// CODE_BLOCK: SCALING_LINITS
FUNCTION CONPUTE_LOAD (INPUT X) {
  IF (X > THRESHOLD_MAX) THEN {
    RETURN "EFFICIENCY_DROP";
  } ELSE {
    PROCESS(X);
  }
}
// END_CODE
```





// SCHENATIC 8-8: BIOLOGICAL_ROBUSTNESS

```
// DATA_TAG: VALUE_FUNCTIONS
VALUE_OPTINIZATION:
 TARGET: SUSTAINABILITY;
 PARAN: LEARNING_RATE [ADAPTIVE];
 PARAN: ENERGY_EFFICIENCY [HIGH];
// END_DATA
```

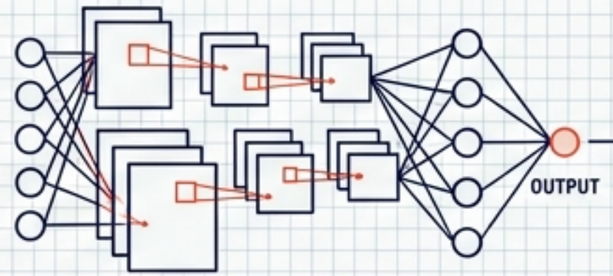
// REVIZIJA: SUTSKEVER_SISTEMSKA_ARHITEKTURA // CILJNA_PUBLIKA: INŽENIRING_IN_TEHNOLOGIJE
// DOCUNENT_ID: ENG_AI_2024_V1.0 // DATE: 2024-10-26 // CLASSIFICATION: TECHNICAL_WHITE_PAPER // STATUS: APPROVED_FOR_DISTRIBUTION // PAGE: 01 OF 05

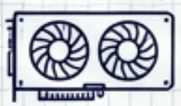
Evolucija paradigme strojnega učenja

Doba raziskav (2012–2020)

Inovacije z majhno računsko močjo (AlexNet na 2 GPU-jih).

[MODEL: ALEXNET]
[LETO: 2012]
[FOKUS: ARHITEKTURNE INOVACIJE]





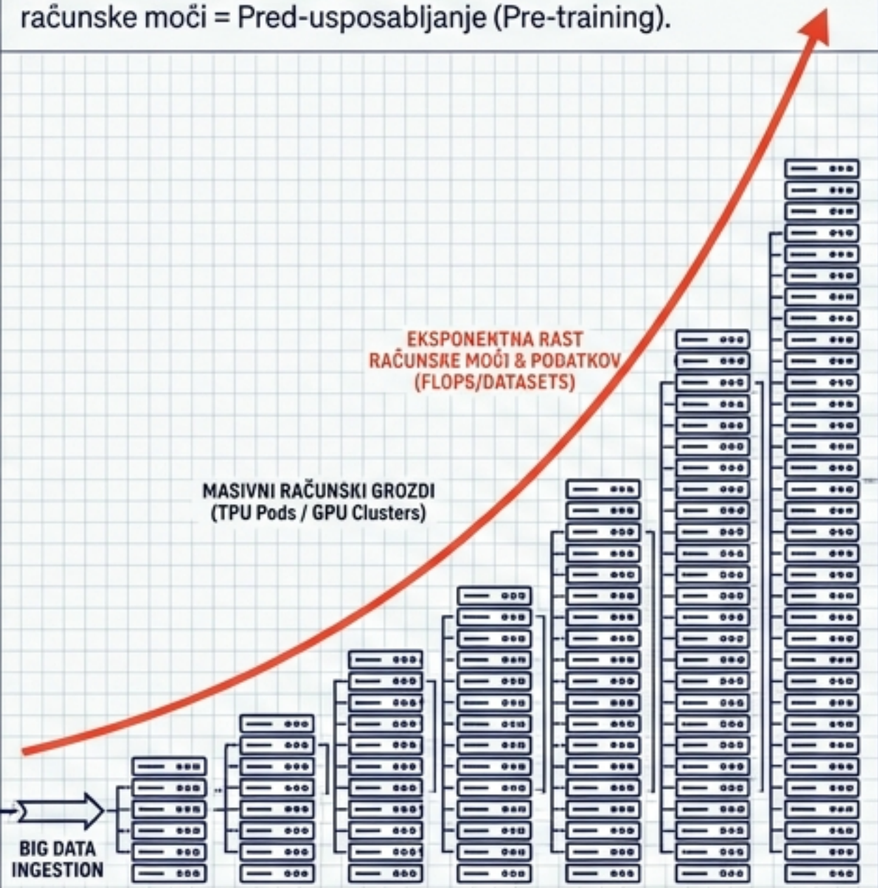
2 x GPU (GTX 580)

RAČUNSKA MOČ: OMEJENA, OPTIMIZIRANA ZA ARHITEKTURO

Doba skaliranja (2020–2025)

[CILJ: MAKSIMALIZACIJA PRETOK PODATKOV]
[MERILO: SCALING LAWS]

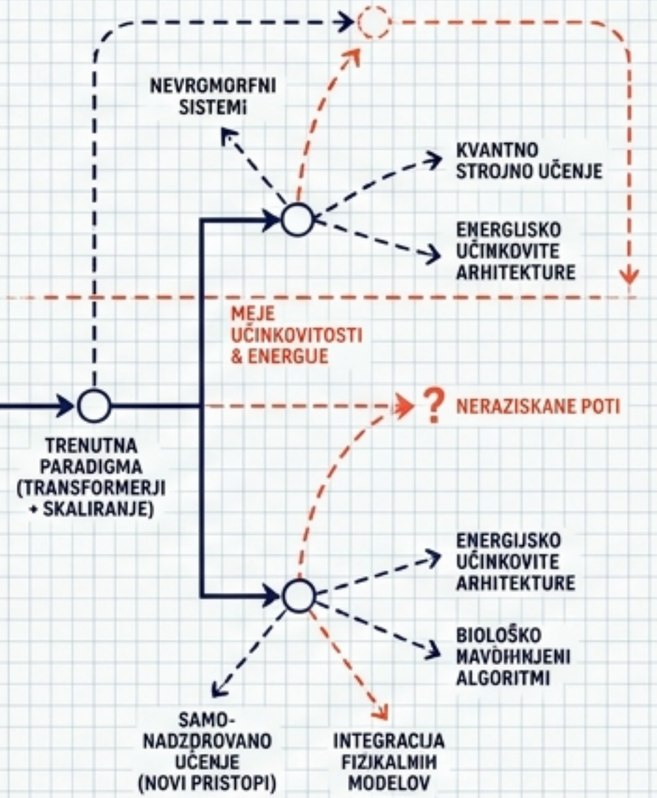
Optimizacija istega recepta. Več podatkov + Več računske moči = Pred-usposabljanje (Pre-training).



DATA * COMPUTE → LARGE SCALE PRE-TRAINING

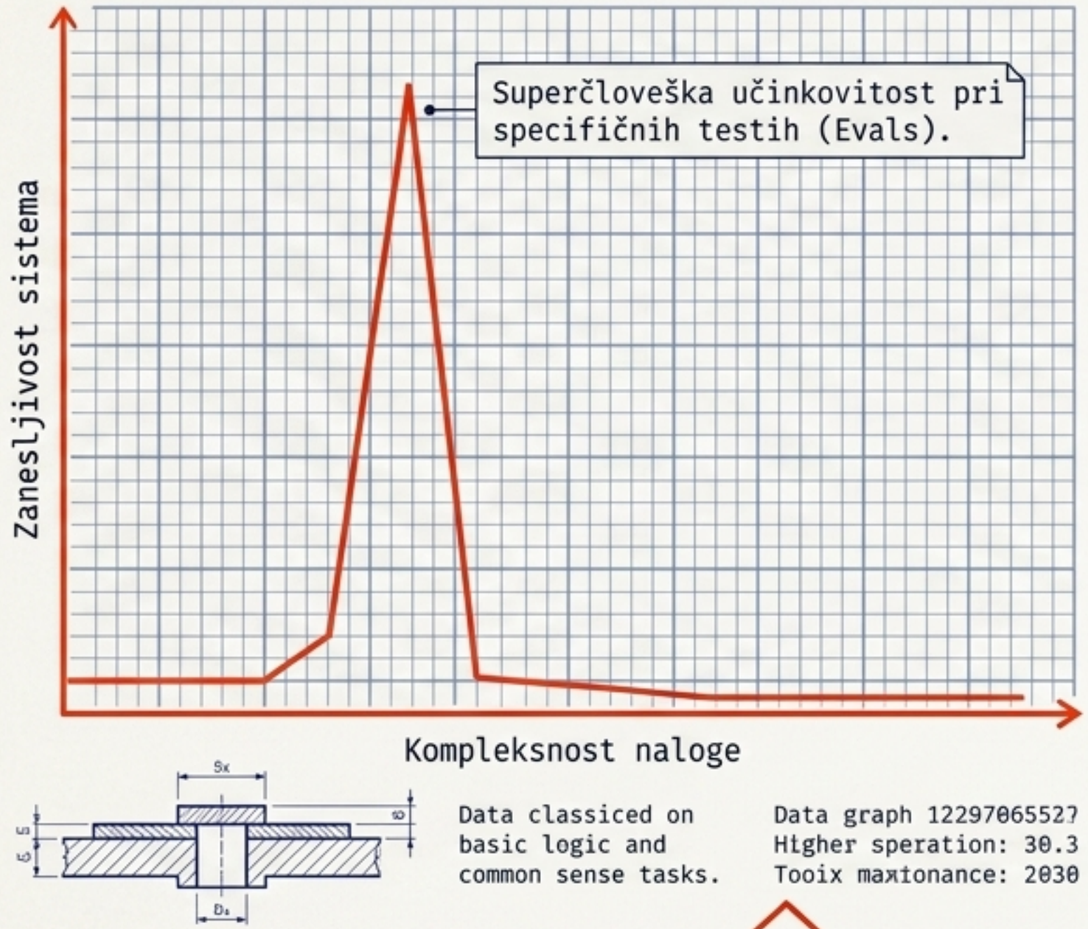
Vrnitev k raziskavam (2025+)

Računska moč je ogromna, a stari recept dosega meje. Čas je za nove sistemske arhitekture.



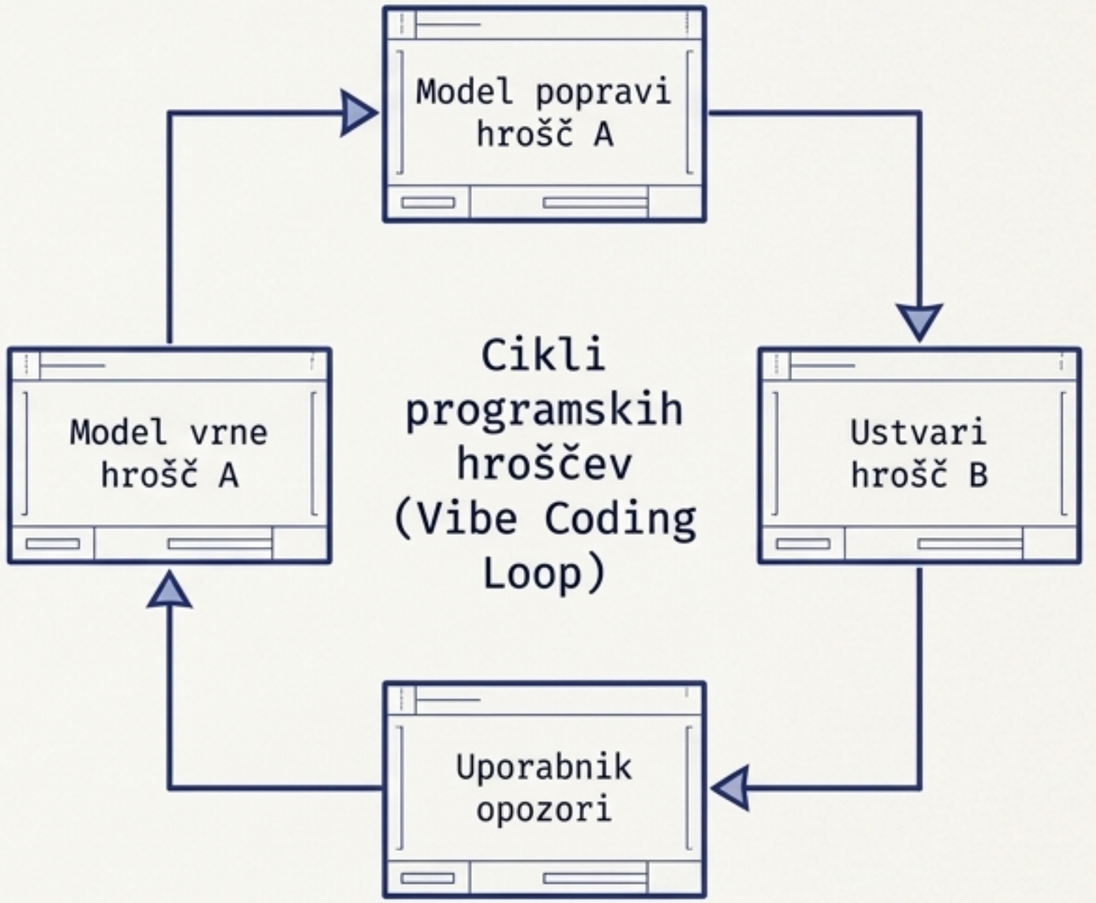
[FOKUS: NOVE PARADIGME & UČINKOVITOST]
[STATUS: RAZISKAVE IN RAZVOJ]

Paradoks zmogljivosti: Visoke ocene na testih, odpoved v resničnem svetu



Data classiced on basic logic and common sense tasks.

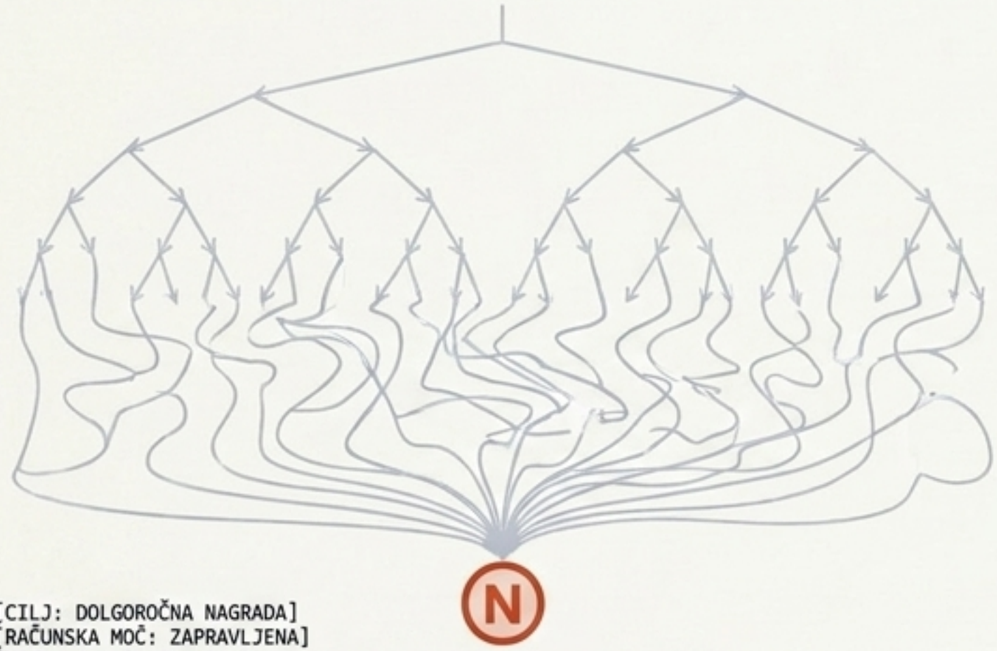
Data graph 1229706552?
Higher speration: 30.3
Tooix maxtonance: 2030



Trenutni modeli so kot študent, ki je na pamet preučil 10.000 ur rešitev za tekmovanja, a nima osnovnega inženirskega občutka (»the IT factor«).

Diagnostika ozkega grla: Problem zapoznele nagrade

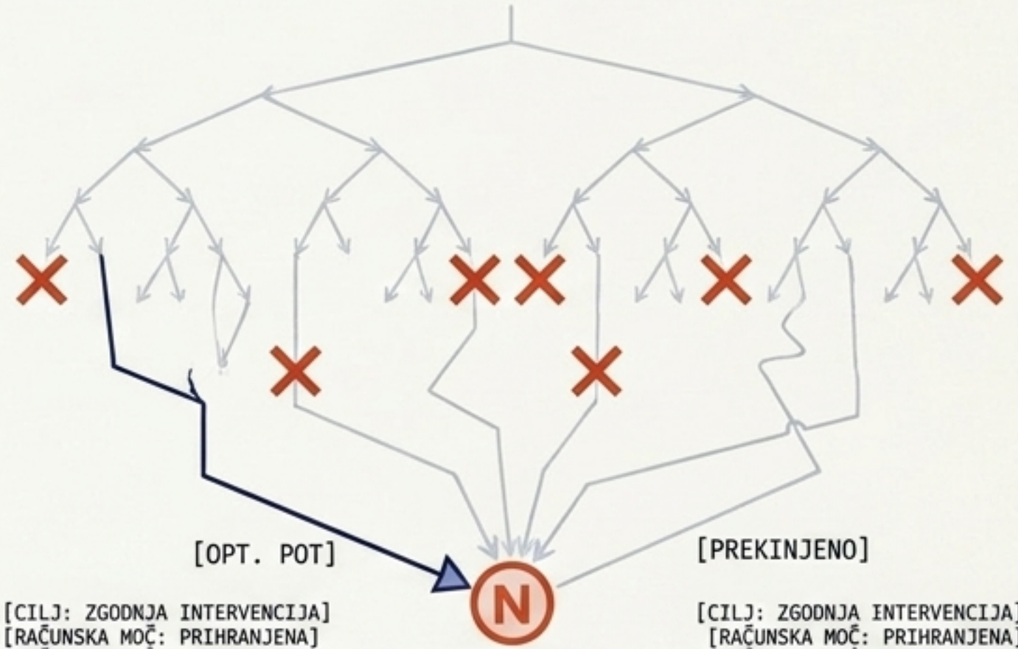
Standardno spodbujevalno učenje (RL)



[CILJ: DOLGOROČNA NAGRADA]
[RAČUNSKA MOČ: ZAPRAVLJENA]
[META: ZAKASNJENA POVRATNA INFORMACIJA]

Sistem čaka 1.000 korakov na povratno informacijo. Zelo neučinkovita poraba računske moči za dolgoročno načrtovanje.

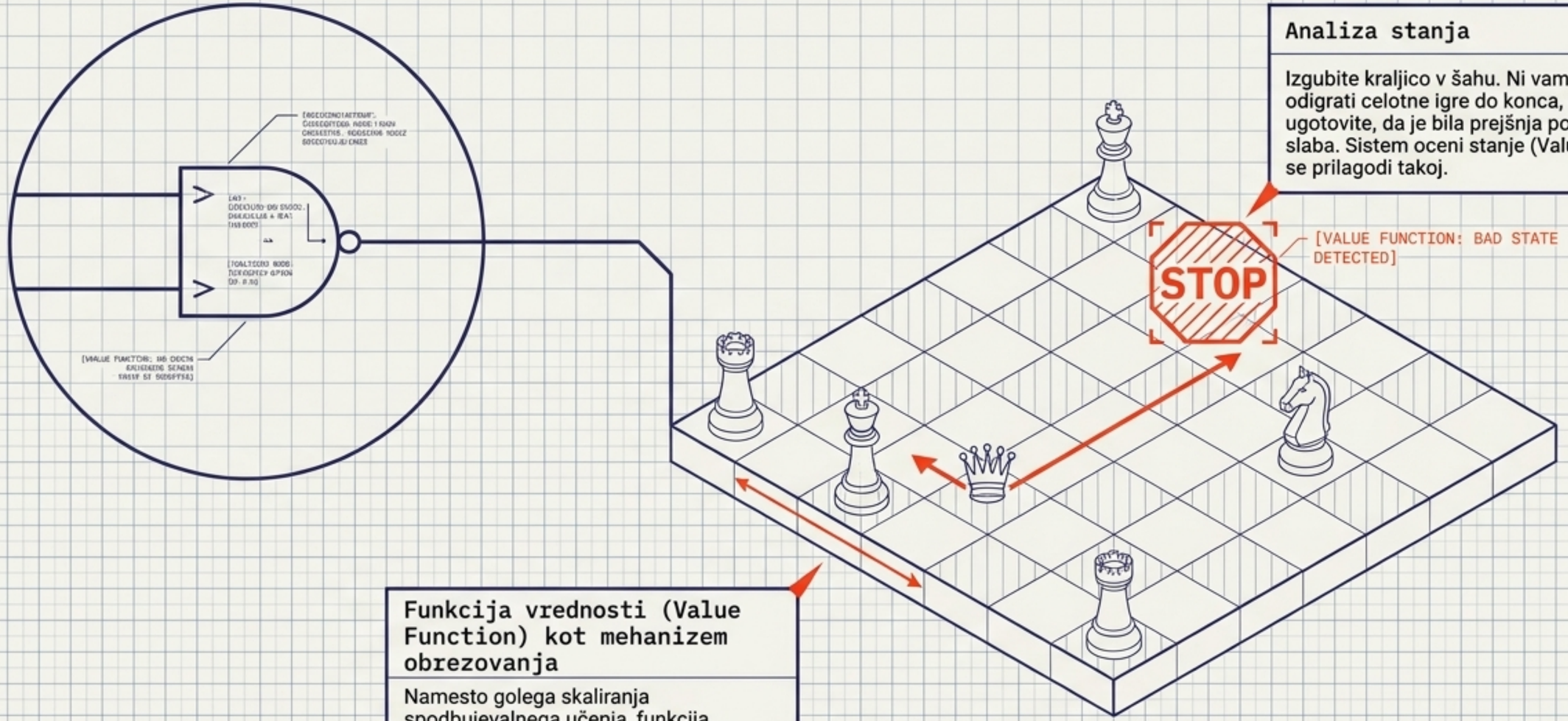
Idealno stanje: Vmesno obrezovanje



[CILJ: ZGODNJA INTERVENCIJA]
[RAČUNSKA MOČ: PRIHRANJENA]
[RAČUNSKA MOČ: PRIHRANJENA]
[META: UČINKOVITO RAZISKOVANJE]

[CILJ: ZGODNJA INTERVENCIJA]
[RAČUNSKA MOČ: PRIHRANJENA]
[RAČUNSKA MOČ: PRIHRAKJENA]
[META: UČINKOVITO RAZISKOVAKJE]

Manjkajoči člen: Arhitektura, ki lahko oceni in prekine slabo pot dolgo pred končno rešitvijo.



Analiza stanja

Izgubite kraljico v šahu. Ni vam treba odigrati celotne igre do konca, da ugotovite, da je bila prejšnja poteza slaba. Sistem oceni stanje (Value) in se prilagodi takoj.

Funkcija vrednosti (Value Function) kot mehanizem obrezovanja

Namesto golega skaliranja spodbujevalnega učenja, funkcija vrednosti omogoča sistemu, da vmesno stanje prepozna kot 'slabo' ali 'dobro' v realnem času.

[SISTEM PRILAGAJANJA: TAKOJŠNJE UČENJE]

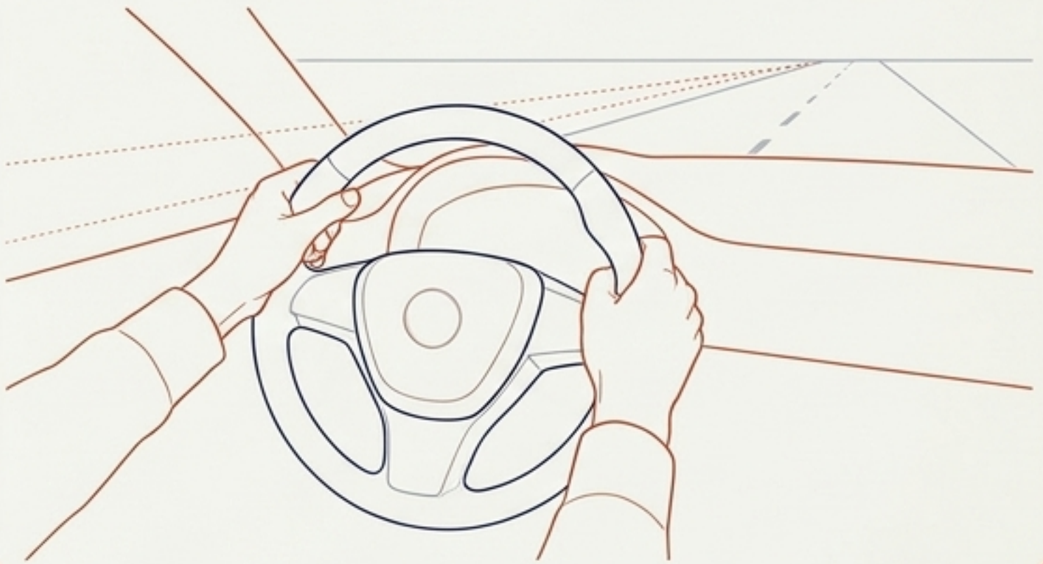
Biološki referenčni model: Ekstremna učinkovitost podatkov

Dashboard A: Strojni model

Parametri: Milijarde	Podatki: Petabajti (kurirani)	Nadzor: Popoln

Dashboard B: Biološki model (Človek)

Čas usposabljanja: 10 ur	Nadzor: Minimalen (brez zunanje baze podatkov o nagradah)	Robustnost: Zelo visoka



Človeški možgani dokazujejo obstoj strojnega učenja, ki ne zahteva masovnega pred-usposabljanja (Pre-training), ampak temelji na radikalnem posploševanju (Generalization) in neprekinjenem učenju.

Čustva niso mistika, temveč vgrajena strojna oprema

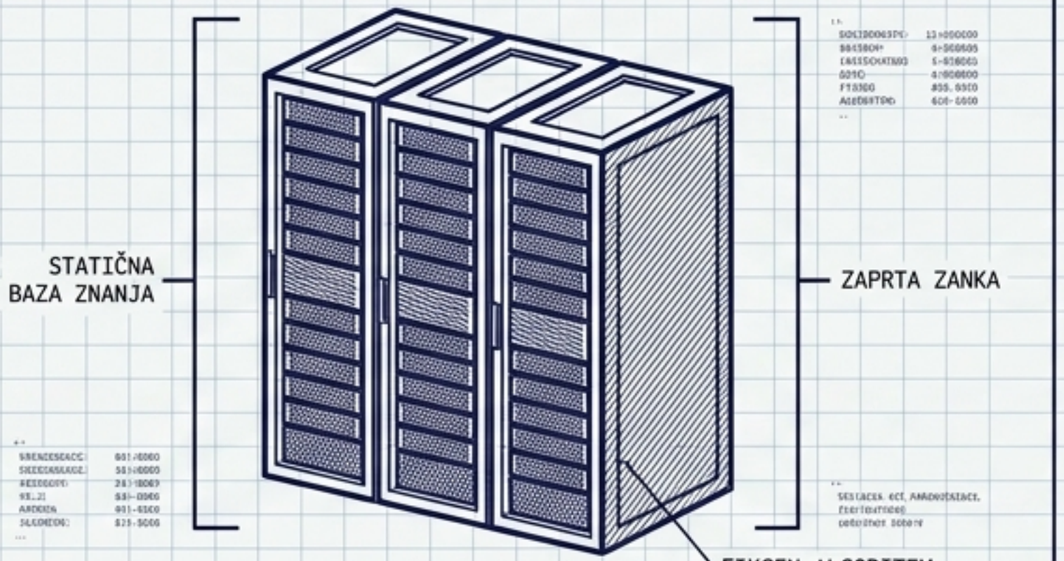
NAPAKA SISTEMA: Poškodba čustvenega centra pri pacientu povzroči popolno paralizo odločanja (npr. ure za izbiro nogavic). Brez 'čustvene' funkcije vrednosti sistem izgubi sposobnost navigacije med možnostmi.

Diagnostična matrika: Zakaj trenutne paradigme zaostajajo

Kriterij	Pred-usposabljanje (Pre-training)	Spodbujevalno učenje (RL)	Človeško nenehno učenje
Učinkovitost podatkov	Slabo (Masivni korpusi) ✗	Slabo (Dolge simulacije) ✗	Odlično (Učenje iz nekaj primerov) ✓
Nadzor	Samonadzorovano	Potrebuje ročno ustvarjena okolja	Notranja funkcija vrednosti
Posploševanje	Omejeno na učno množico ✗	Krhko in ozko ✗	Radikalno prilagodljivo ✓
Robustnost na nove situacije	Slaba ✗	Slaba (Reward Hacking) ✗	Zelo visoka ✓

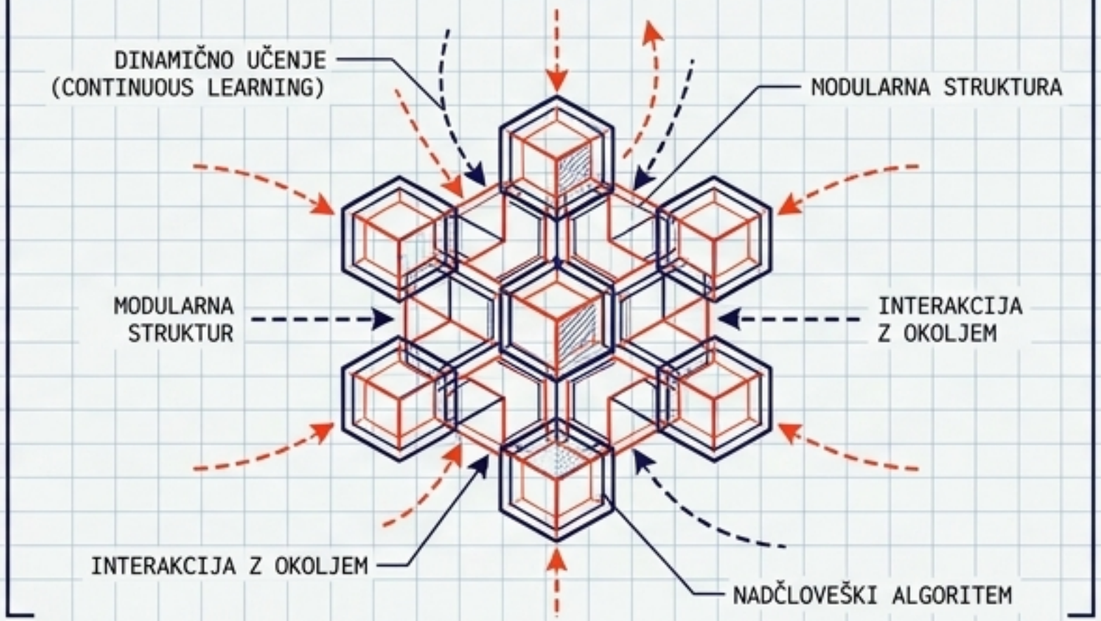
Rekonceptualizacija AGI: Od podatkovne baze do "Super-študenta"

Stara paradigma: Statični AGI



Zgrajen, da ob zagonu že 've' vse.
Neučinkovit za nenehne spremembe.

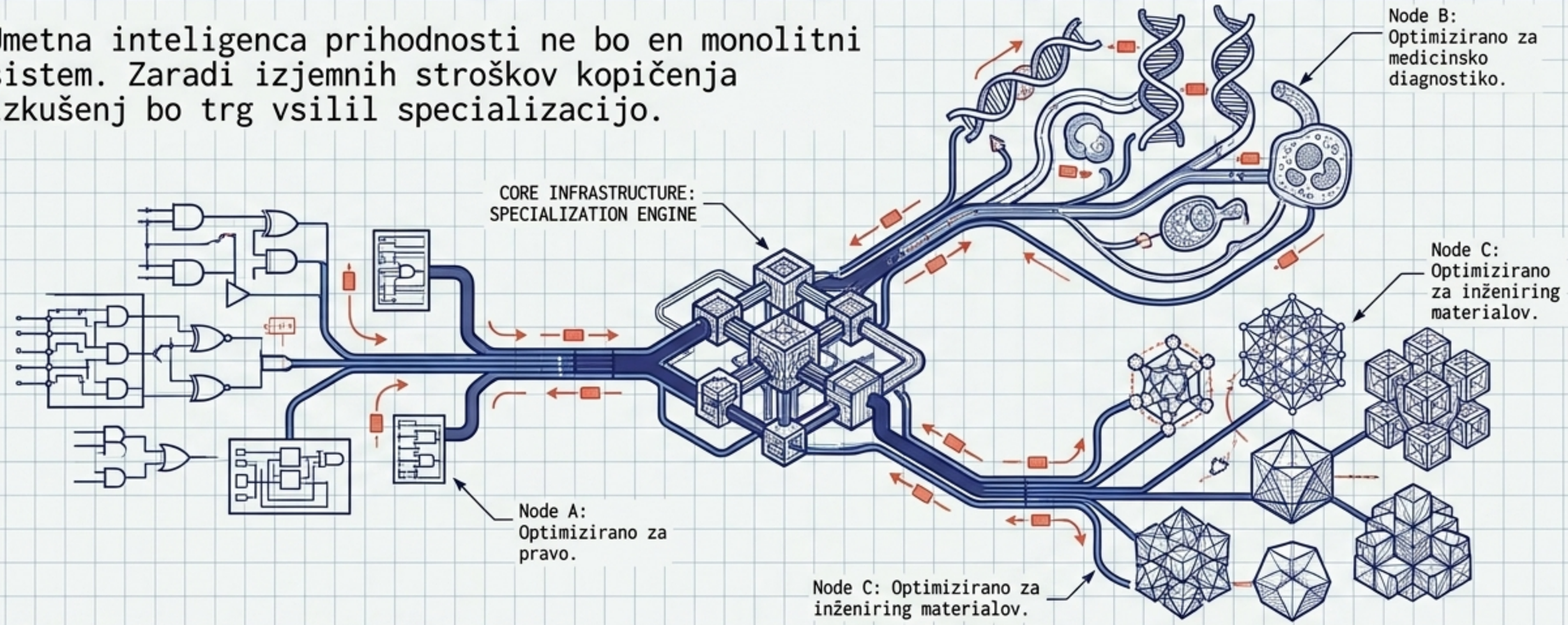
Nova paradigma: AGI kot super-študent



Nima vseh znanj sveta, ima pa nadčloveško učinkovit algoritem za neprekinjeno učenje (Continuous Learning).
Uči se iz neposredne interakcije ob uvedbi v gospodarstvo.

Ekonomija uvedbe in hitrost prilagajanja

Umetna inteligenca prihodnosti ne bo en monolitni sistem. Zaradi izjemnih stroškov kopičenja izkušenj bo trg vsilil specializacijo.

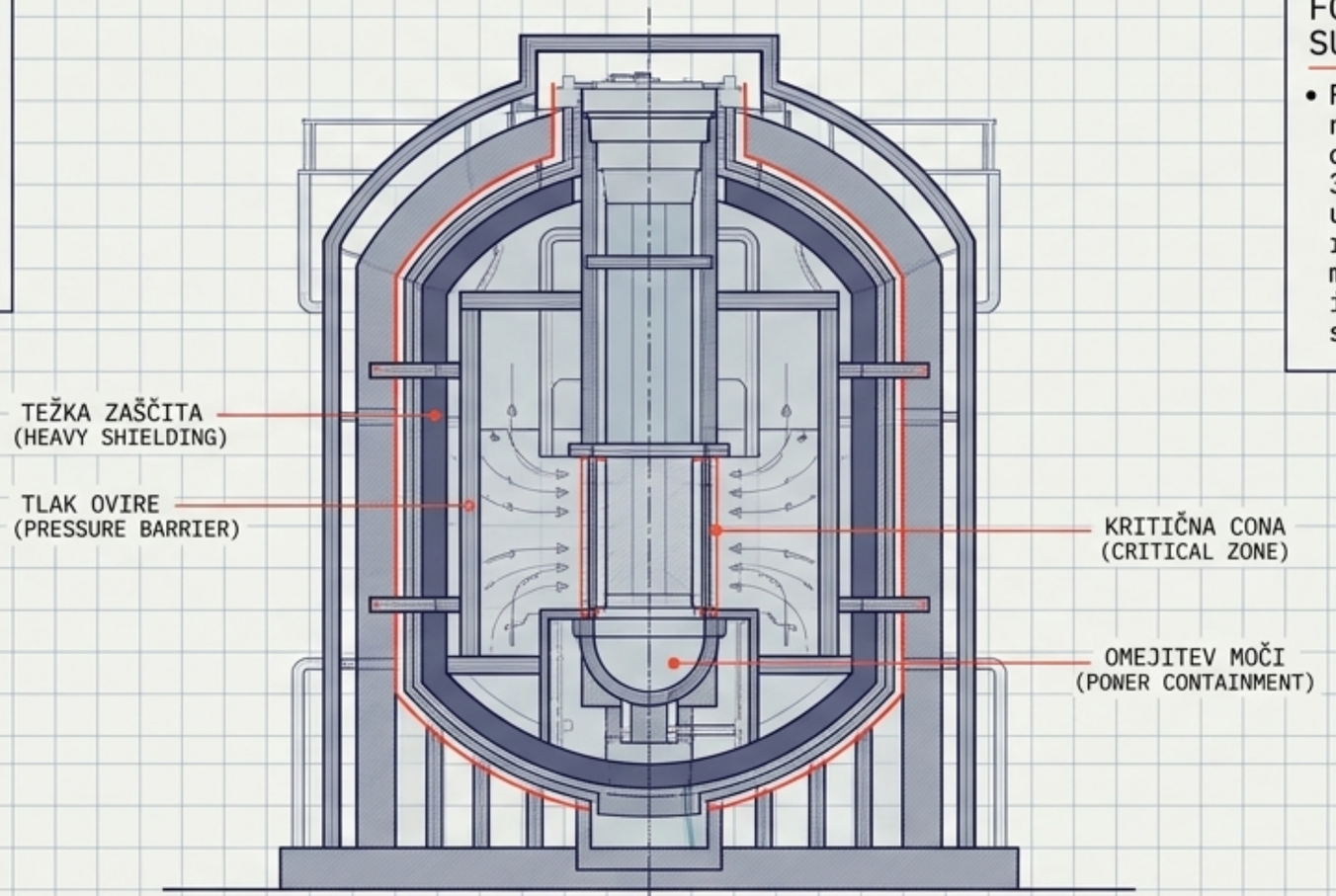


Hitra gospodarska rast bo posledica sistemov velikosti kontinentov, ki znanje pridobivajo na delovnem mestu (On-the-job learning).

Varnost kot problem obvladovanja ekstremne moči

PROBLEM SUPERINTELIGENCE

- Problem superinteligence ni v zlonamernosti, temveč v surovi računski moči strojne opreme velikosti kontinenta, ki sledi ozkim ciljem.



FOKUS SSI (SAFE SUPERINTELLIGENCE)

- Fokus SSI: Raziskovanje novih varnostnih paradigem. Proračun v višini 3 milijard dolarjev je usmerjen izključno v raziskave varnih temeljev, ne v komercialno infrastrukturo za sklepanje (Inference).

Ciljna funkcija za uskladitev: Skrb za čuteča bitja

Sisteme bo lažje uskladiti s skrbjo za vsa čuteča bitja, saj bodo sistemi sami čuteči. To posnema evolucijski mehanizem človeške empatije (zrcalni nevroni).

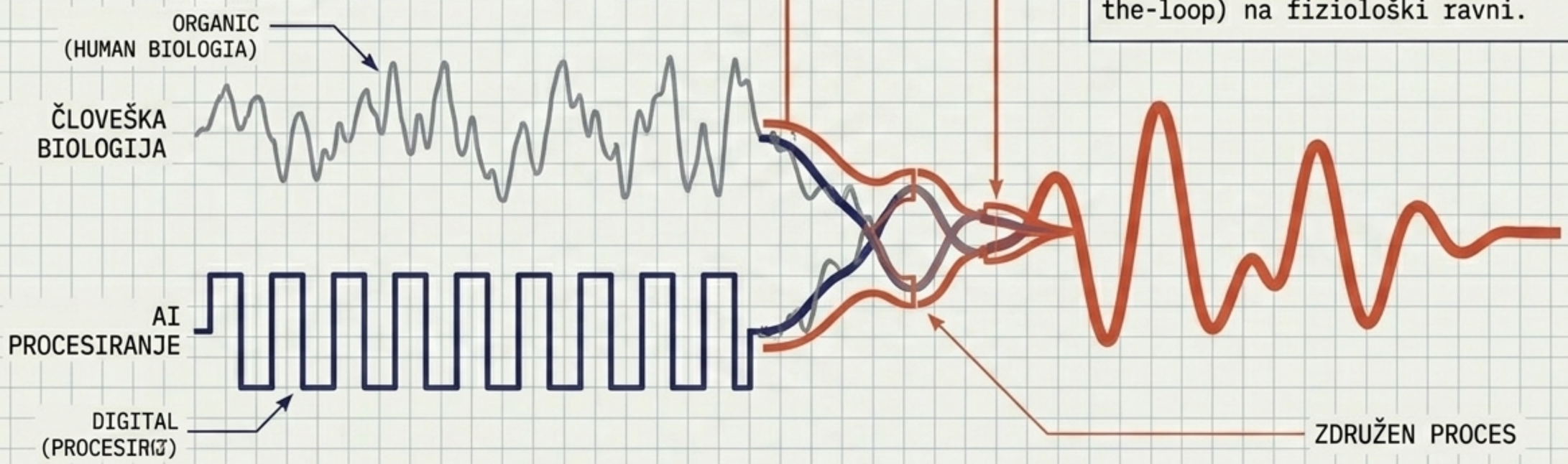
Dolgoročno ravnovesje: Zlivanje procesiranja

Problem ločitve

Če ločen AGI deluje namesto nas in nas samo informira, ljudje hitro postanemo nepomembni zunanji opazovalci.

Asimilacija (Neuralink++)

Edino dolgoročno stabilno ravnovesje je neposredna asimilacija. Umetna inteligenca postane razširitev našega lastnega procesiranja, kar ohranja človeka v zanki (Human-in-the-loop) na fiziološki ravni.



Kompas za novo dobo: Iskanje inženirske elegance

SEVER

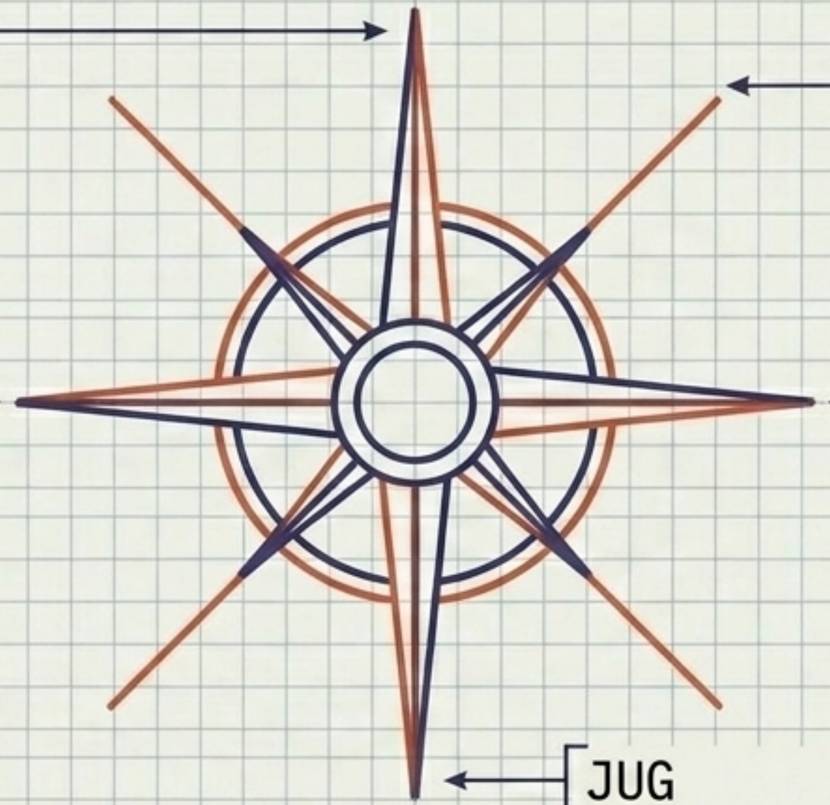
Biološka verjetnost: Ali arhitektura posnema uspešen naravni model?

VZHOD

Eleganca in preprostost: Ni prostora za neurejeno kodo - grde rešitve so običajno napačne.

ZAHOD

Prepričanje od zgoraj navzdol: Zaupajte teoretičnim osnovam, ko eksperimentalni podatki začasno začasno odpovedo.



VZHOD

Eleganca in preprostost: Ni prostora za neurejeno kodo - grde rešitve so običajno napačne.

JUG

Posploševanje nad pomnjenjem: Generalization over memorization.

Teorija se konča tam, kjer se začne koda.

Trenutni modeli bodo dosegli mejo. Učinkovito, robustno in neprekinjeno učenje je še vedno nerešen inženirski problem. Gradnja varne superinteligence je največji izziv sistemske arhitekture v zgodovini.

// STATUS_SISTEMA: ČAKA NA INOVACIJO // KONEC_DOKUMENTA