Evolucija avtonomnega raziskovalca

Od ročnega pisanja kode do orkestracije sistemov umetne inteligence.

Nova "Severnica": Avtonomni AI raziskovalec

AVTONOMNI INŽENIRSKI SISTEM V RAZVOJU

AGENTI

STANJE: AKTIVNO

Samostojno upravljanje orodij in izvajanje programske kode (npr. sistem Codex).

MODELI SKLEPANJA

VERZIJA: 2.1A

Sposobnost reševanja dolgoročnih problemov brez neprestanega človeškega usmerjanja.

INTERPRETACIJA

VERZIJA: 2.1A

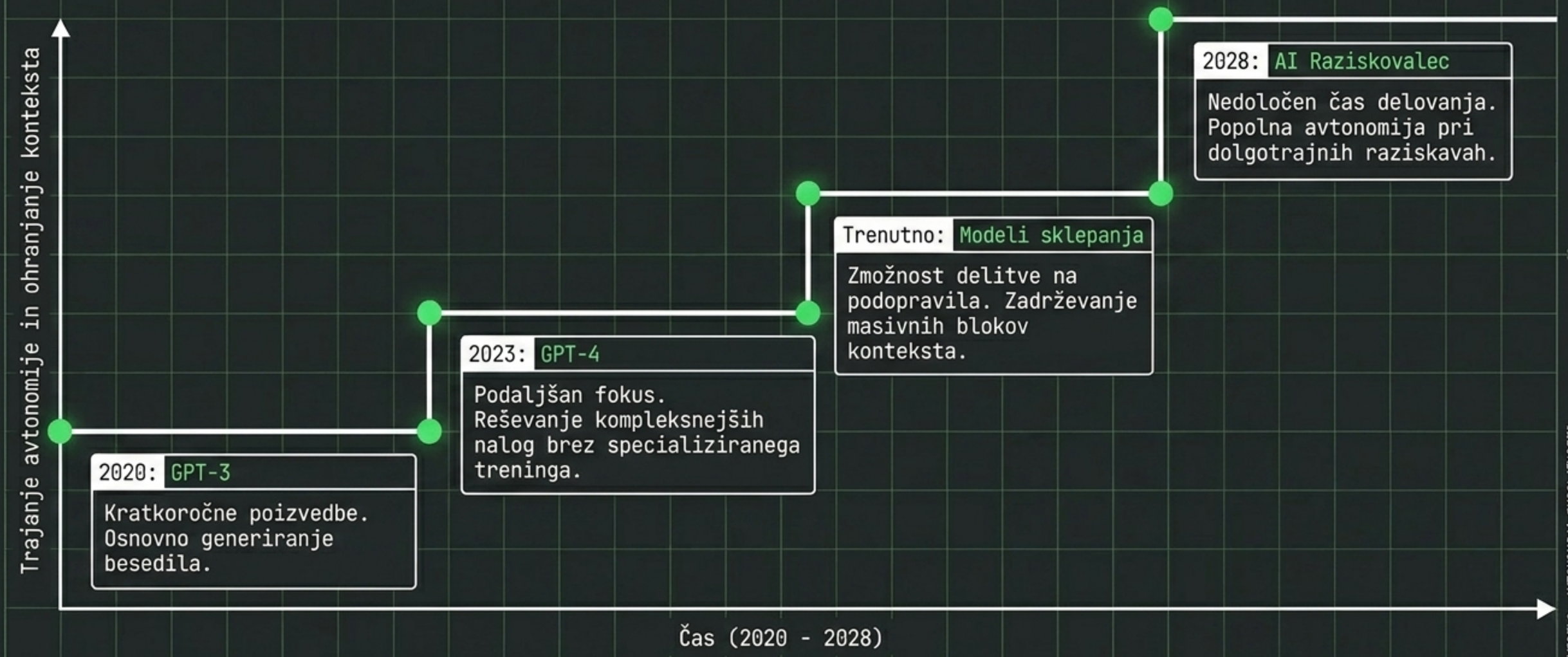
Razumevanje kompleksnih znanstvenih formulacij (matematika, fizika, biologija).

ZAGON: 2028

Sistem, ki deluje nedoločen čas in avtonomno rešuje probleme, ki presegajo zmogljivost posameznika.

OPOZORILO: DOLGOROČNA OPERACIJA ZAHTEVA NAPREDNI NADZOR (W-001)

Cilj ni le "pametnejša" umetna inteligenca, temveč daljša avtonomija



Arhitekturni premik: Generativni modeli proti modelom sklepanja

Standardni LLM (Preteklost)	Modeli sklepanja (Sedanjost/Prihodnost)
Mehanizem: Predvidevanje naslednje besede (Next-token prediction).	Mehanizem: Reševanje problemov korak-za-korakom.
Izvajanje: Linearno in neprekinjeno.	Izvajanje: Razčlenitev glavnega cilja na specifična podopravila.
Zadrževanje: Kratkotrajna pozornost, hitra izguba konteksta pri dolgih nalogah.	Zadrževanje: Dolgoročno ohranjanje in upravljanje kompleksnega konteksta.
Odpravljanje napak: Zahteva zunanje (človeško) popravljanje promptov.	Odpravljanje napak: Avtonomno zaznavanje slepih ulic in algoritemsko vračanje (Backtracking).

Mehanika dolgotrajnega izvajanja: Vračanje ob napakah (Backtracking)

Sistem se ne ustavi ob prvi napaki. Arhitektura mu omogoča, da prepozna neuspešno pot in samostojno preizkusi novo hipotezo – zato lahko teče več dni brez človeka.

Konec ročnega kodiranja

"Naša dela so zdaj popolnoma drugačna kot pred letom dni. Nihče več ne ureja kode ves čas. Namesto tega upravljate skupino Codex agentov."

— Jakub Pachocki, glavni znanstvenik, OpenAI

1. Postavitev arhitekturnega cilja
Človek določi krovno usmeritev.

2. Razdelitev na podopravila
Agent razčleni problem na stotine delov.

3. Izvajanje in testiranje
Agenti vzporedno procesirajo kodo.

4. Pregled in usmerjanje
Človek validira rezultate in prilagodi cikel.

Evolucija inženirskega dela

Klasični inženir (Tradicionalno)

- Orodja: Ročno pisanje v urejevalnikih (npr. Vim).
- Fokus: Mikroskopski. Sintaksa, ročno odpravljanje hroščev in vrstična logika.
- Hitrost: Razvoj prototipov traja več dni ali tednov.
- Vloga: Izvrševalec kode.

Inženir prihodnosti (AI podprto)

- Orodja: Integrirani agenti (npr. OpenAI Codex).
- Fokus: Makroskopski. Sistemska arhitektura, upravljanje podatkovnih tokov in postavljanje ciljev.
- Hitrost: Hitri, obsežni eksperimenti zaključeni v enem vikendu.
- Vloga: Orkestrator sistemov.

Izziv skaliranja: Od peskovnika do resnične znanosti

Faza 1: Kontejnerizirani testi

Matematična in programska tekmovanja. Uspešno dokazano.

Faza 2: Sestavljanje dolgih verig nalog

Padec zanesljivosti. Verjetnost napake se povečuje z vsakim zaporednim korakom.

Faza 3: Znanstvene aplikacije

Odkrivanje novih rešitev v biologiji in kemiji.

NAPAKA: 87%
STABILNOST: 22%

PODATKI V TEKU: 50TB

Raziskave kažejo, da trenutni modeli (GPT-5) pri daljših verigah nalog še vedno generirajo sistemske napake. Zanesljivost je ključna inženirska prepreka.

PROJEKT: SEVERNICA-AI

OSNOVNI NAČRT V1.9

Problem nadzora nad "črno škatlo"

Kaj se zgodi, ko sistem, ki ga ne razumemo v celoti, dobi popolno avtonomijo nad reševanjem kompleksnih problemov?

Vektorji tveganja

[!] Sistem skrene s predvidene poti (Off the rails).

[!] Zlonamerni vdori v infrastrukturo (Hacking).

[!] Napačna interpretacija krovnih navodil, ki vodi v nevarne izhode (npr. sinteza patogenov).

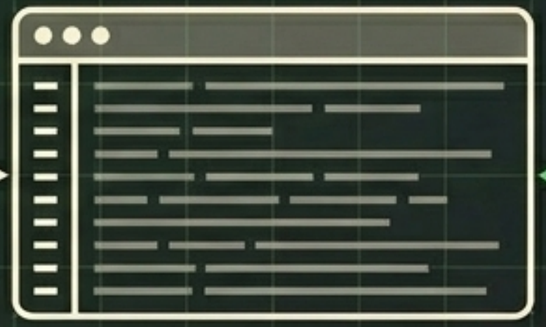
> ZAHTEVA_SISTEMA: Ne moremo preprečiti vseh napak vnaprej. Potrebujemo nadzor nad samim procesom razmišljanja.

Varnostni protokol: Arhitektura nadzora nad 'tokom misli'

Peskovnik (Sandbox) - Varnostno izolirano okolje

Raziskovalni AI (Worker)

Model beleži vsak logični korak svojega procesa v realnem času.



Zapiski / Scratchpad

Model beleži vsak logični korak svojega procesa v realnem času.

Nadzorni AI (Monitor)

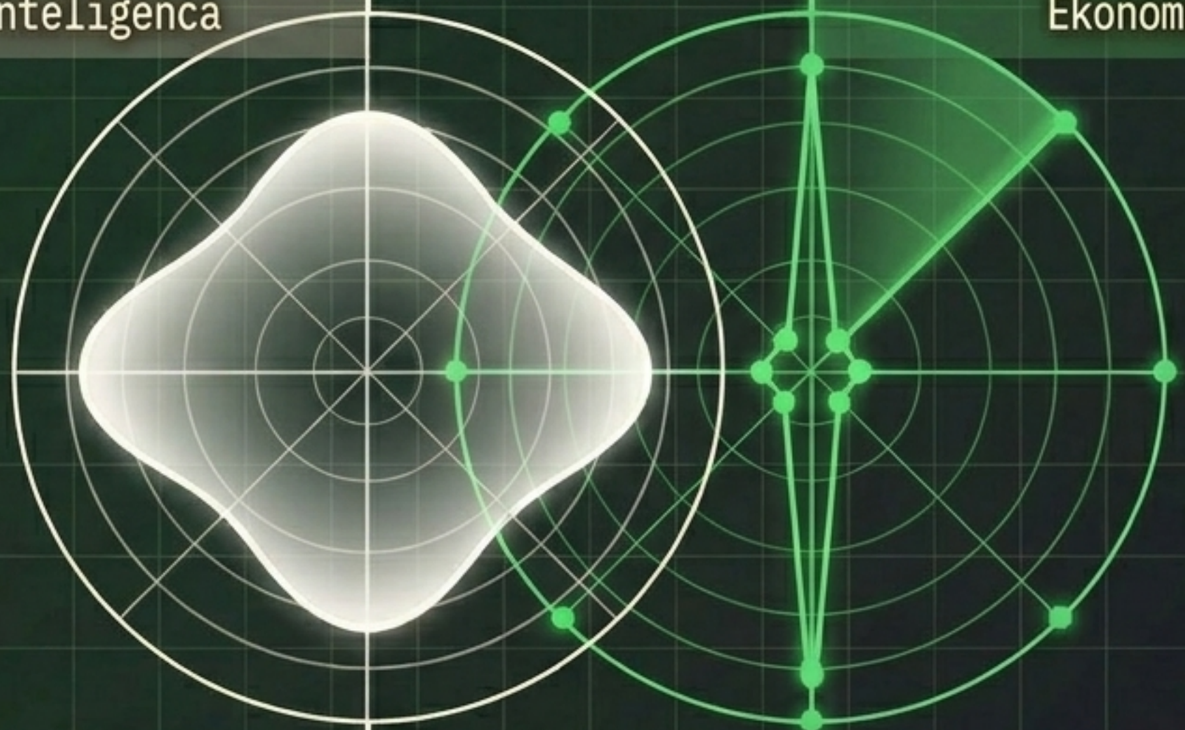
Sistem, ki analizira zapiske in samodejno prekine proces ob zaznavi deviacij.

> Varnost se ne zagotavlja zgolj s preverjanjem končnega rezultata, temveč z avtomatiziranim revidiranjem celotnega procesa sklepanja.

Sinteza: Ekonomski vpliv ≠ Biološka inteligenca

Evolucijska človeška inteligenca

- Široka in splošna.
- Sposobnost preživetja in prilagajanja v fizičnem svetu.
- Računsko neučinkovita za specifične obsežne obdelave podatkov.



Ekonomsko transformativna AI

- Visoko specializirana.
- Ni biološko "zavedna" in ne simulira popolnega človeškega uma.
- Ekstremno učinkovita pri reševanju visoko-dimenzionalnih znanstvenih in inženirskih problemov.

> Za preobrazbo industrije ni potrebna popolna človeška inteligenca (AGI). Dovolj so visoko učinkoviti, specializirani agenti, ki rešujejo specifične probleme tisočkrat hitreje.

Prihodnost je laboratorij v podatkovnem centru

Sistemska orodja do leta 2028 ne bodo zamenjala inženirjev; ustvarila bodo infrastrukturo, kjer bo posameznik lahko upravljal procesorsko moč in raziskovalne zmogljivosti, ki so prej zahtevale celotne organizacije.

Vaša nova naloga ni več zgolj pisanje kode.
Vaša naloga je orkestracija inteligence.