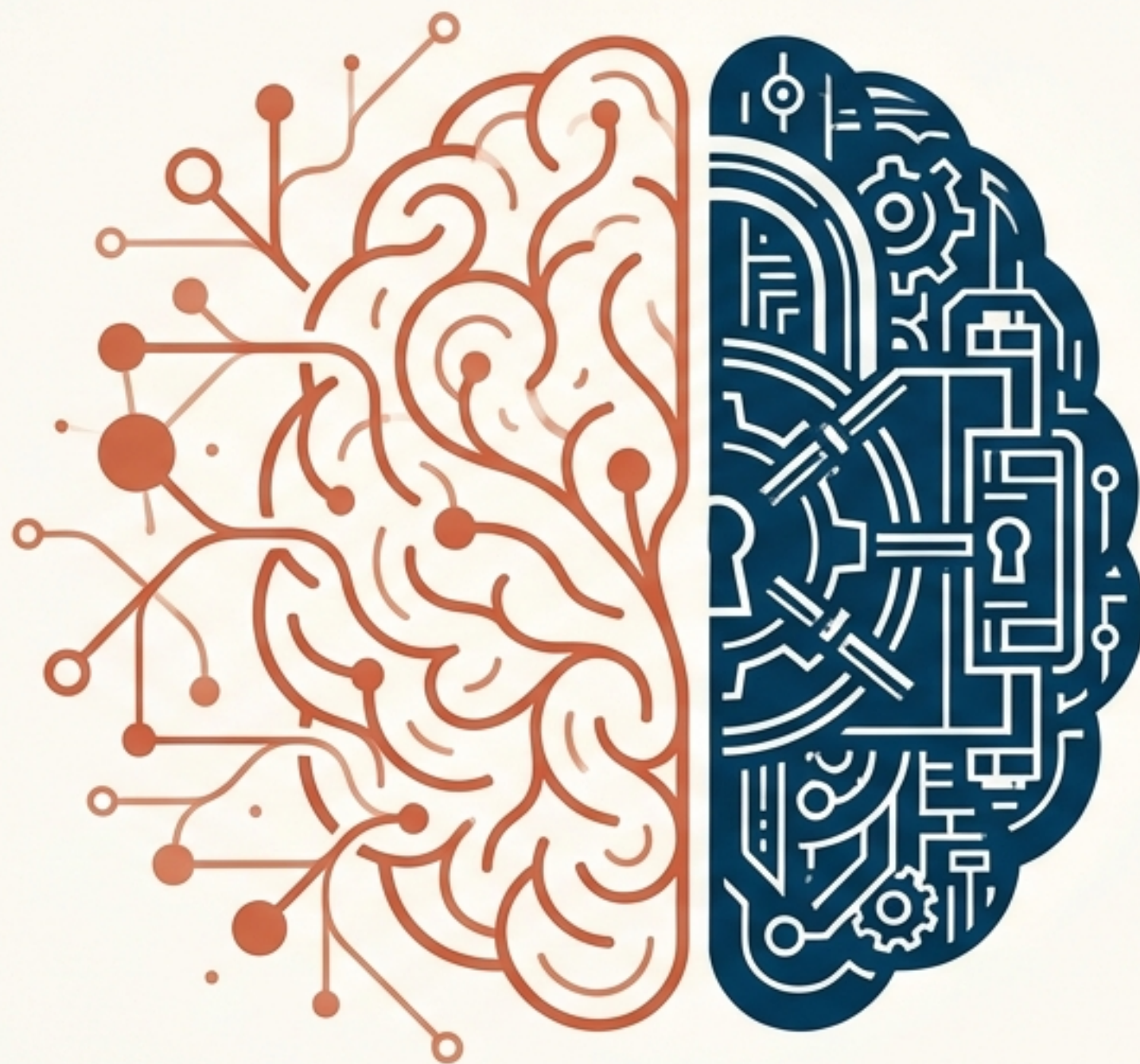


Odprta umetna inteligenca: Inovacija ali tveganje?

Analiza osrednje razprave, ki oblikuje prihodnost tehnologije.



Iskra, ki je sprožila razpravo: Zgodba o AlphaFold3

Google DeepMind objavi AlphaFold3, kar povzroči navdušenje v znanstveni skupnosti.



Odziv na kritike: Po pritisku skupnosti DeepMind objavi izvorno kodo.



Maj 2024

Prvotno razočaranje

November 2024

Omejitve ostajajo

DeepMind ne objavi izvorne kode ali uteži modela. Raziskovalci so omejeni na spletno orodje z do 10 (kasneje 20) napovedmi na dan.



Dostop je omejen na nekomercialno uporabo, za pridobitev uteži modela pa so potrebne akademske reference.



Dve plati razprave: Spodbujanje inovacij proti obvladovanju tveganj

ZA ODPRTOST

Polna transparentnost in neomejena uporaba spodbujata ustvarjalnost, pospešujeta razvoj in preprečujeta koncentracijo moči.

Ključna točka: Preprečiti, da bi bila UI "v rokah nekaj kalifornijskih podjetij."

"Nasprotujem lastniškemu in zaprtemu razvoju, v katerem je nekaj tako pomembnega kot UI v rokah peščice akterjev." — Chirag Shah, Univerza v Washingtonu

ZA PREVIDNOST

Odprtost omogoča zlorabo tehnologije za škodljive namene.

Primeri tveganj: Ustvarjanje deepfake posnetkov, otroške pornografije, razvoj biološkega ali kemičnega orožja.

"Obstajajo nekatere tehnologije, za katere si ne želimo, da bi bile odprtokodne." — David E. Harris, Univerza v Kaliforniji, Berkeley

Analogija: Primerjava umetne inteligence z jedrsko tehnologijo.

Zaplet: Izraz 'odprtost' je izgubil svoj pomen

Koncept 'odprte kode' ni neposredno prenosljiv na UI. UI ni zgolj koda: vključuje podatke za učenje in uteži modela. "Pri UI koda ni *meso*," pravi David Widder.

Paradoks OpenAI: Podjetje, ustanovljeno kot odprto, danes ne deli več svoje tehnologije, kar je privedlo do tožb s strani soustanovitelja Elona Muska.

'Openwashing': Izraz, ki ga je skoval Mark Dingemanse, opisuje podjetja, ki trdijo, da so bolj transparentna, kot v resnici so.

Poskus standardizacije: Definicija Open Source Initiative (OSI)

Vplivna neprofitna organizacija OSI je avgusta 2024 objavila osnutek definicije odprtokodne UI.

Ključni kriteriji OSI

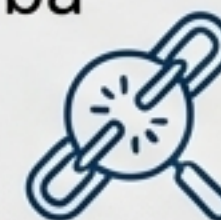
- ✓ Prosto dostopen, spremenljiv in deljiv za kakršenkoli namen.
- ✓ Popoln dostop do izvirne kode in uteži modela.

Identificirana 'ključna luknja'

Definicija ne zahteva objave celotnih podatkovnih zbirk za učenje.

Razlog: Modeli se lahko učijo na zaupnih virih (npr. zdravstveni kartoni).

Zahteva: Razvijalec mora zagotoviti dovolj informacij, da lahko 'usposobljena oseba zgradi bistveno enakovreden sistem.'

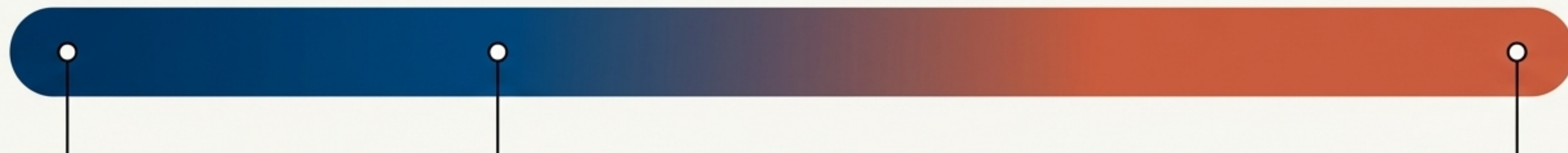


Realnost: Spekter odprtosti modelov

Študija Liesenfelda in Dingemanseja (junij 2024) je ocenila 40 modelov UI.
Ugotovitev: večina tako imenovanih 'odprtih' modelov ne izpolnjuje niti kriterijev OSI.

Popolnoma zaprt

Resnično odprt



ChatGPT

Ocenjen kot najbolj zaprt sistem v študiji.

Llama 3

Četrty najbolj zaprt. Ne izpolnjuje pogojev OSI: Meta ne razkriva podatkov za učenje in zahteva licenco za velike velike uporabnike.

OLMo 7B

Model neprofitne organizacije Ai2, ocenjen kot najbolj odprt.

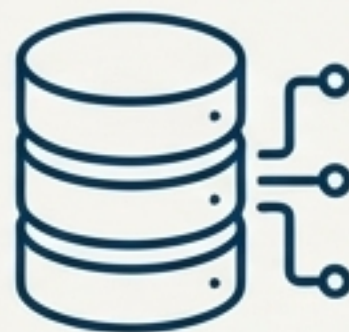
Iluzija dostopnosti: Praktične ovire za resnično odprtost

Argument Davida Widderja: Tudi tehnično popolnoma odprt model je lahko dejansko zaprt za večino uporabnikov.



Ogromna računska moč

Usposabljanje in zagon modelov zahtevata drage računalniške vire.



Dostop do podatkov

Gradnja modelov zahteva dostop do ogromnih podatkovnih zbirk.



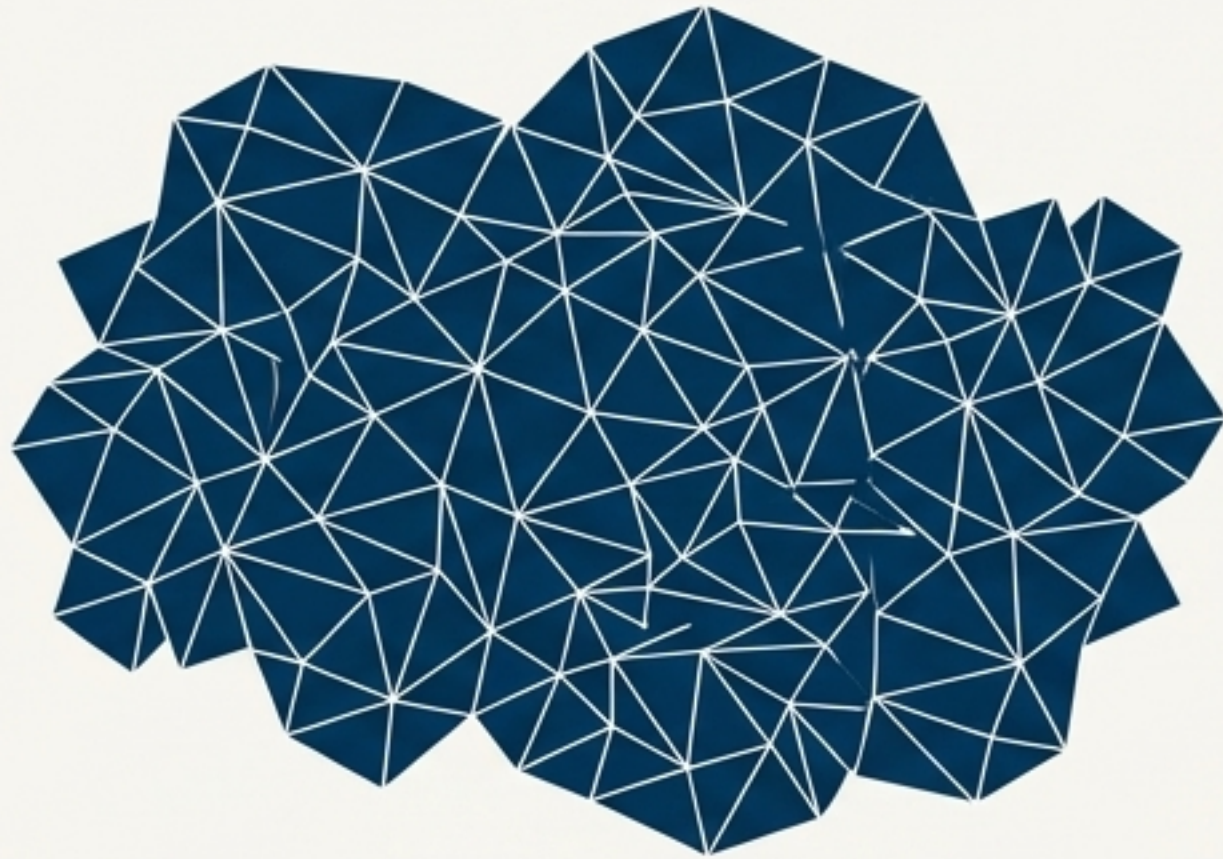
Strokovno znanje

Potrebna je delovna sila z znanjem o UI za gradnjo in fino uravnavanje modelov.

'Ideja, da "odprta UI nekako demokratizira UI", je zmotno prepričanje.'

— David Widder, Cornell Tech

Potencialna rešitev? Vzpon manjših in učinkovitejših modelov



Llama 3: 400+ milijard parametrov



Manjši modeli: Delujejo na prenosniku ali telefonu

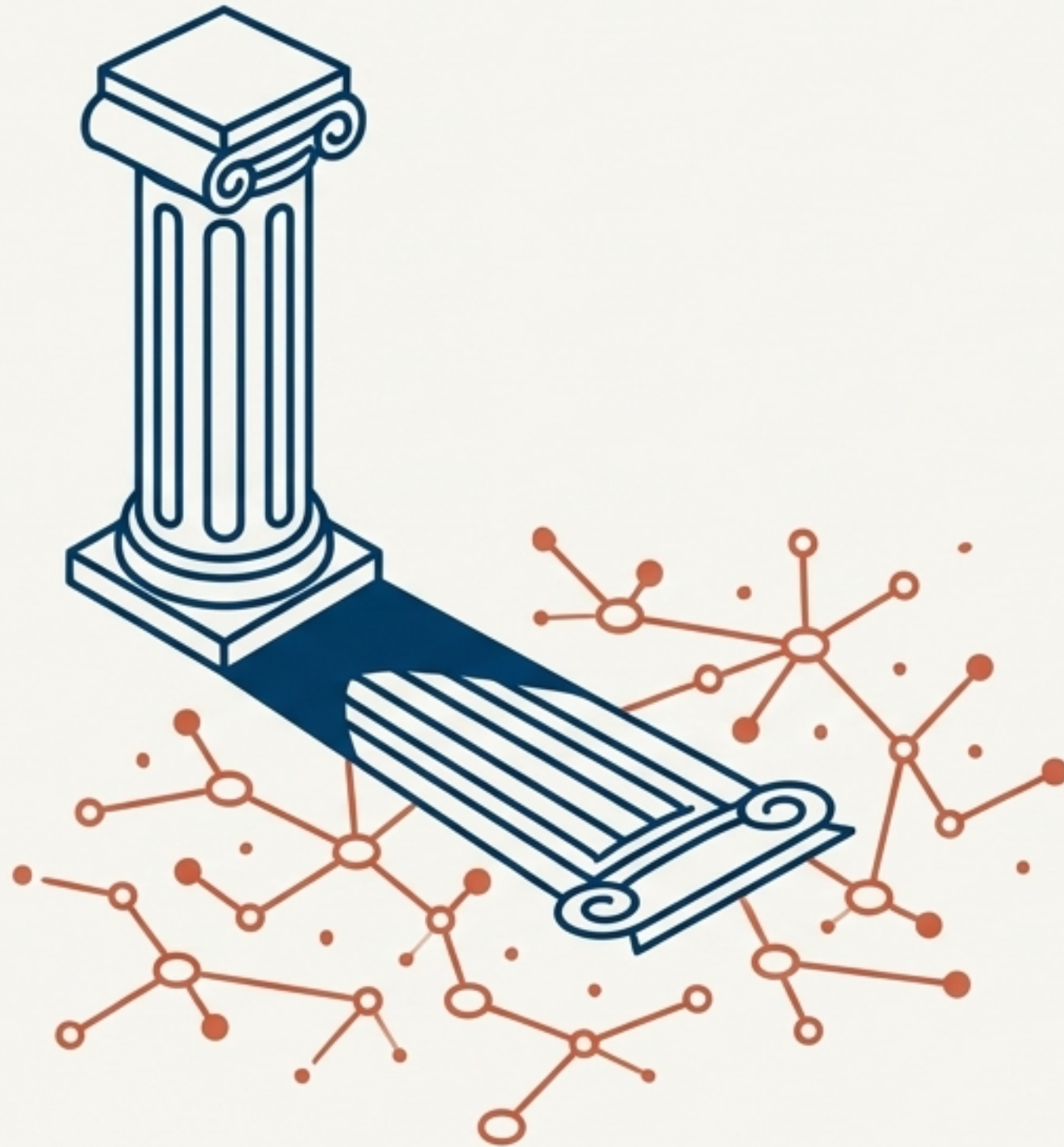
Ključni akterji v razvoju manjših modelov:

Apple, Microsoft in neprofitna organizacija Ai2 so objavili manjše modele, ki vključujejo uteži za učenje. Ai2-jevi modeli izpolnjujejo kriterije OSI.

Protiargument (Widder):

Tudi te manjše modele pogosto razvijajo institucije z obilico virov, zato problem dostopnosti ni v celoti rešen.

Regulatorji vstopajo v arena: Akt EU o umetni inteligenci



- **Pomen:** Akt EU o UI, ki je stopil v veljavo leta 2024, je opisan kot “najpomembnejši zakonodajni akt o UI na svetu danes.”
- **Ključne določbe:**
 - Določa celovita pravila za modele UI za splošne namene (npr. ChatGPT).
 - Zahteva, da razvijalci pojasnijo, kako so usposabljali in testirali svoje sisteme.
- **Pomembna izjema:** Odprtokodni sistemi so izvzeti iz nekaterih zahtev, kar pomeni, da je definicija ‘odprtosti’ ključnega pomena.
- **Globalni trend:** Regulativni zagon se krepi tudi drugod. V ZDA je leta 2024 45 zveznih držav obravnavalo zakone o UI.

Je 'varnost' zgolj regulatorni jarek?

Proti jedrski analogiji
(Chirag Shah):

"UI ni kot jedrska tehnologija. 'Lahko jo zgradite na prenosniku. To ni nekaj, kar bi lahko ustavili na enak način."



Skriti motiv (David Widder):

"Podjetja, kot je OpenAI, podpirajo regulacijo v imenu varnosti, vendar imajo morda skrite motive."

'Velik del retorike o "nevarni" UI je le slabo prikrit poskus zgraditi regulatorni jarek" za zaščito lastnih izdelkov.

Ključna spoznanja

1.



Osrednji konflikt:

Prihodnost UI je ujeta med obljubo odprtih inovacij in nujnostjo obvladovanja varnostnih tveganj.

2.



Kriza definicije:

Pojem 'odprt' je nedorečen in se pogosto zlorablja ('openwashing'), kar otežuje smiselno razpravo in regulacijo.

3.



Praktične ovire so realne:

Resnična dostopnost je omejena z ogromnimi zahtevami po virih, kar postavlja pod vprašaj ideal demokratizacije.

4.



Regulacija prevzema pobudo:

Ker industrija ne najde konsenza, regulatorji, na čelu z EU, postavljajo pravila igre, ki jih bo oblikovala tudi strateška uporaba varnostne retorike.

Zgodba se nadaljuje...

Razprava o odprtokodni UI še zdaleč ni končana.



Tehnološki napredek:

Razvoj manjših, učinkovitejših modelov.



Prizadevanja za standardizacijo:

Delo organizacij, kot je OSI, pri oblikovanju jasnih definicij.



Razvijajoča se globalna regulacija:

Izvajanje Akt EU o UI in podobnih zakonov po svetu.



Vir in reference

Leslie, M. (2025). Open-Source Artificial Intelligence—How Open? How Safe?. *Engineering*, 47, 8–10.

