Generativna umetna inteligenca

Ljupčo Todorovski

Univerza v Ljubljani, Fakulteta za matematiko in fiziko
Institut Jožef Stefan, Odsek za tehnologije znanja (E8)

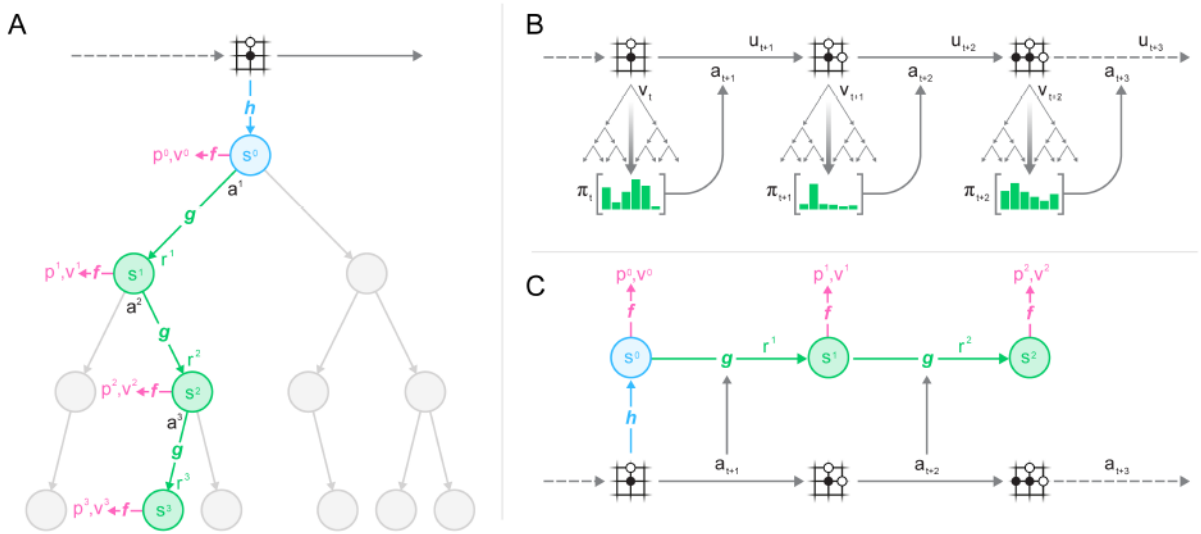
Oktober 2025

Pregled predavanja

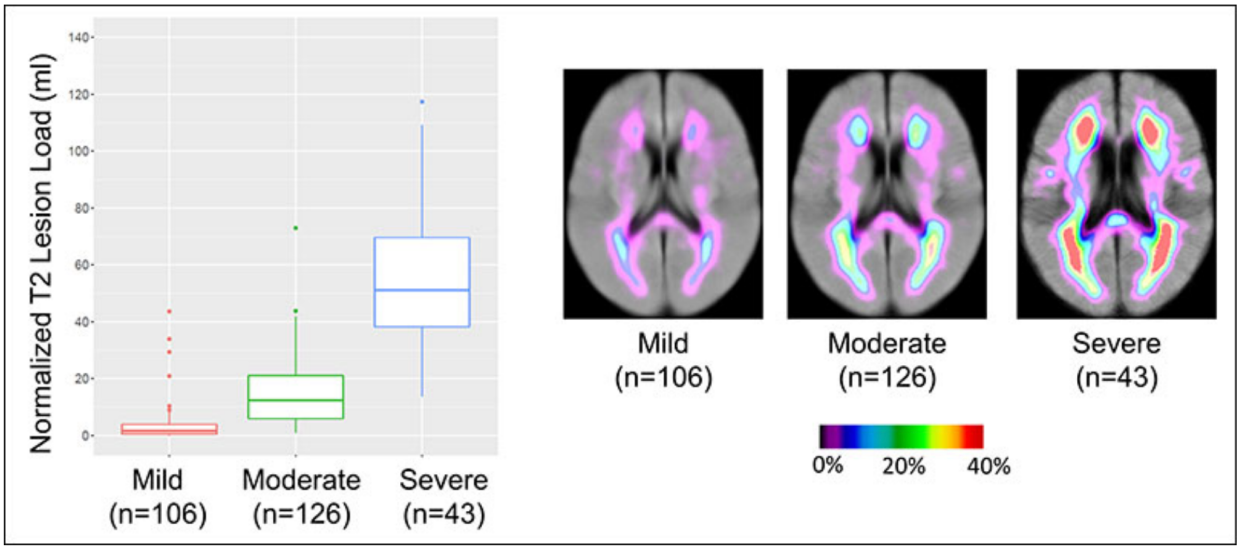
1. Zakaj se računalniki učijo?
   - Nekaj primerov
   - Ena anekdota
   - Zakaj, kaj in kako?
2. Kako se računalniki učijo?
   - Nevronske mreže
   - Učenje nevronskih mrež
   - Nadaljnji razvoj
3. Veliki jezikovni modeli in uporaba v izobraževanju

Razpoznavanje slik

Izbiranje (optimalnih) potez v igrah



Diagnosticiranje bolezni



Splošen vzorec

Podatki $\longrightarrow$ Odločitve

- Primeri že prepoznanih slik $\longrightarrow$ razpoznavanje nove slike
- Odigrane partije šaha $\longrightarrow$ izbira optimalne poteze v podani situaciji
- Podatki o preverjenih diagnozah $\longrightarrow$ diagnoza za novega pacienta

Ključno vprašanje: kako vzpostaviti povezavo med podatki in odločitvami?

John Snow, London, 1854

Od podatkov do modela

Porazdelitev primerov neenakomerna.

Od podatkov do modela

Porazdelitev primerov neenakomerna.

Primeri veliko bolj pogosti okoli ene vodne črpalke (rdeči krogec).

Modeli kot ključen manjkajoči člen verige

Podatki $\longrightarrow$ Modeli $\longrightarrow$ Odločitve

Epidemija kolere

Podatki lokacije primerov okužb in lokacije vodnih črpalk

Model povezava med lokacijami okužb in lokacijami črpalk

Odločitev župan Londona zapre izbrano vodno črpalko

Strojno učenje: vprašanja Zakaj-Kaj-Kako

Zakaj se računalniki učijo?

Da nam (sebi) omogočijo *boljše odločanje*.

Kaj se računalniki učijo?

Iz podatkov se učijo *modele* opazovanega okolja.

Kako se računalniki učijo?

V nadaljevanju predavanja.

Preden gremo na kako, še malo zgodovine o zakaj

- Turing in igra posnemanja (1950)
- Racionalni agenti in baze znanja (1970)
- Strojno učenje: Lenat vs. Feigenbaum

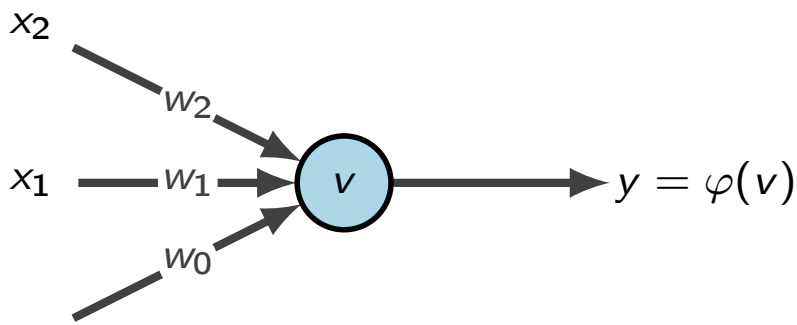
Podatki za učenje modela

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Model naj izračuna vrednost y iz podanih vrednosti x_1 in x_2

- Napovedni spremenljivki x_1 in x_2, ciljna y
- Model m izračuna $\hat{y} = m(x_1, x_2)$: ni nujno, da velja $\hat{y} = y$
- Je torej funkcija $m : \{0, 1\} \times \{0, 1\} \to \{0, 1\}$

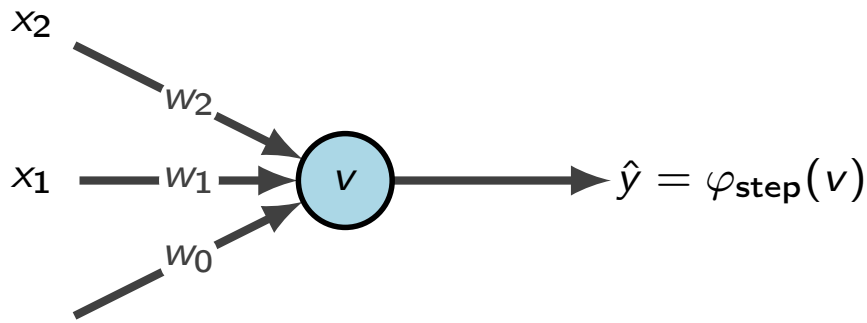
Osnovni gradnik: Nevron



Komponente nevrona: vhodi, stanje, izhod

- *Vhoda* nevrona x_1 in x_2
- *Sinapse* z utežmi w_1 in w_2 in prosta utež w_0
- Sinapse prenašajo signale med nevroni, vhodi in izhodi
- *Stanje* nevrona v
- *Funkcija aktivacije* prevede stanje v v *izhod* nevrona y

Najbolj preprosta nevronska mreža: Perceptron



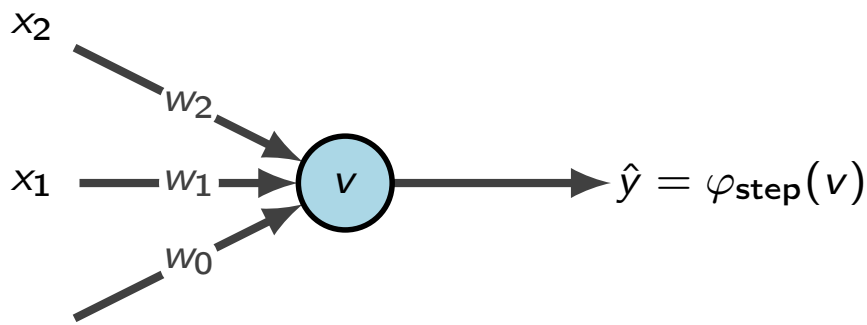
Posebnosti perceptrona

- Stanje je linearna kombinacija vhodov $v = w_0 + w_1 x_1 + w_2 x_2$
- Funkcija aktivacije je »stopnica«

$$\varphi_{\text{step}}(v) = \mathbb{1}_{v>0} = \begin{cases} 1, & \text{če } v > 0 \\ 0, & \text{sicer} \end{cases}$$

- $\boldsymbol{w} = [w_0\ w_1\ w_2]^T$ je vektor uteži, ki so predmet *učenja iz podatkov*

Učenje uteži: Pravilo delta



Za podan primer vrednosti x_1, x_2, y

- S perceptronom izračunamo $\hat{y}$
- Opazujemo napako izračuna, *izgubo* $(y - \hat{y})^2$
- V primeru napake, popravimo uteži s pravilom delta (Rosenblatt 1958)

$$\Delta \boldsymbol{w} = (y - \hat{y})[1\ x_1\ x_2]^T$$

- Uteži premaknemo v smeri negativnega *gradienta funkcije izgube*

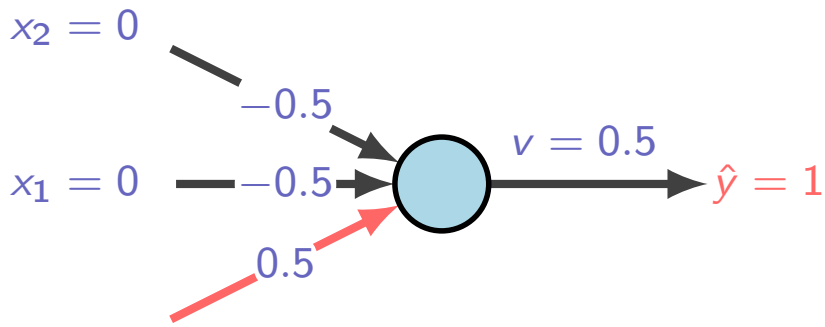
Iterativni algoritem gradientnega sestopa

Iterativno ponavaljamo dokler je napaka različna od 0

- Iz učnih podatkov naključno izberemo primer $(\boldsymbol{x}, y)$
- S perceptronom in pravilom delta izračunamo $\hat{y}$ in popravek $\Delta \boldsymbol{w}$
- Popravimo uteži perceptrona $\boldsymbol{w} = \boldsymbol{w} + \Delta \boldsymbol{w}$

Začnemo z naključnimi vrednostmi $\boldsymbol{w} = [0.5 \ -0.5 \ -0.5]^T$

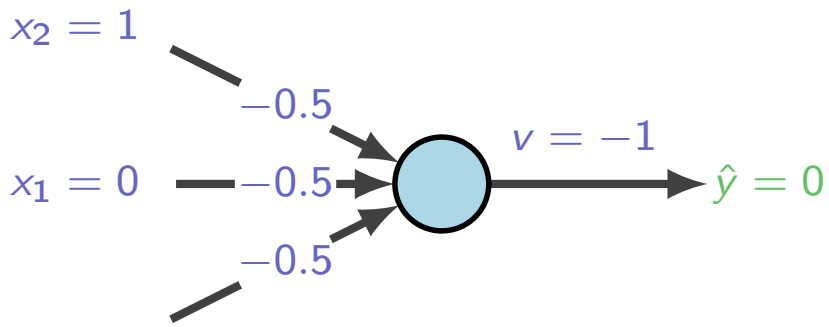
Prvi primer $x_1 = 0$, $x_2 = 0$, $y = 0$



$$\Delta \boldsymbol{w} = (y - \hat{y})[1 \ x_1 \ x_2]^T = [-1 \ 0 \ 0]^T$$
$$\boldsymbol{w} = [-0.5 \ -0.5 \ -0.5]^T$$

Nadaljujemo z $\boldsymbol{w} = [-0.5 \ -0.5 \ -0.5]^T$

Drugi primer $x_1 = 0$, $x_2 = 1$, $y = 0$

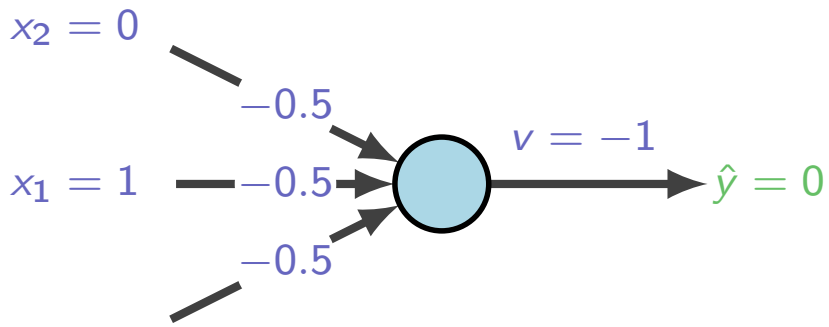


$$\Delta \boldsymbol{w} = (y - \hat{y})[1 \ x_1 \ x_2]^T = [0 \ 0 \ 0]^T$$

$$\boldsymbol{w} = [-0.5 \ -0.5 \ -0.5]^T$$

Nadaljujemo z $\boldsymbol{w} = [-0.5 \ -0.5 \ -0.5]^T$

Tretji primer $x_1 = 1$, $x_2 = 0$, $y = 0$

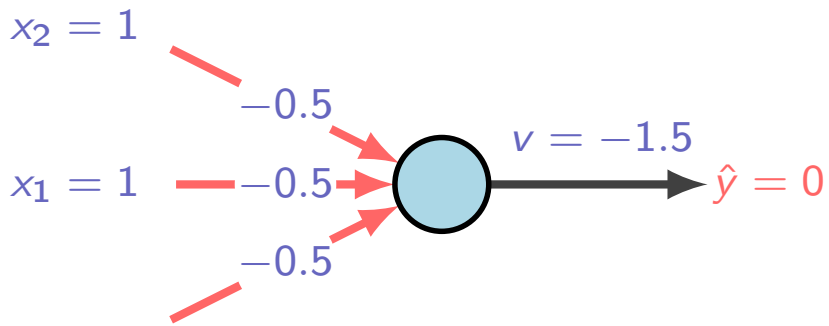


$$\Delta \boldsymbol{w} = (y - \hat{y})[1 \ x_1 \ x_2]^T = [0 \ 0 \ 0]^T$$

$$\boldsymbol{w} = [-0.5 \ -0.5 \ -0.5]^T$$

Nadaljujemo z $\boldsymbol{w} = [-0.5 \ -0.5 \ -0.5]^T$

Četrti primer $x_1 = 1$, $x_2 = 1$, $y = 1$

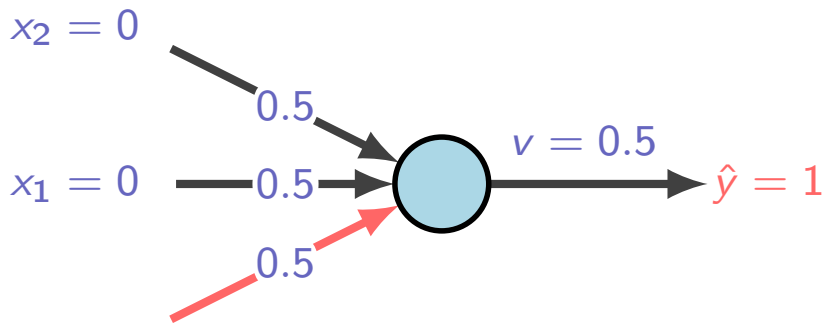


$$\Delta \boldsymbol{w} = (y - \hat{y})[1 \ x_1 \ x_2]^T = [1 \ 1 \ 1]^T$$

$$\boldsymbol{w} = [0.5 \ 0.5 \ 0.5]^T$$

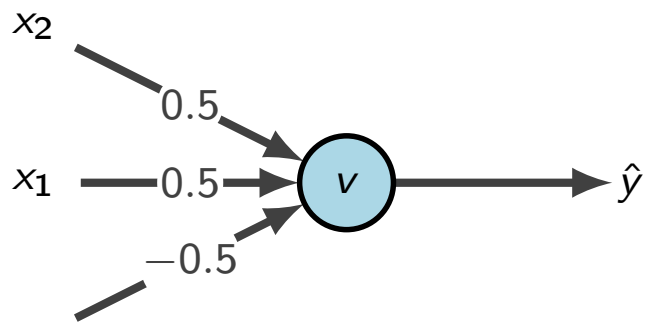
Nadaljujemo z $\boldsymbol{w} = [0.5\ 0.5\ 0.5]^T$

Prvi primer $x_1 = 0,\ x_2 = 0,\ y = 0$



$$\Delta \boldsymbol{w} = (y - \hat{y})[1\ x_1\ x_2]^T = [-1\ 0\ 0]^T$$
$$\boldsymbol{w} = [-0.5\ 0.5\ 0.5]^T$$

Končamo z $\boldsymbol{w} = [-0.5\ 0.5\ 0.5]^T$



Model pravilen za vse primere, preverimo

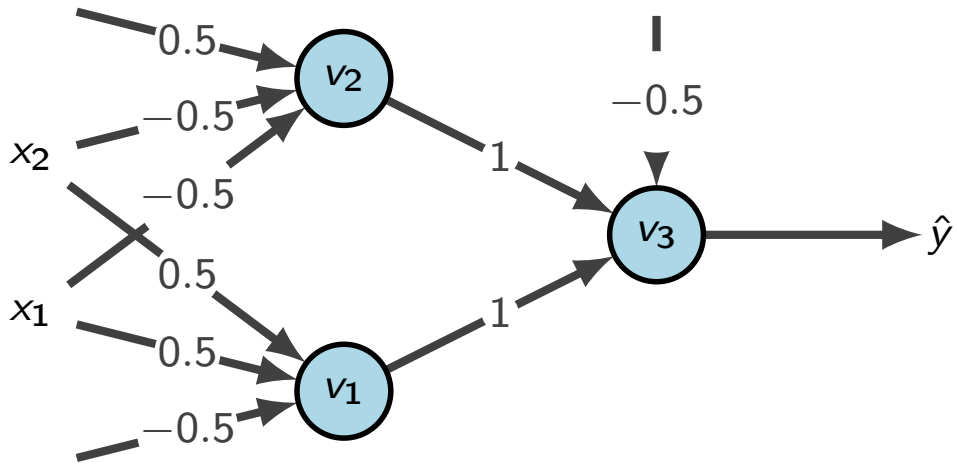
podatki			model		izguba
x_1	x_2	y	v	$\hat{y}$	$(y-\hat{y})^2$
0	0	0	-0.5	0	0
0	1	0	0	0	0
1	0	0	0	0	0
1	1	1	0.5	1	0

Perceptron ima omejeno kapaciteto: Ekskluzivni ali

Podatki za učenje modela

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Za modeliranje rabimo *večplastno* nevronsko mrežo



Časovnica (nadaljnjega) razvoja

Leto	Avtor-ji	Prispevek
1958	Rosenblatt	Perceptron in pravilo delta
1969	Minsky	Omejena kapaciteta, računska zahtevnost
1986	Rumelhart, Hinton, Wiliams	Algoritem vzvratnega razširjanja napake in večplastne nevronske mreže
1998	Lecun in ost.	Konvolucijske nevronske mreže za slike, globoke nevronske mreže
2014	Goodfellow	Generativni modeli, slike in besedila
2024	Hopfield, Hinton	Nobelova nagrada za fiziko

Napovedni in generativni modeli

Danes smo spoznali napovedne modele

- Deterministična različica $y = m(\boldsymbol{x})$
- Verjetnostna različica, model pogojne verjetnosti $P(y|\boldsymbol{x})$

Generativni modeli

- Verjetnostni modeli (pogojne) verjetnosti
- $P(w_0|w_{-1}, w_{-2}, w_{-3}, w_{-4}, \dots)$
- Napoved naslednje besede na osnovi nekaj prejšnjih

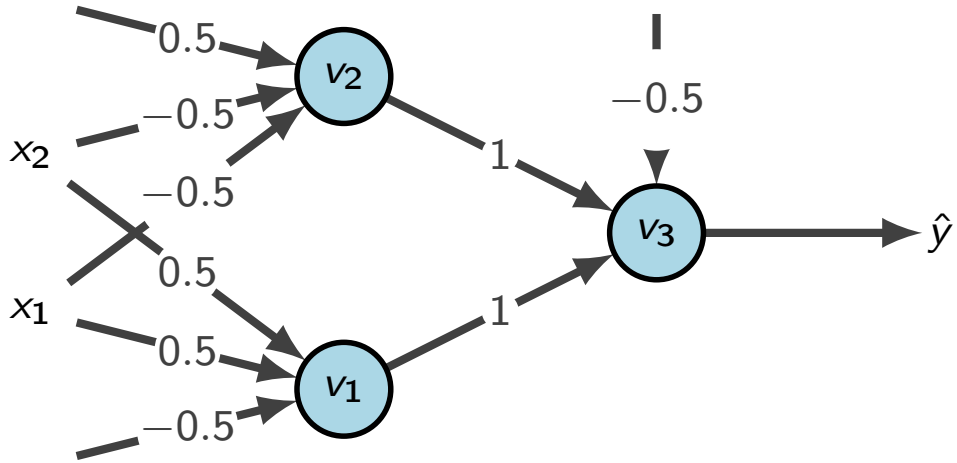
Uporaba generativnih modelov

Razvpiti modeli GPT, osnova za delovanje ChatGPT

- 175 *milijard* parametrov, velikost zapisa 800 GB
- Učna množica dokumentov več kot 500 milijard besed

- Odlično napoveduje naslednjo besedo, prepričljivo piše besedila
- Uporab(lj)en na mnogih področjih (npr. avtomatsko programiranje)
- Pozor: nemožnost ugotavljati pravilnosti (resničnosti) besedila

(Ne)transparentnost modelov: Ekskluzivni ali



Nevronska mreža nič ne nakazuje preproste resnice

- $y = 1$, če velja $x_1 = x_2$
- $y = 0$, sicer

ChatGPT

Je klepetalni agent, ki uporablja jezikovni model GPT

- Slednji je, kot smo videli prej, ogromna nevronska mreža
- Naučena iz vseh besedil dostopnih podjetju OpenAI
- Zato postopek učenja terja ogromno časa in računalniške moči
- Nevronska mreža modela GPT-3.5 ima več sto milijard sinaps
- Naučena je iz besedil, ki vsebujejo več sto milijard besed

- Za različico modela GPT-5 ni uradnih podatkov o velikosti

Torej, je ChatGPT umetna inteligenca?

Ja, ChatGPT je šibka umetna inteligenca.

Tri ravni umetne inteligence

Šibka UI	zmožna reševanja zelo konkretnih nalog, ki jih poda človek
Močna UI	zmožna razumevanja, učenja in splošne uporabe znanja za reševanje poljubnih nalog (nevronske mreže temu niso kos)
Super UI	superiorna vsem vidikom človeške inteligence

Je UI koristna?

Nedvomno je koristna

- Šibka UI nam lahko pomaga hitreje reševati probleme
- Omogoča hiter in enostaven dostop do ogromne količine znanja
- Številne možnosti za pospešek znanstvenega raziskovanja in poslovanja

Nevarnosti in težave: mitske in resnične

Mit: UI predstavlja eksistencialno tveganje za človeštvo

- Šibka UI ni avtonomna, ni samostojna pri odločanju
- Zato mit povzroča nepotrebno paniko
- Odvrača nas od debate o smotrni uporabi (generativne) UI

- Kljub hitremu razvoju, smo še zelo daleč od močne ali super UI

Mit je lahko nevaren, saj odvrača pozornost od resničnih nevarnosti UI.

Resnično pomembne težave in vprašanja: družbeni vpliv

Vpletanje UI v demokratične procese

- Spomnimo se vloge socialnih omrežij v kampanji Brexit
- Takrat so manipulacije javnega mnenja izvajali najeti ljudje
- UI lahko veliko bolj učinkovito manipulira in ustvarja lažne novice

Širjenje predsodkov in diskriminacije

- Povzame predsodke in diskriminacije v besedilih za učenje UI
- In jih zelo učinkovito ter na široko poustvarja in razširja

Resnično pomembne težave in vprašanja: nepreglednost

Slabo razumevanje in nepreglednost delovanja

- Zaradi ogromne velikosti nevronskih mrež
- Ne znamo dobro pojasniti in razumeti njihovo delovanje
- Zato ga tudi težje nadzorujemo

Nagnjenost k napakam in izmišljanju

- UI ne razume besedil in zato nima odnosa do njihove resničnosti
- V jezikovni model integrira tudi neresnice in zavajanja v učnih podatkih

Resnično pomembne težave in vprašanja: Širši vplivi

Zloraba intelektualne lastnine

- Jezikovni modeli se učijo iz avtorskih besedil
- Pripadajo človeštvu, ne zgolj podjetjem, ki razvijajo jezikovne modele
- Kako bodo ta podjetja kompenzirala avtorske pravice?

Vpliv na delovno okolje in ustvarjanje

- Izkoriščanje delavcev za pripravo učnih podatkov za UI
- Nadomeščanje delovnih mest in celih poklicev

Delovanje jezikovnih modelov terja velikanski ogljični odtis.

Edini racionalni odgovor: Regulacija razvoja in uporabe

Regulacija s politikami

- Definirajo pojem zaupanja vredne umetne inteligence
- Zahteve po razvoju UI, ki lahko razloži svoje odločitve
- Debata o omejevanju uporabe jezikovnih modelov in ChatGPT

- Tehnološki razvoj močno prehiteva javno debato o regulaciji

Regulacija na področju izobraževanja

Tukaj lahko ukrepamo pedagogi sami, debata v nadaljevanju.

Koncept odgovorne uporabe orodij UI

Zaenkrat jo večinoma definirajo uredništva znanstvenih revij.

Principa odgovorne uporabe

1. Prevzemanje popolne odgovornosti za napisano besedilo
2. Iskrenost in transparentnost pri navajanju virov

Princip popolne odgovornosti

ChatGPT ne sme biti naveden med avtorji besedila

- Avtorji smo lahko zgolj ljudje
- S tem ne prenašamo odgovornosti na UI in ChatGPT

- Jamčimo, da smo preverili pravilnost samodejno ustvarjene vsebine
- Preprečimo ustvarjanje neresnic za nadaljnje učenje jezikovnih modelov

Princip transparentnosti uporabe

V posebnem dodatku dela razkrijemo interakcijo s ChatGPT

- Obvezno za bistvene dele članka z vsebinskimi prispevki
- Manj bistveno za uvodne dele članka in npr. pregled obstoječega dela

Programska orodja za preverjanje vsebine

- Orodja za kontrolo upoštevanja principa transparentnosti
- Identificirajo dele besedila, ki so bili samodejno ustvarjeni z UI

- Pozor: Težave podobne tistimi z orodji protivirusne zaščite

Pogovor in dogovor namesto prepovedi

Učencem dovolimo uporabo in pri tem

- S primeri jih navajamo zastavljati dobra vprašanja
- Spodbujamo h kritični analizi odgovorov, spet primeri
- Zahtevamo razumevanje rešitev in odgovorov pridobljenih z UI

Ustno preverjanje znanja in reševanje testov brez uporabe UI.

Nekaj nasvetov

Običajen pogovor o vsakem predavanju

- Pripravljam predavanje o [temi X], kako bi ga strukturiral?
- Pomagaj mi s primerom, ki ilustrira [koncept Y].
- Izpelji mi [formulo Z].

Običajen pogovor pred vsako domačo nalogo in/ ali izpitom

- Za sestavljene naloge, preverjanje koliko dobre so rešitve UI
- Popravljanje besedila nalog tako, da so večji izziv za UI
- Sestavljanje nalog tipa: ugotovi napake v rešitvi UI in jih popravi
- Tako nastajajo (ideje za) presenetljivo dobre in zahtevne naloge

Namesto zaključka, nadaljnje branje

- S.J.D. Prince (2023) *Understanding Deep Learning*, The MIT Press. `http://udlbook.com`
- A. Narayanan in S. Kapoor (2024) *AI SNAKE OIL: What Artificial Intelligence Can Do, What it Can't, and How to Tell the Difference*. Princeton University Press. `https://www.normaltech.ai`
- S.J. Russell (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.