Zaim Mehić

prof. dr. Nina Bijedić

Umjetna inteligencija

10. Novembar 2024

Gödel-ov argument protiv jake umjetne inteligencije

Prije nešto više od 80 godina, tačnije 1931. godine na skupu u Königsberg-u , Kurt Gödel je objavio rezultate svog tadašnjeg rada, a koji se danas smatraju jednim od najvećih intelektualnih dostignuća modernog vremena – Gödel-ovi teoremi o nepotpunosti. Čitavih 33 godine trebalo je da se ti teoremi, preciznije prvi teorem, po prvi put dovedu u vezu sa ubrzanim razvojem računarskih mašina, i to kada je John R. Lucas tvrdio da ovaj teorem podrazumjeva da mašine neće moći dostići inteligenciju identičnu onoj koju posjeduje čovjek. Jaka umjetna inteligencija i njen razvoj su u direktnoj suprotnosti sa navedenom tvrdnjom pošto „razvoj jake umjetne inteligencije podrazumjeva kreiranje umjetne osobe: mašine koja ima sve mentalne moći koje i mi imamo, uključujući i fenomenalnu svijest“ [1]. Jedno je sigurno, a to je da još uvijek ne znamo kako bi se, ukoliko bi je i bilo moguće kreirati, jaka umjetna inteligencija ponašala i na koji način bi ona utjecala na čovječanstvo u cjelini. Prema tome, bilo kakav „neoborivi“ argument protiv iste zasigurno bi utjecao na put razvoja umjetne inteligencije i srodnih oblasti. Ovaj esej diskutuje o Gödel-ovom argumentu i njegovoj uspješnosti da ostane relevantan u domenu razvoja jake umjetne inteligencije. Fokus je stavljen na analizu samog Gödel-ovog teorema o nepotpunosti, njegovu promociju od strane John R. Lucas i Roger Penrose, te obradu ovog argumenta od strane pojedinih koji ga osporavaju.

Gödel-ov argument protiv jake umjetne inteligencije zasniva se prvom teoremu o nepotpunosti istoimenog naučnika iz 1931. godine. Prvi teorem o nepotpunosti glasi: „Bilo koji dosljedan formalan sistem u sklopu kojeg se može izraziti određena količina osnovne aritmetike nije kompletan, odnosno, postoje tvrdnje jezika u sistemu F koje se ne mogu dokazati, a ni osporiti u okviru tog sistema F“. [2]Koja tvrdnja bi ustvari mogla da potvrdi ovaj teorem? „Razmotrimo formulu koja kaže: 'Ova formula nije dokaziva unutar sistema'. Ako bi ova formula bila dokaziva unutar sistema, imali bismo kontradikciju: ako bi bila dokaziva unutar sistema, onda ne bi bila nedokaziva unutar sistema, pa bi formula 'Ova formula nije dokaziva unutar sistema' bila lažna. Podjednako, ako bi bila dokaziva unutar sistema, tada ne bi bila lažna, već bi bila istinita, jer u bilo kojem dosljednom sistemu ništa lažno ne može biti dokazano unutar sistema, već samo istine. Prema tome formula 'Ova formula nije dokaziva unutar sistema' nije dokaziva unutar sistema, nego je nedokaziva unutar sistema. Nadalje, ako je formula 'Ova formula nije dokaziva unutar sistema' nedokaziva unutar sistema, onda je istina da je formula nedokaziva unutar sistema, odnosno 'Ova formula je nedokaziva unutar sistema' je istina“. [3]Na ovaj način je J.R. Lucas u svom radu „Mind, Machines and Gödel“ povezao prvi teorem o nepotpunosti sa usporedbom između ljudskog uma i inteligentne mašine. Cijeli

[1] Stanford Encyclopedia of Philosophy. 12. Juli, 2018 – The Gödelian Argument Against „Strong AI“

[2] Stanford Encyclopedia of Philosophy. 11. Nov, 2013 – Gödel's Incompleteness Theorems

[3] Mind, Machines and Gödel – J. R. Lucas, str. 112.

fokus Gödel-ovog teorema jeste pokazati da nema nikakve „kvake“ i da se ovaj rezultat može uspostaviti najrigoroznijom dedukcijom; teorem važi za sve formalne sisteme koji su konzistentni i adekvatni za jednostavnu aritmetiku, te pokazuje da su ti sistemi nepotpuni.

Gödel-ov prvi teorem o nepotpunosti obrađivan je u više navrata sa akcentom na njegovu primjenu kod inteligentnih mašinskih sistema, a njegovu validnost u takvim sistemima ponajviše su izdvajali već pomenuti John R. Lucas, te Roger Penrose. John R. Lucas je između ostalog, postavio i tezu da je „naša ideja mašine takva da je njeno ponašanje u potpunosti određeno načinom na koji je kreirana i na nadolazeće podražaje: nema šanse da će mašina nešto uraditi sama“[4]. Lucas na ovaj način zaključuje da se mašina koja dobije određenu konstrukciju (set pravila) i ulazne informacije mora ponašati na neki tačno određeni način i ne ostavlja prostor ideji o mašinama sa vlastitom svijesti. Sve navedeno ne dovodi do zaključka da je čovjekov um superioran u odnosu na sve mašine, niti je to cilj Gödel-ovog argumenta. Njegov cilj je dati odgovor na pitanje: „Može li mašina biti jednaka ljudskom umu?“. Iako njegovi argumenti za mnoge nisu bili dovoljno uvjerljivi, bili su dovoljni da iniciraju debate koje su proizvele druge impresivne argumente koji su išli u korist ovog teorema, a i one koji nisu.

Fizičar Roger Penrose bio je jedan od neumornih promotera Lucas-ove argumentacije. Pri njegov pokušaj da opravda svog „heroja“ bila je knjiga „The Emperor's New Mind“ prvi put objavljena 1989. godine. U to vrijeme ova knjiga smatrana je najjačim napadom na ideju o jakoj umjetnoj inteligenciji do tada i on je svoju argumentaciju vezao za dvije linije. Prva linija, i jedina koju ćemo spomenuti, ima za cilj dokazati da je matematičko razmišljanje nešto što se ne može enkapsulirati unutar bilo kakvog čisto računarskog modela misli. Kako ova knjiga nije u potpunosti ispunila očekivanja, Robert Penrose objavljuje opširniju i detaljniju tezu u knjizi „Shadows of The Mind“ iz 1999. godine. Jedna od Penrose-ovih pretpostavki u ovoj knjizi jeste da „Matematičari (ljudi) ne koriste algoritam za koji se zna da je ispravan kako bi utvrdili matematičku istinu“[5]. Ovu tezu R. Penrose je izveo iz činjenice da postoje neki izračuni npr. *„Pronaći broj koji nije suma n kvadrata*“ ili *„Pronaći neparan broj koji je suma n parnih brojeva“* za koje se može (ili ne može) lako zaključiti da se njihov izračun nikada neće završiti (ili hoće završiti). Za sistem koji bi obuhvatio sve izračune, a i imao mogućnost da zaključi da li se neki izračun završava ili ne, uspio je (prema vlastitom mišljenju) dokazati da postoji kontradikcija i da postoji nešto što taj sistem ne može dokazati. Ova knjiga također nije izbjegla svoje kritičare, pa je Penrose istima odgovorio sa usavršenom verzijom temeljne Gödel-ove teze postavljene u knjizi „Shadows of The Mind“ u radu „Beyond the Doubting of A Shadow“ koja kaže da „pretpostavimo da se sve metode matematičkog zaključivanja koje su u principu ljudima dostupne možemo obuhvatiti nekim (ne nužno računarskim) ispravnim formalnim sistemom F. Ljudski matematičar, ako mu se predstavi sistem F, mogao bi da da zaključi sljedeće (imajući na umu da fraza „Ja sam F“ ovdje samo skraćeno znači „Sistem F obuhvata sve ljudima dostupne metode matematičkog dokazivanja“):

Iako ne znam nužno da sam F, zaključujem da, ako bih bio F, tada bi sistem F morao biti ispravan, a još važnije, F' bi također morao biti ispravan, gdje je F' sistem F dopunjen dodatnom tvrdnjom „Ja sam F“. Uviđam da iz pretpostavke da sam F slijedi da bi Gödelova izjava G(F') morala biti tačna i, nadalje, da ne bi

[4] Mind, Machines and Gödel – J. R. Lucas, str. 113.
[5] Shadows Of the Mind – Roger Penrose, str. 76

bila posljedica sistema F'. Međutim, upravo sam uvidio da ako bih bio F, tada bi G(F') morala biti tačna, a upravo je ovakvo uviđanje ono što bi F' trebao omogućiti. Pošto sam stoga sposoban uvidjeti nešto što nadilazi moći sistema F', zaključujem da ipak ne mogu biti F. Štaviše ovo se odnosi na bilo koji sistem umjesto F“ [6]

Kritičari ovakve teze koju je usavršio Penrose bili su Geoffrey LaForte,Patrick J. Hayes, Kenneth M. Ford, Selmer Bringsjord, Xiao, itd. U naučnom radu *„Why Gödel's theorem cannot refute computationalism“* autori LaForte, Hayes i Ford predstavili su niz razloga kao što su dvosmislenosti ključnih pojmova („ispravan“, „dokaz“, „algoritam“), zloupotrebe Gödel-ove teoreme, ograničenja ljudskog znanja, itd. Ovu argumentaciju sumirali su zaključkom da ni jedan Gödel-ov argument, uključujući onaj koji propagira Penrose, ne može opovrgnuti komputacionalizam. Oni naglašavaju da se računarska ograničenja identificirana kroz Gödel-ove i Turing-ove rezultate podjednako odnose na ljude kao i na mašine, čime se osporavaju tvrdnje da je ljudsko razmišljanje u suštini nemehnizabilno. Selmer Bringsjord i Hong Xiao u radu „*A refutation of Penrose's Gödelian case against artificial intelligence*“ kažu da Penrose nije uspio poboljšati položaj Lucas-ovih teza niti jednim od svojih argumenata koje je objavio u knjigama „*The Emperor's New Mind*“ i „*Shadows of the Mind*“. Smatraju da Ponrose-in Gödel-ov argument sadrži određene tehničke nedostatke i logičke neistine. Oni tvrde da Penrose-in argument ne demonstrira činjenicu da je ljudski um nekomputabilan i da umjetna inteligencija ne može dostići nivo ljudskog razuma. Stewart Shapiro je u svom radu „*Mechanism, Truth, and Penrose's New Argument*“ kritiku usmjerio na demonstriranje kako Penrose-in novi argument ne „obara“ mehanističku tezu da se ljudski um može uspješno modelirati kao digitalni računar ili Turing-ova mašina.

Zaključno, uzeli smo u obzir pripadnike i pozicije i opozicije kako bismo prikazali razvoj Gödel-ov argument i doprinos vrhunskih članova akademske zajednice njegovom razvoju. Pobornici ove ideje nastojali su pokazati da računari nikada neće uspjeti obuhvatiti svu širinu ljudskog uma i razmišljanja, pošto ljudi, kako oni smatraju, odluke ne donose koristeći set predefinisanih algoritama, nego koristeći ljudsku intuiciju koja nije premostiva u mašinski i računarski svijet sastavljen od tranzistora i složenijih elektronskih sklopova. Oporba je kritikovala način na koji su dokazi formulisani, zaključivši da su neki dokazi netačni, kontradiktorni ili rezultirali paradoksalnim rješenjima. Bilo kako bilo, Gödel-ov argument ostaje važan dio razvoja naučnih polja kao što su matematika, filozofija i umjetna inteligencija.

[6] Beyond the Doubting of a Shadow, 3. The Central New Argument of Shadows

Reference

Beyond the Doubting of a Shadow. (1996). *PSYCHE*.

Lucas, J. (1959). Mind, Machines and Gödel.

Penrose, R. (1994). *Shadows of the Mind.* Oxford University Press.

Stanford Encyclopedia of Philosophhy. (n.d.).