FILOZOFIA
Volume 79, Number 5, 2024

DOI: https://doi.org/10.31577/filozofia.2024.79.5.1

Cesta k inteligencii: Nadmerné zjednodušenie a samokontrola

DANIEL C. DENNETT, Formerly Center for Cognitive Studies, Tufts University, Medford, Massachusetts, USA

DENNETT, D. C.: A Route to Intelligence: Oversimplify and Self-monitor
FILOZOFIA, 79, 2024, No 5, pp. 471 – 482

This article offers some reflections on some features of human consciousness that are in need of an explanation. I wish to see if these features could be understood as solutions to design problems, solutions arrived at by evolution, but also, in the individual, as a result of a process of unconscious self-design. I consider this issue in the context of work in AI on the attempt to design intelligent robots– not "bed-ridden" expert systems, but systems that have to act in real time in the real world. How then does one partition the task of the robot so that it is apt to make reliable real time decisions? Any real, finite deliberator must partition the states of its world in such a way as to introduce the concept of possibility. This idea of partitioning the world into "possible" alternatives that remain "open" is very clearly the introduction of a concept of *epistemic* possibility. The article concludes with some reflections on the Frame Problem.

Keywords: action – anticipation – decision – design – system

Nedávno som premýšľal nad tým, ako by sa dali vysvetliť niektoré vlastnosti ľudského reflexívneho vedomia, ktoré si, podľa mňa, veľmi vyžadujú vysvetlenia. Snažím sa zistiť, či by sa tieto vlastnosti dali chápať ako riešenia pre konštrukčné problémy, riešenia, ku ktorým dospela evolúcia, ale aj ako výsledok procesu nevedomého seba-projektovania u jednotlivca. Snažil som sa o tom premýšľať v kontexte práce v oblasti umelej inteligencie, ktorá sa týka snahy navrhnúť inteligentné roboty – nie expertné systémy „pripútané na

Táto práca bola pôvodne publikovaná ako „A Route to Intelligence: Oversimplify and Self-monitor" v zbierke *From Electrons to Elephants and Elections: Exploring the Role of Content and Context*, eds. Shyam Wuppuluri – Ian Stewart, Cham: Springer 2022, 587 – 595. Pôvodný príspevok bol prednesený v roku 1984 na workshope, ktorý sa konal v Maison Française v Oxforde a ktorý organizoval Jean Khalfa, a zúčastnili sa na ňom René Thom, Richard Gregory, ja a ďalší.

lôžko", ale systémy, ktoré musia konať v reálnom čase a v skutočnom svete. Ak chcete rozmýšľať o niečom, musíte sa pomerne vzdialiť od experimentov a tvrdých empirických údajov; musíte sa dostať do pomerne špekulatívnej roviny. Avšak zdá sa, že snahy ľudí v oblasti umelej inteligencie vedú k presvedčeniam – ak nie k dôkazu – o rôznych konštrukčných prekážkach, ktoré sa javia ako veľké a neodvratné. Ak dokážeme pochopiť, prečo systém – alebo orgán či vzorec správania – musí mať určité vlastnosti alebo určitú štruktúru k tomu, aby mohol plniť svoju úlohu, môže nám to pomôcť klásť správne otázky, alebo nám to aspoň zabráni v tom, aby sme sa pri snahe vysvetliť mechanizmy v mozgu, ktorý je zodpovedný za inteligentné konanie, zaoberali niektorými nesprávnymi otázkami.

Dovoľte mi pripomenúť, že inteligentné konanie v reálnom svete závisí od očakávania, a to dvojakého druhu: jednak zabudované, rýchle, nevedomé modulárne očakávanie, a zároveň, v prípade ľudských bytostí a možno aj u niektorých iných vyšších druhov, ide o niečo, čo sa oveľa viac podobá na dobrovoľné, vedomé vytváranie očakávaní a kalkulácií o budúcnosti.

Existuje dôležitá rodina slovies, ktorá napodiv ešte nevzbudila filozofickú pozornosť. Ústrednými členmi tejto rodiny sú „vyhnúť sa", „zabrániť", „zamedziť", „podporiť" a azda najzákladnejšie zo všetkých je „zmeniť" v prechodnom význame, kedy si myslíme, že sa jedna vec, činiteľ alebo udalosť, „aktívne" mení. Ide o dominantné slovesá *konania*, pri ktorých sa situácia charakterizuje v zmysle racionálneho činiteľa, ktorý sa, ako sa hovorí, chystá „zmeniť chod dejín". Toto je zaujímavá fráza. Všetci chceme mať možnosť zmeniť chod dejín, hoci len v rámci našich malých kútikov sveta. Problém slobodnej vôle je do veľkej miery otázkou toho, či si človek myslí, že môže zmeniť chod dejín, ale samozrejme, táto známa fráza sa aj pri najpovrchnejšej analýze ukáže ako hlboko mätúca. Ak predpokladáte, že sa má chápať vo svojom doslovnom význame, tak je absurdná. Ako by ste mohli zmeniť chod dejín? Z čoho na čo? Ak sú dejiny jednoducho sledom udalostí, ktoré sa skutočne stali, potom ich samozrejme nemôžete zmeniť. Ľudia hovoria, že nemôžete zmeniť minulosť, a to je celkom pravda, ale potom zároveň nemôžete zmeniť ani budúcnosť.

Keď človek uvažuje v tomto duchu, keď premýšľa o uskutočnení takýchto zmien, ktoré chce tak veľmi zrealizovať, musí mať na zreteli *očakávané* dejiny, smer, akým sa dejiny búdu uberať *ceteris paribus*, ako sa budú vyvíjať, ak niekto niečo neurobí, alebo *kým* niekto niečo neurobí, alebo *napriek tomu*, čo niekto urobí. Tieto slovesá pôsobenia nemôžu mať oporu mimo rámca predpokladanej, očakávanej histórie, a to ani vtedy, keď sa používajú na charakteristiku účinkov spôsobených celkom neživými predmetmi. Dovoľte mi,

aby som to ilustroval príkladom, ktorý som si požičal z knihy *Elbow Room* (Dennett 1984b).

Predstavte si, že astronómovia objavia meteoroid, ktorý smeruje k Zemi, a vypočítajú, že v utorok zasiahne Severnú Ameriku, a neexistuje nič, čo by s tým niekto mohol urobiť. Ľudia by, samozrejme, prepadli zúfalstvu, pýtali by sa, či sa dá niečo spraviť, a možno by sa modlili za zázračné vyslobodenie z tejto strašnej katastrofy. A následne predpokladajme, že v predvečer zničenia sa objaví ďalší meteoroid, ktorý sa vyrúti z najtemnejšieho vesmíru na dráhu, ktorá je práve tá správna na to, aby vychýlila prvý meteorit na trajektóriu, po ktorej nás tesne minie, a tak na poslednú chvíľu *zabráni* katastrofe a *predíde* nešťastiu.

Pri takejto príležitosti by nám tieto slová prirodzene napadli. Naznačujem však, že druhý meteoroid bol zázrak – Božia odpoveď na naše modlitby? Nie, predpokladám, že druhý meteoroid tam bol odjakživa a smeroval presne na rovnakú dráhu ako ten prvý; astronómovia si ho jednoducho všimli až v poslednej chvíli. V skutočnosti, ak by si všimli druhý meteoroid vtedy, keď si všimli ten prvý, nikdy by nás nevystrašili, pretože (ako teraz s odstupom času vidia a mohli si to vtedy spočítať) *by nikdy nedošlo ku katastrofe*. Bola to len očakávaná katastrofa – nesprávne očakávaná katastrofa. Zdá sa, že je vhodné hovoriť o odvrátenej katastrofe alebo katastrofe, ktorej bolo zabránené, pretože porovnávame očakávanú históriu s tým, ako sa veci vyvinuli, a nachádzame udalosť, ktorá bola „kľúčovou" udalosťou vzhľadom na odchýlku medzi týmto očakávaním a skutočným priebehom udalostí, a toto nazývame „aktom" odvrátenia alebo zabránenia.

Mark Twain raz povedal: „Som starý muž a videl som veľa problémov, ale väčšina z nich sa nikdy nestala." Toto vychádza zo skúsenosti človeka, ktorý je zvyknutý na život vo svete vyhýbania sa a predchádzania. Toto je svet, v ktorom žije racionálne uvažujúci človek. Takto uvažujúci človek má pohľad na svet, ktorý sa neustále díva dopredu a predvída, akým smerom sa veci budú vyvíjať, pokiaľ niečo neurobí, alebo dokedy niečo neurobí.

Predpokladajme, že chceme navrhnúť robota, ktorý bude žiť v skutočnom svete a bude sa vedieť rozhodovať tak, aby mohol presadzovať vlastné záujmy – nech ho umelo obdarujeme akýmikoľvek záujmami. Inými slovami, chceme navrhnúť prezieravého plánovača. Ako treba štruktúrovať schopnosti – reprezentačné a deduktívne alebo výpočtové schopnosti – takejto bytosti? Problém, ktorému takáto bytosť čelí, je, ako je to už pri umelej inteligencii zvyčajné, problém kombinatorickej explózie. Spôsob, akým nadobúdame

očakávanie je ten, že si vyberáme vzorové trajektórie vecí vo svojom percepčnom svete a takto získané informácie používame na vyvodenie alebo extrapoláciu budúcej trajektórie veci. Nemožno sa rozumne zaoberať ničím, čo nemožno sledovať týmto spôsobom. Keď hovorím o sledovaní, nemám na mysli len sledovanie trajektórií vecí pohybujúcich sa v priestore, ale taktiež trajektórie v čase, akými sú zásoby potravín, ročné obdobia, miera inflácie, relatívna politická sila protivníkov, dôveryhodnosť atď. Existuje nekonečne veľa vecí, ktoré by sa dali sledovať, ale pokus o sledovanie všetkého, o obsiahnutie aktuálnych informácií o všetkom, zaručene vedie k sebadeštruktívnemu záchvatu informačného preťaženia. Bez ohľadu na to, koľko informácií už o danej problematike človek má, vždy ich môže mať viac a často vie, že by ich mohol mať viac, keby si len našiel čas na ich zozbieranie. Vždy je možné uvažovať viac, takže trik spočíva v tom, ako navrhnúť tvora tak, aby robil spoľahlivé, ale nie bezchybné rozhodnutia v časových medziach, prirodzene daných udalosťami v jeho svete, na ktorých mu záleží.

Základným problémom je teda to, čo by sme mohli nazvať problémom Hamleta, ktorý, ako si spomínate, premárnil svoj čas úvahami (alebo to tak aspoň vyzerá), váhaním a odkladaním. O takomto odkladaní hovoril na workshope v roku 1984 René Thom, vo svojom príklade muža na priechode pre chodcov, ktorý si musí vybrať. Človek sa musí rozhodovať v reálnom čase, a to znamená, že ak chce vôbec uspieť, musí odviesť prácu, ktorá nie je úplne ideálna. Preto musí byť od počiatku navrhnutý tak, aby bol úsporný, aby vedel prehliadnuť *väčšinu* dostupných informácií.

Ako možno potom zadeliť úlohy robota tak, aby bol schopný urobiť spoľahlivé rozhodnutia v reálnom čase? Možno urobiť jednu vec, a to vyhlásiť, že niektoré veci vo svete tohto tvora sa považujú za *fixné*; nebude sa vynakladať žiadne úsilie na ich sledovanie, za účelom získania ďalších informácií. Podmienky týchto vlastností budú ustanovené v axiómach, ktoré sú však do systému zabudované *bez akýchkoľvek reprezentačných nákladov*. Človek jednoducho navrhne systém tak, aby pracoval dobre za predpokladu, že svet je taký, aký predpokladá, že vždy bude, a nezabezpečí, aby systém pracoval dobre („správne“) aj za iných podmienok. Systém ako celok funguje tak, *akoby* mal byť svet vždy rovnakým, takže to, či je svet naozaj taký, nie je otázkou, ktorá by mohla byť predmetom rozhodovania. Dobrým príkladom je predpoklad pevnej väzby v ľudskom videní, ktorý opísal Ullman (1979). Pravdepodobne ide o konštrukčnú funkciu, ktorú v priebehu vekov upevnil prirodzený výber. V minulosti boli dôležité veci, ktoré sa pohybovali v našom vizuálnom okolí, zvyčajne zostavy väzieb, ktorých časti sú tuhé (ruky, zápästia, ramená,

lakte atď.), a je možné vytvoriť omnoho efektívnejší vizuálny systém pre tvora pomocou takéhoto sveta tým, že sa do neho jednoducho zapracuje predpoklad pevnosti. Toto umožní veľmi rýchle výpočty pre rýchlu identifikáciu a extrapoláciu budúcností príslušných častí sveta. Zvyšné veci vo svete treba vyhlásiť za *nepovšimnuteľné*, aj keď by v zásade mohli byť povšimnuteľné, ak by sa tým dalo niečo získať. Ide o veci, ktoré síce nie sú fixné, ale ktorých zmeny nemajú priamy vplyv na blaho tvora. Tieto veci sú v našom vnímaní akoby rozmazané a viac sa im nevenujeme. Príkladom z Wimsatta (1980) je rozdiel v kognitívnej stratégii dvoch rôznych dravcov: hmyzožravého vtáka a mravčiara, ktorú obaja potrebujú, aby dokázali sledovať pohybujúci sa hmyz. Hmyzožravý vták sleduje jednotlivé kusy lietajúceho hmyzu a vzorkuje ich trajektórie pomocou rýchlej techniky snímania: ide o veľmi vysokú mieru fúzie mihotania v porovnaní s ľudským zrakom. (Ak by ste takémuto vtákovi ukázali film, videl by ho ako prezentáciu snímok, a nie ako nepretržitý pohyb.) Vták vidí jednotlivé druhy hmyzu *ako* jednotlivcov. Mravčiar nesleduje jednotlivé mravce. Mravčiar vidí roje mravcov ako várky jedlej látky. (Keby som veril, že je vždy vhodné hovoriť týmto spôsobom, povedal by som, že „mravec" bol masový pojem v jazyku mravcov!) Pohlcuje várky mravcov a neplytvá kognitívnymi zdrojmi na to, aby sledoval jednotlivé mravce rovnako tak, ako my nesledujeme jednotlivé molekuly, keď zachytíme „prenikajúci" jednotný zápach v objeme vzduchu, ktorý môže obsahovať niekoľko častíc na miliardu významnej molekuly.

„Hrúbka zrna" nášho vlastného vnímania sa môže líšiť; rozlišovanie detailov je funkciou nášho vlastného kalkulu blahobytu vzhľadom na naše potreby a iné schopnosti. V prípade nášho systému, rovnako ako v prípade iných tvorov, dochádza ku kompromisu pri vynakladaní kognitívneho úsilia a vývoji efektorov rôzneho druhu. U hmyzožravého vtáka teda dochádza ku kompromisu medzi rýchlosťou fúzie mihotania a veľkosťou jeho zobáka. Ak má širší zobák, dokáže na jeden ťah obsiahnuť väčší objem, a teda má väčšiu toleranciu voči chybe pri výpočte polohy jednotlivých častí koristi.

Ak sa potom niektoré veci vo svete považujú za nemenné a iné za nepostrehnuteľné, a tým pádom sa len spriemerujú, ostávajú nám veci, ktoré sa menia a o ktoré sa oplatí zaujímať. Tieto sa delia zhruba na dve časti: sledovateľné a chaotické. Chaotické veci sú tie, ktoré nedokážeme bežne sledovať, a pre účely nášho uvažovania s nimi musíme zaobchádzať ako s náhodnými, nie však v kvantovo-mechanickom zmysle, a dokonca ani v matematickom zmysle (napríklad ako s informačne nestlačiteľnými), ale len v zmysle pseu-

donáhodnosti. Ide o vlastnosti sveta, ktoré sú vzhľadom na vynaložené kognitívne úsilie, ktoré je tvor pripravený vyvinúť, nezachytiteľné; ich budúci stav je nepredvídateľný.

To znamená, že každý reálny, koncový deliberátor si musí rozdeliť stavy svojho sveta tak, aby vedel zaviesť pojem možnosti: je možné, že položka n bude v stave A, a je možné, že položka n bude v stave B alebo v stave C. Získame súbor ekviposibilných (ale nie nevyhnutne ekviprobovateľných) alternatív. Táto myšlienka rozdelenia sveta na „možné" alternatívy, ktoré zostávajú „otvorené", je veľmi jasne zavedením pojmu epistemickej možnosti. Ide o to, čo je možné vzhľadom na vedomosti konkrétneho agenta. Čím viac vedomostí agent získava, o to viac sa môže zmenšiť množina možností. „Kedysi som si myslel, že stav B je možný, ale vzhľadom na to, čo som sa práve dozvedel, si uvedomujem, že možný nie je" (pozri Dennett, 1984b, 151 – 152).

Sellars (1963, 1966) veľmi užitočne rozlišuje medzi tým, čo nazýva zjavným obrazom a vedeckým obrazom. Zjavný obraz je každodenný obraz sveta, svet makroskopických, pevných, farebných predmetov a iných osôb alebo racionálnych činiteľov. Je to svet ľudovej fyziky a ľudovej psychológie. Potom je tu vedecký obraz: svet atómových a subatómových častíc, ktoré sú príliš malé na to, aby ich bolo možné vnímať voľným okom, svet síl a svetelných vĺn. Sellars svoje rozlišovanie robí tak, že sa zameriava na zjavný obraz, ktorý zdieľajú (normálne) ľudské bytosti, ale myslím si, že jeho rozlišovanie môžeme užitočne rozšíriť aj na iné druhy. Sme jediný druh, ktorý vyvinul vedu, a tak máme vedecký obraz sveta, sveta, v ktorom žijeme my a iné druhy, napriek obrovským rozdielom v našich zjavných obrazoch tohto sveta. Predpokladám, že zjavný obraz, ktorému sa druh teší, je určený súborom konštrukčných „rozhodnutí", ktoré rozdeľujú veci v jeho prostredí do kategórií pevných, nepovšimnuteľných, sledovateľných alebo chaotických. (Je dôležité poznamenať, že tento spôsob uvažovania o zjavnom obraze druhu trochu odporuje konotáciám prídavného mena „zjavný" – keďže nepredpokladá nič o vedomí. Vôbec nie je vylúčené, že úplne nevedomý tvor – napríklad náš imaginárny robot – by mal zjavný obraz.)

Prečo sme jediný druh, ktorý si vytvoril vedecký obraz popri našom zjavnom obraze – a trochu v rozpore s ním? O tejto téme už bolo písané často, preto sa pozastavím len pri jednom bode. Konštrukčné princípy, ktoré v prvom rade vytvárajú zjavný obraz, vytvárajú aj prvky neukončenosti, ktoré môžu viesť k jeho rozpadu. Niektoré z inžinierskych „skratiek", ktoré sú nevyhnutné, ak sa chceme vyhnúť kombinatorickej explózii, majú podobu igno-

rovania – posudzovania akoby neexistujúcich – malých zmien vo svete. Sú obdobou „zaokrúhľovacej chyby" pri počítačovom prepočítavaní čísel. A podobne ako pri zaokrúhľovaní, tak aj pri týchto lokálne neškodných zjednodušeniach sa môžu za určitých podmienok naakumulovať veľké chyby.

Ak si potom systém dokáže všimnúť veľkú chybu a diagnostikovať ju (aspoň približne), môže začať konštruovať vedecký obraz. Boli sme napríklad navrhnutí tak, aby sme „priamo" detegovali len tie zmeny, ktoré sa vyskytujú v určitom rozsahu rýchlostí. Mimo nášho priameho zorného poľa sa odohrávajú zmeny, ktoré sa dejú príliš rýchlo alebo príliš pomaly na to, aby sme ich mohli vnímať bez pomoci napríklad časozberných alebo slow-motion fotografií. Nedokážeme vidieť, ako rastlina alebo dieťa rastú z okamihu na okamih. Pohyb Slnka voči Zemi dokážeme vidieť len pri východe alebo západe Slnka, alebo pomocou jednoduchého nástroja, ktorý je rozšírením našich zmyslov – stačí pár tyčí zapichnutých do zeme. Ale v priebehu niekoľkých minút v druhom prípade, alebo mesiacov či rokov v prípade rastlín alebo detí, zachytíme rozdiel: naše očakávania, že nedôjde k žiadnej zmene (nula plus nula plus nula… sa rovná nule), sú vyvrátené. Teraz je minimálnou, nevýraznou reakciou na to jednoducho urobiť korekcie uprostred našich extrapolácií trajektórie a pokračovať ako predtým. Bystrou reakciou je všimnúť si, že tak musíme urobiť (často), a predstaviť si zmeny *príliš malé na to, aby sme ich mohli vidieť*, čím vstúpime do vedeckého sveta postulovaných, neviditeľných javov. Takto dochádza k značnému posunu videnia na základe rôznorodej samokontroly – najmä na základe spozorovania vzorca vo vlastných kognitívnych reakciách.

Dovoľte mi, aby som sa vrátil k zjavnému obrazu nášho anticipujúceho plánovateľa s jeho „otvorenou budúcnosťou" typov epistemicky možných udalostí, ktoré sú pre neho dôležité, ale ktoré bežne nemôže postrehnúť.

Toto sú alternatívy, o ktorých môže uvažovať a musí uvažovať, ak sa chce vo svete udržať. Jedným z popredných druhov epistemicky možných udalostí je kategória vlastného konania agenta. Tie sú pre neho systematicky nepredvídateľné. Môže sa pokúšať sledovať, a tak predvídať rozhodnutia a činy iných agentov, ale (z pomerne zrejmých a dobre známych logických dôvodov, napríklad v probléme Haltingu v informatike) nemôže robiť detailné predpovede svojich vlastných činov, pretože mu hrozí nekonečný regres sebakontroly a analýzy. Všimnite si, že toto neznamená, že náš tvor nemôže robiť nejaké hraničné predpovede svojich vlastných rozhodnutí a činov. A tak môžem spoľahlivo predpovedať rozhodnutia, ktoré urobím v blízkej budúcnosti: zajtra pri raňajkách sa rozhodnem, koľko šálok čaju vypijem,

a práve teraz predpovedám, že sa rozhodnem, že si dám viac ako nulu a menej ako štyri. Ak má byť náš tvor schopný vybrať si z alternatív, ktoré si dokáže predstaviť, akými stratégiami uvažovania by sme ho mali vybaviť? Jednou z funkcií, ktorú chceme zabudovať, je tá, ktorú spomenul Réné Thom: musíme sa chrániť pred možnosťou, že sa proces hodnotenia skončí nerozhodne – klasický problém Buridanovho osla. Lacným spôsobom, ako zabezpečiť toto bezpečnostné opatrenie, je zakomponovať do systému čosi funkčne porovnateľné k hodu mincou: ľubovoľné pseudonáhodné „orákulum", ktoré je k dispozícii pre pomoc pri rozhodovaní, kedykoľvek to systém potrebuje. Fascinuje ma špekulácia Juliana Jaynesa (1976), podľa ktorej rôzne tradície poverčivého rozhodovania a predpovedania, vyskytujúce sa v starovekom svete – hádzanie kostí a žrebov, pozeranie do vnútorností zvierat, konzultácie s veštcami, čítanie čajových lístkov – sú v skutočnosti viac-menej nevedomky vynájdené prvými ľuďmi, aby sa dostali z pozície Buridanovho osla alebo z trochu príbuznej situácie (môžeme povedať Hamletovej) človeka, ktorý jednoducho nevie, ako účinne uvažovať o komplikovanej situácii, a napriek tomu potrebuje konať včas. Keď sú otázky príliš zložité, keď si človek nevie predstaviť žiadne iné okolnosti, ktoré by problém vyriešili, keď jednoducho nevie, ako pokračovať vo svojich úvahách, tu, tak ako v prípade chodca na priechode pre chodcov, môže byť cenné jednoducho sa pohnúť jedným alebo druhým smerom. Z dlhodobého a priemerného hľadiska nezáleží na tom, ktorým smerom sa pohnete, pokiaľ sa dostanete zo stavu rozhodovacej neistoty a pohnete sa. Jaynes predpokladá, že tieto rituály mali za následok to, že rozhodovali za ľudí, ktorí sa nevedeli rozhodnúť sami za seba. Išlo teda o účelové barličky alebo protézy. Spomínam ich tu preto, lebo sú názorným príkladom niečoho, čo nebolo navrhnuté a prenesené geneticky prírodným výberom, ale čo bolo skôr kultúrnym artefaktom, nevedome navrhnutým jednotlivcami.

Pre niekoho je slovné spojenie „nevedome navrhnuté" oxymoron, ale to, čo pod tým mám na mysli, je celkom prosté: niektorí jednotlivci sa náhodne pustili do týchto zvláštnych správaní bez toho, aby mali na zreteli nejaký cieľ, ale zistili, že im prinášajú príjemné výsledky, a tak sa za istých okolností stali populárnymi správaniami. A tak boli rituály podrobené ďalšiemu zdokonaľovaniu a následne sa zachovali prostredníctvom kultúrneho prenosu. Takto prenášaná stratégia správania pravdepodobne nemá žiadny špecifický, organický (neurónový) riadiaci systém (počítačovo povedané, žiadny „špecializovaný hardvér"), ale je skôr len softvérom, súčasťou „virtuálneho stroja" ľudského rozhodovania, ktorý je formovaný kultúrnymi a inými faktormi prostredia, a rôzne implementovaný v jednotlivých riadiacich štruktúrach.

Najzákladnejším problémom, ktorému čelí dizajnér takéhoto deliberátora, je to, čo umelá inteligencia nazýva rámcový problém. Keďže som tento nevyriešený problém obšírne opísal na inom mieste (Dennett 1984a), poznamenám tu len to, že ho môžeme vnímať ako problém efektívneho riadenia zjavného obrazu plánovača, a to tak, aby prijaté informačné alebo reprezentačné skratky priniesli anticipácie, ktoré sú včasné a spoľahlivé. Nazýva sa to rámcový problém, kvôli takzvaným rámcovým axiómam, ktoré sa zrejme musia použiť na systematické stanovenie druhov konštánt účinku, ktoré sa predpokladajú v každom konkrétnom zjavnom obraze. Aké sú napríklad (hrubé, spoľahlivé, normálne) účinky premiestnenia jednej veci na inú? Môžeme takéto chápanie kodifikovať do definovania axióm pre typ akcie *presunúť x na y*?

Toto by mala byť pomerne bazálna činnosť v repertoári každého zaujímavo schopného agenta a okamžite ju rozpozná každý, kto pozná slávny „blokový svet" umelej inteligencie – imaginárny stolový svet pozostávajúci z niekoľkých farebných, rôzne tvarovaných blokov, ktorými môže pohybovať a ukladať ich na seba pomocou rovnako imaginárneho robotického ramena (pozri napríklad SHRDLU, in Winograd, 1972). Tento svet je v porovnaní s reálnym svetom každého, aj veľmi prostého tvora, zarážajúco jednoduchý. Ale aj v tomto zmenšenom svete sa črtá rámcový problém. Zvážte niektoré z potrebných rámcových axióm:

(1) Ak z ≠ x, potom ak presuniem x na y, tak ak z bolo predtým na w, tak z bude následne na w.
(2) Ak je z modré, potom ak presuniem x na y, z je potom modré.
(3) Ak je z červené, potom ak presuniem x na y, z je potom červené.

Naozaj potrebujeme samostatné, nezávislé axiómy pre všetko, čo sa nemení? Ak áno, definícia každého typu akcie bude musieť obsahovať klauzuly pre každý predikát, ktorý je možné použiť v opise stavov, v bezhlavom množstve axióm – zrejme inžinierska obludnosť. Nemôžeme mať nejaké všeobecnejšie, základné axiómy, napríklad v tom zmysle, že farby vecí sa nemenia?

(4) (Pre všetky x) (Ak je x červené, x zostane červené).

To nebude stačiť, pretože jedným z typov akcií, ktoré môžeme chcieť zahrnúť do repertoáru, je *zafarbiť x na červeno*, čo však vylučuje (4) a jej príbuzné pod hrozbou kontradikcie. Nevyriešeným problémom je to, ako vytvoriť systém reprezentácie poznania sveta, ktorý by bol dostatočne jednoduchý a účinný na to, aby sa vyhol kombinatorickej explózii, a zároveň dostatočne pružný

a citlivý na to, aby sa dokázal spamätať aspoň z niektorých hlúpych dôsledkov jeho zámerného zjednodušenia.

Nikto zatiaľ nemá dobré riešenie rámcového problému, a už vôbec nie ja, ale tvrdím, že jedným z prvkov každého dobrého riešenia budú úrovne sebapovšimnutia. Na záver stručne opíšem dva príklady toho, čo tým myslím. Kedysi som mal psa, ktorý rád aportoval hodené tenisové loptičky, ale keď mal na trávniku dve loptičky a nedokázal ich obe naraz udržať v papuli, rýchlo ich prehadzoval sem a tam, jednu pustil, aby mohol chytiť druhú, potom uvidel pustenú loptičku a hneď znova otvoril papuľu, aby po ňu išiel, a tak ďalej. Urobil to tak možno dvadsať alebo tridsaťkrát, zrejme sa riadil nejakým príliš jednoduchým pravidlom, *získať* je lepšie ako *udržať*. Toto zlé pravidlo mal v sebe viac-menej zabudované – nikdy sa ho neodnaučil –, ale nezomrel na následky jeho dodržiavania. To znamená, že týmto pravidlom nebol natoľko posadnutý, aby sa ním riadil, až pokým nepadne mŕtvy od hladu. Po tých niekoľkých desiatkach opakovaní sa v ňom niečo preklopilo a prestal. Nemusel vedieť, prečo prestal. Mal minimálny ochranný mechanizmus – nejako citlivý na „nadmerné" opakovanie vlastnej reakcie – ktorý ho zastavil a umožnil mu vybrať sa na nejakú sľubnejšiu cestu.

Podobný prípad nedávno opísal Geoffrey Hinton v prednáške na MIT o architektúre Boltzmannovho stroja, ktorú vyvinul spolu s Terrym Sejnowskim (Hinton – Sejnowski, 1983a; Hinton – Sejnowski 1983b). Boltzmannove stroje sú výkonnými riešiteľmi problémov v určitých tradične náročných problémových oblastiach, ale majú svoje charakteristické slabiny. Typický problém si graficky predstavte ako úlohu nájsť najnižšie miesto – globálne minimum – v rozsiahlom teréne s mnohými priehlbinami – lokálnymi minimami. (Toto je, samozrejme, len „lezenie do kopca" obrátené naruby!) Boltzmannove stroje sú efektívnymi hľadačmi globálnych miním za rôznych podmienok, ale môžu sa zaseknúť v nezvyčajných terénoch.

Uvažujme o teréne, ktorý pretína strmá roklina, ktorá sa v dolnej časti mierne spúšťa ku globálnemu minimu. Keď Boltzmannov stroj počas svojho prieskumu „vstúpi" do takejto rokliny, v podstate sa sám seba pýta: „Ktorým smerom sa mám vydať, aby som mohol zostúpiť?" a lokálne hľadá najstrmšie klesanie. Iba na samom dne rokliny je „viditeľný" mierny sklon smerom k východisku; na všetkých ostatných miestach bude línia pádu (v lyžiarskom žargóne) približne v pravom uhle k tomuto smeru. Pri miernom prevýšení sa Boltzmannov stroj ocitne niekde na opačnom svahu rokliny, znova si položí svoju otázku a vydá sa na opačný svah. V rokline bude oscilovať tam a späť,

neuvedomujúc si pritom márnosť svojho pátrania. Boltzmannov stroj uväznený v takomto prostredí stráca svoju normálnu rýchlosť a účinnosť a stáva sa príťažou pre každý organizmus, ktorý sa naň spolieha.

Ako pri tejto príležitosti poznamenal Hinton, to, čo človek v takejto situácii chce, je, aby si systém dokázal „všimnúť", že sa dostal do takéhoto repetitívneho cyklu, a aby sa sám preorientoval na iný postup. Konštrukčné riešenie, ktoré by sme mohli uprednostniť, nespočíva v zavrhnutí myšlienky Boltzmannovho stroja, preto, že má túto slabinu, ale v kompenzácii takejto slabiny pomocou nejakej ad hoc stratégie kontroly a riadenia. Myslím si, že práve táto stratégia sa ukáže ako endemická pri návrhu inteligentných riadiacich systémov.

Preložil Dominik Kulcsár

Literatúra

DENNETT, D. C. (1984a): Cognitive Wheels: The Frame Problem of AI. In: Hookway, C. (ed.): *Minds, Machines and Evolution*. Cambridge: Cambridge University Press, 129 – 151.

DENNETT, D. C. (1984b): *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, Massachusetts: MIT Press, Bradford Books.

HINTON, G. E. – SEJNOWSKI, T. J. (1983a): Analyzing Cooperative Computation. In: *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*. Rochester, New York (May 1983).

HINTON, G. E. – SEJNOWSKI, T. J. (1983b). Optimal Perceptual Inference. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington DC (June 1983).

JAYNES, F. J. (1976): *The Origins of Consciousness in the Breakdown of the Bicameral Mind*. Boston: Houghton Mifflin.

SELLARS, W. (1963): *Science*. London: Routledge & Kegan Paul.

SELLARS, W. (1966): Fatalism and Determinism. In: Lehrer, K. (ed.): *Freedom and Determinism*. New York: Random House.

ULLMAN, S. (1979): *The Interpretation of Visual Motion*. Cambridge, Massachusetts: MIT Press.

WIMSATT, W. (1980): Randomness and Perceived Randomness in Evolutionary Biology. *Synthese*, 43, 287 – 329.

WINOGRAD, T. (1972): *Understanding Natural Language*. New York: Academic Press.

Daniel C. Dennett
Formerly Center for Cognitive Studies
Tufts University
115 Miner Hall
Medford, MA 02155
USA
ORCID ID: https://orcid.org/0000-0003-1181-3093