NICK BOSTROM

SUPERINTELIGENȚA

CĂI, PERICOLE, STRATEGII

LITERA®

Superintelligence
Paths, Dangers, Strategies
Nick Bostrom

Ediție publicată pentru prima dată în limba engleză în 2014
Ediție publicată prin înțelegere cu Oxford University Press

LITERA®
Editura Litera
O.P. 53; C.P. 212, sector 4, București, România
tel: 021 319 6390; 031 425 1619; 0752 548 372;
e-mail: comenzi@litera.ro

Ne puteți vizita pe
www.litera.ro

Superinteligența
Direcții, pericole, strategii
Nick Bostrom

Traducere din limba engleză:
Doru Căstăian

Editor: Vidrașcu și fiii
Redactor: Isabella Prodan
Corector: Georgiana Enache
Copertă: Flori Zahiu
Tehnoredactare și prepress: Ana Vârtosu

Descrierea CIP a Bibliotecii Naționale a României
BOSTROM, NICK
Superinteligența. Direcții, pericole, strategii /
Nick Bostrom; trad.: Doru Valentin Căstăian –
București: Litera, 2016
Index
ISBN 978-606-33-0281-7

I. Bostrom, Nick
II. Căstăian, Doru Valentin (trad.)
612.82

Cuprins

Liste cu figuri, tabele și casete

Listă cu figuri

Listă cu tabele

Listă cu casete

Povestea neterminată a rândunicilor

În plin sezon de cuibărire, după zile întregi de muncă neobosită, rândunicile s-au așternut, în amurg, la un taifas, trăgându-și sufletul:

– Suntem atât de mici și de neajutorate! Imaginați-vă cât de ușoară ar fi viața noastră, dacă ne-ar ajuta o bufniță să ne facem cuibul!

– Da, spuse o alta. Iar apoi bufnița ar putea avea grijă de bătrâni și de pui!

– Ne-ar putea da sfaturi și ar putea să stea cu un ochi pe pisică! spuse o a treia.

Atunci Pastus, cea mai vârstnică dintre păsări, zise:

– Să trimitem iscoade în cele patru zări, poate găsește pe undeva un pui de bufniță abandonat sau măcar niște ouă. Un pui de cioară ar fi la fel de bun sau poate unul de nevăstuică. Acesta ar fi cel mai grozav lucru care ni s-ar putea întâmpla, exceptând, desigur, inaugurarea Pavilionului Grânelor Interminabile, din curtea din spate.

Stolul începu să se agite, iar rândunicile ciripeau în toate direcțiile. Doar Scronkfinckle, o rândunică având un singur ochi, ceva mai liniștită de felul ei, nu părea foarte impresionată de aceste spuse înțelepte. Ea zise:

– Cum să facem una ca asta? N-ar trebui, înainte să aducem o astfel de vietate printre noi, să ne interesăm cum se domesticește, cum se îmblânzește?

Pastus îi răspunse:

– Să îmblânzim o bufniță pare ceva extraordinar de greu! Ne va fi greu până și să găsim un ou! Să începem de aici.

După ce vom izbuti să creștem bufnița, ne vom ocupa și de problema aceasta.

– Dar planul e greșit! spuse Scronkfinckle, însă nu-l mai auzi nimeni, întrucât stolul își luase deja zborul, hotărât să pună în aplicare ideea lui Pastus.

Doar două, trei rândunele rămăseseră pe urmă. Împreună, începură să se gândească cum puteau fi bufnițele îmblânzite sau domesticite. Dar și-au dat seama repede că Pastus avusese dreptate: sarcina se dovedea prea dificilă, mai ales că nu aveau o bufniță pe care să exerseze. Cu toate acestea, au continuat să caute o soluție, temându-se să nu se întoarcă stolul, cumva, cu un ou și să le prindă nepregătite.

Nu știm cum se va termina povestea, dar autorul îi dedică această carte lui Schronkfinckle și discipolilor lui.

PREFAȚĂ

Craniul tău găzduiește în interiorul lui ceva care te ajută să citești. Acest lucru, creierul uman, are anumite abilități pe care creierele animalelor nu le au. Abilitățile respective ne-au dat, în definitiv, poziția dominantă în natură. Restul animalelor au mușchi mai puternici și colți mai ascuțiți, noi avem creiere mai inteligente. Modestul nostru avantaj în privința inteligenței generale ne-a făcut să dezvoltăm limbajul, tehnologia și organizarea socială complexă. Avantajele speciei noastre s-au adunat de-a lungul timpului, întrucât fiecare generație a construit pe ceea ce i-au lăsat predecesorii ei.

Dacă, într-o zi, vom realiza creiere artificiale care să le depășească intelectual pe cele umane, superinteligența aceasta ar putea deveni extrem de puternică. În plus, așa cum soarta gorilelor depinde, acum, mai mult de oameni, decât de gorile, soarta speciei noastre va ajunge să depindă, în același fel, mai degrabă de acțiunile superinteligenței artificiale. Avem totuși un avantaj. Noi suntem cei care construim mașinile. În principiu, am putea realiza o superinteligență care să protejeze valorile umane. Am avea motive întemeiate să facem asta. În practică, problema controlului – adică modul în care am controla ce va face superinteligența – pare destul de dificilă. De asemenea, cel mai probabil, șansa noastră, din acest punct de vedere, va fi unică. Dezvoltarea unei superinteligențe ostile va face imposibilă înlocuirea sau modificarea preferințelor acesteia. Soarta ne-ar fi pecetluită.

Prin această carte, caut să înțeleg provocarea pe care ar putea-o constitui superinteligența și modul în care am putea reacționa față de ea. Aceasta este, probabil, cea mai importantă și mai amplă provocare cu care s-a confruntat vreodată umanitatea. Și, indiferent că vom avea succes sau nu, va fi, probabil, ultima acțiune de acest nivel pe care o vom avea de înfruntat.

Cartea de față nu susține că suntem în pragul unei revoluții a inteligenței artificiale, nici nu spune când ar putea avea aceasta loc. Probabil că acest lucru se va petrece la un moment dat, în acest secol, dar nu știm sigur. Primele două capitole au în vedere câteva căi către superinteligență și ridică o serie de probleme cu privire la plasarea lor în timp. Cea mai mare parte a cărții va fi, însă, despre ceea ce se va întâmpla ulterior. Studiem, de asemenea, cinetica unei dezvoltări explozive a inteligenței, formele și abilitățile acesteia, precum și alegerile strategice pe care le are la dispoziție un agent superinteligent care a dobândit un avantaj decisiv. Ne vom îndrepta, după aceea, atenția asupra problemei controlului și vom analiza ce e de făcut pentru a modela condițiile inițiale în vederea supraviețuirii și a obținerii unor rezultate benefice. Spre finalul cărții, vom încerca să realizăm o imagine de ansamblu, ca rezultat al investigațiilor noastre. Vom da anumite soluții pentru ceea ce trebuie făcut acum, cu scopul de a evita pericolele de mai târziu.

Nu a fost o carte ușor de scris. Sper ca drumul pe care îl înfățișează aceasta îi va ajuta pe ceilalți cercetători să descopere noi frontiere mai ușor și mai agreabil, pe care să le depășească încrezători și cu forțe proaspete, în vederea unor eforturi care să ne amplifice nivelul înțelegerii. (Iar, dacă drumul înfățișat este puțin anevoios și cu hârtoape, sper ca aceia care vor citi cartea să nu subestimeze *ex ante* ostilitatea terenului.)

Nu a fost o carte ușor de scris. Am încercat să fac în așa fel încât să fie simplu de citit, neștiind dacă am reușit în totalitate nici măcar acest lucru. Atunci când am elaborat-o, m-am gândit la mine, pe când eram mai tânăr, și am încercat să scriu o carte așa cum mi-ar fi plăcut mie atunci să citesc. Poate că publicul ei nu va fi, așadar, numeros. Dar cred că informațiile prezentate ar trebui să fie accesibile cât mai multor oameni, dacă îi interesează cu adevărat și nu întâmpină fiecare idee nouă cu unul dintre clișeele disponibile în cultura lor. Cititorii neinstruiți temeinic în domeniu nu trebuie să fie descurajați de micile referiri matematice sau de vocabularul specializat, pentru că vor reuși întotdeauna să deducă informațiile din context. (Invers,

cititorii care vor explicații mai tehnice, vor putea să afle mai multe din notele de subsol.)[1]

Multe dintre punctele de vedere din această carte sunt, probabil, greșite.[2] Se poate, de asemenea, să existe considerații de importanță critică pe care să nu fi reușit să le am în vedere și care să invalideze unele dintre concluziile mele. Am izbutit, într-o oarecare măsură, să punctez diversele grade ale incertitudinii față de anumite puncte de vedere exprimate în carte, folosind termeni precum „posibil", „ar putea", „poate", „ar putea foarte bine să...", „foarte probabil", „aproape sigur". Fiecare dintre acești cuantificatori a fost plasat cu intenție și cu multă atenție. Totuși, acești indicatori ai modestiei epistemice, distribuiți textual, nu sunt de ajuns: trebuie suplimentați prin admiterea de principiu a incertitudinii și a caracterului failibil. Nu este o falsă modestie: deși cred că există șanse bune ca volumul meu să fie substanțial eronat și să inducă opinii greșite, cred, totodată, că alternativele din literatura de specialitate sunt cu mult mai proaste – aici intrând și viziunea actuală, a ipotezei nule, conform căreia, în acest moment, putem ignora, fără riscuri și în mod rezonabil, posibilitatea apariției superinteligenței.

[1] Nu toate notele conțin, totuși, informații importante.
[2] Nu știu care.

Pentru ajutorul acordat la pregătirea manuscrisului, le mulțumesc lui Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, Kristian Rönn, Susan Rogers, Andrew Snyder-Beattie, Cecilia Tili și lui Alex Vermeer. Vreau să-i mulțumesc în special editorului meu, Keith Mansfield, pentru încurajările susținute de pe parcursul proiectului.

Le cer scuze tuturor celor pe care am omis să-i menționez aici.

Nu în ultimul rând, le mulțumesc celor care au finanțat acest proiect, familiei și prietenilor: fără ajutorul vostru, această carte n-ar fi existat.

1. Perfecționări trecute și abilități prezente

Să începem prin a privi în urmă. Istoria, înțeleasă la cea mai largă scară, pare să indice existența unor moduri diferite de dezvoltare, fiecare dintre ele fiind mai rapid decât cel de dinaintea lui. Această tendință ne face să credem că este posibil un nou mod de dezvoltare (unul chiar mai rapid). Și, totuși, nu punem mare preț pe această observație – cartea de față nu este despre „dezvoltare tehnologică" ori despre „creștere exponențială" și nici nu cuprinde idei definite generic prin eticheta „excentricitate". În cele ce urmează, vom trece în revistă istoria inteligenței artificiale. Vom evalua, apoi, potențialul prezent al acestui domeniu. Și, în fine, vom vedea câteva opinii ale experților și ne vom recunoaște ignoranța în ceea ce privește dezvoltarea viitoare a fenomenului.

Tipuri de progres și istoria la scară mare

Cu doar câteva milioane de ani în urmă, strămoșii noștri încă stăteau atârnați de crengile copacilor din pădurile Africii. Din perspectivă evoluționistă și la scară geologică, detașarea lui *Homo sapiens* de ultimul strămoș pe care l-am avut în comun cu celelalte maimuțe mari s-a petrecut rapid. Am deprins mersul pe două picioare, ne-am ales cu degete opozabile și am suferit mici modificări ale volumului și structurii creierului – crucială schimbare! –, care au dus la un mare salt în materie de abilități cognitive. În consecință, oamenii pot gândi abstract, dezvoltă raționamente complexe și înmagazinează și transmit informații culturale de la o generație la alta mai bine decât orice altă specie de pe planetă.

Aceste abilități le-au permis oamenilor să realizeze tehnologii tot mai performante, ceea ce i-a determinat pe strămoșii noștri să lase în urmă pădurea tropicală și savana. În special

după descoperirea agriculturii, populația totală a lumii a crescut și, odată cu ea, și densitatea. Creșterea demografică a facilitat apariția unui număr mai mare de idei; densitatea sporită a făcut posibilă răspândirea mai rapidă a acestora, unii indivizi dezvoltând abilități specializate. Aceste perfecționări au sporit *rata progresului* în materie de productivitate economică și de capacitate tehnologică. Ceea ce s-a întâmplat ulterior, în special după Revoluția Industrială încoace, a generat o a doua schimbare globală, comparabilă în ceea ce privește progresul.

Modificările ratei progresului au avut urmări importante. Cu câteva sute de mii de ani în urmă, în timpul preistoriei umane (sau hominide), progresul a fost atât de lent, încât capacitatea umană de producție a avut nevoie de aproximativ un milion de ani pentru a crește îndeajuns cât să poată asigura simpla subzistență a unui milion de indivizi în plus. În jurul anului 5000 î.Hr., după revoluția agricolă, rata progresului a crescut până în punctul în care același spor demografic a putut fi susținut în doar două secole. În zilele noastre, de după Revoluția Industrială, economia mondială crește, în medie, până la același grad, la fiecare 90 de minute.[1]

Actuala rată a progresului va produce, și ea, rezultate spectaculoase, dacă se menține îndeajuns de mult. Dacă economia

[1] Un venit la nivelul subzistenței astăzi este de aproximativ 400 de dolari (Chen și Ravallion, 2010). Astfel, un milion de venituri reprezintă 400 000 000 de dolari. PIB-ul omenirii este de 60 000 000 000 000 și în ultimii ani a crescut constant cu o rată de 4% (rata compusă de creștere din 1950, conform lui Maddison – 2010). Aceste numere par să susțină cifra prezentată în text, care este, desigur, doar o aproximare la un anumit ordin de magnitudine. Dacă ne uităm la rata de creștere a populației lumii, descoperim că, astăzi, populația lumii crește cu aproximativ un milion la fiecare săptămână și jumătate. Dar această, observație subestimează rata de creștere a economiei, din moment ce venitul pe cap de locuitor crește și el. În jurul anului 5000 d.Hr., imediat după revoluția agricolă, populația lumii creștea cam cu un milion la 200 de ani – o accelerare substanțială în raport cu creșterea de un milion la un milion de ani în preistoria umanității. Chiar și așa, este impresionant că o creștere economică netă care lua 200 de ani acum 500 de ani ia doar 9 minute astăzi și că o creștere demografică care lua în trecut două secole ia astăzi o săptămână și jumătate. Vezi, de asemenea, Maddison, 2005.

mondială va continua să crească în același ritm în care a făcut-o în ultimii cincizeci de ani, omenirea va fi de 4,8 de ori mai bogată până în 2050 și de 34 de ori până în 2100.[1]

„Conversația noastră centrată pe tehnologie și pe modificări continue ale vieții umane mă face să mă gândesc că va apărea un punct singular în istoria speciei noastre dincolo de care istoria umană, așa cum o știm, nu va putea continua." (Ulam 1958)

Cu toate acestea, perspectiva menținerii unei căi cu o creștere exponențială sigură pălește în fața a ceea ce ar aduce o nouă creștere spectaculoasă a ratei progresului, similară cu aceea din vremea revoluției agrare și a celei industriale. Economistul Robin Hanson estimează, bazându-se pe date economice și demografice, că dublarea economiei din Pleistocen, a societății de vânători-culegători, s-ar realiza în aproximativ 224 000 de ani; pentru perioada agrară de 909 ani, iar pentru perioada industrială de 6,3 ani.[2] (În modelul lui Hanson, epoca actuală presupune o combinație între modelele de creștere din perioada agrară și cea industrială – economia mondială, per ansamblu, nu progresează încă în baza ratei de dublare de 6,3 ani.) Dacă s-ar produce o trecere la un model diferit de progres, de magnitudine similară cu cele două anterioare, s-ar ajunge la un regim al progresului în care economia mondială s-ar dubla, ca mărime, în numai două săptămâni.

O astfel de creștere pare neverosimilă, după standardele actuale. Observatorii din celelalte epoci posibil să fi acceptat la fel de greu ideea ca economia din timpurile respective să se dubleze de mai multe ori pe parcursul unei vieți de om. Și totuși, această stare extraordinară este, pentru noi, astăzi, ordinară.

Ideea unei singularități tehnologice ce stă să se producă este, astăzi, popularizată la scară largă, începând cu eseul fundamental al lui Vernon Vinge și continuând cu lucrările lui Ray Kurzweil și ale altora.[3] Cu toate acestea, termenul „singularitate" a fost

[1] O astfel de creștere spectaculoasă poate sugera apariția unei singularități, așa cum sugera John von Neumann într-o conversație cu matematicianul Stanislaw Ulam.

[2] Hanson, 2000.

[3] Vinge, 1993; Kurzweil, 2005.

folosit într-o manieră confuză, cu mai multe sensuri, și a ajuns să dezvolte o aură laică (și aproape milenaristă) de conotații tehno-utopice.[1] Din moment ce majoritatea semnificațiilor și conotațiilor respective sunt irelevante pentru demersul nostru, este preferabil să ne dispensăm de termenul „singularitate", în favoarea unuia mai riguros.

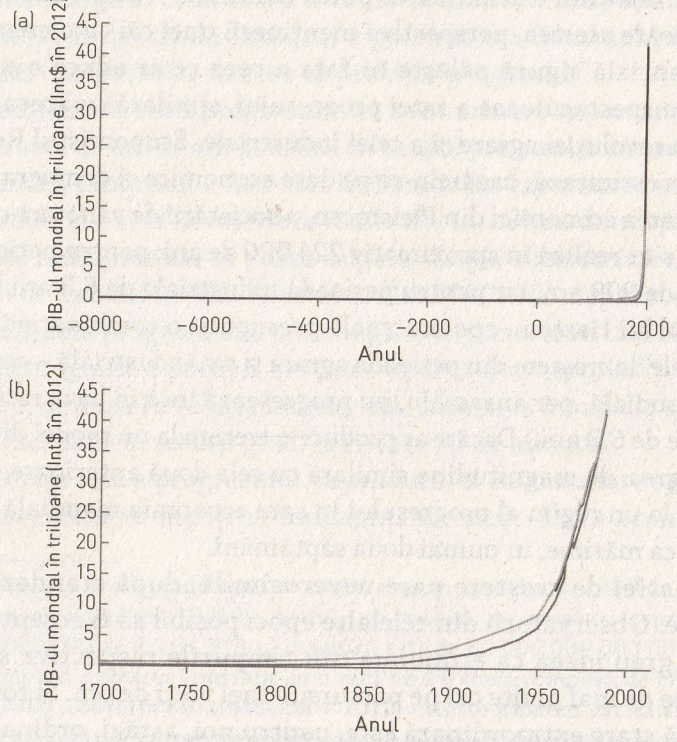

Figura 1. Istoria pe termen lung a PIB-ului mondial. Ilustrată linear, istoria economiei arată ca o dreaptă orizontală îmbrățișând axa x, până în punctul în care pornește, brusc, în sus. (a) Chiar și concentrându-ne numai pe ultimii 10 000 de ani, tiparul își menține, în esență, unghiul de 90°. (b) Abia în ultimii aproximativ 100 de ani, curba se ridică sesizabil deasupra nivelului zero. (Liniile diferite ale tiparului corespund unor seturi de date diferite, care ne oferă estimări ușor diferite.[2])

[1] Sandberg, 2010.

[2] Van Zanden (2003); Maddison (1999, 2001); De Long (1998).

În privința „singularității", sensul care ne interesează aici prevede posibilitatea *unei explozii a inteligenței*, mai exact, perspectiva unei superinteligențe de tip artificial. Unii oameni cred în diagrame ale progresului precum vedem în figura 1, care arată iminența unei noi schimbări drastice în modelul progresului, comparabile cu revoluția agrară și cu cea industrială. Acestora le este greu să conceapă un scenariu în care ritmul de dublare a economiei mondiale să fie redus la câteva săptămâni, fără a implica realizarea unor creiere mai rapide și mai eficiente decât cele biologice, cunoscute nouă. Totuși faptul de a lua în serios eventualitatea unei revoluții a inteligenței artificiale nu trebuie să țină de studierea curbelor de creștere sau de extrapolarea datelor economice din trecut. După cum vom vedea, avem motive mai temeinice pentru a face asta.

Marile speranțe

Mașinile similare oamenilor în privința inteligenței generale – adică mașini care să aibă bună judecată, abilitatea înnăscută de a învăța și rațiune și care să poată desluși provocări legate de procesarea informațiilor din domenii diverse, naturale și artificiale – au fost un deziderat încă din anii 1940, odată cu inventarea computerelor. În acel moment, apariția unor astfel de mașini era preconizată undeva în următoarele două decenii.[1] De atunci încoace, rata temporală privind producerea lor a scăzut, anual, cu câte un an; astfel că, astăzi, futurologii care iau în serios apariția inteligenței artificiale de ordin general cred că ne despart doar câteva decenii de acest eveniment.[2]

Două decenii reprezintă o perioadă convenabilă pentru predicatorii schimbării radicale: suficient de scurtă pentru ca acest fenomen să devină relevant și atractiv și, totuși, suficient de lungă cât

[1] Două declarații optimiste reluate des în anii 1960: „Mașinile vor fi capabile, în aproximativ 20 de ani, să facă tot ceea ce pot face oamenii" (Simon, 1965, 1996); „în decurs de o generație [...] problema creării inteligenței artificiale va fi rezolvată" (Minsky, 1967, p. 2). Pentru o trecere exhaustivă în revistă a predicțiilor cu privire la IA, vezi Armstrong și Sottala, 2012.

[2] Vezi, de exemplu, Baum et. all, 2011.