Nick Bostrom este profesor la Facultatea de Filosofie de la Universitatea Oxford și director fondator al Institutului despre Viitorul Umanității și al Programului asupra Impactului Tehnologiei Viitorului din cadrul Oxford Martin School. Este autorul a peste 200 de lucrări. A predat la Yale și a făcut studii postdoctorale la Academia Britanică. Bostrom are studii de fizică, neuroștiință computațională și logică matematică, dar și de filosofie.

NICK BOSTROM

SUPERINTELIGENȚA

CĂI, PERICOLE, STRATEGII

Traducere din limba engleză
DORU VALENTIN CĂSTĂIAN

LITERA®
București
2019

Superintelligence
Paths, Dangers, Strategies
Nick Bostrom

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

Carte Pentru Toți este parte a Grupului Editorial Litera
O.P. 53; C.P. 212, sector 4, București, România
tel.: 021 319 63 93; 0752 101 777

Superinteligența. Direcții, pericole, strategii
Nick Bostrom

Traducere din limba engleză:
Doru Valentin Căstăian

Editor: Vidrașcu și fiii
Redactor: Isabella Prodan
Corectori: Georgiana Enache, Ionel Palade
Copertă: Flori Zahiu
Tehnoredactare și prepress: Mihai Suciu

Descrierea CIP a Bibliotecii Naționale a României

BOSTROM, NICK

Superinteligența. Direcții, pericole, strategii / Nick Bostrom; trad.: Doru Valentin Căstăian – București: Litera, 2019

Index

ISBN 978-606-33-4078-9

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

Alexander Tamas, Max Tegmark, Roman Yamposkyi și Eliezer Yudkowsky.

Pentru comentariile lor foarte detaliate, le mulțumesc următorilor: Milan Cirkovic, Daniel Dewey, Owain Ewans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan și Carl Shulman. Pentru sfaturile privind munca de cercetare, vreau să le mulțumesc lui Stuart Armstrong, Daniel Dewey, Eric Drexer, Alexandre Erler, Rebecca Roache și Anders Sandberg.

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.

Capitolul 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 mondială va continua să crească în același ritm în care a făcut-o în ultimii 50 de ani, omenirea va fi de 4,8 de ori mai bogată până în 2050 și de 34 de ori până în 2100.[2]

„Conversația noastră centrată pe tehnologie și pe modificări continue ale vieții umane mă face să mă

[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.

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

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.[1] (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.

[1] Hanson, 2000

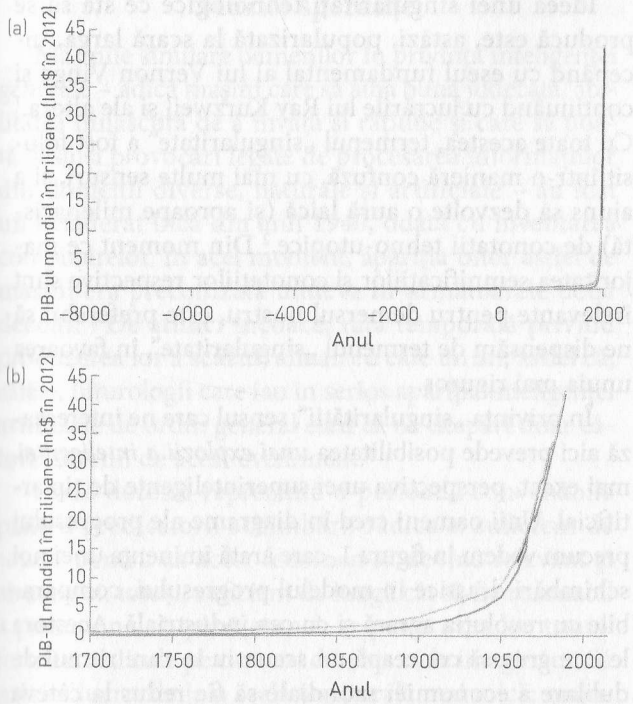

Figura 1. Istoria pe termen lung a PIB-ului mondial. Ilustrată linear, istoria economiei arată ca o dreaptă orizontală îmbrățișând axa x, până în punctul în care pornește, brusc, în sus. (a) Chiar și concentrându-ne numai pe ultimii 10 000 de ani, tiparul își menține, în esență, unghiul de 90. (b) Abia 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.[1])

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

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.[1] Cu toate acestea, termenul „singularitate" a fost folosit într-o manieră confuză, cu mai multe sensuri, și a ajuns să dezvolte o aură laică (și aproape milenaristă) de conotații tehno-utopice.[2] 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.

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.

[1] Vinge, 1993; Kurzweil, 2005

[2] Sandberg, 2010

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 să fie rezonabil a presupune că, până atunci, ar putea avea loc transformări radicale pe care azi de-abia ni le imaginăm. Să privim lucrurile la o scală temporală mai mică: majoritatea tehnologiilor care vor influența lumea în cinci sau zece ani sunt deja folosite, în stoc limitat, iar cele care își vor lăsa amprenta asupra lumii în mai puțin de

[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.

cinsprezece ani probabil că există, şi ele, deja, ca prototipuri de laborator. Două decenii constituie, de asemenea, durata tipică a carierei unui specialist în astfel de predicţii, presupunând punerea în joc a reputaţiei prin realizarea unei predicţii curajoase.

Faptul că, în trecut, unii au făcut estimări nerealiste cu privire la apariţia inteligenţei artificiale nu înseamnă că această formă de inteligenţă este imposibilă, că nu va apărea nicicând.[1] Principalul motiv pentru care progresul a fost mai lent decât spera lumea ţine de faptul că elaborarea unei maşinării inteligente s-a dovedit mai dificilă decât au preconizat primii care au avut în vedere acest proiect. Aşadar, nu se ştie cât de mari sunt aceste dificultăţi şi nici cât ne va lua pentru a le depăşi. Uneori, o problemă care pare extrem de complicată poate avea o soluţie foarte simplă (deşi situaţia inversă este, probabil, mai comună).

În capitolul următor, vom inspecta căile posibile care pot duce la o inteligenţă artificială comparabilă cu cea umană. Înainte, permiteţi-mi totuşi să subliniez că, oricât ne-am poticni între momentul actual şi până în cel în care vom avea o maşinărie cu o inteligenţă de nivel uman, nici măcar acest ultim punct nu va fi unul final. Următoarea staţie, prima după realizarea maşinăriei, va fi aceea a inteligenţei artificiale de nivel suprauman. Trenul s-ar putea să nu oprească, poate

[1] Asta ar putea sugera, totuşi, că cercetătorii care se ocupă de IA ştiu mai puţine decât cred despre cronologia evoluţiei – ceea ce poate avea rezultate contrare: ei fie vor subestima, fie vor supraestima momentului apariţiei inteligenţei artificiale.

nici măcar să încetinească, în staţia umanităţii. Posibil o s-o traverseze ca vântul.

Matematicianul I.J. Good, statistician-şef în echipa lui Alan Turing, în timpul celui de-al Doilea Război Mondial, a formulat, pentru prima dată, aspectele esenţiale ale acestui scenariu. Într-un pasaj des citat, ce datează din 1965, el susţine:

> Prin maşină ultrainteligentă înţelegem o maşină care poate depăşi, de departe, activitatea intelectuală a oricărei fiinţe umane, oricât de inteligentă ar fi aceasta din urmă. Din moment ce construirea unei astfel de maşini depinde tocmai de activitatea intelectuală, o maşină ultrainteligentă ar putea realiza, de una singură, maşini chiar mai eficiente; asta ar duce, fără îndoială, la o „explozie de inteligenţă" care ar lăsa cu mult în urmă inteligenţa umană. Astfel, inventarea unei maşini ultrainteligente va fi ultimul lucru pe care trebuie să-l facă omul, cu condiţia ca aceasta să fie suficient de docilă, încât să ne arate cum s-o ţinem sub control.[1]

Pare evident că o asemenea dezvoltare a inteligenţei va fi asociată cu riscuri existenţiale majore şi că urmările vor trebui prevăzute cu maximă atenţie, chiar şi ştiind că probabilitatea ca ele să se concretizeze ar fi redusă (nefiind însă cazul). Şi, totuşi, pionierii inteligenţei artificiale, deşi postulează iminenţa realizării uneia de nivel uman, s-au gândit mai puţin la posibilitatea apariţiei unei inteligenţe supraumane. Ca şi cum latura mentală cu care aceştia fac speculaţii

[1] Good, 1965, p. 33

s-ar fi epuizat într-un asemenea grad odată cu conceperea posibilității existenței unei mașinării care să atingă inteligența umană, încât, pur și simplu, nu mai pot concepe și corolarul – posibilitatea ca mașinile să devină superinteligente.

Pe de altă parte, precursorii inteligenței artificiale nu au luat în calcul posibilitatea ca întreprinderea lor să implice riscuri.[1] Nici măcar nu au verbalizat – dar să se gândească serios la această chestiune! – pericolul implicat de mințile artificiale sau posibilele consecințe etice pe care le-ar putea avea existența unor stăpâni cibernetici: o lacună uimitoare, mai ales pe fundalul unei epoci ce nu s-a remarcat prin gândire critică la adresa domeniului tehnic.[2] Nu putem decât să sperăm

[1] O excepție este Norbert Wiener, care a avut câteva presimțiri cu privire la posibilele consecințe. Iată ce scria în 1960: „Dacă vom folosi, pentru a ne atinge scopurile, agenți mecanici cu care nu vom putea interfera foarte bine după pornire, pentru că totul decurge foarte rapid și este irevocabil și nu vom avea suficiente informații pentru a interveni înainte ca acțiunea să fie finalizată, atunci trebuie să fim cu adevărat siguri că scopul atribuit mașinăriei este exact scopul dorit de noi, și nu o palidă aproximare a lui" (Wiener, 1960). Ed Fredkin a vorbit, și el, despre îngrijorările sale privind superinteligența, într-un interviu redat în McCorduck, 1979. Până în 1970, și Good a scris despre riscuri, sugerând crearea unei asociații care să lupte împotriva acestora (Good, 1970). A se vedea și articolul său de mai târziu (Good, 1982), în care conturează câteva dintre ideile „normativității indirecte", pe care le vom discuta în Capitolul 13. În 1984, Marvin Minsky scria, la rându-i, despre câteva îngrijorări principale (Minsky, 1984).

[2] Cf. Yuodkovski (2008 a). Cu privire la importanța stabilirii unor repere etice pentru viitoarele tehnologii înainte ca acestea să apară, vezi Roache (2008).

că, până în momentul în care această întreprindere va deveni cu adevărat fezabilă, vom avea atât capacitatea tehnică pentru a susține ampla dezvoltare informatică, cât și competența necesară pentru a-i asigura efectele durabile.

Dar, înainte să vedem ce va urma, să aruncăm o privire scurtă peste istoria inteligenței artificiale.

Perioade de speranță și perioade de disperare

În vara anului 1956, la Colegiul Dartmouth, zece oameni de știință interesați, cu toții, de rețelele neuronale, de teoria roboticii și de studiul inteligenței, au hotărât să participe la un seminar de șase săptămâni. Proiectul de vară de la Dartmouth, cum este cunoscut, este văzut cu precădere ca evenimentul în urma căruia a apărut domeniul de cercetare a inteligenței artificiale. Mulți participanți aveau să fie considerați, mai târziu, drept părinți ai domeniului. Spiritul optimist care îi caracteriza pe aceștia transpare din propunerea făcută Fundației Rockefeller, care finanța proiectul:

> Propunem desfășurarea unui studiu realizat de zece oameni, care să dureze două luni [...] Studiul pleacă de la presupoziția că orice aspect al învățării și orice caracteristică a inteligenței poate fi, în principiu, descrisă atât de precis, încât să fie simulată de o mașină. Vom încerca să găsim modalități prin care să determinăm mașinile ce utilizează limbajul și realizează abstracțiuni și concepte să rezolve probleme care astăzi le sunt rezervate oamenilor, precum