IT & C

ISSN 2821 - 8469, ISSN – L 2821 - 8469, Volumul 2, Numărul 1, Martie 2023

Riscuri și provocări în inteligența artificială: Cutii negre și actorii de amenințare

Nicolae Sfetcu

Sfetcu, Nicolae (2023), Riscuri și provocări în inteligența artificială: Cutii negre și actorii de amenințare, *IT & C*, **2**:1, 41-47, DOI: 10.58679/IT21269, https://www.internetmobile.ro/riscuri-si-provocari-in-inteligenta-artificiala-cutii-negre-si-actorii-de-amenintare/

Publicat online: 19.02.2023

Riscuri și provocări în inteligența artificială: Cutii negre și actorii de amenințare

Nicolae Sfetcu
nicolae@sfetcu.com

Risks and Challenges in Artificial Intelligence: Black Boxes and Threat Actors

Abstract

Artificial intelligence has created unprecedented opportunities, but also new risks. The exponential increase in the capabilities of artificial intelligence models enables previously unattainable levels of value and generalization. However, the opacity of these models has also increased, and their black-box nature makes it difficult, even for experts, to explain the rationale behind their conclusions. This may represent a technological and social critical point, as the risk is real, as evidenced by recent episodes, of training systems being compromised by discrimination biases that have learned from training data. It is therefore possible that learning from digital traces of past decisions can lead to existing invisible biases being incorporated into the resulting patterns, perpetuating them.

Keywords: artificial intelligence, risks, challenges, black boxes, threats

Rezumat

Inteligența artificială a creat oportunități fără precedent, dar și noi riscuri. Creșterea exponențială a capabilităților modelelor de inteligența artificială permite atingerea unor niveluri de valoare și generalizare neatinse până acum. Cu toate acestea, opacitatea acestor modele a crescut, de asemenea, iar natura lor de cutie neagră face dificilă, chiar și pentru experți, explicarea justificării concluziilor lor. Acest lucru poate reprezenta un punct critic din punct de vedere tehnologic și social, deoarece riscul este real, după cum demonstrează episoadele recente, ale sistemelor de antrenament care sunt compromise de părtiniri și prejudecăți de discriminare, care au învățat din datele de instruire. Prin urmare, este posibil ca învățarea din urmele digitale ale deciziilor trecute să poată duce la încorporarea prejudecăților invizibile existente în modelele rezultate, perpetuându-le.

Cuvinte cheie: inteligența artificială, riscuri, provocări, cutii negre, amenințări

IT & C, Volumul 2, Numărul 1, Martie 2023, pp. 41-47
ISSN 2821 - 8469, ISSN – L 2821 – 8469, DOI: 10.58679/IT21269
URL: https://www.internetmobile.ro/riscuri-si-provocari-in-inteligenta-artificiala-cutii-negre-si-actorii-de-amenintare/

Modelele de învățare automată sunt adesea cutii negre. Logica unui model de inteligență artificială (IA) de luare a deciziilor poate fi ascunsă chiar și dezvoltatorilor și experților.

Riscuri:

- Învățarea din date părtinitoare (fie din întâmplare, fie din necesitate);
- Modelele instruite pentru, de exemplu, viziunea conducerii și recunoașterea obstacolelor, pot moșteni erori dăunătoare pentru siguranță;
- Modelele de profilare pentru, de exemplu, justiția predictivă, pot moșteni discriminarea grupurilor vulnerabile cunoscute (sau noi).

Direcția de urmărit pentru atenuarea riscurilor este o IA explicabilă și de încredere, cheia unei colaborări eficiente om-mașină în luarea deciziilor cu mize mari.

IA a creat oportunități fără precedent, dar și noi riscuri. Creșterea exponențială a capabilităților modelelor IA permite atingerea unor niveluri de valoare și generalizare neatinse până acum. Cu toate acestea, **opacitatea** acestor modele a crescut, de asemenea, iar natura lor de **cutie neagră** face dificilă, chiar și pentru experți, explicarea justificării concluziilor lor. Acest lucru poate reprezenta un punct critic din punct de vedere tehnologic și social, deoarece riscul este real, după cum demonstrează episoadele recente, ale sistemelor de antrenament care sunt compromise de părtiniri și **prejudecăți** de discriminare, care au învățat din datele de instruire. Prin urmare, este posibil ca „învățarea din urmele digitale ale deciziilor trecute să poată duce la încorporarea prejudecăților invizibile existente în modelele rezultate, perpetuându-le”, așa cum s-a raportat în lucrarea care a lansat studiul mineritului de date conștient de discriminare în 2008 (23) .

Parafrazându-l pe Frank Pasquale, autorul cărții *The black box society* (1), am văzut algoritmi din ce în ce mai opaci răspândiți, folosiți pentru a deduce **trăsături intime** ale indivizilor.

Acești algoritmi generează modele de clasificare și predicție a trăsăturilor comportamentale ale indivizilor, cum ar fi scorul de credit, riscul asigurării, starea de sănătate, înclinația către infracțiuni, preferințele și orientările personale, folosind datele personale difuzate în mediul digital de către cetățeni, cu sau uneori fără conștientizarea lor . Astfel de sisteme automate de luare a deciziilor sunt adesea **cutii negre** care, respectând caracteristicile utilizatorilor, prezic o clasă, o judecată, un vot și sugerează decizii; dar fără a explica motivul predicției sau recomandării propuse. Nu este doar o chestiune de transparență. Modelele sunt instruite pe exemple reconstituite pe baza urmelor digitale ale activităților utilizatorilor, cum ar fi mișcări, achiziții, căutări online, opinii exprimate pe rețelele de socializare. Drept urmare, modelele moștenesc părtinirile și defectele - **prejudecăți** - care sunt ascunse în datele de instruire, ascunzându-le, la rândul lor, în algoritmii de decizie care riscă să sugereze alegeri neloiale, discriminatorii sau pur și simplu greșite, posibil fără conștientizarea factorul de decizie și subiectul deciziei finale. Dacă o casetă de chat, o IA care conversează cu utilizatorii de pe rețelele de socializare, învață din exemplele greșite, de exemplu, rostirea rasistă, va fi rasistă la rândul ei - și creatorii săi vor trebui să o elimine rapid și în tăcere. Multe cazuri care s-au întâmplat deja, precum cel al botului Twitter Tay lansat de Microsoft în 2016 și închis rapid după ce a ajuns la un comportament ofensator și rasist din conversațiile online (2), ne avertizează că delegarea alegerilor către algoritmii cutiei negre poate fi o idee rea.

Inteligența artificială (AI) permite luarea automată a deciziilor și facilitează multe aspecte ale vieții de zi cu zi, aducând cu ea îmbunătățiri ale operațiunilor și numeroase alte beneficii. Cu toate acestea, sistemele AI se confruntă cu numeroase amenințări la adresa securității cibernetice, și AI în sine trebuie să fie securizat, deoarece au fost deja raportate cazuri de atacuri rău intenționate, de ex. Tehnicile AI și sistemele bazate pe AI pot duce la rezultate neașteptate și pot fi modificate pentru a manipula rezultatele așteptate (3) (4) (5). Prin urmare, este esențial să înțelegem peisajul amenințărilor AI și să existe o bază comună și unificatoare pentru înțelegerea potențialului amenințărilor și, în consecință, pentru a realiza evaluări specifice ale riscurilor. Acestea din urmă vor sprijini implementarea unor măsuri și controale de securitate direcționate și proporționale pentru a contracara amenințările legate de AI.

Cutii negre

Compas, deținut de Northpointe Inc., este un model predictiv al riscului de recidivă penală, folosit până de curând de diferite curți de justiție din SUA în sprijinul deciziilor judecătorilor cu privire la cererile de eliberare. Jurnaliștii de la Propublica.org au colectat mii de cazuri de utilizare ale modelului și au arătat că are o puternică părtinire rasistă: negrii care nu vor mai comite o crimă vor primi un risc dublu comparativ cu albii în aceleași condiții (6). Modelul, dezvoltat cu tehnici de învățare automată (Machine Learning), probabil a moștenit părtinirea prezentă în sentințele istorice și este afectat de faptul că populația penitenciară americană reprezintă mult mai mult negrii decât albii;

Primele trei agenții de risc de credit din Statele Unite, Experian, TransUnion și Equifax sunt adesea discordante. Într-un studiu de 500.000 de cazuri, 29% dintre solicitanții de credit au avut o evaluare a riscurilor cu diferențe de peste 50 de puncte între cele trei companii, ceea ce poate însemna diferențe de zeci de mii de dolari în interesele generale. O astfel de variabilitate largă sugerează ipoteze de evaluare foarte diferite, precum și opace, sau un arbitrar puternic (7);

În anii 1970 și 1980, Școala de Medicină a Spitalului St. George din Londra a folosit software pentru a filtra cererile de locuri de muncă, care s-a dovedit ulterior a fi extrem de discriminatoriu față de femei și minoritățile etnice, deduse prin prenume și locul nașterii. Discriminarea algoritmică nu este un fenomen nou și nu se datorează neapărat învățătii automate (8);

Un clasificator bazat pe învățarea profundă (Deep Learning) poate fi foarte precis în ceea ce privește datele de antrenament și, în același timp, complet nefiabil, de exemplu, dacă a învățat din date de calitate proastă. Într-un caz de recunoaștere a imaginilor menite să distingă lupii husky într-un set mare de date, cutia neagră rezultată a fost disecată de cercetători doar pentru a afla că decizia de a clasifica o imagine ca „lup” s-a bazat exclusiv pe zăpada din fundal (9) ! Vina, desigur, nu este a învățării profunde, ci a alegerii accidentale a exemplelor de antrenament în care, evident, fiecare lup fusese fotografiat pe zăpadă. Deci, un husky în zăpadă este clasificat automat ca lup. Transmutând acest exemplu asupra sistemului de viziune al mașinii noastre cu conducere automată: cum putem fi siguri că va putea recunoaște corect fiecare obiect din jurul nostru?

Diverse studii, precum cel menționat în nota (10), arată că textele de pe web (dar și de pe media în general) conțin părtiniri și prejudecăți, precum faptul că numele albilor sunt mai des asociate cu cuvinte cu o sarcină emoțională pozitivă, în timp ce numele persoanelor de culoare

sunt mai des asociate cu cuvinte cu o sarcină emoțională negativă. Prin urmare, modelele instruite pe texte pentru analiza sentimentelor și opiniilor sunt foarte susceptibile de a moșteni aceleași prejudecăți;

Jurnaliștii de date Bloomberg (11) au arătat cum modelul automat, utilizat de Amazon pentru a selecta cartierele orașelor americane pentru a oferi „livrare în aceeași zi" gratuit, are o prejudecată etnică. Software-ul, fără știrea companiei, exclude în mod sistematic din ofertă zonele locuite de minorități etnice în multe orașe, inclusiv în vecinătate. Amazon a răspuns la ancheta jurnalistică că nu era la curent cu această practică, deoarece modelul de învățare automată era total autonom și își baza alegerile pe activitatea anterioară a clienților. Pe scurt, este vina algoritmului.

Actorii de amenințare

Există diferite grupuri de actori ai amenințărilor care ar putea dori să atace sistemele AI folosind mijloace cibernetice (12).

Infractorii cibernetici sunt motivați în primul rând de profit. Infractorii cibernetici vor avea tendința de a utiliza AI ca instrument pentru a conduce atacuri, dar și pentru a exploata vulnerabilitățile sistemelor AI existente (13). De exemplu, ar putea încerca să pirateze chatbot-uri activate cu AI pentru a fura cardul de credit sau alte date. Alternativ, aceștia pot lansa un atac ransomware împotriva sistemelor bazate pe inteligență artificială utilizate pentru gestionarea lanțului de aprovizionare și depozitare.

Personalul companiei, inclusiv angajații și contractanții care au acces la rețelele unei organizații, pot implica fie pe cei care au intenții rău intenționate, fie pe cei care pot dăuna companiei în mod neintenționat. De exemplu, persoanele din interior rău intenționate ar putea încerca să fure sau să saboteze setul de date utilizat de sistemele AI ale companiei. În schimb, persoanele care nu sunt rău intenționate pot corupe accidental un astfel de set de date.

Actorii de stat național și alți atacatori sponsorizați de stat sunt, în general, avansați. Pe lângă dezvoltarea modalităților de a folosi sistemele AI pentru a ataca alte țări (inclusiv industriile și infrastructurile critice), precum și utilizarea sistemelor AI pentru a-și apăra propriile rețele, actorii din statele naționale caută în mod activ vulnerabilități în sistemele AI pe care le pot exploata. Acest lucru ar putea fi un mijloc de a provoca prejudicii unei alte țări sau ca un mijloc de colectare de informații.

Alți actori de amenințare includ teroriștii, care încearcă să provoace daune fizice sau chiar pierderi de vieți. De exemplu, teroriștii pot dori să spargă mașini fără șofer pentru a le folosi ca armă.

Hacktiviștii, care au tendința de a fi în mare parte motivați ideologic, pot încerca, de asemenea, să pirateze sistemele AI pentru a arăta că se poate face. Există un număr tot mai mare de grupuri preocupate de potențialele pericole ale AI și nu este de neconceput că ar putea pirata un sistem AI pentru a obține publicitate. Există, de asemenea, actori de amenințări nesofisticați, cum ar fi hackerii amatori (haxori), care pot fi motivați penal sau ideologic. Aceștia sunt, în general, persoane necalificate care folosesc scripturi sau programe pre-scrise pentru a ataca sistemele, deoarece le lipsește expertiza necesară pentru a le scrie pe ale lor. Dincolo de actorii tradiționali ai amenințărilor discutați mai sus, devine din ce în ce mai necesar să se includă și concurenții ca actori ai amenințărilor, deoarece unele companii au din ce în ce mai mult intenția de a-și ataca rivalii pentru a câștiga cotă de piață. (14)

Bibliografie

1. F. Pasquale, *The black box society*, Harvard University Press 2015
2. Tay (bot) https://en.wikipedia.org/wiki/Tay_(bot) [19-4-2020].
3. A se vedea https://www.idgconnect.com/news/1506124/deepfakes-ai-deceives, septembrie 2020
4. A se vedea https://thenewstack.io/camouflaged-graffiti-road-signs-can-fool-machine-learning-models/, septembrie 2017
5. A se vedea https://www.media.mit.edu/publications/adversarial-attacks-on-medical-machine-learning/, martie 2019
6. Machine Bias https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [19-4-2020].
7. R. Carter – H.V. Auken., *Small firm bankruptcy*, Journal of Small Business Management 2006, 44: pp. 493-512
8. S. Lowry – G. Macpherson, *A blot on the profession*, British medical journal (Clinical research ed.) 1988, pp. 657
9. M.T. Ribeiro et al., *"Why should I trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016, pp. 1135-1144
10. A. Caliskan Islam et al., *Semantics derived automatically from language corpora necessarily contain human biases*, arXiv preprint arXiv:1608.07187 2016
11. *Amazon Doesn't Consider the Race of Its Customers. Should It?* https://www.bloomberg.com/graphics/2016-amazon-same-day/ [19-4-2020].
12. Având în vedere natura largă a sistemelor AI și implementarea lor în diverse sectoare, listarea este generică și nu implică o clasare a probabilității ca un actor de amenințare să atace sistemele AI.

13. Pe măsură ce AI-as-a-service câștigă teren, astfel de sisteme vor fi din ce în ce mai disponibile pentru actorii care nu se pricep la tehnică.
14. Vezi, Sailio, M.; Latvala, O.-M.; Szanto, A. Cyber Threat Actors for the Factory of the Future. *Appl. Sci. 2020*, 10, 4334 https://www.mdpi.com/2076-3417/10/12/4334/htm

Sursa: Sfetcu, Nicolae (2022). *Introducere în inteligența artificială*, MultiMedia Publishing, ISBN 978-606-033-659-4, https://www.telework.ro/ro/e-books/introducere-in-inteligenta-artificiala/