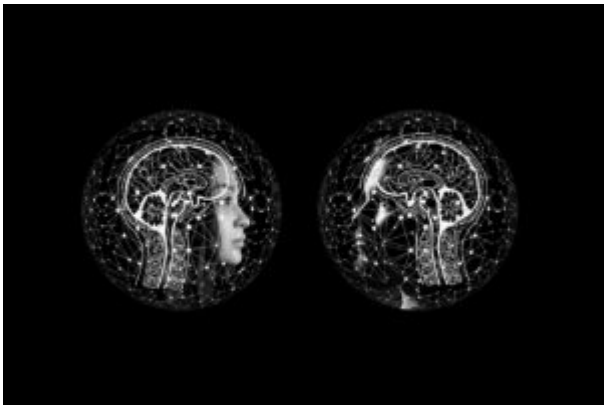
SAM BARON: FILOZOFII AU STUDIAT „CONTRAFACTUALELE” TIMP DE DECENII. NE VOR PUTEA AJUTA ELE SĂ DESLUȘIM MISTERELE INTELIGENȚEI ARTIFICIALE?

Autor: Redacția Syntopic | 8 iunie 2023



Un articol de Sam Baron pentru The Conversation

Articol original: Philosophers have studied ‘counterfactuals’ for decades. Will they help us unlock the mysteries of AI?

Traducere realizată de Roxana Dragomir, studentă la Facultatea de Filozofie, Universitatea din București, pentru stagiul de practică organizat în parteneriat cu redacția Syntopic.

Inteligența artificială este din ce în ce mai des utilizată în întreaga lume pentru a ne ajuta să luăm decizii în viața noastră, fie că este vorba de decizii privind împrumuturile acordate de bănci[1], de diagnostice medicale[2] sau de previziuni ale forțelor de ordine din SUA privind probabilitatea ca un infractor să recidiveze[3].

Cu toate acestea, multe sisteme de inteligență artificială sunt ca niște cutii negre: nimeni nu înțelege cum funcționează. Acest lucru a dus la o cerere pentru „inteligență artificială explicabilă” [un set de instrumente și scheme], astfel încât să putem înțelege de ce un model de inteligență artificială a produs un anumit rezultat și ce prejudecăți ar fi putut juca un rol.

Inteligența artificială explicabilă este o ramură în plină dezvoltare a cercetării în domeniul inteligenței artificiale. Dar ceea ce este poate mai puțin cunoscut este rolul pe care filosofia îl joacă în dezvoltarea ei.

Mai exact, o idee numită „explicație contrafactuală” este adesea prezentată ca o soluție la problemele de tip „cutie neagră”. Dar, odată ce înțelegi filozofia din spatele ei, poți începe să înțelegi de ce nu prea reușește să fie, totuși, o soluție reală.

De ce contează explicațiile

Atunci când inteligența artificială este utilizată pentru a lua decizii care schimbă vieți, persoanele afectate merită o explicație a modului în care s-a ajuns la acea decizie. Acest lucru a fost recunoscut recent prin intermediul Regulamentului general privind protecția datelor al Uniunii Europene[4], care sprijină dreptul unei persoane la explicații.

Nevoia de explicații a fost evidențiată și în cazul Robodebt din Australia[5], în care a fost utilizat un algoritm pentru a prezice nivelul de îndatorare al persoanelor care primesc asigurări sociale. Sistemul a făcut multe greșeli, încadrând persoane fără datorii pe lista celor îndatorate.

Abia după ce algoritmul a fost explicat pe deplin a putut fi identificată greșeala – dar până atunci pagubele fuseseră deja făcute. Rezultatul a fost atât de păgubitor încât a dus la înființarea unei comisii regale[6] în august 2022.

În cazul Robodebt, algoritmul în cauză a fost destul de simplu și a putut fi explicat, însă nu ar trebui să ne așteptăm ca acest lucru să fie întotdeauna așa și în viitor. Modelele actuale de inteligență artificială care utilizează învățarea automată pentru a procesa datele sunt mult mai sofisticate.

Cutia neagră cea mare și strălucitoare

Să presupunem că o persoană pe nume Sara solicită un împrumut. Banca îi cere să furnizeze informații, inclusiv starea civilă, gradul de îndatorare, venitul realizat, economiile, adresa de domiciliu și vârsta.

Banca introduce apoi aceste informații într-un sistem de inteligență artificială cu ajutorul căruia se calculează un scor de creditare. Scorul obținut este mic și este folosit pentru a o descalifica pe Sara de la acordarea împrumutului; nici Sara și nici angajații băncii nu știu de ce sistemul a calculat un scor atât de scăzut.

Spre deosebire de cazul Robodebt, algoritmul utilizat aici poate fi extrem de complicat și nu poate fi explicat cu ușurință. Prin urmare, nu există o modalitate simplă de a ști dacă s-a făcut o greșeală, iar Sara nu are nicio modalitate de a obține informațiile de care are nevoie pentru a contesta decizia.

Acest scenariu nu este în întregime ipotetic: este probabil ca deciziile privind împrumuturile să fie externalizate către algoritmi din SUA și există un risc real ca acești algoritmi să conțină distorsiuni, prejudecăți[7]. Pentru a atenua riscul, trebuie să încercăm

să explicăm cum funcționează acești algoritmi.

Abordarea contrafactuală

În linii mari, există două tipuri de abordări[8] pentru a explica rezultatele obținute de inteligența artificială. Una implică demontarea unui sistem și studierea componentelor sale interne pentru a discerne modul în care acesta funcționează. Dar, de regulă, acest lucru nu este posibil din cauza complexității absolute a multor sisteme de inteligență artificială.

Cealaltă abordare constă în a lăsa sistemul nedeschis și, în schimb, a-i studia informațiile de intrare și de ieșire, căutând tipare. Metoda „contrafactuală” se încadrează în această abordare.

Contrafactualele sunt afirmații despre ce s-ar fi întâmplat dacă lucrurile s-ar fi desfășurat diferit. Într-un context de inteligență artificială, asta implică luarea în considerare a modului în care rezultatul unui sistem de inteligență artificială ar fi putut fi diferit dacă ar fi primit intrări diferite. Se presupune apoi că putem folosi această explicație pentru a explica de ce sistemul a produs rezultatul pe care l-a produs.

Să presupunem că banca alimentează sistemul său de inteligență artificială cu informații diferite (manipulate) despre Sara. Pe baza acestor informații, banca determină că cea mai mică schimbare de care Sara ar avea nevoie pentru a obține un rezultat pozitiv ar fi creșterea venitului său.

Banca poate apoi să folosească această informație ca explicație: împrumutul Sarei a fost refuzat pentru că venitul ei era prea mic. Dacă venitul ei ar fi fost mai mare, i s-ar fi acordat un împrumut.

Astfel de explicații contrafactuale[9] sunt luate în considerare în mod serios[10] ca o modalitate de a satisface necesitatea unei inteligențe artificiale explicabile, inclusiv în cazul cererilor de împrumut și al utilizării inteligenței artificiale pentru a face descoperiri științifice[11].

Cu toate acestea, așa cum au susținut cercetătorii, abordarea contrafactuală este inadecvată[12].

Corelație și explicație

Atunci când luăm în considerare modificările aduse intrărilor unui sistem de IA și modul în care acestea se traduc în ieșiri, reușim să adunăm informații despre corelații. Dar, așa cum spune vechea zicală, corelația nu înseamnă cauzalitate.

Motivul pentru care acest lucru reprezintă o problemă este acela că cercetările din filozofie sugerează faptul că relația de cauzalitate este strâns legată de explicație[13].

Pentru a explica de ce a avut loc un eveniment, trebuie să aflăm ce l-a cauzat.

În acest sens, ar putea fi o greșeală ca banca să-i spună Sarei că împrumutul i-a fost refuzat pentru că venitul ei era prea mic. Tot ceea ce poate spune cu siguranță este că venitul și scorul de credit sunt corelate – iar Sara rămâne în continuare fără o explicație pentru rezultatul său slab.

Este nevoie de o modalitate de a transforma informațiile despre contrafactuale și corelații în informații explicative.

Viitorul inteligenței artificiale explicabile

Cu timpul, ne putem aștepta ca IA să fie folosită și mai mult pentru decizii de angajare, cereri de viză, promovări și decizii de finanțare de stat și federale, printre altele.

Lipsa de explicații pentru aceste decizii amenință să crească substanțial nedreptatea pe care o vor resimți oamenii. La urma urmei, fără explicații nu putem corecta greșelile făcute atunci când folosim IA. Din fericire, filozofia ne poate ajuta.

Explicația a fost un subiect central al studiului filosofic[14] în ultimul secol. Filosofii au conceput o serie de metode pentru a extrage informații explicative dintr-o mare de corelații și au dezvoltat teorii sofisticate despre modul în care funcționează explicația.

O mare parte din această lucrare s-a axat pe relația dintre contrafactualitate și explicație. Eu însumi am elaborat lucrări[15] în acest sens. Dacă ne inspirăm din concepțiile filosofice, am putea dezvolta abordări mai bune pentru o inteligență artificială explicabilă.

Cu toate acestea, pe această temă, nu există încă o suprapunere suficientă între filozofie și informatică. Dacă dorim să abordăm în mod direct nedreptatea, vom avea nevoie de o abordare mai integrată, care să combine lucrările din aceste domenii.

NOTE

1. James Eyers, 2021, “Banks warned using AI in loan assessments could ‘awaken a zombie”, Financial Review, https://www.afr.com/companies/financial-services/banks-warned-using-ai-in-loan-assessments-could-awaken-a-zombie-20210615-p5814i. ↑
2. 2022, Healthcare Outlook, “Top 10 AI-based Medical Diagnostic Tools to try in 2022”, https://www.healthcareoutlook.net/top-10-ai-based-medical-diagnostic-tools-to-try-in-2022/. ↑

3. Karen Hao, 2019, “AI is sending people to jail—and getting it wrong”, MIT Technology Review, https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/. ↑
4. https://gdpr.eu/what-is-gdpr/. ↑
5. David Mariuz, 2020, “Robodebt was a fiasco with a cost we have yet to fully appreciate”, The Conversation, https://theconversation.com/robodebt-was-a-fiasco-with-a-cost-we-have-yet-to-fully-appreciate-150169. ↑
6. Emilie Gramenz, 2022, “Robodebt victim breaks down at royal commission while recalling ‘intrusion’ of ‘rude’ human services staff”, ABC News Australia, https://www.abc.net.au/news/2022-12-16/qld-robodebt-scheme-government-royal-commission-victim/101780890. ↑
7. Sian Townson, 2020, “AI Can Make Bank Loans More Fair”, Harvard Business Review, https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair. ↑
8. Adam Zewe, 2022, “Explained: How to tell if artificial intelligence is working the way we want it to”, MIT News, https://news.mit.edu/2022/explained-how-tell-if-artificial-intelligence-working-way-we-want-0722. ↑
9. Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR (October 6, 2017). Harvard Journal of Law & Technology, 31 (2), 2018, https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf. ↑
10. Julian Fell, Ben Spraggon, and Matt Liddy, 2022, “Wrenching open the black box”, ABC News Australia, https://www.abc.net.au/news/2022-12-12/robodebt-algorithms-black-box-explainer/101215902. ↑
11. Alexander Whiteside, 2022, “Molecular counterfactuals method helps researchers explain AI predictions”, Chemistry World, https://www.chemistryworld.com/news/molecular-counterfactuals-method-helps-researchers-explain-ai-predictions/4015381.article. ↑
12. Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, Joaquim Jorge, 2021, “Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications”, Arxiv, Cornell University, https://doi.org/10.48550/arXiv.2103.04244. ↑
13. Scriven, M. (1975). Causation as Explanation. Noûs, 9(1), 3–16. https://doi.org/10.2307/2214338. ↑
14. James Woodward, Lauren Ross, 2021, “Scientific Explanation”, Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/scientific-explanation/. ↑
15. Sam Baron, Mark Colyvan, David Ripley, 2017, “How Mathematics Can Make a

Difference”, Philosophers’ Imprint, January 2017, Volume 17, No. 3, pp. 1-29, http://hdl.handle.net/2027/spo.3521354.0017.003. ↑

Imagine: Pixabay