Razložljiva umetna inteligenca: kako naprej?

Explainable AI: What next?

Ana Farič[†]
Kognitivna znanost
Univerza v Ljubljani, Pedagoška fakulteta
Slovenija
af27987@student.uni-lj.si

Ivan Bratko
Umetna inteligenca
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Slovenija
bratko@fri.uni-lj.si

Povzetek

Prispevek povzema in ocenjuje stanje metod in raziskav na področju razložljive umetne inteligence. Pregled vsebuje predlagane definicije razlage in lastnosti dobrih razlag. Podan je grob pregled številnih obstoječih pristopov za generiranje razlage, primeri konkretnih avtomatsko generiranih razlag in nekatere empirične ugotovitve, kako uporabniki sprejemajo te razlage. Število raziskav na tem področju se je v zadnjih letih močno povečalo, pri čemer pa razni avtorji uporabljajo različne definicije in kriterije. Kljub veliki količini raziskav, so nekateri vidiki razložljivosti in tehnični pristopi deležni premalo pozornosti, med drugim: razlaga zaporedij odločitev, upoštevanje uporabnikovega predznanja ter induktivno logično programiranje.

Ključne besede

Umetna inteligenca, XAI, razložljivost

Abstract

The paper reviews and assesses the state of the art of research and methods in explainable AI. The review includes proposed definitions of what is an explanation, and what are properties of good explanations. We give a rough overview of numerous existing approaches for generating explanations, concrete examples of explanations and some empirical findings of their acceptance by users. The amount of research in this area has recently increased significantly, but different authors use different definitions and criteria. Despite numerous projects in this area, some aspects of explainability and technical approaches are receiving little attention: explaining sequences of decisions, taking into account user's background knowledge, and inductive logic programming.

1 Uvod

Modeli strojnega učenja postajajo z uspehom globokega učenja in nevronskih mrež vseprisotni. Večina od nas se z njimi srečuje na vsakodnevni ravni, v obliki sistemov za priporočanje glasbe in filmov npr. Taki sistemi brez posredovanja človeka izračunajo za nas najboljše priporočilo, morebitna neustrezna priporočila pa nimajo bistvenih (negativnih) posledic za nas. Nasprotno imajo lahko napačne odločitve v domenah (kot je npr. zdravstvo) odločilne posledice za konkretna življenja ljudi. Če v nekaterih domenah zadošča zgolj točna napoved sistema, to ne zadostuje povsod v družbi in znanosti nasploh [5].

Uporabnost modelov strojnega učenja je vodila v njihovo splošno uporabo pred razvojem kakovostnega konceptualnega okvirja, ki bi omogočal razumevanje njihovega delovanja. Znan je t. i. problem črnih škatel (ang. *black box problem*), ki pomeni, da delovanje modelov strojnega učenja ostaja za uporabnike nerazumljivo. Prav pomanjkanje razumevanja omejuje nadaljnjo in bolj praktično uporabo modelov v ostalih pomembnih domenah odločanja. Potreba po razlagi je vodila v razvoj tehnik in pristopov razložljive umetne inteligence (XAI; ang. *eXplainable Artificial Intelligence*), ki se posveča nalogi razlaganja kompleksnih modelov strojnega učenja [34].

Namen članka je pregled trenutnega stanja XAI področja in analiza pomanjkljivosti.

2 Kaj sploh je razlaga?

Razložljivost je izmuzljiv pojem ne samo na področju umetne inteligence (UI), pač pa širše na področju filozofije in drugih družboslovnih znanosti. Na področju UI se operira s koncepti, kot so vzročnost, informativnost, razumevanje, gotovost, zaupanje, transparentnost ipd. [5] Termin ‘razložljiva umetna inteligenca’ je l. 2019 kot del svojega programa uporabila DARPA [17]. Od takrat je postal zelo popularen, ne gre pa za nov pojav. Kvečjemu gre za imenovanje dolgoletnih prizadevanj, kjer se raziskovalci trudijo prebiti do odgovora na vprašanje, zakaj je sistem prišel do določene napovedi [19].

Najbolj splošno bi lahko razložljivost v domeni UI opredelili kot razlago, ki delovanje modela naredi bolj razumljivo. Seveda je to zelo splošna opredelitev, v poskusih bolj natančnega definiranja pa si raziskovalci niso zedinjeni. [12] opredelita razložljivost kot sposobnost predstaviti nekaj v človeku razumljivih terminih. [5] pravijo, da mora model nuditi razlago za svoje delovanje in napovedi v obliki vizualizacije pravil in vpogleda v potencialne spremenljivke, ki bi lahko povzročile perturbacije modela. Po [29] razložiti pomeni predstaviti besedilne ali vizualne elemente, ki omogočajo kvalitativno razumevanje odnosa med komponentami in napovedjo modela.

Ena od nekonsistentnosti v XAI literaturi je uporaba pojma interpretabilnost, ki je včasih sinonim razložljivosti, drugič ločen

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

pojem, tretjič ena od kategorij razložljivosti. [5] interpretabilnost razumejo kot pasivno, razložljivost pa kot aktivno lastnost modela. Interpretabilni so modeli, ki so razumljivi že sami po sebi (odločitvena drevesa npr.), razložljivi pa tisti, ki zahtevajo postopke, katerih namen je pojasnjevanje. Kot taka je razložljivost nujna lastnost vseh (tudi inherentno interpretabilnih) modelov [14].

Očitno je pomanjkanje konsenza o glavnih konceptih. Problem je, ker vsaka definicija nastopa znotraj specifičnega konteksta, odvisnega od naloge, sposobnosti in pričakovanj raziskovalca. Opredelitve razložljivosti so tako pogosto vezane na specifično domeno. Posledično XAI področje še ni enotno glede definicije razlage, specifičnih ciljev in kriterijev, ki naj bi jim zadostovali modeli, da bi bili razumljivi [5].

3 Metode razlag

Danes obstajajo številne metode razlag. Problem nastane pri njihovi klasifikaciji, ker ima praktično vsak avtor specifično definicijo razložljivosti, iz katere izhaja.

Ena splošnih kategorizacij je delitev na lokalne in globalne razlage. Lokalne so razlage, središčene okoli posameznega primera, kjer pa ostane delovanje modela kot celote nepojasnjeno. Na drugi strani globalne razlage pomagajo razumeti celoten model, so pa pogosto osnovane na približnih vrednostih [3][18][21][34].

Splošna je delitev na model-specifične in agnostične razlage. Prve s tehnikami, kot so relevantnost atributov, vizualizacija ali simplifikacija, pridobijo določene informacije o postopku napovedovanja in so uporabne za vsako vrsto modela [5]. Model-specifične razlage so uporabne zgolj za specifične vrste modelov (npr. maksimizacija aktivacije, ki jo opišejo [16]) [18].

[5] ločijo besedilne, vizualne, lokalne, razlage s primeri, s simplifikacijo in relevantnost atributov. [11] opredelijo tri glavne kategorije razlag: osnovane na funkciji, na primerih in pojasnjevanju atributov. [34] ločita razlage atributov in razlage primerov. [1] razlage razdelijo glede na uporabljeno metodologijo in ločijo med razlagami, ki slonijo na vzvratnem razširjanju (ang. *backpropagation*) in razlagami s perturbacijami. [16] ločijo: 1) odločitvena drevesa; 2) razlage, osnovane na pravilih; 3) razlage pomembnosti atributov, ki predstavijo težo in pomembnost atributov, ki jih je pri svoji napovedi upošteval model. Primer je znana metoda LIME, primerna predvsem za razlago klasifikacije besedil in slik (slika 2) [29]; 4) zemljevidi pomembnosti, ki izpostavijo ključne aspekte predmeta, ki je analiziran. Primer je metoda CAM (slika 1) [36]; 5) PDP (*Partial Dependence Plot*), kjer grafično prikažemo odnos med odločitvijo modela in vhodnimi podatki; 6) razlaga s prototipi, kjer z napovedjo dobimo primer, podoben našemu; 7) maksimizacija aktivacije, kjer opazujemo, kakšni vzorci vhodnih podatkov maksimizirajo aktivacijo določenega nevrona oz. nivoja.

Slika 1 (levo): razlaga CAM metode na način prikaza področij slike, ključnih za klasifikacijo umivanja zob [36]. Slika 2 (desno): razlaga LIME metode. Na levi je izvorna slika, na desni razlaga za klasifikacijo električne kitare [29].

[21] predstavi pojem formalne razložljivosti, zasnovan na logiki, kjer so razlage posledično bolj zanesljive in držijo globalno. Pristop temelji na računanji t. i. *prime implicants* (ang.), kar omogoča logične reprezentacije delovanja modela.

4 Kakšna je dobra razlaga?

Če je eden od ključnih ciljev XAI področja izboljšanje zaupanja v sisteme UI je nujno, da se pozornost usmeri k uporabnikom teh sistemov [35]. Dobre razlage bodo tiste, ki bodo upoštevale, komu so namenjene [5]. To pomeni upoštevanje predznanja, ki ga imajo uporabniki. Opazen je trend, kjer razvijalci metod razlag tega ne upoštevajo dovolj. [30] opredelita tri skupine uporabnikov (razvijalci in raziskovalci, eksperti in laiki), ki zahtevajo različne vrste razlag.

Med raziskovalci ni strinjanja o kriterijih za dobro razlago. V nadaljevanju navajamo nekaj primerov kriterijev. [3] opredelita tri:

- Eksplicitnost: razlaga je takojšnja in razumljiva;
- Zvestoba (ang. *faithfulness*): ocene relevantnosti odražajo resnično pomembnost;
- Stabilnost: za podobne vhodne podatke veljajo podobne razlage.

[11] poudarjajo:

- Robustnost: ob spremembi vhodnega podatka se ustrezno spremeni tudi razlaga:
- Zvestobo: razlaga ponazarja dejansko odločanje modela;
- Kompleksnost: kognitiven napor, potreben za razumevanje razlage;
- Homogenost: zmožnost razlage za pravilno razlago delovanja modela glede na različne skupine (v praksi se to po navadi nanaša na skupine, ki se ločijo glede na občutljive atribute).

[4] opredelita 4 aksiome, katerim naj bi zadostile dobre razlage:

1) morajo biti informativne;
2) ne smejo vsebovati nepotrebnih informacij;
3) razlage razredov morajo pojasniti posamezne primere, hkrati pa morajo biti splošno uporabne;
4) razlaga mora vsebovati samo informacije, ki vplivajo na napoved.

5 Ocenjevanje razlag

Ocenjevanje razlag je najmlajše področje s široko paleto pristopov [30]. Za razliko od točnosti, je kriterije kot so varnost, nediskriminacija in razložljivost težje kvantificirati [12].

Ocenjevanja se (najbolj splošno) lahko lotimo na dva načina: 1) človeško ocenjevanje ali 2) uporaba računskih metod, ki merijo, kako dobro razlaga dejansko razloži delovanje modela. Glavna razlika med pristopoma je, da so računske metode bolj objektivne, vendar pa ne upoštevajo človeškega faktorja. Drugače rečeno, ne kvantificirajo človeškega razumevanja.

Prednost človeške ocene je subjektivnost in večja deskriptivnost. Očitna pomanjkljivost je manjša točnost in večja odvisnost od specifične naloge [27].

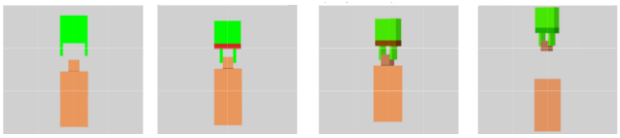
1.4.1. Računske razlage

[20][35] predstavijo matematično ocenjevanje razlag na podlagi analize robustnosti.

Matematično opredeljena mera nezvestobe ponazarja, kako dobro se razlaga ujema z modelom. Če spremenimo vhodni podatek pričakujemo, da se bo spremenila tudi napoved [34].

1.4.2. Človeško ocenjevanje

[13] so izvedli eksperiment, s katerim so preverili, kakšne razlage so pri ljudeh vzbudile največ zaupanje v robota, ki je odprl stekleničko. Robot se je naučil odpirati stekleničke iz človeških demonstracij, pri čemer je bilo ključno učenje zaporedij položaja rok in potrebne sile. Z rokavico s senzorji so zajeli podatke o sili in položaju rok v 64 človeških demonstracijah s tremi različnimi stekleničkami. Sledilo je kompleksnejše učenje, da bi bil robot svoje znanje sposoben posplošiti. Implementiran je bil haptični model, ki je robotu pomagal določiti potrebno silo, čeprav nima človeških rok. Ker odpiranje stekleničke poteka v več korakih (potiskanje, odvijanje itd.), je bil implementiran še t. i. (ang.) *symbolic action planner* in pomeni pravila o zaporedju potrebnih akcij. S kombinacijo takega učenja je robot postal precej dober v odpiranju novih stekleničk. Udeleženci so bili razdeljeni v 5 skupin. Vsaka je videla posnetek robota, ki opravlja nalogo, ter eno od možnih razlag (simbolično: v realnem času so udeleženci videli z eno besedo opisano akcijo, ki naj bi razlagala, kaj robot na posnetku dela (*approach – grasp – push – twist – ungrasp – move – grasp – push …*); besedilno: po ogledu posnetka robota, so udeleženci prebrali kratko besedilo, o tem, kako je robot opravil nalogo (*I succeeded to open the bottle because I pushed on the cap three times and twisted the cap twice*); oz. haptično razlago (slika 3): vizualizacija sile prijema v vsakem trenutku odpiranja stekleničke) oz. kombinacijo haptične in simbolične razlage. Največ zaupanja je spodbudila simbolična razlaga.



Slika 3: haptična razlaga.

[27] so izvedli eksperiment, kjer so udeleženci označili relevantna področja slike, ki je po njihovem mnenju bilo najbolj reprezentativno za določen razred objektov (mačka in pes npr.). Rezultat je zemljevid pomembnosti, ki prikazuje področja slike, ki so jim udeleženci posvečali največ pozornosti (spodnja vrsta na sliki 4). Te rezultate so primerjali z zemljevidi pomembnosti metode Grad-CAM (spodnja vrsta na sliki 4). Zemljevidi so si morda podobni, vseeno pa je statistično testiranje pokazalo pomembne razlike. Distribucija relevantnih atributov je bila pri Grad-CAM metodi bolj uniformna, udeleženci so v primeru živih bitij kot ključne bolj označevali obraze. Prav to so ugotovitve, ki nam lahko pomagajo razumeti, kako dobre so razlage.

Slika 4: zgornja vrsta prikazuje zemljevida pomembnosti Grad-CAM metode, spodnja zemljevida udeležencev [27].

6 Kako naprej?

V tem razdelku opozorimo na nekatere razmeroma slabo raziskane probleme in premalo uporabljene pristope za XAI.

6.1 Tehtanje med točnostjo in razložljivostjo

[31] v članku z zgovornim naslovom »*Stop explaining black box ML models for high stakes decisions and use interpretable models instead*« izraža determinirano stališče. Zavzema se za uporabo metod učenja, ki dajejo naučene modele, ki so sami po sebi razumljivi. Za take se smatrajo npr. odločitvena drevesa. Nasprotuje metodam učenja, katerih rezultati so v principu težko razumljivi. Med te štejemo posebno metode globokega učenja, ki sicer dosegajo visoko napovedno točnost v primerjavi z drugimi metodami učenja, toda ne zastonj: vsaj za ceno razumljivosti in potrebnega velikega števila podatkov za učenje. Pri tem gre Rudin morda res predaleč s svojim optimističnim stališčem, ki implicitno predpostavlja možnost izgradnje elegantnih in razumljivih modelov za vsako problemsko domeno, s čimer zadane ob princip kompleksnosti Kolmogorova.

Glede možnosti obstoja enostavnih modelov in razlag velja vsaj ena teoretična omejitev, ki jo definira kompleksnost Kolmogorova, ki določa, koliko spominskega prostora potrebujemo za najkrajši možni zapis danega objekta v računalniku. Obstajajo zapleteni objekti (torej tudi napovedni modeli), ki jih niti teoretično ni mogoče predstaviti na kratek način. V takih primerih tudi razlaga ne more biti kratka in enostavna. Res pa je, da smo v praksi še zelo daleč od te teoretično dosegljive meje, torej imamo veliko prostora za izboljšanje. Ko zadenemo ob zid Kolmogorova, pa je še vedno možen kompromis, da za boljšo razložljivost žrtvujemo nekaj točnosti [8]. Primer tehnične izvedbe tega tehtanja med točnostjo in razumljivostjo v učenju odločitvenih dreves je [6].

6.2 Navezava razlage na uporabnikovo predznanje

Če bo razlaga dobra, je odvisno od njenega uporabnika, konkretno od uporabnikovega predznanja o problemski domeni. Če je to kvalitetno, zadošča en sam namig. Če je razumevanje domene slabo, je potrebna podrobna in daljša razlaga. Tudi sama formulacija razlage je odvisna od obstoječega znanja na obravnavanem področju. Celo povsem pravilna in jedrnata razlaga je za eksperta na področju uporabe lahko nesprejemljiva in nenaravna. Kot primer omenimo, da so se nekateri primeri razlag, ki jih generirajo naučeni modeli v medicinskih domenah kljub svoji diagnostični točnosti zdravniku zdeli povsem

nenaravni [8]. V enem od primerov je sistem razložil, da gre za vnetni revmatizem, ker ima pacient med drugim več kot dva prizadeta sklepa na roki. To diagnostično pravilo je dejansko točno. Vendar pa je zdravnik vztrajal, da mora imeti pacient prizadete sklepe na vseh petih prstih na roki, ker vnetni revmatizem tipično vpliva na vse sklepe. Ekspertno mnenje je bilo v tem primeru zelo jasno, čeprav je res, da bo pravilo vodilo do pravilne diagnoze v vsakem primeru; če gre za katerokoli število vnetih sklepov med 2 in 5. Ustreznost razlage je odvisna ne le od klasifikacijske točnosti, temveč (tudi) od predznanja, ki ga ima uporabnik o tej obliki revmatizma.

Obstoječe metode razlage ta vidik pogosto ignorirajo. Problem je tudi v tem, da ne omogočajo naravne uporabe predznanja. V tem pogledu je zelo obetaven pristop k strojnemu učenju t. i. induktivno logično programiranje (ILP), ki temelji na uporabi matematične logike. Že sama osnovna formulacija problema učenja v ILP vsebuje uporabo predznanja: dani so učni primeri E in predznanje BK (*background knowledge*), naloga učenja pa je sestaviti logično formulo H (hipoteza) tako, da primeri E logično sledijo iz BK in H.

Pristop ILP je skromno zastopan v obstoječih raziskavah iz strojnega učenja in razložljivosti. Lep primer njegove ustreznosti so raziskave, opisane v [2][28]. Te zasledujejo ne le osnovni cilj XAI (razlage odločitev strojnega učenja), temveč tudi cilj t. i. »ultra-razložljivost«. Ta strožji kriterij strojnega učenja je definiral [26] (ang. *ultra strong criterion for ML*). Strojno učenje je ultra-razložljivo, če je ne le razložljivo, temveč uporabniku omogoča tudi *operativno uporabo za lastno reševanje* novih problemov. Npr. da strojno naučeno znanje lahko uporabi za lastno reševanje določenih matematičnih problemov ali igranje šaha.

6.3 Razlaga zaporedij odločitev v planiranju

Večina XAI metod generira razlago *posameznih* odločitev oz. klasifikacij. Pri razložljivem planiranju pa gre za razlago množice odločitev (npr. zaporedja akcij, ki robota vodi do cilja). Posebej za razlago planov se je formiralo področje razložljivega planiranja [10].

Razlaga planov je navadno zahtevnejša od razlage v klasifikacijskih problemih. Treba je razložiti, kako so posamezne akcije odvisne od drugih, da skupaj rešijo nalogo. Primer razlage zaporedja odločitev je razlaga šahovskih partij, kjer je treba razložiti celo zaporedje potez ali drevo odločitev, ki definira uspešno strategijo. Primer, opisan v [9], so težko razumljive in briljantne poteze šahovskega programa AlphaZero.

Razlaga planov je aktualna tudi na področju vodenja sistemov. Lep primer razlage naučenega plana vodenja je v [32]. Gre za klasično nalogo iz teorije vodenja sistemov: vodenje sistema voziček-palica. Na vozičku je vrtljivo vpeta palica. Palica je postavljena približno vertikalno, vendar se, če ne ukrepamo, prevrne na tla. S potiskanjem vozička levo oz. desno je treba loviti ravnotežni položaj palice okrog vertikale, obenem pa doseči, da se voziček horizontalno premakne iz začetnega položaja do cilja. Naučena strategija vodenja je lepo razložljiva. Najprej nekoliko presenetljivo potisnemo voziček v nasprotno stran od cilja, s čimer dosežemo, da se palica nagne proti cilju. To omogoči potiskanje vozička v smer proti cilju, hkrati pa se ohranja ravnotežje palice, ko je ta nagnjena naprej v smeri cilja.

7 Zaključek

Področje XAI se je v zadnjih 5 do 10 letih močno razraslo. Mnogi zato predpostavljajo, da je bil to tudi začetek področja. V resnici je aktivno zavedanje, da naj bi bilo strojno učenje razložljivo, obstajalo že prej 40 leti. Že takrat so obstajale raziskave o razložljivih modelih. Kljub sedanji količini raziskav in nedvoumnih uspehih se še vedno kaže, da pogrešamo nekatere ključne odgovore. Npr., že pred desetletji se je v sklopu istih prizadevanj pojavilo zavedanje, da potrebujemo formalne mere za ocenjevanje kvalitete razlag. Take sprejete mere še ni. Raziskovalci pri ocenjevanju razlag uberejo različne pristope, odvisne od raznih kriterijev (konteksta, domene, uporabnikov itd.).

Glede vprašanja, kaj je sprejemljiva razlaga, se v pomanjkanju boljših splošnih in principielnih kriterijev v sedanji praksi uporablja predpostavka, da so nekateri modeli razložljivi kar po definiciji, torej razložljivi sami po sebi. Mednje npr. navadno štejemo odločitvena drevesa ali pravila če-potem. Toda tudi ta kriterij je arbitraren. Kaj, če je odločitveno drevo zelo veliko, npr. da ima milijon vozlišč?

V prispevku smo opozorili tudi na počasen napredek pri razvoju metod za razlago zaporedij odločitev. Sem sodi razlaga planov za reševanje nalog, ki imajo eksplicitno definirane cilje. Plan je lahko zaporedje akcij ali pa tudi množica akcij, ki so delno urejene v času. Tu je treba razložiti tudi to, kako se akcije med seboj dopolnjujejo in na kakšen način skupaj dosežejo cilj. S tem so povezani izzivi, ki jih predstavimo spodaj.

En možen pristop, ki upošteva principe planiranja v UI je upoštevanje odvisnosti med akcijami. Nekatere akcije v planu neposredno dosežejo kakšnega od ciljev plana. Druge akcije pa ne dosežejo nobenega danega cilja neposredno, njihova funkcija je, da dosežejo pogoje, ki morajo biti uresničeni, da je možno izvesti druge akcije v planu. Taka razlaga plana je seveda povsem logična. Navadno pa vsebuje preveč podrobnosti. Če plan vsebuje nekoliko večje število akcij, npr. nekaj 10, postane tako podrobna razlaga spet težko razumljiva in za uporabnika neprivlačna. V tem primeru bi za sprejemljivo razlago treba plan razbiti v hierarhično strukturo, definirano s podcilji plana. Odkrivanje smiselnih podciljev pa je lahko težavno. Poseben izziv je, kako poiskati take podcilje, ki rezultirajo v razlagi, ki je za človeka čim bolj naravna.

Zahvala

Prispevek je nastal v okviru ciljnega raziskovalnega projekta V2-2272 Opredelitev okvira za zagotavljanje zaupanja javnosti v sisteme umetne inteligence in njihove uporabe, ob podpori Javne agencije za raziskovalno in inovacijsko dejavnost Republike Slovenije in Ministrstva za digitalizacijo.

Literatura

[1] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion, 99*. DOI: https://doi.org/10.1016/j.inffus.2023.101805.

[2] Ai, L., Langer, J., Muggleton, S. H. and Schmid, U. 2023. Explanatory machine learning for sequential human teaching. *Machine Learning Journal, 112*, 3591-3632. DOI: https://doi.org/10.1007/s10994-023-06351-8.

[3] Alvarez-Melis, D and Jaakkola, T. S. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. 32nd Conference

on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.

[4] Amgoud, L. and Ben-Naim, J. 2022. Axiomatic Foundations of Explainability. Proceedings of the Thirty-First International joint Conference on Artificial Intelligence (IJACAI-22), 636-642.

[5] Barredo Arrieta, A., Diaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gill-López, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv: 1910.10045v2*. DOI: https://doi.org/10.48550/arXiv.1910.10045.

[6] Bohanec, M. and Bratko, I. 1994. Trading accuracy for simplicity in decision trees. *Machine Learning Journal, 15*, 223-250. DOI: https://doi.org/10.1007/BF00993345.

[7] Brasse, J., Broder, H. R., Förster, M., Klier, M. and Sigler, I. 2023. Explainable artificial intelligence in information systems: A Review of the status quo and future research directions. *The International Journal of Networked Business, 33*(1). DOI: https://dx.doi.org/10.1007/s12535-023-00644-5.

[8] Bratko. I. 1997. *Machine learning: between accuracy and interpretability*. V: Learning, Networks and Statistics (ed. Della Riccia, G.), Vienna: Springer.

[9] Bratko, I. 2018. AlphaZero: what's missing? *Informatica, 42*(1).

[10] Chakraborti, T., Sreedharan, S. and Kambhampati, S. 2020. The Emerging Landscape of Explainable Automated Planning & Decision Making. *arXiv:2002.11697*. DOI: https://doi.org/10.48550/arXiv.2002.11697.

[11] Chen, Z., Subhash, V., Havasi, M., Pan, W. and Doshi-Velez, F. 2022. What Makes a Good Explanation?: A Harmonized View of Properties of Explanations. *Progress and Challenges in Building Trustworthy Embodies AI (TEA 2022) co-located with NeurIPS 2022.*

[12] Doshi-Velez, F. and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608*. DOI: https://doi.org/10.48550/arXiv.1702.08680.

[13] Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H. and Zhu, S. C. 2019. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics, 4*(3), 1-13.

[14] Gilpin, L. H., Yuan, B. Z., Bajwa, A., Specter, M and Kagal, L. 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00068*. DOI: https://doi.org/10.48550/arXiv.1806.00069.

[15] Grice, H. P. 1975. Logic and conversation, syntax and semantics. *Speech Acts 3*, 41-58.

[16] Guidotti, D., Monreale, A., Ruggieri, S. and Turini, F. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys, 51*(5), 1-42. DOI: https://doi.org/10.1145/3236009.

[17] Gunning, D., Vorm, E., Wang, J. Y. and Turek, M. 2021. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters, 2*(4). DOI: https://doi.org/10.1002/ail2.61.

[18] Hall, P., Ambati, S. and Phan, W. 15.3.2017. *Ideas on interpreting machine learning.* O'Reilly. https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/.

[19] Holsinger, A., Saranti, A., Molnar, C., Biecek, P. and Samek, W. 2022. Explainable AI Methods – A Brief Overview. xxAI 2020, LNAI 13200, 13-38. DOI: https://doi.org/10.1007/978-3-031-04083-2_2.

[20] Hsieh, C. Y., Yeh, C. K., Liu, X., Ravikumar, P., Kim, S., Kumar, S. and Hsieh, C. J. 2021. Evaluations and Methods for Explanation through Robustness Analysis. *arXiv:2006.00442*. DOI: https://doi.org/10.48550/arXiv.2006.00442.

[21] Ignatiev, A., Narodytska, N. and Marques-Silva, J. 2019. On Validating, Repairing and Refining Heuristic ML Explanations. *arXiv: 1907.02509.* DOI: https://doi.org/10.48550/arXiv.1907.02590.

[22] Kolmogorov, A. 1963. On Tables of Random Numbers. *The Indian Journal of Statistics, Series A, 25*, 369-375.

[23] Krishnan, M. 2020. Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology, 33*(3), 487-502. DOI: http://dx.doi.org/10.1007/s13347-019-00372-9.

[24] Lombrozo, T. 2006. The structure and function of explanation. *Trends in Cognitive Science, 10*(10), 464-470. DOI: https://doi.org/10.1016/j.tics.2006.08.004.

[25] Malle, B. F. 2004. How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. MIT Press.

[26] Michie, D. 1998. Machine Learning in the next five years. *Proceedings of the 3rd European working session on learning*, 107-122.

[27] Mohseni, S., Block, J. E. and Ragan, E. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. *26th International Conference on Intelligent User Interfaces (IUI'21)*. DOI: https://doi.org/10.1145/3397481.3450689.

[28] Muggleton, S., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A. & Besold, T. 2018. Ultra-strong machine learning: Comprehensibility of programs learned with ILP. *Machine Learning, 107*, 1119-1140. DOI: https://doi.org/10.1007/s10994-018-5707-3.

[29] Ribeiro, M. T., Singh, S. and Guestrin, C. 2016. "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*. DOI: https://doi.org/10.1145/2939672.2939778.

[30] Ribera, M. and Lapedriza, A. 2019. Can we do better explanations? A proposal of User-Centered Explainable AI. *IUI Workshops '19.*

[31] Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. DOI: https://doi.org/10.1038/s42256-019-0048-x.

[32] Šoberl, D. and Bratko, I. 2023. Transferring a Learned Qualitative Cart-Pole Control Model to Uneven Terrains. *International Conference on Discovery Science*, 446-459. DOI: http://dx.doi.org/10.1007/978-3-031-45275-8_30.

[33] Tjoa, E. and Guan, C. 2020. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *arXiv: 1907.07374.* DOI: http://doi.org/10.48550/arXiv.1907.07374.

[34] Yeh, C. K. and Ravikumar, P. 2021. Objective criteria for explanations of machine learning models. Applied AI Letters published by John Wiley & Seons Ltd. DOI: https://doi.org/10.1002/ail2.57.

[35] Zhang, Z., Xu, L., Yilmaz, L. and Liu, B. 2021. A Critical Review of Inductive Logic Programming Techniques for Explainable AI. *arXiv: 2112.15319*. DOI: https://doi.org/10.48550/arXiv.2112.15319.

[36] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. 2016. Learning Deep Features for Disciminative Localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921-2929. DOI: 10.1109/CVPR.2016.319.