Dôveryhodnosť výpočtových modelov v umelej inteligencii a robotike*

Igor Farkaš

Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave
igor.farkas@fmph.uniba.sk

Abstrakt

S búrlivým rozvojom umelej inteligencie v rôznych oblastiach súvisí aj zvyšovanie kritérií na navrhované a využívané riešenia. Dôveryhodnosť ako pojem prevzatý z psychológie rezonuje v súčasnej umelej inteligencii a robotike. Týka sa nielen programov implementovaných v počítači, ktoré produkujú požadované výstupy, napr. na displeji, ale aj programov ovládajúcich robotické systémy, ktoré priamo zasahujú do prostredia alebo interagujú s človekom. Požiadavka dôveryhodnosti sa javí samozrejmá z pohľadu človeka, ktorý sa potrebuje na umelý systém spoľahnúť. Zložitejšou otázkou je, čo tento pojem znamená v matematickom zmysle a ako dôveryhodnosť zabezpečiť a merať. Na tieto aspekty poukazujeme v príspevku, kde spomenieme dôležité koncepty, v kontexte súvisiacich pojmov, a so zvláštnym dôrazom na humanoidné roboty interagujúce s človekom.

1 Úvod

Umelá inteligencia (UI) sa teší v ostatnej dekáde vysokej popularite v rôznych aplikačných oblastiach. Úspešné modely UI typicky fungujú vďaka strojovému učeniu založenému na hlbokých neurónových sieťach, ktoré umožnili nachádzať úspešné riešenia rozmanitých úloh ako spracovanie obrazových dát (klasifikácia do tried), úlohy v prirodzenom jazyku či sekvenčné rozhodovacie úlohy (napr. hranie hier proti ľudským súperom) (Schmidhuber, 2015).

Napriek svojim úspechom majú mnohé súčasné UI systemy aj rôzne nedostatky, hlavne to, že sú zraniteľné voči nepostrehnuteľným, tzv. adverzariálnym útokom, nedávajú objektívne výsledky (sú zaujaté voči nedostatočne zastúpeným triedam dát v klasifikačných úlohách), no najmä, sú málo zrozumiteľné, a to nielen pre bežných ľudí ale aj odborníkov. Tieto nedostatky zhoršujú používateľskú skúsenosť a narúšajú dôveru ľudí vo všetky systémy UI. K negatívnemu povedomiu o UI prispieva aj zábavný priemysel, ktorý už desiatky rokov chrlí filmy s emocionálne nabitou tematikou, že svet ovládne umelá inteligencia, najčastejšie v stelesnená v neohroziteľných humanoidných robotoch alebo iných ničivých agresoroch.

*podporené projektmi VEGA 1/0373/23 a APVV-21-0105.

2 Dôveryhodnosť systémov UI

Je mimoriadne dôležité podnikať účinné kroky opačným smerom a zvyšovať objektívnu informovanosť v spoločnosti. Budovanie dôveryhodných systémov UI je zložitý proces, ktorý bude musieť podchytiť rôzne súvisiace aspekty ako sú robustnosť, dostatočná generalizácia, vysvetliteľnosť, transparentnosť, reprodukovateľnosť a ďalšie. Li a spol. (2023) ponúkajú zjednocujúci pohľad na v súčasnosti dostupné, ale roztrieštené prístupy k dôveryhodnej UI, a organizujú ich systematicky. Identifikujú aj kľúčové príležitosti a výzvy pre budúci vývoj dôveryhodných systémov UI.

Koncept dôveryhodnosti je človeku intuitívne zrejmý, nakoľko ju posudzujeme u ľudí podľa ich vonkajšieho správania a konania, hoci berieme do úvahy aj názory iných ľudí. Najspoľahlivejšie však dospejeme k názoru na základe vlastných skúseností. Limitáciou tohto prístupu je, že človeku do jeho mysle nevidíme, preto nás môže niekedy jeho správanie prekvapiť alebo človek oklamať.

V prípade požadovania dôveryhodnosti u systémov UI platia podobné princípy, avšak sú tu aj rozdiely: (1) vlastnosti systému UI môžeme ovplyvniť pri jeho dizajne, a (2) do vnútra systému môžeme nahliadnuť, keďže ide o výpočtový model.

Podľa existujúcej literatúry, dôveryhodný systém musí spĺňať viacero vlastností. Nezávislá expertná skupina na vysokej úrovni pre umelú inteligenciu, zriadená Európskou komisiou (EK) v roku 2018 vypracovala Etické usmernenia pre dôveryhodnú UI, ktorá by mala byť:

1. **zákonná**, čiže by sa mala riadiť celým platným právom a právnymi predpismi;
2. **etická**, čiže by mala zabezpečiť súlad s etickými zásadami a hodnotami, a
3. **odolná**, aby nebola zraniteľná voči zneužitiu, a bola bezpečná.

Každá zložka je sama o sebe potrebná, ale nestačí na dosiahnutie dôveryhodnej UI. V ideálnom prípade pôsobia všetky tri zložky vo vzájomnom súlade a prekrývajú sa vo svojom fungovaní. Ak medzi týmito zložkami v praxi vzniknú rozpory, spoločnosť by sa mala pokúsiť o ich zosúladenie.

Zákonný aspekt sa zdá byť zrejmý, spomenieme len, že EK navrhla štyri stupne regulácie UI (podľa miery rizika; tzv. AI Act), podľa ktorých sa bude musieť nasadzovanie jednotlivých systémov UI riadiť. Samozrejme možno očakávať pribúdanie zákonov týkajúcich sa UI.

Etický aspekt zahŕňa viacero ďalších požiadaviek, ako sú transparentnosť, vysvetliteľnosť, bezpečnosť a férovosť. Tieto pojmy už stihli získať v literatúre dosť pozornosti, najviac asi požiadavka vysvetliteľnosti (Barredo a spol., 2020), ktorá vnáša dôveru v systém, ak vieme vysvetliť jeho správanie. V prípade hlbokých neurónových sietí sú vysvetlenia v princípe zložité, a úroveň vysvetlenia závisí aj od toho, komu je určené (užívateľovi, pacientovi, expertovi). Expert rozumie aj matematickým vzorcom, zatiaľ čo bežný človek uprednostní vysvetlenie v prirodzenom jazyku alebo na obrázkoch.

Odolnosť (robustnosť) neurónových sietí je relatívne nový problém, ktorý bol zistený pred necelými 10-imi rokmi. (Szegedy a spol., 2014) zistili, že minimálna cielená perturbácia vstupov (teda útok) prezentovaných dobre natrénovanej a fungujúcej neurónovej sieti spôsobí jej spoľahlivé oklamanie (čiže sieť si bude "myslieť", že vidí niečo úplne iné). Jedným zo smerov, ktoré by mohli pomôcť riešiť tento problém je, že zavedieme do modelu mechanizmus pozornosti (Farkaš a spol., 2022). Existujúce modifikácie učenia pomocou tzv. adverzariálneho tréningu trochu pomáhajú liečiť symptómy, že takto zdokonalené modely vedia odolať aspoň niektorým útokom, nie sú však (aspoň zatiaľ) principiálnym riešením (Zhao a spol., 2022).

3 Dôveryhodnosť UI v robotike

Dôveryhodnosť UI implementovanej vo fyzických robotoch je veľmi dôležitá, pretože narastá význam úlohy robotov, nielen humanoidných, ktoré budú interagovať s človekom, napr. v priemysle, v domácnosti a pod. (Kok a Soh, 2020). V tomto kontexte ešte viac akcentuje aspekt bezpečnosti človeka (vyjadrený už prvým Asimovým zákonom robotiky), preto sa častejšie začína využívať pri vývoji systémov virtuálna realita na interakciu robota a človeka (Dianatfar a spol., 2021), a na počítačové simulácie robotov.

Výskum v tejto oblasti je v počiatkoch. Jedným zo smerov je požadovaná transparentnosť robota, ktorú môžeme navonok posudzovať podľa jeho správania, keďže každý robot koná vo svojom prostredí. Človek sa prirodzene snaží predikovať správanie iných ľudí, za čím sú (vedomé i nevedomé) mechanizmy prebiehajúce na viacerých úrovniach súčasne (Bubic a spol., 2010). Na motorickej úrovni sa človek snaží predikovať napríklad pohyb ruky alebo tela, na mentálnej úrovni zase úmysel druhého človeka (využívajúc teóriu mysle). Pri skúmaní UI v robotoch máme výhodu, že máme prístup k obom úrovniam - viditeľnej (motorickej) i skrytej (mentálnej), pretože môžeme analyzovať bežiace algoritmy UI.

V prípade motorickej úrovne, ktorá často prezrádza zámery človeka, je jedným zo zaujímavých konceptov čitateľnosť (legibility) pohybu, čo znamená ako rýchlo vieme správne predpovedať cieľ pohybu pozorovaním počiatočnej trajektórie, napríklad robotického ramena. K tomu sa pripája sprevádzajúci pohľad očí, ktorý napomáha predikcii cieľa motorického pohybu. V snahe o bezproblémovú a hladkú interakciu medzi človekom a robotom je dôležité vybaviť roboty náležitými perceptuálnymi vlastnosťami ako schopnosť sledovať a predikovať pohyby ruky človeka a jeho pohľadu. Je známe, že antropomorfné vlastnosti robota napomáhajú lepšej interakcii medzi človekom a robotom. Všetky tieto procesy možno chápať ako výpočtové na oboch stranách (Thomaz a spol., 2013) a ich analýza, modelovanie a testovanie sú otázkou operacionalizácie a behaviorálnych experimentov.

Literatúra

Barredo, A. a spol. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bubic, A., von Cramon, D. Y. a Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4.

Dianatfar, M., Latokartano, J. a Lanz, M. (2021). Review on existing VR/AR solutions in human–robot collaboration. *Procedia CIRP*, 97:407–411.

Farkaš, I., Cimrová, B., Štefan Pócoš a Bečková, I. (2022). Pozornosť ako biologicky inšpirovaný koncept pre vysvetliteľné, robustné a efektívne strojové učenie. V *Kognícia a umelý život XXI*, str. 34–38.

Kok, B. a Soh, H. (2020). Trust in robots: Challenges and opportunities. *Current Robotics Rep.*, 1:297–309.

Li, B. a spol. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9):1–46.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Szegedy, C. a spol. (2014). Intriguing properties of neural networks. V *International Conference on Learning Representations*.

Thomaz, A., Hoffman, G. a Cakmak, M. (2013). Computational human-robot interaction. *New Foundations and Trends*, 4(2-3):105–223.

Zhao, W., Alwidian, S. a Mahmoud, Q. H. (2022). Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8).