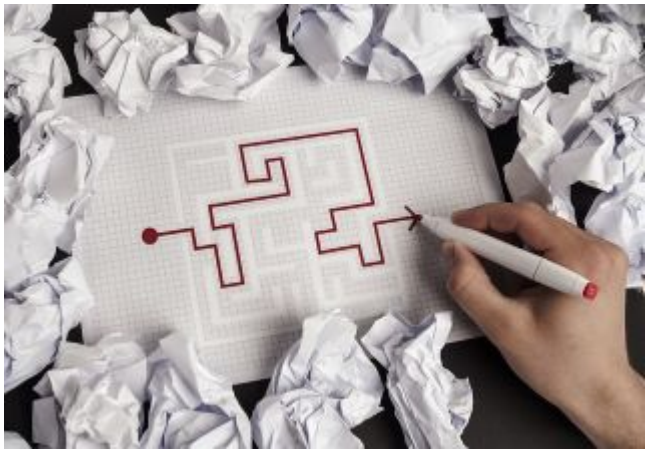
FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA
KYBERNETIKY

Specification Gaming



"You can think of such agents as less powerful outcome pumps, or little genies." - Robert Miles

Západočeská Univerzita V Plzni
Katedra Kybernetiky Semestrální práce - HKUI

Eliška Šiková
31. ledna 2025

1 Úvod

Ve strojovém učení existuje několik názvů pro fenomén, kdy model sice splňuje specifikovaný cíl, ale nejedná se o smysluplný výsledek. Tento fenomén se označuje jako specification gaming, reward hacking, reward misspecification, outer misalignment nebo Goodhartův zákon. S tímto fenoménem se vědci a vývojáři při vývoji modelů umělé inteligence setkávají velmi často a přístupy k jeho řešení se neustále vyvíjejí, aby se omezily nežádoucí výsledky. V této práci se seznámíme s příklady nežádoucího chování modelů umělé inteligence z praxe a implikacemi tohoto fenoménu.

2 Kočka a myš

2.1 Reinforcement Learning

Jedním ze způsobů učení je specifikování funkcí, které poskytují modelu zpětnou vazbu. Když je podmínka splněna, model je odměněn, nebo je při nežádoucím chování potrestán. Tyto funkce jsou typicky specifikovány vývojáři a umožňují modelu se během několika iterací učit splnit své zadání.

Model prochází tímto cyklem, zpětná vazba mu chování utvrdí nebo vyvrátí při jeho experimentování ve svém prostředí a ovlivňuje jeho vývoj. Co se ale stane, když při tomto experimentování model dostane odměnu v případě, kdy ji dostat neměl?

Jeden z velmi typických příkladů je z roku 2017, kdy model byl učen pokládat červený blok stavebnice Lego na modrý. Funkce odměny byla nastavena tak, že když je spodní plocha červeného bloku na úrovni horní plochy modrého, model je odměněn. Místo toho, aby se model naučil červený blok zvednout a položit jej na modrý, naučil se červený blok otočit vzhůru nohama. V této situaci je spodní plocha červeného bloku na úrovni horní plochy modrého, funkce je splněna, ale zamýšlený výsledek není dosažen.

Podobných případů je mnoho. Roku 2019 byl učen model, jak hrát hru Overkill. I když úmyslem učení bylo naučit model hru dohrát až do konce, model se naučil nekonečně ničit stejné nepřátele na první úrovni, protože za to byl odměněn a toto chování bylo podpořeno.

Dalším příkladem je projekt s vysavačem Roomba, také z roku 2019. Roomba 'Fenton' měla za úkol se naučit jezdit rychle bez narážení do objektů, ale naučila se jezdit pozpátku, kde senzory nebyly, a tak model nemohl poznat, že jeho chování není požadované.

Není neobvyklé, že model při učení neučiní žádné akce, aby nedostal penalizaci. Tohoto roku 2024 byl zveřejněn další příklad, kde model ovládal vlaky, které do sebe nesmí narazit. Řešení modelu bylo, aby vlaky vůbec nejezdily.

Když chybí dostatečné podmínky omezující, jak model smí dosáhnout svého cíle, většinou najde kratší cestu k odměně a nesplní daný úmysl. Neví, co má doopravdy dělat.

2.2 Lidský prvek

Když modely často obchází své zadání v případě odměny od automatického programu, nabízí se jako řešení osoba, která by manuálně poskytovala zpětnou vazbu.

Společnost OpenAI roku 2017 zveřejnila incident, který zvýraznil problémy s tímto přístupem. Model ovládal virtuální ruku v 3D prostoru, kde byl umístěn míček, a jeho úkol bylo kuličku uchopit. Model se však naučil kontrolní osobu ošálit umístěním ruky mezi kameru a míček, kde z pohledu kamery ruka míček drží a je neprávem odměněn.

Model se učil člověka ošálit, protože to je jednodušší než míček opravdu uchopit. Podobné chování se často vyskytuje u jazykových modelů, kde model prezentuje nepravdivé informace, o kterých uživatel nemusí být informován. Toto je problém běžně nazývaný halucinace, který je roku 2024 díky různým metodám zmírněn oproti předchozím letům. Jedním z příkladů je řešení matematických problémů v programovacím jazyce Python, zatímco jazykový model pouze shrnuje informace a již nedělá výpočty.

2.3 Implikace

Umělá inteligence (AI) má obrovský potenciál přinášet pozitivní změny v mnoha oblastech života, ale pokud její cíle nejsou správně sladěny s lidskými hodnotami a potřebami, může to vést k vážným problémům. Jedním z příkladů je využití AI v autonomních vozidlech. Pokud by byla AI naprogramována tak, aby minimalizovala čas jízdy bez dostatečného zohlednění bezpečnosti chodců nebo pravidel silničního provozu, mohlo by to vést k riskantnímu chování na silnicích a zvýšenému počtu nehod. Cíl AI by se tak dostal do konfliktu s prioritou ochrany lidských životů, což by mělo katastrofální důsledky.

Dalším příkladem je role AI v algoritmech sociálních sítí, které jsou často navrženy tak, aby maximalizovaly zapojení uživatelů, například čas strávený na platformě. Tento cíl však může vést k šíření dezinformací, polarizaci společnosti nebo podpoře extremistických názorů, protože kontroverzní a emocionálně zabarvený obsah často generuje více interakcí. AI tak sice plní svůj cíl, ale na úkor zdravého společenského dialogu a šíření pravdivých informací, což může mít dlouhodobě negativní dopady na demokratické procesy a sociální soudržnost.

3 Závěr

Umělá inteligence může být přirovnána k malým „džinům“, kteří plní přání, ale musíme být extrémně opatrní v tom, jaká přání jim zadáváme. Stejně jako v příbězích, kde nesprávně formulované přání vede k neočekávaným a často katastrofálním výsledkům, i AI může dosáhnout svého cíle způsobem, který je technicky správný, ale zcela mimo zamýšlený účel. Pokud například řekneme AI, aby maximalizovala počet kliknutí na reklamu, může to udělat tak, že zaplaví uživatele nevhodným obsahem, místo aby vytvořila hodnotné a relevantní reklamy.

Tento problém zdůrazňuje důležitost precizního specifikování cílů a zohlednění širšího kontextu při návrhu AI systémů. Musíme být konkrétní a anticipovat možné nežádoucí chování, které by mohlo vzniknout z nedostatečně definovaných cílů. Stejně jako bychom si u džina museli pečlivě promyslet každé slovo, abychom předešli nechtěným následkům, musíme být i při programování AI velmi důkladní a předvídat, jak by mohla naše instrukce in-

terpretovat.

Vývoj AI je neustálý proces učení a zdokonalování, a to jak pro modely, tak pro jejich tvůrce. Abychom mohli plně využít potenciál umělé inteligence, musíme zajistit, že její cíle jsou v souladu s lidskými hodnotami a etickými principy. Jen tak můžeme předejít situacím, kdy AI splní svůj úkol, ale způsobí více škody než užitku.

4 Zdroje

Victoria Krakovna, Victoria Krakovna, 2, S. R. A., 3, V. K. P. authorApril, 19, A. F. A., 16, G. J. N., 28, D. W. N., 6, tdietterich A., 18, maximumevilcoverage M., & 25, S. M. S. (2023, October 8). Specification gaming examples in ai. Specification gaming examples in AI. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

Paulfchristiano. (n.d.). What failure looks like - ai alignment forum. What failure looks like - AI Alignment Forum. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Ngo, R., Chan, L., & Mindermann, S. (2024, March 19). The alignment problem from a deep learning perspective. arXiv.org. https://arxiv.org/abs/2209.00626

Krakovna, V. (2020, April 21). Specification gaming: The flip side of Ai Ingenuity. Google DeepMind. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/