Potenciál

Interpretace tématu „Umělá inteligence“

Daniel Soják

1. Úvod

Umělá inteligence je velké téma technologického pokroku za poslední roky, ambice na její následující vývoj pouze rostou. Všichni se s touto technologií v každodenním životě potkáváme. Je to náš asistent, učitel, prohlížeč, řidič i spousta dalších věcí. Chyby a nedostatky, které doprovázeli starší modely, postupně nahrazují vylepšené a funkčnější verze a mnoho z nás si díky tomu může zjednodušit život.

Věřím tedy, že i v oblasti kybernetické bezpečnosti a její perspektivní budoucnosti má AI potenciál uplatnit se jak v defenzivních, tak i ofenzivních aplikacích.

2. Hostování AI na zařízení oběti

Hostovat AI model je dnes, mimo serverovny plné výpočetních zařízení, možné i na výkonnějších herních počítačích. Pokud náročnost modelů klesne díky optimalizaci nebo jinému faktoru a výkony osobních zařízení se budou neustále zvyšovat tak, jak tomu bylo doposud, je reálná možnost při kyberútoku využít i umělé inteligence.

Tato skutečnost otevírá dveře řadě novým příležitostem uplatnění AI jako řídící prvek malwaru na zařízení oběti. Takový algoritmus je náročný na výkon, a proto si jako inspiraci můžeme vzít typ malwaru známý pod jménem cryptominer, který odpovídá fundamentálnímu cíli skrytého hostování AI – nenápadně provádět náročné výpočetní operace na zařízení netušící oběti.

AI by měla oproti obvyklým programům možnost při napadení analyzovat povahu zařízení i oběti, což by umožňovalo odhadnout nejvhodnější strategii a postup útoku pro dosažení optimální účinnosti s minimálním rizikem odhalení.

V případě již napadeného zařízení může lokálně hostovaná AI podle vlastního uvážení napadat další zařízení a replikovat sebe nebo méně sofistikovaný malware například za účelem vybudování botnetu.

Zajímavý aspekt umělé inteligence, jenž stojí za zmínění, je její záhadné interní fungování, které je pro nás příliš objemné a komplikované na pochopení[1]. Při forenzní analýze napadeného počítače by tedy mohlo být náročnější zjistit fungování a cíl

[1] https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast

malwaru. Dynamická analýza v simulovaném prostředí sandboxu se ale nabízí jako jedno z možných řešení této problematiky.

3. Nová generace sociálního inženýrství

Jaké riziko přináší generativní modely nám bylo demonstrováno už v minulých letech. Takzvaný deepfake je velice názorným příkladem napodobení jiné osoby použitím adaptačních a generativních schopností AI. K jejich kalibraci může stačit jen pár minut videa osoby s různými výrazy a úhly pohledu[2]. Odpověď na tuto technologii byl další model umělé inteligence tentokrát specializované na detekci neautentických médií.

Toto byl doposud problém vztahující se jen na populární veřejně vystupující osoby, avšak pokud útočník získá přístup k zařízení s kamerou a mikrofonem na dostatečně dlouhou dobu, vytvořit přesvědčivý deepfake ze získaných záběrů není příliš náročné.

Navíc díky ukradeným záznamům zpráv a jiných lokálních dat může útočník naučit AI napodobovat chování osoby až do detailů jako opakované gramatické chyby a používaný slovníček. Důvěryhodnost lze posílit ještě lepší znalostí důvěrných témat a vzpomínek zmíněných ve zprávách, které se AI může sama naučit.

Oběti v takové situaci je potom takřka odcizena digitální identita a útočník se může pokusit manipulovat s rodinou, přáteli i kolegy. I když slabým článkem tedy nemusíme být my, toto riziko nás přesto ohrožuje. Útok skrze zprávy, hovory a videohovory, kde se umělá inteligence vydává za známého, nám hrozí všem.

4. Rapidní vývoj pentestingu

V dnešní době hledání zranitelností funguje na bázi vyhýbání se chybám a problémům které již známe. Existují programy, které procházejí malware databáze a na základě nich napadají testované systémy. Nové exploity jsou nacházeny každý den a společnosti za jejich nalezení a nahlášení vyplácí velice objemné finanční odměny, jelikož nenahlášené exploity, které jsou poté zneužity, mohou společnosti stát značně víc než jen peníze.

Studie, která testovala schopnost AI vymýšlet neobvyklá a kreativní využití každodenních předmětů v porovnání s lidskými účastníky[3] naznačuje, že strojové učení by mohlo mít schopnost vymýšlet nové a inovativní exploity i obrany proti nim, a to značně rychleji, než jsou toho přirozeně schopni lidé.

Využít by se mohlo i takzvané generativní adversariální sítě, což je typ adversariálního trénování[4] (anglicky GAN adversarial training), kdy dva modely trénují na vzájemných

[2] https://aislackers.com/deepfake-how-easy-is-it-to-make-and-detect-it/
[3] https://pubmed.ncbi.nlm.nih.gov/37709769/
[4] https://aws.amazon.com/what-is/gan/

výstupech a snaží se porazit jejich oponenta v nastavených trénovacích podmínkách. Využitím v kontextu kybernetické bezpečnosti to znamená, že jeden AI agent usiluje o nabourání se do systému spravovaného druhým agentem. Tímto trénováním se oba agenti budou neustále zdokonalovat a vyvíjet své schopnosti. Takto trénovaná AI může asistovat při penetračních testech a konfiguraci obranných prostředků. Má také potenciál odhalit nové techniky a jejich následný vývoj pro útok i obranu nejrůznějších systémů.

5. Závěr

Kybernetická bezpečnost i umělá inteligence jsou velice široké a proměnlivé pojmy, o jejichž budoucnosti můžeme při nejlepším polemizovat. Příležitosti pro jejich vzájemné ovlivňování s největší pravděpodobností porostou, a tak se třeba uplatní i jeden z nápadů vymezených v tomto článku.

Zpracovávání této tématiky bylo velice zajímavé, a proto bych chtěl poděkovat pořadatelům soutěže za tuto kreativní příležitost a Vám za přečtení mých vizionářských nápadů pro budoucnost.