ATTACCO VS. DIFESA: COME I MODELLI DI INTELLIGENZA ARTIFICIALE VENGONO UTILIZZATI DA ENTRAMBE LE PARTI

L'intelligenza artificiale (IA) sta diventando sempre più importante nel campo della cybersicurezza, sia per i difensori che per gli aggressori. Entrambi gli schieramenti stanno investendo in ricerca e sviluppo per migliorare le proprie capacità di IA. Questa corsa agli armamenti dell'IA è destinata a continuare, rendendo sempre più difficile per le organizzazioni rimanere al sicuro

A cura di Alex Galimi (*)

Tutti parlano di Intelligenza Artificiale: la tecnologia che sta rivoluzionando diversi settori come la sanità, i servizi finanziari e l'industria manifatturiera e che sarà uno dei temi caldi degli anni futuri. In ambito cybersicurezza, i vendor del settore sono stati probabilmente tra i primi ad adottare l'intelligenza artificiale per affrontare la lotta alla criminalità informatica. Purtroppo, non sono solo i "buoni" a investire in tecnologie emergenti come questa. Per questo è importante osservare l'evoluzione di questa tecnologia, concentrandosi in particolare su due tipi di modelli:

l'IA categorizzante e quella generativa.
Conosciuta anche come modello classificatore, segmentatore o descrittivo, l'intelligenza artificiale categorizzante, come dice il suo nome, prende un input e lo categorizza. Un'applicazione ovvia per i difensori di rete è quella di classificare input come connessioni di rete, comportamenti o file in "buoni" e "cattivi". Si tratta di una continuazione logica del classico elenco di blocchi e permessi, con l'unica differenza che la decisione buono/cattivo viene presa automaticamente anziché creata manualmente.
Tuttavia, occorre prestare attenzione a un paio di aspetti: in primo luogo, gli sviluppatori di queste soluzioni dovrebbero evitare di stabilire in modo esclusivamente binario se un input è buono o cattivo. C'è una grande differenza tra un file giudicato dannoso all'1% e uno giudicato dannoso al 49%; per questo motivo potrebbe essere utile fornire categorie definibili dall'utente, come ad esempio "potenzialmente dannoso" o "indesiderato", per aggiungere maggiore variabilità in termini di affidabilità. In secondo luogo, gli utenti hanno bisogno di conoscere non solo la risposta, ma anche il modo in cui un modello di IA la elabora. Purtroppo, la maggior parte delle IA attuali non è autoriflessiva, cioè non è consapevole di sé stessa e del proprio processo decisionale. L'IA di oggi è quindi come un esperto ben addestrato, valuta i fatti che gli vengono presentati sulla base del proprio istinto in modo da giungere a una conclusione. Per ridurre questi rischi, potrebbe essere una buona idea memorizzare le decisioni dell'IA per una futura analisi forense ed eventualmente anche i dati grezzi su cui si sono basate tali decisioni. Ciò consentirebbe a un essere umano esperto di ricostruire o riconfermare la decisione dell'IA.

Indirizzare il successo dell'intelligenza artificiale generativa

Il motivo per cui tutti i consigli di amministrazione parlano di IA è dato dal fatto che modelli generativi, come ChatGPT, debuttano sulla scena con una straordinaria capacità di interagire con gli utenti attraverso il linguaggio naturale. Il modello calcola o "genera" un output da un enorme pool di dati di addestramento, combinati con il contesto attuale (domande o cronologia delle chat). Tuttavia, è importante ricordare che, nonostante la fluidità nella lingua dell'utente e l'apparente capacità di comporre battute, poesie e altre opere d'arte, questi modelli non "capiscono" realmente i contenuti che apprendono. Di conseguenza, tutto ciò che viene prodotto è fondamentalmente solo un ottimo "remix" dei contenuti su cui il modello è stato addestrato, anche se si tratta di un patrimonio di conoscenze a cui nessun essere umano potrebbe attingere nel corso della propria vita.
A differenza dell'IA categorizzante, la forza dei modelli generativi non sta nel prendere decisioni, ma nel riassumere e presentare informazioni e fatti in un dialogo. Se addestrati sui dati giusti - ad esempio input che coprono le connessioni di rete, le interazioni, la conformità e i requisiti aziendali - potrebbero diventare assistenti informatici estremamente efficaci. Un modello di intelligenza artificiale generativa potrebbe essere in grado di consigliare impostazioni di sistema ottimizzate, o suggerire priorità per una strategia di conformità. Con le giuste informazioni tempestive, potrebbe persino essere in grado di fornire un'analisi delle cause degli attacchi.
È importante ricordare che la qualità linguistica dei risultati non è un indicatore della qualità del contenuto effettivo. I falsi positivi sono un problema comune per alcuni modelli generativi, soprattutto quando questi vengono addestrati solo su pochi dati. Alcuni attacchi zero-day potrebbero passare inosservati proprio per questo motivo.
Il nuovo assistente di Trend Micro basato sull'intelligenza artificiale generativa, Companion, è diverso: questo è addestrato su dati strettamente controllati che fanno parte della proprietà di Trend Micro, ed è progettato per essere utilizzato dagli analisti SecOps (Security Operations indica la fusione collaborativa tra sicurezza IT e operazioni IT, ndr). Questi ultimi sono spesso sovraccaricati da avvisi di minacce e hanno difficoltà a gestire il carico di lavoro a causa della carenza di competenze. La soluzione aiuterà gli utenti, di tutti i livelli di competenza, a essere più produttivi grazie a:

- Spiegazione e contestualizzazione degli avvisi
- Gestione e raccomandazione di azioni
- Decodifica di script complessi
- Sviluppo e test di query di ricerca complessi

Utilizzo dell'IA categorizzante per scopi malevoli

Purtroppo, ciò che funziona per i team di sicurezza IT può essere utilizzato anche da aggressori con obiettivi dannosi, ed è risaputo che i cybercriminali sono sempre pieni di risorse. Infatti, un'importante fuga di dati che l'anno scorso ha colpito il famigerato gruppo ransomware Conti, ha rivelato che questo spendeva 6 milioni di dollari annui in strumenti, servizi e stipendi, destinandone la maggior parte in ricerca e sviluppo.
In questo gioco al gatto e al topo, solo lo studio

dell'uso malevolo dell'IA può permettere alla cybersecurity di progettare prodotti e misure di mitigazione migliori.
Per i team di sicurezza informatica, i modelli di intelligenza artificiale categorizzante costituiscono la punta di diamante della difesa informatica. I cybercriminali, invece, li utilizzano principalmente per selezionare strategicamente le vittime e pianificare nuovi attacchi. Questo primo utilizzo non si allontana troppo dai processi di marketing aziendale pensati per trovare un target ottimale: si inizia definendo l'obiettivo e si cerca successivamente di abbinarlo a un modello preciso.
Nel caso in cui non volessero automatizzare gli attacchi, con lo scopo di rivolgerli a un ampio gruppo di utenti come un "direct mailing" nel marketing, i criminali informatici potrebbero anche voler definire le "vittime ottimali". Si tratta di individui maggiormente disposti a pagare, meno propensi a sporgere denuncia e che richiedono uno sforzo minore in termini di attacco. La fase successiva consiste nel cercare di trovare questi obiettivi, utilizzando le fonti di dati disponibili come: le ricerche su Shodan, l'OSINT (Open Source Intelligence, disciplina dell'intelligence che si occupa della ricerca, raccolta e analisi di informazioni da fonti aperte, ndr) derivata dai social network, i dati di attacchi precedenti e le informazioni trapelate o rubate. In questo caso, grandi quantità di dati e insiemi di dati complessi possono essere elaborati dall'intelligenza artificiale per trovare potenziali vittime.
Teoricamente, l'intelligenza artificiale potrebbe essere utilizzata anche durante un attacco per determinare quale attività post compromissione - come la crittografia, il ricatto o il furto di dati - prometta i maggiori profitti. Un modello di IA di categorizzazione, eseguito localmente, potrebbe aiutare in questo caso.
I limiti dell'IA generativa
Come spiegato in precedenza, l'IA generativa non è creativa, in quanto tutto ciò che produce è un remix, a volte complesso, di contenuti su cui è stata addestrata. Ciò significa che può mostrare agli utenti exploit già pronti, con tanto di spiegazioni e commenti, ma non sarà in grado di generare exploit zero-day dal nulla. Può però aiutare in altri modi, ad esempio un modello di IA generativa è stato usato per scrivere il codice dei plugin distribuiti durante una competizione Pwn2Own. ChatGPT è stato adoperato invece per la creazione di varianti polimorfiche di malware, sebbene, anche in questo caso si sia trattato solo di riscritture di contenuti noti appresi durante l'addestramento.
Nonostante il limitato uso che potrebbe farne un criminale informatico esperto, questo strumento potrebbe contribuire a democratizzare l'accesso a determinate conoscenze tra i criminali meno esperti, a patto che sappiano cosa chiedere.
È nella creazione di contenuti che i modelli di IA generativa si rivelano davvero efficaci. Questi strumenti sono in grado di produrre contenuti di phishing accattivanti e altamente leggibili e senza errori grammaticali. Il possibile utilizzo dell'intelligenza artificiale generativa per produrre contenuti altamente convincenti per attacchi BEC (Business Email Compromise) e altri attacchi di impersonificazione, è una preoccupazione già sollevata da Europol.
È anche il motivo per cui aziende come Trend Micro continuano a svolgere ricerche intensive nel settore dell'IA, analizzando le azioni degli avversari e sviluppando misure di mitigazione per rilevare e bloccare i loro sforzi in modo più efficace. La corsa agli armamenti dell'IA è appena iniziata.

Keywords: AI, Cybersecurity, ChatGPT, Trend Micro, Alex Galimi, Pwn2Own, OINST, SecOps

www.trendmicro.com

(*) Alex Galimi
Technical Partner Manager Trend Micro Italia

ATTACK VS. DEFENSE: HOW ARTIFICIAL INTELLIGENCE MODELS ARE BEING USED BY BOTH SIDES

Artificial intelligence (AI) is becoming increasingly important in cybersecurity for both defenders and attackers. Both sides are investing in research and development to improve their AI capabilities. This AI arms race is set to continue, making it increasingly difficult for organizations to remain secure.

By Alex Galimi (*)

Everyone is talking about Artificial Intelligence-the technology that is revolutionizing several sectors such as healthcare, financial services and manufacturing and will be one of the hot topics in the years to come. In the cybersecurity arena, vendors in the industry have arguably been among the early adopters of artificial intelligence to tackle cybercrime. Unfortunately, it is not only the "good guys" who are investing in emerging technologies like this. That is why it is important to watch the evolution of this technology, focusing in particular on two types of models: categorizing AI and generative AI.

Also known as a classifier, segmenter or descriptive model, categorizing artificial intelligence, as its name implies, takes an input and categorizes it. An obvious application for network defenders is to categorize input such as network connections, behaviors, or files into "good" and "bad." This is a logical continuation of the classic list of blocks and permissions, with the only difference being that the good/bad decision is made automatically rather than created manually.

However, attention should be paid to a couple of issues: first, developers of these solutions should avoid making a purely binary determination of whether an input is good or bad. There is a big difference between a file judged to be 1% malicious and one judged to be 49% malicious; therefore, it might be useful to provide user-definable categories, such as "potentially malicious" or "unwanted," to add more variability in terms of reliability.

Second, users need to know not only the answer, but also how an AI model processes it. Unfortunately, most of today's AIs are not self-reflective, that is, they are not aware of themselves and their decision-making process. Thus, today's AI is like a well-trained expert, evaluating the facts presented to it based on its own instincts in order to reach a conclusion. To reduce these risks, it might be a good idea to store the AI's decisions for future forensic analysis and possibly even the raw data on which those decisions were based. This would allow an experienced human to reconstruct or reconfirm the AI's decision.

Addressing the success of generative artificial intelligence

The reason all boards are talking about AI is because generative models, such as ChatGPT, are debuting on the scene with an extraordinary ability to interact with users through natural language. The model calculates or "generates" an output from a huge pool of training data, combined with the current context (questions or chat history). However, it is important to remember that despite the fluency in the user's language and the apparent ability to compose jokes, poems and other works of art, these models do not really "understand" the content they learn. As a result, all that is produced is basically just a very good "remix" of the content that the model has been trained on, even though it is a wealth of knowledge that no human being could draw on in their lifetime.

Unlike categorizing AI, the power of generative models is not in making decisions, but in summarizing and presenting information and facts in a dialogue. If trained on the right data-such as inputs covering network connections, interactions, compliance, and business requirements-they could become extremely effective computing assistants. A generative AI model might be able to recommend optimized system settings, or suggest priorities for a compliance strategy. With the right timely information, it might even be able to provide root cause analysis of attacks.

It is important to remember that the linguistic quality of the results is not an indicator of the quality of the actual content. False positives are a common problem for some generative models, especially when they are trained on only a few pieces of data. Some zero-day attacks may go undetected for this very reason.

Trend Micro's new generative AI-based assistant, Companion, is different: this one is trained on tightly controlled data that is part of

Trend Micro's property, and is designed to be used by SecOps analysts (Security Operations indicates the collaborative fusion of IT security and IT operations, ed.). The latter are often overloaded with threat alerts and find it difficult to manage the workload due to skills shortages. The solution will help users, of all skill levels, to be more productive through:
- Explanation and contextualization of alerts
- Managing and recommending actions
- Decoding complex scripts
- Developing and testing complex search queries

Using categorizing AI for malicious purposes

Unfortunately, what works for IT security teams can also be used by attackers with malicious targets, and it is well known that cyber criminals are always resourceful. In fact, a major data leak that hit the infamous ransomware group Conti last year revealed that it was spending $6 million annually on tools, services and salaries, spending most of it on research and development.

In this cat-and-mouse game, only the study of the malicious use of AI can enable cybersecurity to design better products and mitigation measures.

For cybersecurity teams, categorizing AI models is the cutting edge of cyber defense. Cyber criminals, on the other hand, use them primarily to strategically select victims and plan new attacks. This first use does not stray too far from corporate marketing processes designed to find an optimal target: they start by defining the target and then try to match it to a precise model.

In case they do not want to automate attacks, with the aim of targeting a large group of users as a "direct mailing" in marketing, cybercriminals may also want to define "optimal victims." These are individuals who are more willing to pay, less likely to press charges, and require less effort in terms of the attack. The next step is to try to find these targets, using available data sources such as: research on Shodan, OSINT (Open Source Intelligence, an intelligence discipline that deals with research, collection and analysis of information from open sources, ed.) derived from social networks, data from previous attacks, and leaked or stolen information. In this case, large amounts of data and complex data sets can be processed by artificial intelligence to find potential victims.

Theoretically, AI could also be used during an attack to determine which post-compromise activity-such as encryption, blackmail, or data theft-promises the greatest profits. A categorization AI model, run locally, could help here.

The limitations of generative AI

As explained earlier, generative AI is not creative, as all it produces is a remix, sometimes complex, of content it has been trained on. This means that it can show users ready-made exploits, complete with explanations and comments, but it will not be able to generate zero-day exploits out of thin air. It can help in other ways, however; for example, a generative AI model was used to write code for plugins distributed during a Pwn2Own competition. ChatGPT, on the other hand, was used to create polymorphic variants of malware, although, again, these were only rewrites of known content learned during training. Despite its limited use by experienced cybercriminals, this tool could help democratize access to certain knowledge among less experienced criminals, provided they know what to ask for.

It is in content creation that generative AI models really prove effective. These tools are capable of producing engaging and highly readable phishing content without grammatical errors. The possible use of generative AI to produce highly compelling content for BEC (Business Email Compromise) and other impersonation attacks is a concern already raised by Europol.

It is also why companies such as Trend Micro continue to do intensive research in the field of AI, analyzing the actions of adversaries and developing mitigation measures to detect and block their efforts more effectively. The AI arms race has just begun.

(*) Alex Galimi, Technical Partner Manager Trend Micro Italy

Keywords: AI, Cybersecurity, ChatGPT, Trend Micro, Alex Galimi, Pwn2Own, OINST, SecOps