Protecting AI: We Built the Brains, but What About Helmets?

Mark Campbell, EVOTEK

Artificial intelligence models introduce attack surfaces that current defenses simply do not protect. But much like helmets, new tools, and techniques are emerging to safeguard the smarts.

Most companies are evolving from consumers of products with embedded artificial intelligence (AI) into producing their own custom models to service their customers' needs in a smarter, automated, and adaptable fashion. This shift of new technologies, skills, and processes causes many companies to throw ever-growing resources into the smart application development race. However, bad actors are finding novel ways to exploit and subvert this fresh new architecture before the "wet paint" signs are even removed.

AI models are susceptible to nefarious bias injected during model training that can produce models with nascent blind spots or hypersensitivities. Once deployed, these models may trigger unintended actions. Alternatively, models that "learn in the wild" can be manipulated through input flooding to "teach" them that ice hockey is Ecuador's most popular sport or left-handed cod fishermen are terrible credit risks.

AI FOR SECURITY

AI is no newcomer to the cybersecurity arena. For years, AI has ushered in shockingly effective techniques for user and entity behavior analytics, threat hunting, and intrusion detection. Since the advent of continuous development/continuous deployment (CI/CD), security experts have deployed many tools with built-in AI to protect application development against the insertion of programmatic backdoors, insecure coding habits, or compromised open source libraires.

Conversely, AI has become a handy weapon for adversaries to thwart perimeter defense, generate unique signature malware, crack passwords, customize social engineering attacks, and make lateral movement almost undetectable. This measure–countermeasure arms race will continue for

Digital Object Identifier 10.1109/MC.2021.3055927
Date of current version: 17 November 2021

0018-9162/21©2021IEEE

EDITOR **MARK CAMPBELL**
EVOTEK; mark@evotek.com

the foreseeable future. But amid this cybersecurity fervor, a new battleground emerged mostly unnoticed (unnoticed by the good guys at least).

SECURITY FOR AI

In today's typical enterprise technical landscape, there is no group, tool, or procedure directly responsible for securing machine learning processes, data, or models. Even if a company practices otherwise perfect security hygiene, AI introduces attack vectors that current defenses simply do not address. As such, the development of smart applications presents an attacker with an irresistibly soft underbelly to attack.

AI models must be trained, and training requires data—a lot of it. Today's data scientist has access to cheap, huge, and highly specialized training data repositories. An immense sample data set is extracted from the selected data corpus to train a base model to perform complex tasks like credit scoring, facial recognition, or natural language generation. These trained models are verified and deployed into production environments. Some production models, such as targeted advertisement, anomaly detection, or even your autonomous vacuum cleaner, continue to learn after deployment. Every step in this process offers an attacker avenues of exploitation (Figure 1).

Data poisoning

Should an attacker gain access to the source data corpus, it can nefariously poison the training data. Imagine an attacker labeling images of pigeons as passenger airliners and the havoc this would wreak on an AI-assisted air traffic control system. Note also that the attacker in this case need not breach the perimeter of the AI development company. The poisoned data are brought right in through the front door and injected into base model's training process. Commonly, the model verification stage uses a data corpus subset to verify the model correctly handles previously unseen data. Of course, data poisoning also infects the verification data, so pigeons still look like jumbo jets to the verified model.

Data biasing

Another clandestine data-tampering technique does not require creating erroneous information but instead an attacker manipulates the extracted data set selection criteria. In our smart air traffic control example, the extracted training data could be skewed to omit all Airbus aircraft. The model will still train and verify without issue but will be stumped in the real world when approached by the first Air France flight.

Data theft

Training data can be an incredibly valuable treasure for adversaries to target. The process of compiling, collating, and cleaning proprietary business data so it can be used to train an AI model results in a very concentrated pool of intellectual property. The sampling process further refines this down to the "really important stuff," making it invaluable to a bad actor who can now steal a bar of silver rather than a truckload of ore. Most AI development and training environments have far less security than production systems and data repositories, so there is a much softer perimeter to penetrate and an easier path of exfiltration.

Model theft

If a preselected training data set is silver, a trained model is gold. Should an interloper gain access to the model repository and evade detection, they can extract the trained model and all subsequent improvements on it. The filched model can be easily replicated and deployed by a competitor or examined for vulnerabilities and later exploited.

Model hijacking

A bad actor with access to the machine learning operations (MLOps) deployment pipeline can inject its own model. The surrogate model, nominally a manipulated version of the original, behaves almost identically except for key behaviors triggered by special circumstances known to the attacker. This wolf in sheep's clothing could, for instance, be a smart video surveillance system that doesn't report incursions on Thursdays between 8:00 and 8:30 or a smart retail

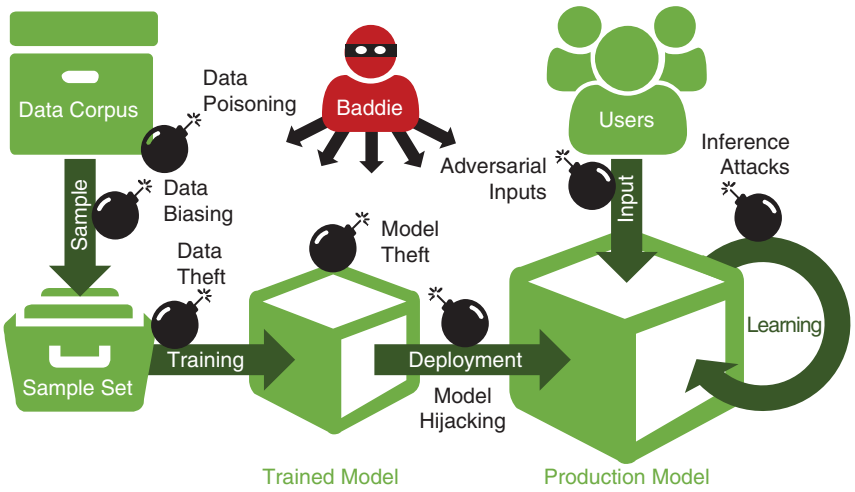

FIGURE 1. An AI model development life and threats.

system that gives a 99% customer loyalty discount on any purchases with one rechargeable AAA battery and a snorkel.

Adversarial inputs

An attacker may deploy an evasion attack where carefully crafted input tricks the AI model into misclassifying malicious data as benign. This is quite common in malware and spam attacks where the adversary crafts payloads to blind the classifier to their malicious intent, thereby evading security measures. Other recent examples include adversarial glasses or stealth T-shirts that fool or even blind facial and human recognition systems.[1]

For smart models that learn and adapt after deployment, attackers can feed them synthetic inputs to manipulate learning. Attackers with a stolen model can analyze its architecture, internal structure, and parameters, then perform "white-box" attacks by feeding the deployed model input that teaches it bad habits.[2] However, if the adversary does not have a copy of the model, a black-box technique (known as an inference attack) can yield a rudimentary understanding of how the model adapts to various inputs and then flood it with data to induce model drift and exploitable behavior.[3]

In a disturbing discovery, researchers have shown that the coupling of adversarial inputs with data poisoning attacks has a "mutually reinforcement" effect, amplifying the deception probability greater than the sum of the two attacks separately.[4]

AI HARDENING

Application security is one of the fastest growing cybersecurity areas.[5] However, the most common forms of application security—namely *DevSecOps*, static and dynamic application security testing, interactive application security testing, run-time application self-protection, and software composition analysis (SCA)—are ineffective against AI model attacks since they are designed to protect code, configuration, and application rather than training data or model behavior.

So, how can companies protect their AI models? There are several techniques to thwart model attackers now emerging as features in application security products.

Data set protection

Most companies today use SCA products and techniques to vet the provenance of software modules used or added to their software applications. Analogous techniques can be employed to verify the lineage of training data sets and alert security systems to external data sets that have been modified, potentially poisoned, or have an abnormal data distribution and possible biasing. Training data sets can also be stored in databases and repositories as immutable data structures to prevent insertion, modification, or deletion (any attempt to violate this triggers a security alert). Additionally, traditional data loss prevention systems can be employed to detect training data sets or models exfiltration.

Model protection

AI models can be protected by adopting proven software paradigms such as Docker Enterprise from the application container space. This type of framework, adapted to handle containerized AI models, provides signing, scanning, registration, tracking, and logging of AI models while providing role-based access controls, version control, and authentication and authorization across organizations or companies.[6] This type of rigid control framework is required to prevent model hijacking.

Model observability

Commercial offerings are also emerging to monitor model behavior. Observability tools like Fiddler AI watch models operate in training and production and evaluate results over time to detect model drift, skew, or error—indicators of adversarial input, data biasing, or model hijacking, respectively.

Observability tools report on anomalous behavior and can identify the root cause (for example, model defect, training data bias, or alteration) and even generate adversarial examples. "Fiddler helps determine the cause of drift—is it a real behavior change, a system error, or nefarious actions?"[7]

Adversarial defense

Several defense techniques are designed to detect and thwart adversarial input attacks, including the following:

- *Adversarial training*: After a model is trained for its intended purpose, a second training phase is conducted using adversarial inputs to ensure the model handles these correctly. Of course, it is a nontrivial challenge to develop an expansive adversarial training data set, but even minimal adversarial training is better than one. (Side note: adversarial training is also being effectively applied to nonsecurity cases like mud on an autonomous car's camera, background noise in natural language processing, or interference in signal processing.[8])
- *Input modification*: This technique strips adversarial noise from incoming production data. While easier said than done, there are several analysis techniques that can detect overly noisy input and, if this noise cannot be removed, alert security systems or operators to the potential threat.
- *Adversarial detection*: After the primary training phase, input modification techniques can be employed on training data. If certain training data are determined to be noisy, they are cleaned (denoised) and run back through the model. Drastic differences between noisy and clean data results indicate possible data poisoning and adversarial action.
- *Null class*: One of the strengths of AI is its ability to handle previously unseen data or input on the fuzzy edge between classifications. However, attackers

can flood a model with noisy, new, or gray-area data until they find cases that induce adverse behavior. AI models can introduce a "null" or "I don't know" classification for data that is too noisy, indistinct, or vaguely classifiable. This avoids attacks that exploit border data uncertainties, thus limiting an AI model's ability to make a "good enough" guess in unclear situations.

CHALLENGES TO PROTECTING AI

Awareness is the largest challenge to protecting AI models today. Many companies embarking on their AI model development odyssey are not cognizant of the threats awaiting them along the journey. Of those who grasp the looming threats, most ignore the danger for now and bolt on protection solutions at some point in the future when better tools and techniques emerge. Many companies weigh the benefit of deploying AI against the risk and impact of an attack, while others simply brush off the threat.

One of today's unresolved technical concerns, however, is while making AI models less susceptible to adversarial inputs or data poisoning, the training data set's structure and training data distribution is revealed, ironically, opening the door for other forms of attack.[9] A current trend in AI models is increasing transparency and explainability in decision making. While these are admirable ethical, social, and personal features, they open a rich vein of vulnerabilities for adversaries to mine.[10]

"There's no bible for protecting AI models … yet," comments Shahrzad Ahmadi, CEO of Klever Ai, "but regulatory agencies are identifying procedures and standards to protect AI in every stage of the process from data input through algorithm development, training, testing, and production results."[11] While this is a daunting task, governments—most notably the EU's Assessment List for Trustworthy Artificial Intelligence[12] and the U.S. National Security Commission on Artificial Intelligence[13]—are defining frameworks for creating more hardened AI models.

Attackers today have a distinct advantage over companies deploying AI models. Whether through obliviousness, a lack of protective tools, or a deliberate risk–reward tradeoff, most companies are not prepared to fend off a concerted AI model attack. Undoubtedly, we will see headlines of major exploits in the coming years (or months), each with serious impact on its victims. But help is on the way. New tools and techniques are emerging every week to shore up the dam holding back the tide of bad actors. We built the brains. Now we need to build the helmets.

REFERENCES

1. ilmoi, "Evasion attacks on machine learning (or 'adversarial examples')," Towards Data Science, July 14, 2019. [Online]. Available: https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1
2. G. Boesch, "Adversarial machine learning," VISO.ai Deep Learning, June 26, 2021. [Online]. Available: https://viso.ai/deep-learning/adversarial-machine-learning/
3. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security Privacy*, San Jose, CA, 2017, pp. 3–18. doi: 10.1109/SP.2017.41.
4. R. Pang et al., "A tale of evil twins: Adversarial inputs versus poisoned models," in *Proc. 2020 ACM SIGSAC Conf. Comput. Commun. Security (CCS '20)*, pp. 85–99. doi: 10.1145/3372297.3417253.
5. "Global application security market to gain USD 9779.8 million and enhance at a CAGR of 16.1% during 2020–2027 timeframe Exclusive COVID-19 impact analysis (264 pages) report," Research Dive, New York, 2021. [Online]. Available: https://www.globenewswire.com/news-release/2021/05/03/2221475/0/en/Global-Application-Security-Market-to-Gain-USD-9779-8-Million-and-Enhance-at-a-CAGR-of-16-1-during-2020-2027-Timeframe-Exclusive-COVID-19-Impact-Analysis-264-pages-Report-by-Resear.html
6. "Integrated container security at every step of the application lifecycle," Docker. [Online]. Available: https://www.docker.com/products/security (accessed Sept. 9, 2021).
7. M. Campbell, Interview with K. Gade, Aug. 24, 2021.
8. J. Chu, "Algorithm helps artificial intelligence systems dodge 'adversarial' inputs," MIT News, Mar. 8, 2021. [Online]. Available: https://news.mit.edu/2021/artificial-intelligence-adversarial-0308
9. L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," in *Proc. 2019 ACM Conf. Comput. Commun. Security*, London, pp. 241–257. doi: 10.1145/3319535.3354211.
10. A. Burt, "The AI transparency paradox," Harvard Business Review, Dec. 13, 2019. [Online]. Available: https://hbr.org/2019/12/the-ai-transparency-paradox
11. M. Campbell, Interview with S. Ahmadi, Sept. 13, 2021.
12. "Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment," European Commission, Brussels, Belgium, July 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
13. J. Wolff, "How to improve cybersecurity for artificial intelligence," Brookings, Washington, D.C., 2020. [Online]. Available: https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/

MARK CAMPBELL is the chief innovation office for EVOTEK, San Diego, California, 92121, USA. Contact him at mark@evotek.com.