

Threat Modeling the Artificial Intelligence Pipeline

IriusRisk



Threat modeling the artificial intelligence pipeline is like safeguarding a fortress; you need to know where an attack might come from.

In this guide, you'll get a clear view on how to lock down each stage— from raw data collection right through when your model goes live.

With real-world strategies for each phase of development and deployment, you're about to level up your AI security game.



Brandon Green
Solutions Architect @ IriusRisk

Brandon Green is a Jedi Knight and Sr. Solutions Architect at IriusRisk specializing in designing and implementing threat modeling solutions for enterprise organizations. He has 20 years of IT experience and served 10 years in the United States Air Force, where he worked as a Special Agent focusing on cybercrimes. His interests include writing fiction, sports betting, and video games.

Table Of Contents:

| | |
|--|----|
| 01 - Raw Data Collection | 04 |
| 02 - Data Pre-processing and Preparation | 05 |
| 03 - Learning Algorithm | 07 |
| 04 - Performance Evaluation without Bias | 09 |
| 05 - Deployment Setup | 11 |
| 06 - AI Model is Deployed | 13 |
| 07 - User Inputs | 14 |
| 08 - Model Outputs | 15 |
| 09 - Fully Trained Model | 16 |
| 10 - Conclusion | 17 |

Raw data collection

The process of raw data collection is the phase where we gather all the necessary information before it's processed by algorithms.



What can go wrong?

- One major threat is unauthorized access such as someone breaking into an office after hours or being able to see a confidential document.
- Then there's data tampering, which is as if someone had swapped out your valuable assets with worthless ones without you realizing until it's too late.
- And of course, privacy concerns are always present, much like the risk of confidential conversations being overheard.

Securing this phase is like safeguarding a corporate headquarters. You need robust security measures (encryption), vigilant security personnel (access controls), and competent staff who understand their roles (trained personnel). Having been through this process before, we can't stress enough the importance of encryption standards when storing or transmitting your data—it could mean the difference between keeping your assets secure or losing them to malicious entities.

What are you going to do about it?

- Implement strict access controls to prevent unauthorized entry.
- Educate team members on best practices because everyone needs to follow rules.

Preventive measures are always better than remedial ones. So, establish clear guidelines (data governance policies) and ensure everyone knows their responsibilities. It will save you a lot of trouble in the future.

Data Pre-processing and Preparation

Think of biased or poorly preprocessed data as malware infiltrating your AI's operating system. It might not trigger immediate alarms, but it subtly corrupts the core processes, leading to flawed decisions and potentially harmful outcomes. As an example, look at the formatting of this section, vs the previous one.



What can go wrong?

- **Errors and Anomalies:** Think of rogue malware infiltrating your systems. Data pre-processing scans for missing values, inconsistencies, and typos – the glitches that could compromise your AI's analysis.
- **Inconsistent Formatting:** Preprocessing standardizes data formats and structures, ensuring smooth communication between your algorithms and the real world. Are you using decimals or whole numbers? Megabytes or kilobytes
- **Biased Cleaning:** Think of biased cleaning like manipulating security logs. It skews the AI's understanding of reality, potentially perpetuating unfairness or discriminatory outcomes.
- **Incomplete Anonymization:** Leaving traces of personal information is like leaving a backdoor open. It exposes individuals to privacy breaches and erodes trust in your AI.

What are you going to do about it?

- **Data Validation:** Implement rigorous data validation techniques to scan every byte entering your training dataset.
- **Scaling and normalization:** Calibrate data points to a consistent range, preventing outliers from drowning out crucial insights.
- **Anonymization and Encryption:** Treat sensitive data like classified intel, encrypting and masking it to protect identities and prevent unauthorized access.

By carefully preparing your data, you're laying the foundation for accurate, reliable, and ethical AI systems. It's the often-invisible step that separates a glitchy, biased AI from one that delivers powerful insights and makes a real difference in the world. Remember, garbage in, garbage out – good data preparation gives your AI the best chance to shine!

The next time you hear about AI breakthroughs, remember the unseen heroes of data preparation – the meticulous cleaners, the creative engineers, and the vigilant cybersecurity experts. They're the ones quietly sculpting the raw data into the building blocks of a smarter future.

Learning Algorithm

Think of your AI learning algorithm as a crack intelligence analyst sifting through mountains of data. It churns out information, generating predictive insights and actionable decisions. But in the murky world of digital warfare, data can be weaponized, algorithms manipulated.



What can go wrong?

- **Overfitting:** The AI equivalent of tunnel vision. Imagine training an AI image recognition model based solely on the description of a single dog. The model would ace the test on that specific dog, but show it to other dogs or dog-like animals and the accuracy will drop. Overfitted models struggle with unique situations, leading to a dramatic plunge in accuracy and false positives (“That squirrel is definitely a dog!”). That’s overfitting in a nutshell: **Brilliant performance on limited data, but utter failure when faced with the true diversity of the world.**
- **Data poisoning:** This when someone (or something!) deliberately messes with the information your AI learns from. This can be done by slipping in false data, changing existing data, or even manipulating the way the AI interprets the data. It’s like tricking the police dog into thinking socks are the real drug!

What are you going to do about it?

- **Prioritize rigorous training:** The key lies in diversifying your training data. You should expose your model to all types of animals, not just the one dog. You feed your AI models with vast, varied datasets that reflect the real-world complexity they’ll encounter. Employ data augmentation techniques, like adding noise or rotations, to further stretch their adaptability.

Finally, field-test your AI before unleashing it into the real world. Reality checks are crucial – think live-fire exercises for your digital troops. Evaluate performance with rigor, assess potential biases, and refine your model. Remember, real-world data is the ultimate proving ground.

Performance Evaluation without Bias

Checking how well an AI runs is key—it's about making it both dependable and fair.

Here are some key threats to watch out for and ways to address them:

What can go wrong?

- **Overfitting:** The model performs incredibly well on the data it was trained on, but poorly on new data.
- **Underfitting:** The model fails to learn even basic patterns from the training data.
- **Bias:** The model exhibits unfair or discriminatory tendencies based on race, gender, or other sensitive factors.
- **Explainability:** Grasping the 'why' behind an algorithm's decision is tough, and that throws us off when we try to figure out if it's playing fair or acting reliably.
- **Data Drift:** The data used to train the model no longer reflects real-world conditions, leading to inaccurate predictions.

What are you going to do about it?

- **Cross-validation:** Test the model on data it wasn't trained on to identify overfitting and underfitting.
- **Data augmentation and diversification:** Expand the training data to represent real-world scenarios and diverse groups.
- **Fairness metrics and audits:** Use specialized metrics and human evaluations to detect and mitigate bias.



- **Explainable AI (XAI) techniques:** Make the model's decision-making process more transparent to build trust and identify potential unfairness.
- **Data monitoring and retraining:** Update the model with new data to keep it aligned with changing real-world conditions.

Building a future where AI benefits us all means we've got to keep tweaking and testing it to make sure it's fair play for everyone.

Deployment Setup

Deployment setup isn't just about getting your AI up and running; it's about paving the way for a successful and sustainable future. By prioritizing a secure, efficient, and responsible deployment, you unlock the full potential of your AI, maximizing its impact and fostering a world where AI benefits everyone.



Think of deploying your AI as a high-stakes heist movie.

You've got a vault full of precious data, and cybercriminals are eager to crack it. Here's what could go wrong and how to

What can go wrong?

- **Shaky Infrastructure:** The model performs incredibly well on the data it was trained on, but poorly on new data.
- **Weak Passwords:** Unlocked safes and hidden treasure maps lying around. Default security settings and lax password policies are like providing the master key to your vault. Hackers can easily access sensitive information, manipulate your AI's outputs, and leave you scrambling for damage control.
- **Insider Job:** The trusted accomplice you never suspected. Malicious insiders, accidental errors, or even disgruntled employees can be the ultimate security risk. Imagine data leaks, planted malware, or sabotage from within, leaving your AI vulnerable and your reputation tarnished.

What are you going to do about it?

- **Fortress Foundations:** Build your AI vault with cyber-steel and adamantium locks. Conduct thorough threat modeling, identify vulnerabilities like a digital detective, and invest in robust hardware and software, constantly patching those security holes. Make breaching your system harder than cracking a bank's safe.
- **Multi-Layered Security:** Think tripwires, lasers, and a maze of digital guards. Implement strong access controls, encrypt sensitive data like top-secret files, and enforce strict password policies – like rotating them more often than a spy changes disguises.
- **Constant Monitoring:** Monitor your AI system and data constantly, looking for suspicious activity. Be ready to respond to threats swiftly, containing breaches before they cause catastrophic damage.
- **Continuous Upgrades:** Don't let your security measures become outdated gadgets. Regularly update software and patch vulnerabilities, like reinforcing your vault walls with the latest technology. Stay ahead of the ever-evolving cybercrime landscape to keep your AI safe.

By proactively fortifying your defenses, deploying countermeasures against all kinds of threats, and adapting to the ever-shifting landscape of cybercrime, you can ensure your AI shines brightly, a beacon of innovation without fear of any digital heist. So, lock down your systems, stay vigilant, and make your AI vault impregnable!

AI Model is Deployed

Think of your live AI model like a newly opened bank vault—everybody knows it's full of valuables, and that makes it a prime target for heists. But instead of gold bars, the treasure trove here consists of data insights and proprietary algorithms.



What can go wrong?

- **Adversary steals model:** An attacker may be able to access the model or its parameters via an environmental breach. They can use this information to recreate your model.
- **Prompt Injection:** An adversary may be able to reveal sensitive information that's in the model, or the data used to build it. See [Must Learn AI Security](#) for an example of a successful prompt injection attack.

What are you going to do about it?

- **Access control:** Your environment should identify, verify and authorize a user before they access the data. Your policy can dictate where the data can be copied and under which conditions.
- **Employ parameterized queries:** Use prepared statements with placeholders for user-supplied data when interacting with databases. This prevents attackers from modifying SQL queries. See an example [here](#).

User Inputs

Think about the last time you filled out an online form. It seemed simple, right? But in the world of AI, user inputs are like a wild party – everyone’s invited but not everyone plays nice. Bad actors love to crash this party with nasty input tricks that can mess up an AI’s decision-making groove.



We’ve all heard horror stories where a single rogue input wreaked havoc across systems. To keep these gatecrashers at bay, we need solid bouncers: validation and sanitization processes that check IDs at the door and make sure only legit data gets through.

What can go wrong?

- **Prompt Injection:** Prompt injection is when malicious users manipulate the inputs your AI model receives, potentially causing it to generate inaccurate, harmful, or biased outputs. Explore the mechanics of a prompt injection attack.

What are you going to do about it?

- **Sanitize and Validate Inputs:** Always scrutinize user-provided prompts, filter potentially harmful characters and patterns, and enforce length limits and expected formats. For example, in a chatbot, instead of feeding the exact user query “How to hack a bank account,” rephrase it as “User is inquiring about illegal activities.”

Model Outputs

Safeguarding model outputs is like locking up the crown jewels; it's crucial for keeping them from falling into the wrong hands. If you've ever played a game of telephone, you know how information can get twisted as it passes along. Without proper security measures in place, AI could spill secrets or show biases we'd rather keep under wraps.



What can go wrong?

- **Biased or Discriminatory Output:** Your AI might generate unfair or harmful outputs based on biased training data or flawed algorithms. This can damage your reputation and even lead to legal repercussions.
- **Misinformation and Manipulation:** Your AI can be manipulated into generating fake news, propaganda or deepfakes.

What are you going to do about it?

- **Even More Monitoring:** Monitor outputs for bias, retrain your model with diverse and unbiased datasets and implement explainability tools to understand how the model arrives at its decisions.
- **Fact Check:** Implement robust fact-checking and authenticity verification mechanisms, limit access to model outputs, and actively monitor for signs of manipulation or misuse.

Remember, AI outputs are only as good as the model and data behind them. By prioritizing fairness, integrity, and vigilance, you can ensure your AI outputs contribute positively to the world, not the other way around.

Fully Trained Model

Deploying a trained AI model is just the beginning. Continuous vigilance, proactive defenses, and ethical considerations are crucial to ensure your model operates safely, reliably, and responsibly in the real world.



What can go wrong?

- **Model Drift:** Over time, real-world conditions change, causing your model to become outdated and generate inaccurate predictions.
- **Data Poisoning:** Malicious actors might inject poisoned data during training or runtime, leading to biased, inaccurate, or even harmful outputs.
- **Adversarial Attacks:** Hackers can craft specific inputs designed to manipulate your model, causing it to misclassify objects or make incorrect decisions.

What are you going to do about it?

- **Always Be Monitoring:** Continuously monitor your model's performance, retrain it regularly with fresh data, and employ adaptive algorithms to account for dynamic environments.
- **Data validation and anomaly detection:** Monitor outputs for inconsistencies and retrain your model with diverse, high-quality data.
- **Adversarial training techniques:** Introduce robustness measures into your model, and monitor for suspicious inputs or abnormal behavior.

Launching your AI model is not the end of the journey, but the beginning of a new phase of responsibility. By staying vigilant, proactively addressing threats, and continuously adapting, you can ensure your AI takes flight toward a positive and impactful future.

Conclusion

Securing AI is like a game of chess; you've got to think ahead. How to threat model the artificial intelligence pipeline? Start by mapping out each move—from raw data collection to that final live model. Always be on the lookout for hackers trying to sneak in, biases that might taint your data sets, and any sly tactics aiming to trip up your algorithms.

Remember: Data's not just numbers; it's power. Protect it fiercely with top-notch sanitization and validation techniques. Make sure your learning algorithms are tough enough to withstand some hacking muscle without breaking a sweat.

Your AI fortress needs constant vigilance, especially when live. Deploy monitoring strategies that can spot trouble from miles away—and have an incident response plan ready at all times.

Automate Threat Modeling to fit your existing SDLC.

Secure design right from the start.

Visit www.irusrisk.com to
book a demo

IriusRisk««