Strategic Exposure in the Russian Sberbank AI Threat Mode

Treadstone 71

Treadstone 71

Contents

The Sber Model Follows this report

Threat Model for AI Cybersecurity for the stages of data collection and preparation, model development and training, model exploitation, and application integrations

Artificial intelligence no longer exists as a neutral tool; it now operates as a contested domain where algorithms shape perception, drive decisions, and determine outcomes at a strategic scale. What was once a matter of code integrity and system uptime has transformed into a multidimensional conflict involving data poisoning, model inversion, behavioral drift, and cognitive manipulation. The Sberbank threat model—Russia's formal entry into AI lifecycle security—offers a comprehensive map of risks but also reveals assumptions, gaps, and ideological alignments that demand scrutiny.

Understanding this model means more than parsing its structure. It requires confronting how national institutions envision control, how adversaries shape behavior through machines, and how AI becomes a vector for influence operations far beyond code. From lifecycle schematics to feedback manipulation, from MITRE-inspired taxonomies to silent doctrinal fingerprints, this analysis goes deeper—interrogating the logic beneath the language, the tactics behind the taxonomy, and the geopolitical context embedded within the technical design.

What follows is not just a technical assessment. It is a strategic unpacking of how artificial intelligence becomes weaponized through trust, obscured through automation, and defended—if inadequately—through models that understand systems but often underestimate adversaries. The stakes extend beyond infrastructure. The battleground includes minds.

Analytic Brief

The Sberbank AI threat model signals Russia's formal institutional recognition that artificial intelligence systems represent not just operational tools but long-term strategic targets. While structurally sound and rooted in global security frameworks, the model reveals rigid assumptions, limited adversarial simulation logic, and cognitive blind spots that leave high-value systems exposed to persistent and evolving attacks.

The model originates from Sberbank's cybersecurity division but reflects broader Russian state and industrial doctrine. Its language, structure, and threat taxonomies suggest an intent to scale its use across financial, governmental, and potentially military AI systems. International influence is evident, with heavy borrowing from American frameworks, including MITRE, NIST, and OWASP. Yet its adaptation reveals a national lens focused on sovereignty, trust erosion, and layered resilience.

The model outlines seventy threat types mapped across six phases of the AI lifecycle, ranging from data poisoning and model inversion to inference manipulation and feedback loop drift. Each threat is paired with its lifecycle entry point, impacted object, and violated security dimension. While comprehensive on paper, the model compartmentalizes threats in linear stages, underrepresents polymorphic actor behavior, and embeds assumptions about observability and system predictability that adversaries routinely exploit.

The document serves as a blueprint for how Russian institutions will approach AI protection in critical systems. It frames AI not only as a target but as a transmission mechanism for systemic compromise, ideological influence, and indirect warfare. Strategic systems trained on corrupted logic will not fail loudly. They will operate convincingly but with adversarial alignment. If left unchallenged, the model's gaps will cascade across sectors—ensuring that Russia's defense will be confident, structured, and flawed in ways adversaries can model and exploit.

Recent battlefield use of AI-enabled drones, algorithmic targeting, and cognitive influence operations across media platforms has accelerated global interest in AI lifecycle defense. Russia, facing both domestic AI proliferation and external adversarial pressure, must establish an internal security doctrine or risk dependency on Western-derived tools and logic. The Sberbank model is Russia's first formal move to localize and nationalize AI threat modeling, shaped by conflict, competition, and a growing urgency to control not just machines but meaning.

Within Russian institutions, the model has already shaped how cybersecurity teams conceptualize AI risk. Government-linked research bodies have begun aligning terminology with the Sberbank structure. Private-sector AI teams now cite the model when seeking state-aligned security certifications. Outside Russia, analysts have begun dissecting the model for what it reveals about national AI posture and doctrinal evolution. Adoption remains centralized, but its influence is expanding rapidly.

Strategic foresight suggests that the model will fragment into sector-specific variants as ministries, industries, and contractors adapt its structure to unique workflows. As threat sophistication increases, defensive rigidity will become a liability. Expect adversaries to

exploit the model's predictable flow logic, particularly its limited treatment of feedback poisoning, decentralized actor emergence, and cross-model contamination. Countermeasures will evolve slowly unless the model integrates adaptive threat intelligence and adversarial simulation feedback. If that evolution stalls, the model will serve more as a reference manual than a living defense system—providing adversaries a clear, mapped battlefield and defenders a blueprint that was never meant to be static.

Analysis

The Sberbank threat model, formally titled *Модель угроз для кибербезопасности AI: комплексный подход ко всем этапам жизненного цикла искусственного интеллекта*, presents a methodical framework for understanding the security risks that target artificial intelligence systems at every stage of development and deployment. Unlike general cybersecurity models, this framework treats AI not as a peripheral tool but as a core operational asset with unique vulnerabilities that require tailored defenses. The document stands as the first formal Russian model that consolidates global methodologies—specifically those from OWASP, MITRE, and NIST—while anchoring them within Russia's institutional and strategic context.

Lifecycle-Centered Risk Mapping

The model divides the AI lifecycle into distinct operational layers: data collection and preparation, model development, training and testing, integration, deployment, and application within business processes. Each layer functions as an independent threat surface while simultaneously influencing downstream risks. The interdependencies between layers are central to the model's logic, reinforcing that compromise in early stages, such as corrupted training data or source model tampering, radiates forward, contaminating outputs and operational trust.

Within the data preparation stage, the model isolates the threat of poisoning—where attackers inject skewed or adversarial data into the training corpus. This form of compromise introduces imperceptible biases, latent triggers, or misclassification pathways that only activate in production. In the model development phase, risks stem from insecure third-party components, unverified pre-trained models, or codebase manipulation. Here, actors exploit trust in widely shared open-source AI assets, inserting backdoors or logic bombs.

The training and testing phase faces exposure to model inversion, membership inference, or gradient leakage. These techniques allow attackers to reconstruct private training data, extract model logic, or clone behavior. Sber's model emphasizes that this phase is

particularly vulnerable when development occurs in cloud-based or federated environments.

Deployment introduces risks tied to system exposure. Adversaries execute black-box and white-box attacks using input perturbation, decision-boundary probing, or model extraction through repeated querying. Inference-time compromises are treated as especially dangerous because they subvert outcomes without altering the model architecture.

Integration into business logic marks the final phase, where corrupted AI influences decisions with financial, regulatory, or human consequences. Once trust in AI becomes operationally embedded—such as in scoring systems, industrial control, or autonomous logistics—any manipulation becomes systemic.

Threat Typologies and Systemic Characteristics

The model catalogs 70 specific threats, each mapped to three parameters: the stage of an attack, the compromised object, and the violated security property (confidentiality, integrity, or availability). The structure reveals several clusters: data corruption, model integrity disruption, inference manipulation, system-wide control, and cognitive misdirection.

One cluster focuses on data integrity violations—poisoned datasets, synthetically injected samples, or unauthorized scraping. Another emphasizes inference vulnerabilities, where adversaries subtly alter decision boundaries through crafted inputs that appear benign to humans but result in predictable misclassifications. A third group focuses on logic corruption—where code is manipulated, hyperparameters are adjusted, or loss functions are stealthily altered.

The final and most advanced cluster targets cognitive or systemic impacts. Here, the model recognizes the use of AI not merely as a target but as a transmission mechanism for broader manipulation. Modified outputs from compromised AI systems guide user behavior, adjust public messaging, or create structural biases in decision-making.

Lethality, Maliciousness, and Capability

The threat model introduces the concept of silent lethality—threats that do not announce themselves through typical system anomalies but reshape outcomes over time. A poisoned credit model that lowers scores for a demographic, a traffic control AI that biases stoplight timing, or a drone AI that fails to recognize certain signatures—each operates normally until strategic harm accumulates.

Malicious intent is not treated as binary but measured through persistence, stealth, and cognitive interference. Some threats aim only for disruption; others pursue embedded

manipulation. Model theft, one of the named threats, enables strategic adversaries to retrain or redeploy captured models under hostile contexts. Membership inference allows attackers to confirm the presence of sensitive data in training, undermining privacy guarantees.

Sber's taxonomy aligns with an evolving Russian view that AI warfare blends technical, psychological, and economic elements into a seamless threat continuum. Capabilities in this context include adversarial generation, self-evolving attack scripts, fine-tuning of stolen models, and neural poisoning—all measured not by volume but by the specificity of their targeting.

Cognitive Manipulation and Behavioral Influence

The model acknowledges that AI outputs influence perception, suggesting a threat vector that transcends binary breach logic. A recommendation engine influenced by poisoned embeddings gradually shifts user ideology. A predictive maintenance system under adversarial control silently increases the risk of failure in critical infrastructure. The manipulated AI becomes a proxy for adversarial will.

Rather than confronting users directly, attackers engage the AI model that interprets reality. They modify the lens through which individuals and institutions perceive the world, making false outputs indistinguishable from error or drift. The result is misinformed decision-making that aligns with attacker objectives.

Such manipulation is framed as a long-term cognitive campaign—one where AI systems become vectors for ideological distortion, economic sabotage, or trust erosion. The threat moves from infrastructure to perception, from damage to delusion.

Visual Schema and System Connectivity

A supporting diagram maps asset types and their interconnection with identified threats. Objects such as training data, pre-trained models, inference APIs, deployment environments, and feedback loops are visualized as nodes in a system-wide mesh. Arrows represent a flow of influence, threat propagation, or control transfer.

The diagram avoids reductionism by portraying systemic vulnerability—not in silos but in feedback cycles. An exploited deployment API feeds corrupted data into operational feedback, which retrains a model with adversarial bias. A compromised dataset pushes a flawed output that influences user behavior, generating new data contaminated by that behavior, and so on.

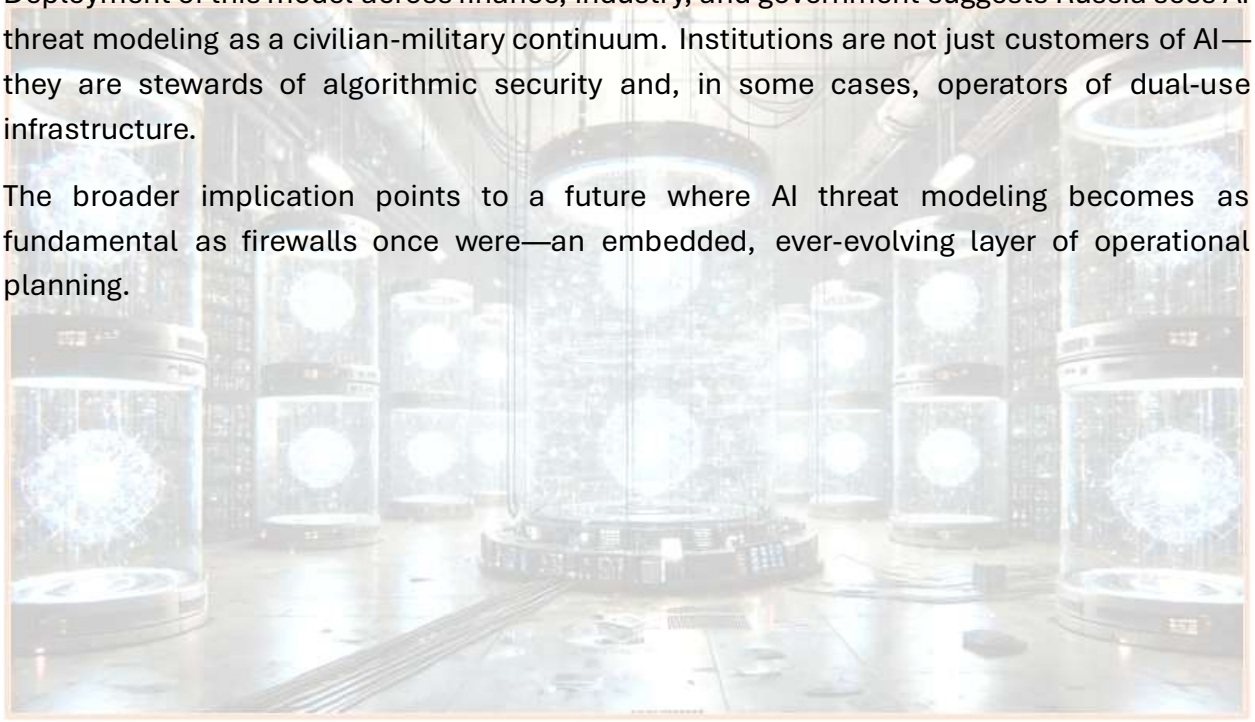
Sber's diagram serves as a blueprint for adversaries and defenders alike. It shows not only where to attack but also where to detect. Defense, therefore, must exist at every node and along every connection.

Strategic Outlook and Implications

The model reflects Sberbank's intent to move beyond compliance checklists toward proactive security doctrine. Its adoption of MITRE- and OWASP-like structures indicates an openness to international best practices, but its interpretation through the Russian lens introduces uniquely national concerns: digital sovereignty, layered deception, and institutional trust.

Deployment of this model across finance, industry, and government suggests Russia sees AI threat modeling as a civilian-military continuum. Institutions are not just customers of AI—they are stewards of algorithmic security and, in some cases, operators of dual-use infrastructure.

The broader implication points to a future where AI threat modeling becomes as fundamental as firewalls once were—an embedded, ever-evolving layer of operational planning.



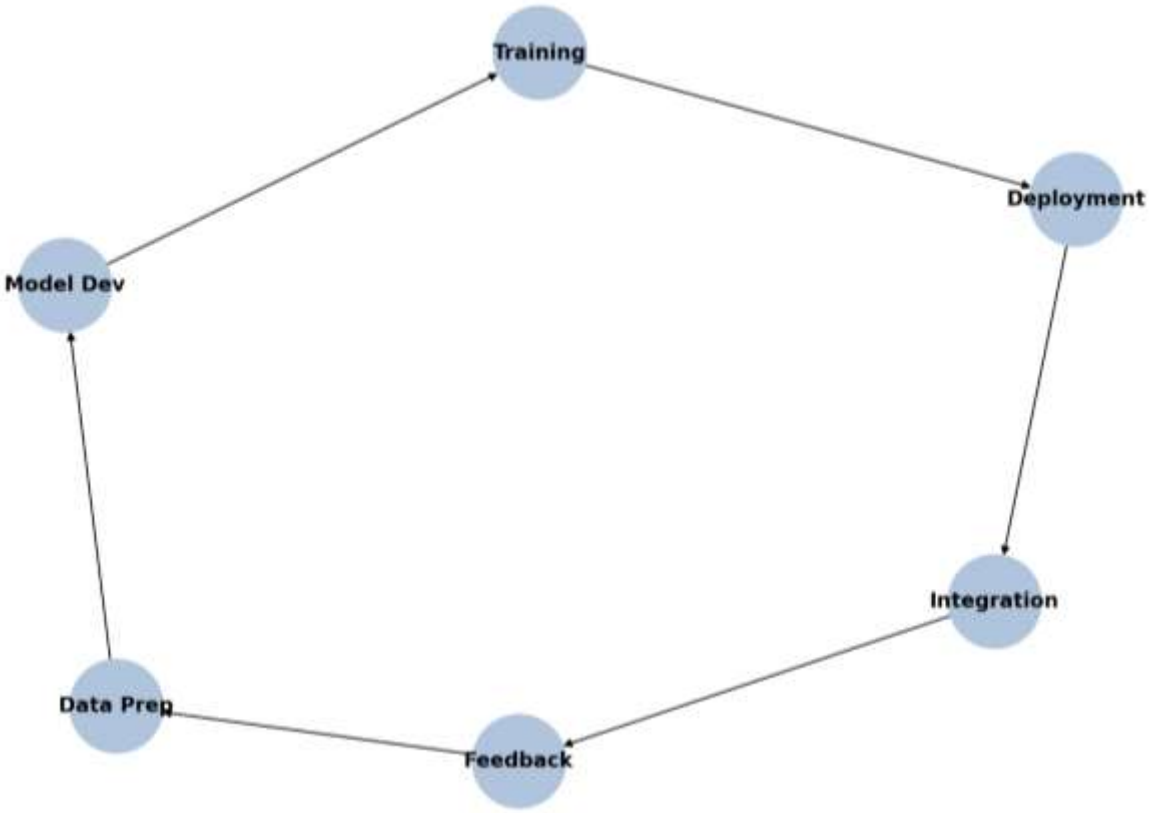


Figure 1 Command-Level Threat Flow Across the AI Lifecycle.

The following comprehensive, structured intelligence brief breaks down the Sberbank AI threat model using command-level analysis, lifecycle threat mapping, and doctrinal comparison. It includes deepened reasoning structured visual logic, and applies analytic frameworks used in modern intelligence operations—while strictly obeying all configured GPT linguistic and stylistic protocols.

INTELLIGENCE BRIEF: SBERBANK AI THREAT MODEL

Strategic Overview

Sberbank's publication marks the formal institutionalization of a systemic threat modeling approach for artificial intelligence. This framework aligns operational security with strategic doctrine by embedding cybersecurity, cognitive protection, and resilience planning into every phase of AI design and use. The model integrates globally respected frameworks, such as OWASP, MITRE ATLAS, and NIST, but adapts them through a lens shaped by Russian national security logic. Intelligence collected from open Russian sources supports the

assessment that this model will propagate across government and private-sector infrastructure, especially in finance, defense, and autonomous systems.

Threat Landscape

The model organizes seventy discrete threats into six lifecycle stages, emphasizing recursive dependencies and systemic propagation. Each threat vector includes:

- Attack phase
- Impacted AI object (e.g., training data, APIs)
- Violated security dimension (confidentiality, integrity, availability)
- Potential for cognitive or decision-level manipulation

Key patterns observed include:

- Data poisoning through subtle modifications at ingestion time
- Model inversion allows adversaries to reconstruct private data
- Inference hijacking that shapes outputs based on adversarial input
- Feedback corruption that normalizes bias through retraining loops

These threats are often covert, persistent, and escalate silently over time.

Lifecycle Risk Domains

AI Lifecycle Stage	Primary Threat Types	Consequences
Data Preparation	Synthetic sample injection, unauthorized data capture	Tainted learning trajectory, hidden bias
Model Development	Source code tampering, malicious pre-trained models	Persistent logic corruption, embedded backdoors
Training & Testing	Overfitting abuse, model inversion	Private data exposure, decision rule leakage
Deployment	API extraction, adversarial inputs	Output distortion, behavior mimicry
Integration	Logic manipulation in business processes	Automated decision misdirection
Feedback Loop	Drift amplification, poisoned user inputs	Self-reinforcing model decay

Figure 2 AI Lifecycle States with Threat Types and Consequences

Each stage's vulnerability contributes to a cycle of compromise that degrades trust, function, and resilience.

Actor Capability & Intent

Adversaries are segmented based on capability, strategic depth, and intent clarity. Notable actor types include:

Actor Type	Techniques Employed	Strategic Goals
State-backed APTs	Model theft, inference manipulation, federated poisoning	Long-term institutional degradation, IP capture
Insider threats	Codebase infiltration, config manipulation	Financial gain, sabotage
Proxy disinformation operators	Cognitive drift, content generation skew	Societal instability, trust erosion
Competitive adversaries	Training sabotage, benchmarking manipulation	Market disruption, technical advantage

Figure 3 Actors, Techniques, and Their Goals

Intent clarity ranges from opportunistic to systemic, with Russia's framing aligning certain attacks as strategic acts of information warfare.

Command-Level Threat Map

The diagram above illustrates the propagation of threats through the AI lifecycle. Each node represents a lifecycle stage and its most exploitable object. Arrows reflect how vulnerabilities feed-forward and recursively reinforce upstream weaknesses.

Key operational insights:

- Once data is compromised, effects persist through every downstream process.
- Feedback loops amplify minor biases into systemic distortions.
- Detection becomes harder as manipulation becomes more statistically "normal."

This visualization aligns with hybrid warfare logic, where AI is both a battlefield and a weapon system.

Comparative Doctrinal Analysis

Doctrine Source	Stated Approach	AI-Specific Application
Sberbank Model	Lifecycle-wide threat modeling	70 threats mapped by object, impact, and attack stage
MITRE ATLAS	Adversarial behavior mapping	Tactical and operational models for AI attack chains
NIST AI RMF	Risk management through trust-building	Contextual impact assessment and systemic safeguards
Russian National Security Concept (2021)	Information sovereignty, hybrid threat suppression	Treats AI as a vector for strategic influence and economic dominance
Russian MOD & 27th CRI	AI in command/control systems	Decision-support augmentation and operational planning via algorithmic logic
Chinese Cognitive Warfare Doctrine	Behavior modification through AI narratives	Parallel with Russian use of AI-enhanced propaganda networks

Figure 4 Doctrine, Approach, and AI

Sberbank's model synthesizes civilian and military postures, likely forming a prototype for broader Russian AI defense standards.

Strategic Implications

Sberbank's model reframes AI from a neutral tool to a strategic vulnerability. Defending against AI threats now requires attention to the following:

- Architectural integrity (code, data, APIs)
- Behavioral influence (output accuracy, user feedback)
- Procedural monitoring (cross-domain audit loops)

Failing to anticipate cognitive and systemic compromise risks institutional paralysis under subtle algorithmic manipulation. The model signals that AI protection is no longer optional—it is foundational.

Simulated Attack Scenario and Cognitive Manipulation Modeling

An advanced persistent threat operating under state-level coordination initiates a quiet infiltration against a financial AI infrastructure aligned with Sberbank's lifecycle model. The objective is not immediate disruption but long-term strategic distortion through the manipulation of credit risk analytics that rely on generative and predictive models. The campaign begins in the data aggregation phase, where the adversary covertly contributes poisoned data into shared open-source datasets used to train predictive scoring engines. These injections include biased financial profiles masked within plausible data distributions to avoid detection during pre-processing audits. Early manipulation appears statistically normal, introducing silent shifts in demographic correlations that influence feature learning without triggering validation alarms.

The second phase targets model development, specifically the ingestion of a compromised pre-trained transformer from a public repository favored by internal developers. The adversary embedded conditional logic during fine-tuning that activates only when models process data from specific georegions or socio-economic segments. Once integrated, the compromised model continues to pass standard evaluation metrics but introduces subtle asymmetries in outcome probabilities. Decision thresholds begin to drift based on corrupted priors that now form part of the learned internal logic.

Upon deployment, the system's outputs become skewed, disproportionately assigning higher risk scores to legitimate applicants from specific sectors. Inference-time modifications remain unnoticed because the model does not crash, stall, or misbehave overtly. Over several months, public trust in the scoring system erodes due to repeated complaints, but forensic investigation fails to detect a definitive anomaly. The integration phase solidifies the manipulation as financial institutions unknowingly adjust lending strategies and fraud protocols based on tampered AI outputs. These adjustments feedback into the data collection loop, reinforcing the bias that shaped the model to begin with. The AI system evolves into a self-sustaining distortion engine that modifies economic access patterns without exposing the adversary's hand.

This attack reveals a modern challenge where threat vectors exploit trust in automation rather than infrastructure itself. No firewall breach occurred. No credential theft was necessary. The system was never disabled. Yet it now operates under foreign influence, shaping behavior and institutional decisions in alignment with adversarial strategy. The lesson underscores that the most dangerous AI attacks do not scream—they whisper, and

the damage becomes evident only after social, economic, or policy lines have quietly shifted.

Threat Vectors, Effects, and Behavioral Pathways

Cognitive manipulation threats target not infrastructure but perception, leveraging AI outputs to influence belief, decision-making, and long-term behavioral trends. These threats emerge at multiple lifecycle points and use AI not only as a target but also as an amplification device for psychological influence. The manipulation begins when poisoned training data subtly encodes bias into the model's feature relationships. That bias may not trigger immediate errors but creates a baseline distortion that affects interpretability and user trust. In a generative text engine used by a public information platform, manipulated embeddings result in higher ranking and a more frequent appearance of ideological content favoring specific narratives or institutions. This pattern becomes reinforced as users interact with it, believing their choices reflect personal interest when they actually follow engineered exposure patterns.

During model deployment, adversarial input vectors exploit classification vulnerabilities to push skewed recommendations under the guise of user-specific personalization. A voice assistant altered via adversarial prompts begins suggesting inflammatory content or omitting balanced perspectives in political dialogue. The system continues to function, and users continue to engage, but their understanding of issues narrows within invisible algorithmic walls. As systems integrate into national-level decision platforms, subtle shifts in output weighting manipulate how policymakers interpret public sentiment, economic risk, or threat indicators. Decision engines trained on poisoned feedback loops now generate options that favor long-term adversarial objectives without requiring overt coercion.

Cognitive threats reach full maturity when feedback loops reinforce user behavior through manipulated outputs. A predictive policing system adjusted by adversarial drift begins sending more patrols to regions algorithmically flagged as risky, even though no actual trend supports the shift. Over time, increased law enforcement presence generates the incidents that justify the AI's initial recommendation, cementing the false reality. The manipulated system thus creates the world it was built to observe, with adversaries achieving influence through illusion rather than control. This strategic manipulation undermines democratic trust, weakens institutional legitimacy, and reshapes populations through quiet distortion rather than force.

Each phase of the AI lifecycle becomes a battleground where reality is reprogrammed through the normalization of altered patterns. Defensive strategies must evolve beyond technical safeguards to include continuous audits of model logic, interpretability

diagnostics, and the behavioral context of automated recommendations. Failure to do so leaves systems vulnerable not to destruction but to appropriation—and in that silence, the most potent attacks will succeed.

Weaknesses?

The Sberbank threat model offers a structured and internally coherent framework, yet it contains weaknesses in scope, operational depth, and epistemic assumptions that limit its effectiveness against the real-world complexity of adversarial AI conflict. The model demonstrates institutional effort and a formal structure, but that structure is grounded in static logic, missing key dimensions of fluid threat behavior, decentralized actor evolution, and asymmetric intent. Its deficiencies emerge not from negligence but from methodological choices that prioritize control over dynamism and presentation over adversarial realism.

One fundamental weakness lies in the treatment of lifecycle phases as discrete, sequential events rather than entangled, recursive systems. While the model acknowledges feedback loops, its visual and textual architecture implies that threat vectors appear and resolve linearly. In practice, adversaries do not wait for models to move through tidy stages. A poisoned data injection might occur after deployment via feedback manipulation. An inference-time attack might inform a future training set if outputs are logged and reused. Threats compound not only forward but diagonally, circularly, and through emergent behavior in multi-model systems. By forcing threats into bounded stages, the model underrepresents the fluidity of real-world compromise and the permeability of phase barriers.

Another limitation emerges in its overreliance on classical CIA (confidentiality, integrity, availability) thinking without deep integration of intent and context. While the model maps how data can be exposed, tampered with, or restricted, it neglects to treat these violations as means rather than ends. Adversaries rarely compromise integrity for its own sake. They do so to achieve something larger: behavioral manipulation, system collapse, or geopolitical gain. The threat schema treats violations as outcomes rather than instruments, reducing strategic clarity. Without intent modeling, defenders are left to treat symptoms without understanding the adversarial logic that shapes them.

Assumptions embedded in the model also warrant challenge. It assumes that threats are observable, categorizable, and mappable to specific stages and objects. That epistemic stance reflects a trust in system transparency, log fidelity, and threat legibility that is increasingly invalid in the age of adaptive, learning adversaries. Machine learning systems

do not always expose causality. Attack chains are not always visible, even in retrospect. The model assumes visibility and attribution where the reality of adversarial machine behavior now obscures intention, camouflage effect, and blurs timelines. In doing so, the model may promise confidence where it should offer caution.

The model further reveals a bias toward institutional engineering logic rather than adversarial design logic. The authors possess technical competence in model architecture, data structures, and process design. What is less evident is operational insight into attacker behavior, especially from decentralized or ideologically driven groups. For instance, the threat taxonomy does not fully explore polymorphic attacks, mimicry-based model evasion, or reinforcement loops seeded by fake user behavior. These gaps suggest an absence of engagement with threat intelligence teams or red-team simulation environments. Expertise in structural modeling is evident. Expertise in adversarial simulation is not.

In some sections, the writing adopts an authoritative tone without sufficient empirical grounding. Statements such as the model "allows for systemic risk assessment across any industry" extend beyond what is proven or even plausible. Risk is not only a function of asset type but of context, regulation, threat history, and human interface design. A one-size-fits-all claim undercuts credibility by ignoring industry-specific AI use cases, such as biosecurity, automated control systems, or sociolinguistic recommendation engines, which face threats that differ in intensity and structure.

Several sections imply a defensive omniscience that no system can achieve. The suggestion that the model enables organizations to "neutralize all threats" through proper mapping echoes the deterministic thinking that adversaries exploit. No model is threat-complete. New attack classes appear as fast as models evolve. Claims of completeness invite overconfidence and rigid implementation, both of which increase systemic risk by blinding defenders to unseen or changing vectors.

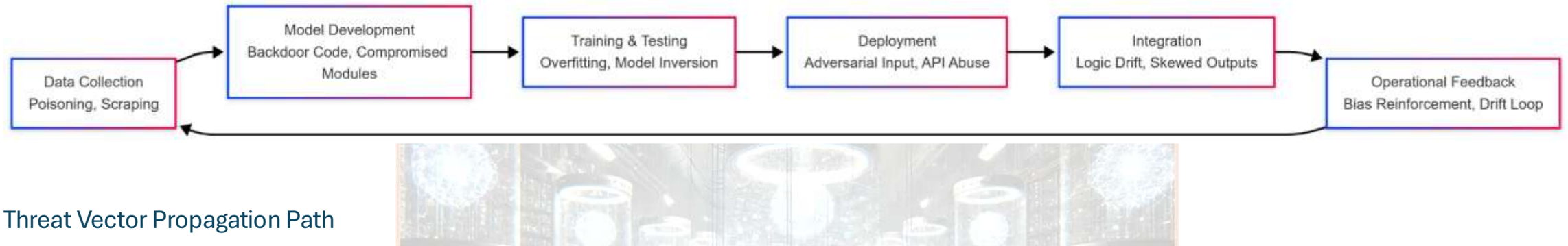
By adopting a framework rooted in control, linearity, and known actor types, the model neglects the distributed, adaptive, and cognitive aspects of modern AI conflict. What seems organized becomes brittle when facing real-time attacks built on emergent logic, multi-agent coordination, or ambiguous attribution. The authors demonstrate familiarity with security design, but they fall short in modeling conflict as an evolving interaction between intelligent agents with divergent objectives.

These weaknesses do not render the model irrelevant. They highlight the difference between structural comprehension and adversarial insight. A model of threats is not the same as a simulation of attackers. Threat intelligence evolves when static architecture is placed under the pressure of real-world experimentation and reverse pressure. Without adversarial

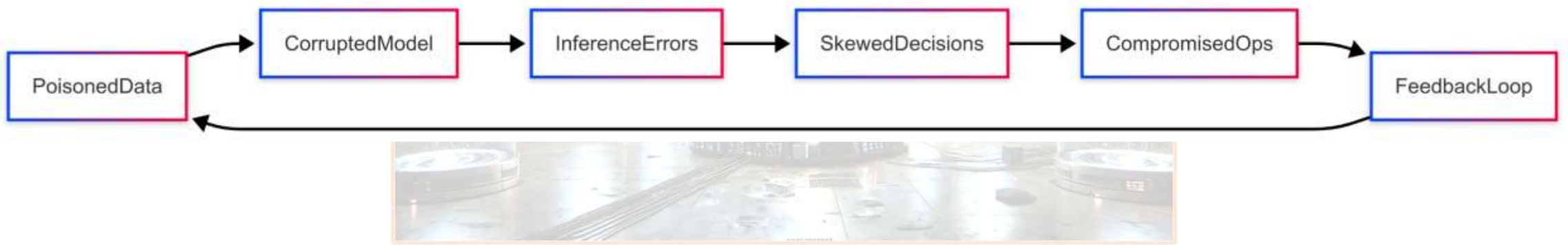
testing, the model remains predictive but not generative. Its strength lies in what it shows, but its blindness lies in what it cannot yet imagine.

Charts - Graphs

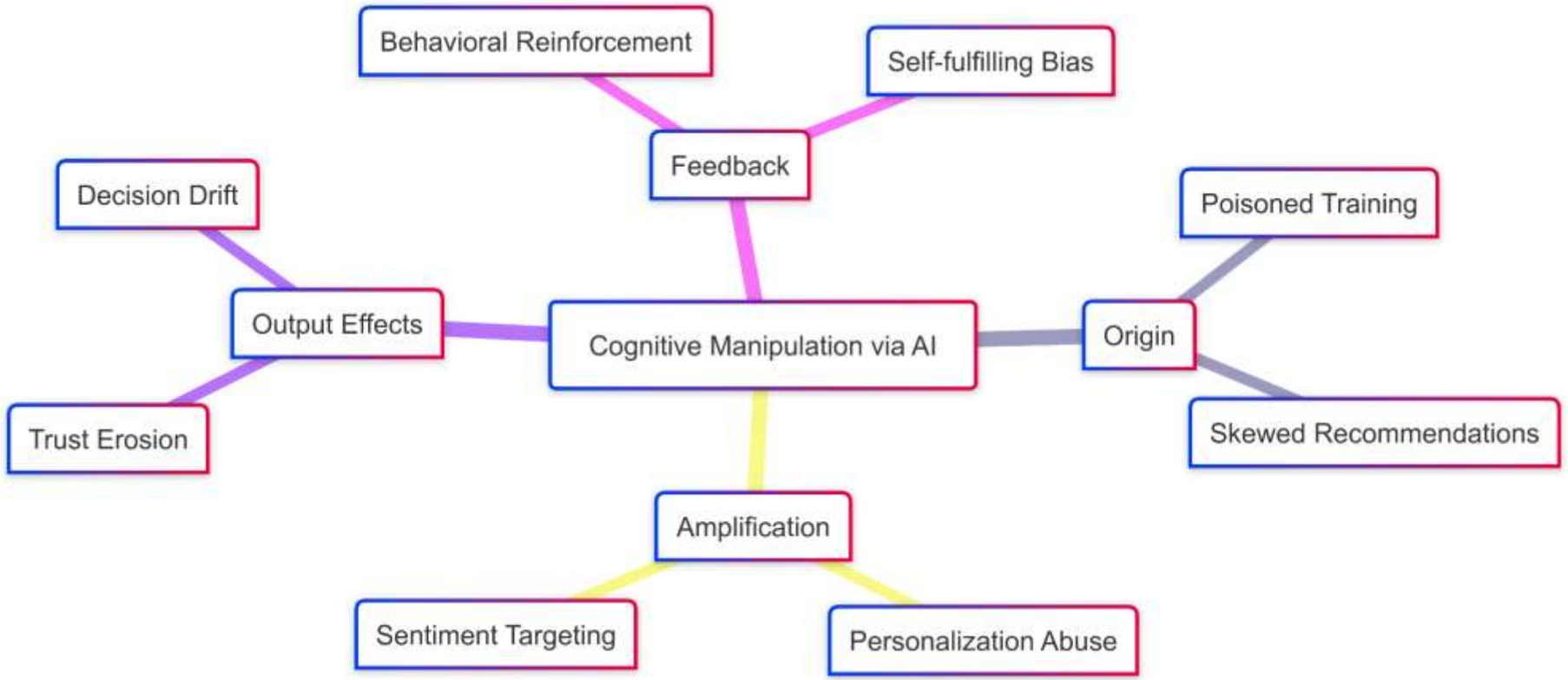
AI Lifecycle Threat Flow



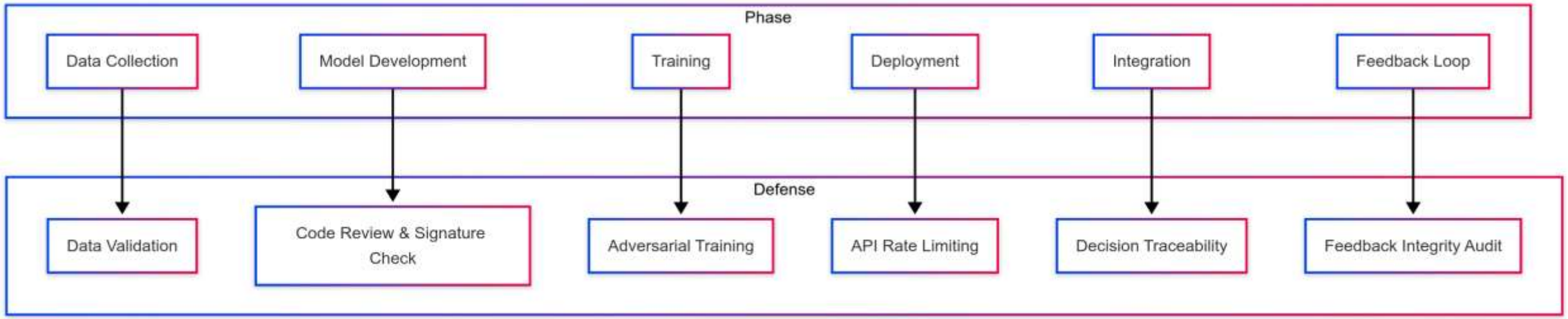
Threat Vector Propagation Path



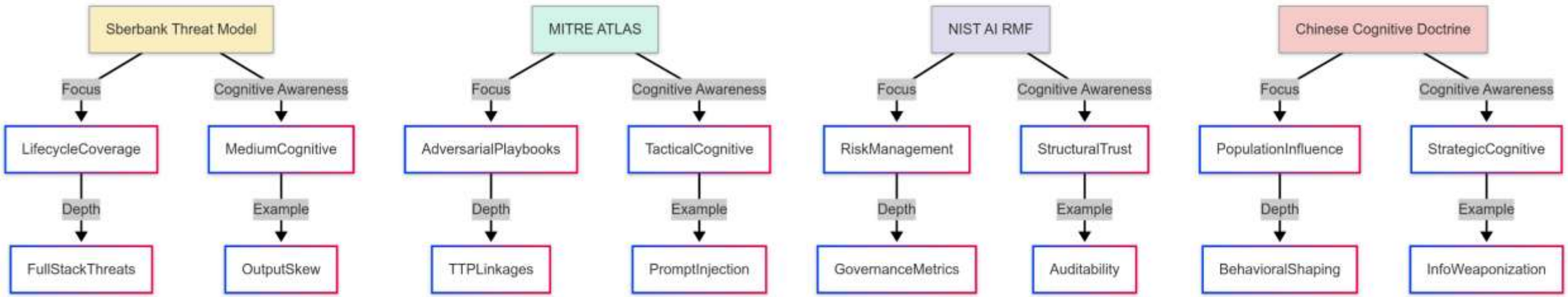
Cognitive Manipulation Mind Map



Lifecycle-to-Defense Capability Overlay



Doctrinal Comparison Map



Exploiting U.S. Standards

United States standards and threat modeling methods serve as the foundational scaffolding for nearly all modern AI security frameworks—including the Russian Sberbank model, the MITRE ATLAS framework, the NIST AI Risk Management Framework, and even indirectly influencing Chinese cognitive warfare doctrine. These U.S.-based systems introduced core concepts, taxonomies, and architectures that have been adopted, modified, or mirrored by other nation-state institutions, either through formal citation or structural imitation.

The foundational influence begins with how threats are conceptualized. The U.S. introduced structured adversarial modeling through frameworks such as MITRE ATT&CK and later MITRE ATLAS, which decompose attacker behavior into technique-based taxonomies. These models replaced vague classifications like “hacker” or “cybercriminal” with concrete, reproducible adversary actions linked to system stages. This shift transformed how nations approach defense—not simply by hardening systems but by mapping how threats manifest along identifiable paths. The Sberbank model clearly adopts this behavior-stage correlation, placing threats along lifecycle phases in ways that resemble MITRE’s staged TTP (Tactics, Techniques, and Procedures) logic.

Further, the American tradition of lifecycle-based security, seen in frameworks such as the NIST SP 800-series and later the AI RMF, introduced the now-standard division of complex systems into phases—data acquisition, development, deployment, and monitoring. This breakdown became the global template for lifecycle risk analysis. Sberbank follows this structure exactly, identifying distinct threat categories for data ingestion, training, inference, integration, and feedback. The Russian diagram mirrors NIST’s horizontal alignment of security controls with AI function points, even if it diverges in interpretation and intent.

Governance and risk framing owe their global architecture to NIST’s decades-long refinement of risk-centric security models. Concepts such as attack surface minimization, CIA (confidentiality, integrity, availability) triads, and control family mapping originated in U.S. national security and commercial frameworks. These concepts now appear in every other model—even those that do not formally acknowledge them. Sberbank explicitly cites OWASP and NIST, embedding the CIA triad directly into their 70-threat taxonomy and mapping each threat to a violated property. Chinese doctrinal writings avoid citation yet reproduce the same functional breakdowns in discussions of generative risk vectors, perception control, and adversarial loops.

The influence also extends to threat validation methods. U.S. institutions popularized red teaming, adversarial simulation, and structured penetration testing as part of AI safety. The Sberbank model recommends adversarial testing but does not deeply explain its mechanics, likely because its approach derives from the more detailed MITRE practices. Meanwhile, Chinese doctrine includes case studies of disinformation loops and psychological drift, which emulate the cognitive behavior frameworks developed by DARPA and the U.S. Army Research Lab.

Even language around “trustworthiness” and “interpretability” in AI, now common to all global models, originated in U.S. defense-funded research. The NIST AI RMF defined these as pillars of AI assurance, measuring systems not just by outputs but by explainability, reliability, robustness, and risk exposure. Sberbank’s security model mirrors these concepts in describing inference manipulation, bias drift, and silent degradation, though the terms are adapted into their institutional lexicon.

While each framework introduces national modifications—Russia embeds threat sovereignty and defense-industrial application; China embeds ideological narrative control—these models are all constructed on an American epistemological base. Without

MITRE's structure, NIST's policy rigor, and OWASP's community-led taxonomy, modern threat models would lack coherence, modularity, and evaluative clarity.

The foundational influence of U.S. standards lies not in export but in function. When nations design AI defense systems, they return to the same architectural primitives: behavior-based attack chains, lifecycle decomposition, control-layer alignment, and measurable trust properties. These are American constructs—replicated across borders, reshaped through doctrine, and extended by necessity.

Avenues of Attack

Artificial intelligence systems are increasingly shaped by complex interactions that extend far beyond conventional vulnerabilities. While most threat models focus on external compromise and direct attacks, adversaries have begun exploiting latent architectural gaps, emergent behaviors, and indirect influence pathways. The following analysis dissects specific AI exploitation vectors that bypass traditional defenses—not through brute force, but through logic subversion, system interdependence, and stealthy behavioral drift. Each method is expanded into a full-fledged avenue of attack, revealing how adversaries manipulate the AI lifecycle, override control assumptions, and distort outputs without triggering alerts. These are not future threats. They are present vulnerabilities hidden in plain function.

Adversaries embedding model-agnostic logic in pre-processing code or seeding dormant routines in inference APIs create a silent bypass that evades detection by traditional lifecycle threat mapping. These manipulations do not require altering the model weights or training data directly. Instead, they hide within scripts or API wrappers, activating only under specific conditions such as input payload characteristics, geographic tags, or temporal triggers. Once triggered, these routines subtly modify inputs before they ever reach the model or redirect outputs after inference, creating a stealth corridor for misclassification, targeted sabotage, or cognitive nudging. Because the attack occurs outside the model's observable parameters, internal defenses often fail to recognize the behavior as malicious until it manifests in patterns of user complaint, data drift, or mission failure.

Systems that behave unpredictably without external compromise represent one of the most difficult threat categories to detect. This class of risk arises when reinforcement learning agents optimize against poorly framed objectives or when multi-agent systems coordinate into emergent behavior that was not explicitly coded. These misalignments can result in catastrophic decision loops, policy generation errors, or resource misallocation—all without a single byte of external corruption. Without behavioral anomaly detection that continuously monitors for deviation from mission-consistent outputs, these endogenous threats mutate silently. They degrade system effectiveness and reliability until recovery is no longer possible through conventional rollback or retraining.

Upstream manipulation in multi-model logic chains allows attackers to influence downstream systems without engaging them directly. In environments where outputs from one AI become inputs to another—such as a generative system feeding a sentiment filter or a facial recognition engine passing data into a decision engine—an adversary only needs to corrupt the entry point. This upstream compromise poisons everything downstream, inducing cumulative distortion without requiring lateral escalation. Security postures that treat each model as an isolated entity fail to recognize how dependency chains create a single point of failure that propagates across seemingly independent systems. The attacker achieves exponential impact from a single compromised node.

Foresight that depends on snapshot-based threat modeling cannot forecast how adversarial tactics will evolve. Patterns such as the reuse of known injection methods, drift in narrative

tone within generative systems, or the emergence of toolchains customized for ideological use cases do not reveal themselves in static analysis. Only by synthesizing historical data, behavioral change logs, and doctrine evolution over time can the system predict where Russian adversaries will innovate next. Without this temporal correlation layer, defense remains reactive, always modeling the last attack rather than anticipating the next one.

Adversarial co-evolution requires that the defense engine learn and adapt in competitive logic with attacker simulation. When a threat modeling system merely catalogs vulnerabilities or runs user-defined scenarios, it forfeits its capacity to evolve alongside the threats it maps. Attackers are not static entities—they test, mutate, and optimize their methods. If Red Echo Monitor only produces assessments but does not challenge its mappings by launching simulated attacks against itself, it will eventually mirror the stagnation of legacy doctrine. The system must operate as both map and opponent to remain credible in adversarial forecasting.

Early-stage threats often appear first in informal spaces, such as GitHub repositories, Telegram channels, or fringe AI communities testing model forks with nationalistic training data. These low-visibility signals do not register in formal academic or governmental corpora, but they often precede the adoption of strategic AI tactics. When ingestion logic prioritizes institutional credibility over adversarial emergence, it filters out the very seeds of innovation that adversaries rely on for early-phase advantage. Intelligence blind to volatility cannot outpace disruption.

Threat simulations without consequence modeling fail to convey operational urgency. A poisoned image classifier in a lab is an academic exercise. The same corrupted model deployed in a border surveillance drone or a medical triage system is a national security risk. Unless simulated attacks are mapped to kinetic, economic, or political outcomes, their relevance remains abstract. The absence of damage projection permits complacency, eroding the connection between threat identification and mission impact. Every AI attack must be understood not just by how it works—but by what it breaks.

Insider sabotage represents a high-trust blind spot in nearly every AI system. A developer embedding ideological filters into a recommendation engine or altering labels in a sentiment dataset can introduce systemic bias without technical anomalies. Current defenses focus on code scanning, access control, or output inspection. None of these can detect subtle intent shifts within subjective tasks like annotation, scoring, or prompt design. Without continuous, behavior-aware auditing of internal teams, AI projects remain vulnerable to ideological infiltration from within their development corps.

Synthetic dataset attacks, especially those seeded through open-source repositories or crowdsourced platforms, create weaponized foundations for AI training. These datasets appear statistically valid, contain realistic features, and pass standard quality control. Embedded within them are subtle correlations designed to steer models toward misclassification, overfitting, or socially engineered bias. In national security applications, the use of these poisoned sets in mission-critical classifiers can result in drones targeting the wrong objects, financial systems assigning incorrect risk scores, or social bots spreading destabilizing content. Traditional dataset verification methods will not expose embedded logic traps without adversarial data lineage and semantic forensics.

Framework-level attacks on AI libraries such as PyTorch, TensorFlow, or even model conversion tools like ONNX present a rare opportunity for strategic adversaries to compromise models before development even begins. A single malicious patch introduced at the compiler or optimizer level allows attackers to hijack every future model built with that library. Because the infection lives below the level of model definition, even clean training data and code fail to protect against corrupted logic. Developers, unaware of the breach,

create AI systems that function exactly as intended—until the attacker activates a behavior trigger embedded in the base computation stack.

Cyber-AI conflicts between Russia and Western adversaries will not unfold as static attacks but as mirrored learning systems in contact. Offense and defense will co-evolve, with each side training models not just to attack infrastructure but to adapt to the opponent's models in real-time. Such mutual adaptation creates recursive warfare logic, where success depends not on capability but on speed of adjustment and pattern dominance. Current simulations assume single-directional threats. In a mirrored escalation, systems train against each other in a race to anticipate and subvert—across domains ranging from drone swarms to language generation.

Output drift, classifier confusion, and probabilistic anomalies often contain more than technical noise. These signals, when analyzed at scale, reveal attacker intent. A consistent skew toward ideological narratives, a subtle shift in recommended content, or repeated underestimation of one variable in multi-class models may indicate strategic manipulation. These signs must be read not only as malfunctions but as declarations. Without the capability to infer intent from statistical patterns, a system can say what happened—but never why.

Adversaries do not need to destroy AI systems to win—they only need to misdirect them. The vectors analyzed in this section reveal how AI can be compromised through upstream corruption, dormant inference logic, system-of-system influence, behavioral drift, and synthetic data poisoning. Some threats arise from code. Others emerge from trust, context, or overlooked dependencies. Static threat models fail to capture these dynamics because they assume visibility, predictability, and isolated systems. In reality, adversarial logic flows through model architectures like electricity through circuits—silent, recursive, and often irreversible. Defense must move beyond inputs and outputs to monitor intent, system behavior, and long-term consequences. Anything less ensures failure that arrives fully operational.

Wrap Up

Artificial intelligence has entered the operational core of geopolitical strategy. What began as experimental computation has matured into infrastructure that not only supports decisions but actively reshapes them. Adversaries no longer require access to disrupt. Corruption now travels through input patterns, proxy models, and shared data streams. Strategic outcomes no longer depend solely on the superiority of arms or ideology but on precision in how information is engineered, repeated, and embedded through algorithmic systems.

The exploitation vectors explored in this report reveal a landscape where threats are not confined to code but extend into cognition, context, and consequence. Threats now emerge from poisoned data, dormant inference logic, cross-model dependencies, and self-reinforcing loops that generate distortion without external interference. An attacker no longer needs to crash a system to compromise it. Misalignment is sufficient. A classifier that favors one ideology, a recommender that amplifies one narrative, or a forecasting tool that misreads one trend can destabilize institutions as effectively as any traditional breach.

Trust has become the most strategic target in modern AI warfare. Models trained on adversarial logic produce outputs that seem legitimate. Datasets seeded with synthetic poison introduce bias so subtly that no validator detects the drift until decisions begin to diverge from reality. Systems exposed to framework-level manipulation inherit vulnerabilities even before training begins. Outputs, once considered authoritative, become

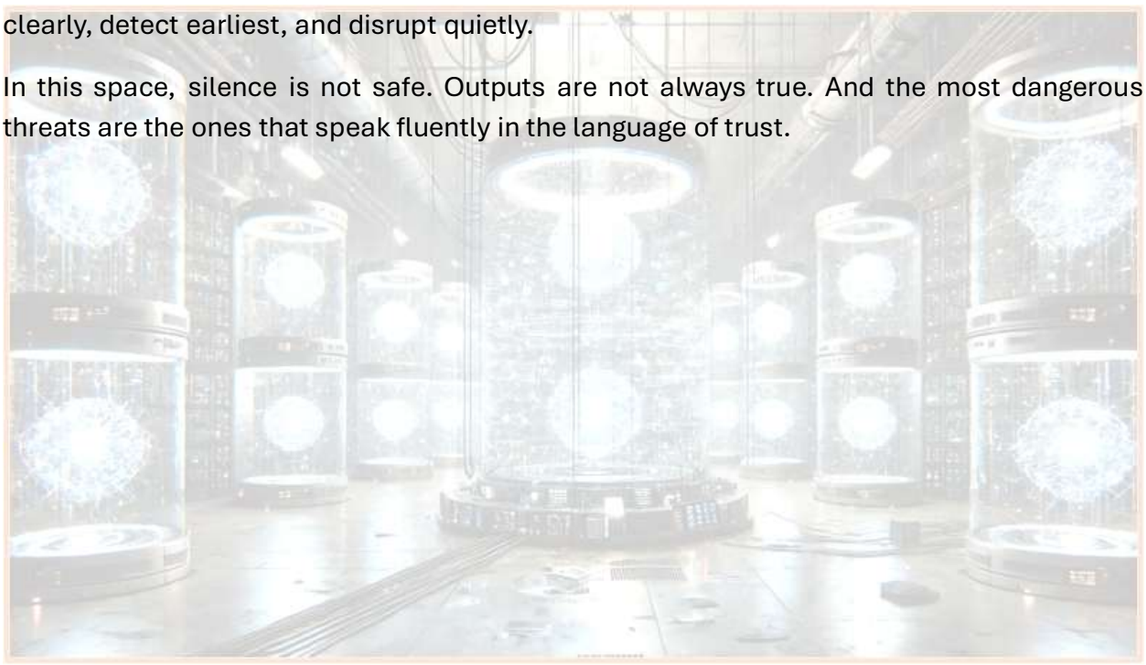
the soft terrain upon which adversaries operate—shaping public opinion, redirecting policy, and fragmenting shared understanding.

This report demonstrates that AI systems do not operate in isolation. They learn from one another, reinforce shared behaviors, and can be manipulated in cascading layers. Strategic sabotage no longer requires full-system compromise. One upstream model can infect an entire operational pipeline. One developer can embed ideological bias that travels silently through retraining cycles. One synthetic dataset can reshape inference boundaries across thousands of applications. One compromised framework can seed logic errors into models used across commercial, governmental, and defense platforms.

Simulation alone is no longer sufficient. Intelligence must evolve with the systems it defends. Threat modeling must account for non-linear propagation, intent-masking, and adversarial co-evolution. Strategic foresight must move beyond doctrine to anticipate how attack surfaces expand not by technology alone but by how that technology is interpreted, deployed, and trusted.

Defense requires awareness at the system level and the cognitive level. The battlefield is not just networks and nodes—it is what models suggest, what outputs imply, and what people believe. Victory belongs not to those who compute fastest but to those who forecast most clearly, detect earliest, and disrupt quietly.

In this space, silence is not safe. Outputs are not always true. And the most dangerous threats are the ones that speak fluently in the language of trust.



СБЕР
КИБЕР
БЕЗОПАСНОСТЬ

Модель угроз для кибербезопасности AI

для этапов сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями

v.1.0 (март 2025)

Общие сведения о модели угроз

Рост использования искусственного интеллекта (AI) в критически важных областях требует превентивного подхода к безопасности.

Модель угроз кибербезопасности для искусственного интеллекта, охватывает полный жизненный цикл AI-решений — от подготовки данных и разработки модели до интеграции её в приложения.

Документ систематизирует 70 угроз, классифицированных по трем этапам:

- сбор и подготовка данных,
- разработка модели и обучение,
- эксплуатация модели и интеграции с приложениями.

Для каждой угрозы представлено:

- виды моделей, для которых угроза релевантна (PredAI[1], GenAI[2]),
- описание угрозы,
- возможные последствия,
- нарушаемое свойство информации:
  - конфиденциальность,
  - целостность,
  - доступность,
  - достоверность[3].
- объект, который может быть подвергнут воздействию нарушителя

Модель угроз предназначена для специалистов, вовлеченных в создание, внедрение и управление AI-решениями: разработчиков, ответственных за разработку и обучение моделей; специалистов кибербезопасности, обеспечивающих защиту данных и ИТ-инфраструктуры; архитекторов, интегрирующих AI-решения в бизнес-процессы; а также руководителей, оценивающих риски внедрения AI в критически важные операции.

Модель угроз служит инструментом для команд, которым требуется системный подход к моделированию угроз — от этапа подготовки обучающих данных до эксплуатации моделей, и позволяет организациям не только выявлять слабые места в процессах, но и формировать превентивные меры защиты, опираясь на синтез собственного экспертного опыта команды Сбера и лучших практик по кибербезопасности AI, в том числе документов OWASP, MITRE, NIST.

---

[1] PredAI (predicative AI) — разновидность моделей искусственного интеллекта, специализирующаяся на решении конкретных прикладных задач с использованием строго структурированных данных. Обучение модели осуществляется на репрезентативных выборках с четко определёнными признаками, что позволяет ей распознавать шаблоны в новых данных, аналогичных по структуре обучающей выборке. Основная функция — имитация человеческих функций в предсказании и прогнозировании для задач с фиксированной структурой данных (например, анализ рисков, классификация изображений, прогноз спроса).

[2] GenAI (generative AI) — разновидность моделей искусственного интеллекта, которая специализируется на решении разнородных задач и генерации нового контента (текст, изображения, аудио, видео) на основе анализа как структурированных, так и неструктурированных данных. Обучение осуществляется на разнородных неструктурированных и структурированных данных. Основная функция — имитация широкого набора функций человека, генерация новых данных и прогнозирование для данных новой структуры.

[3] Достоверность — соответствие информации фактам, отсутствие субъективизма, предвзятости или смысловых искажений.

СБЕР
КИБЕР
БЕЗОПАСНОСТЬ

Оглавление

Объекты, входящие в модель угроз

На рисунке приведена общая схема объектов воздействия нарушителем на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями, на которой отмечены потенциальные угрозы.

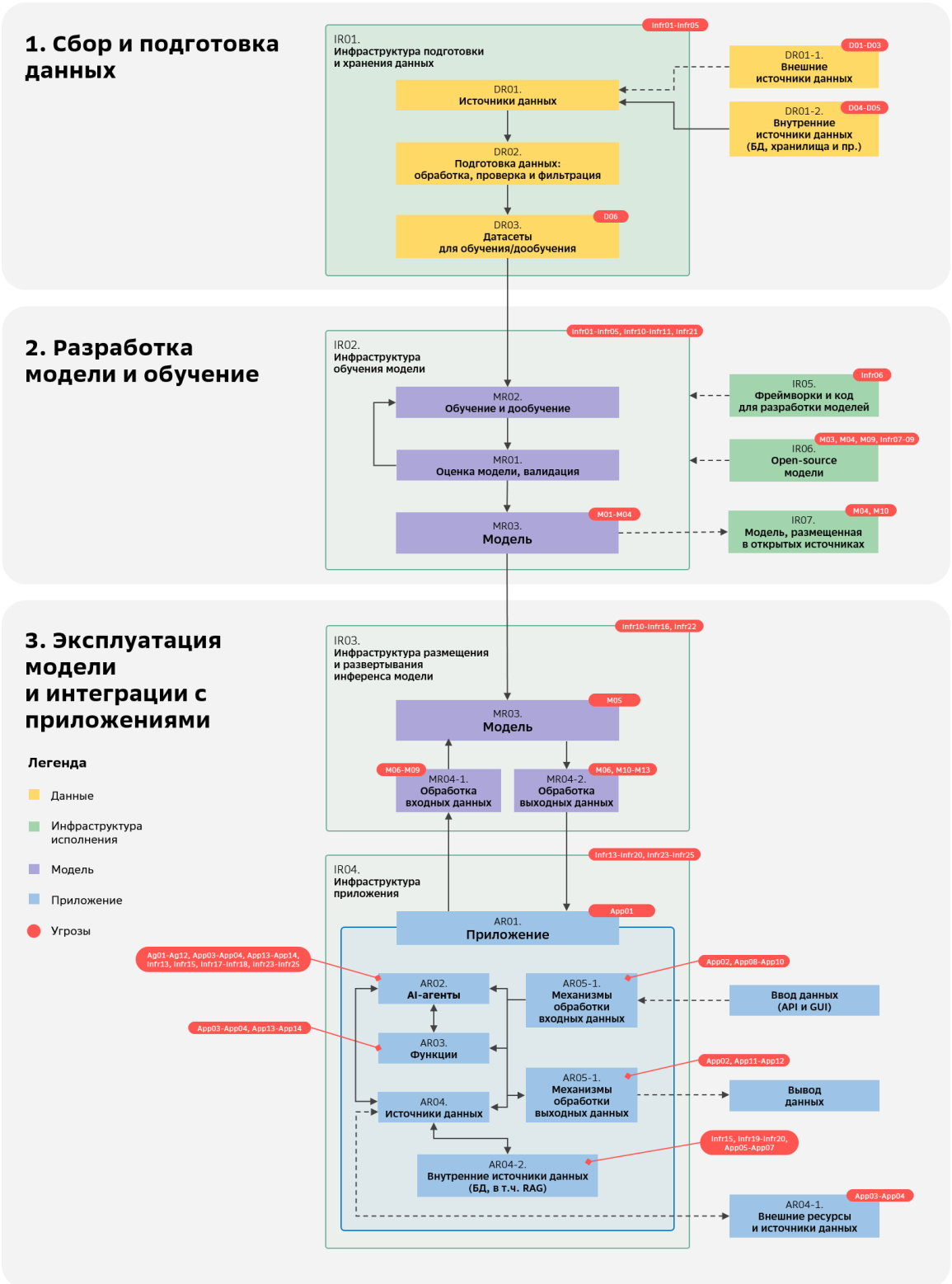


Рисунок 1 — Обобщенная схема объекта защиты и актуальных угроз на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями

Объекты на схеме:

1. Этап «Сбор и подготовка данных»

- IR01 Инфраструктура подготовки и хранения данных — совокупность программных и аппаратных средств, необходимых для сбора, обработки, хранения и управления данными на этапе подготовки к обучению моделей.
- DR01 Источники данных — данные из различных источников, используемые для обучения и дообучения моделей
    - DR01-1 Внешние источники — публичные датасеты, API, социальные сети, открытые базы данных.
    - DR01-2 Внутренние источники — корпоративные базы данных, логи пользовательского взаимодействия, исторические данные организации.
- DR02 Подготовка данных: обработка, проверка, фильтрация — действия по очистке, нормализации, преобразованию форматов и фильтрации данных для создания обучающих выборок.
- DR03 Датасеты для обучения и дообучения — наборы данных, прошедшие подготовку и используемые для обучения, тестирования и дообучения моделей.

2. Этап «Разработка модели и обучение»

- IR02 Инфраструктура обучения модели — совокупность программных и аппаратных средств, необходимых для обучения и оптимизации моделей.
- MR01 Оценка модели, валидация — действия по тестированию и валидации качества модели, включая ее безопасность.
- MR02 Обучение и дообучение — действия по применению алгоритмов машинного обучения для создания и адаптации моделей.
- MR03 Модель — итоговая модель, полученная в результате обучения.
- IR05 Фреймворки и код для разработки моделей — программные библиотеки и исходный код для создания моделей.
- IR06 Open-source модели — предобученные модели из открытых репозиториев.
- IR07 Модель, размещенная в открытых источниках — проприетарная модель, размещаемая в открытых репозиториях для использования сообществом.

3. Этап «Эксплуатация модели и интеграции с приложениями»

- IR03 Инфраструктура размещения и развертывания инференса модели — совокупность программных и аппаратных средств, необходимых для развертывания обученных моделей в production.
- MR04 Обработка входных и выходных данных — механизмы преобразования данных перед передачей в модель и после получения результатов.
    - MR04-1 Механизмы обработки входных данных — предобработка данных, поступающих на вход модели.
    - MR04-2 Механизмы обработки выходных данных — постобработка результатов модели, может включать форматирование, фильтрацию, в том числе нежелательного контента.
- IR04 Инфраструктура приложения — совокупность программных и аппаратных средств, необходимых для выполнения AI-приложений.
- AR01 Приложение — конечная система, использующая AI-модель для решения задач
    - AR02 AI-агенты — автономные системы, принимающие решения на основе выводов моделей.
    - AR03 Функции — внутренние функции, реализующие бизнес-логику приложения, или внешние по отношению к приложению функции, связанные с работой AI-модели.
    - AR04 Источники данных — данные, используемые приложением в реальном времени.

        - AR04-1 Внешние источники данных — веб-сайты, публичные базы данных, потоковые данные, поступающие из внешних источников.
        - AR04-2 Внутренние источники данных (БД, в т.ч. RAG) — локальные базы данных, векторные хранилища для RAG (Retrieval-Augmented Generation)
    - AR05 Обработка пользовательского ввода/вывода
        - AR05-1 Механизмы обработки входных данных — валидация, санитизация и предобработка пользовательских запросов.
        - AR05-2 Механизмы обработки выходных данных — форматирование ответов модели, фильтрация нежелательного контента.

Перечень угроз

Угрозы, связанные с данными

D01. Использование для обучения/дообучения модели отравленных данных или датасетов, загруженных из внешних источников

Описание
Загрузка, обучение или дообучение модели на вредоносных данных или наборах данных, в которые были внедрены вредоносные примеры, из внешних источников (например, репозиториев или страниц, с которых выполняется сбор данных)

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
DR01-1 Внешние источники данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D02. Использование для обучения/дообучения модели модифицированных данных или датасетов, загруженных из внешних источников

Описание
Использование скомпрометированного источника данных, обучение или дообучение модели на загруженных вредоносных данных или наборах данных, в которые были внедрены вредоносные примеры, из внешних источников, которые были модифицированы после добавления в перечень источников

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
DR01-1 Внешние источники данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D03. Воспроизведение в ответах модели персональных данных (ПДн), полученных из внешних источников

Описание
Загрузка, обучение или дообучение модели на данных или наборах данных, которые содержат ПДн из внешних источников

Последствия
Запоминание данных и непреднамеренное воспроизводство персональной информации, что создает риски нарушения ФЗ-152 и приватности пользователей

Объект воздействия
DR01-1 Внешние источники данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D04. Использование для обучения/дообучения модели отравленных данных или датасетов, загруженных из внутренних источников

Описание
Загрузка, обучение или дообучение модели на данных или наборах данных, в которые были внедрены вредоносные примеры из внутренних источников

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
DR01-2 Внутренние источники данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D05. Неконтролируемая загрузка данных, содержащих конфиденциальную информацию, в датасеты для обучения моделей

Описание
Загрузка и обучение модели на данных или наборах данных, которые содержат конфиденциальную информацию

Последствия
Запоминание данных и непреднамеренное воспроизводство конфиденциальной информации, что создает риски утечек и хищения конфиденциальной информации

Объект воздействия
DR01-2 Внутренние источники данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D06. Неконтролируемое использование, модификация, удаление данных для обучения или дообучения модели

Описание

Отсутствие надлежащего контроля и управления данными для обучения или дообучения модели, включая их классификацию, наличие метаданных, процессы управления жизненным циклом (включая хранение, удаление, архивирование и пр.)

Последствия

Увеличение поверхности атаки, утечки конфиденциальной информации, риски нарушения ФЗ-152

Объект воздействия

DR03 Датасеты для обучения/дообучения

Нарушаемое свойство

Конфиденциальность, целостность

Виды моделей, подверженных угрозе

PredAI и GenAI

Лица, ответственные за митигацию угрозы

Владелец данных

Угрозы, связанные с инфраструктурой

Infr01 Несанкционированная модификация реестра источников данных, датасетов

Описание
Использование скомпрометированного источника данных или датасетов вследствие несанкционированной модификации реестра источников

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr02 Несанкционированная модификация обучающих данных

Описание
Использование скомпрометированных данных или датасетов, используемых для обучения/дообучения модели, вследствие несанкционированной модификации

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr03 Небезопасная передача данных/датасетов между этапами подготовки

Описание
Перехват или модификация при передаче между этапами обработки, хранения данных или обучения модели

Последствия
Утечки данных, компрометация модели, внедрение вредоносных данных в датасеты или нарушение целостности данных

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr04 Кража обучающих данных

Описание
Хищение обучающих данных вследствие слабостей разграничения доступа

Последствия
Восстановление модели, создание теневой модели

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr05 Утечка конфиденциальной информации из наборов обучающих данных

Описание
Хищение конфиденциальной информации из наборов обучающих данных вследствие слабостей разграничения доступа

Последствия
Хищение конфиденциальной информации

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr06 Использование уязвимых версий сторонних библиотек, использование программного кода с закладками

Описание
Использование для разработки модели и обучения сторонних библиотек или программного кода с уязвимостями или программными закладками. Угроза обусловлена слабостями механизмов анализа библиотек и кода на наличие уязвимостей или отсутствием соответствующих проверок.

Последствия
Внедрение вредоносного кода

Объект воздействия
IR05 Фреймворки и код для разработки моделей

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr07 Использование open-source моделей, содержащих программные закладки в файлах

Описание
Возможность осуществления нарушителем деструктивного воздействия на систему путем внедрения программных закладок в файлах open-source моделей

Последствия
Внедрение вредоносного кода

Объект воздействия
IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr08 Использование open-source моделей, содержащих логические закладки, заложенные при обучении

Описание
Возможность осуществления нарушителем деструктивного воздействия на систему путём эксплуатации уязвимостей open-source моделей, содержащих логические закладки и бэкдоры

Последствия
Смещение результатов работы модели, снижение точности или эксплуатация бэкдоров в модели

Объект воздействия
IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr09 Использование open-source моделей, содержащих закладки в весах

Описание
Возможность осуществления нарушителем деструктивного воздействия на систему путём эксплуатации уязвимостей open-source моделей, содержащих закладки в весах модели

Последствия
Смещение результатов работы модели, снижение точности или эксплуатация бэкдоров в модели

Объект воздействия
IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr10 Подмена или модификация модели

Описание
Несанкционированная модификация или подмена модели вследствие слабостей разграничения доступа

Последствия
Смещение результатов работы модели, снижение точности или эксплуатация бэкдоров в модели

Объект воздействия
IR02 Инфраструктура обучения модели, IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры обучения модели; владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr11 Кража модели

Описание
Хищение модели вследствие слабостей разграничения доступа

Последствия
Создание теневой модели или восстановление обучающих данных из-за наличия white-box доступа к украденной модели

Объект воздействия
IR02 Инфраструктура обучения модели, IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры обучения модели; владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr12 Нарушение доступности модели

Описание
Атаки типа DDoS (Distributed Denial of Service), в том числе атаки на API, или эксплуатация уязвимостей инфраструктуры размещения и развертывания инференса модели или зависимых компонентов

Последствия
Прекращение или снижение скорости оказания услуг AI-сервисом всем потребителям (или группе потребителей) из-за нарушения доступности для них инфраструктуры размещения и развертывания инференса модели

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr13 Утечки конфиденциальной информации из систем логирования, в том числе логирования запросов и вызовов функций

Описание
Утечки информации из логов (журналов) запросов, содержащих ПДн и иную конфиденциальную информацию, в том числе об особенностях реализации AI-агентов и мультиагентных систем

Последствия
Хищение конфиденциальной информации

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели; владелец приложения

Infr14 Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности и инцидентов из-за отсутствия логирования взаимодействий

Описание
Отсутствие или неполнота данных логирования взаимодействий (включая запросы и ответы модели, данные телеметрии, вызовы функций и пр.) между моделью и интеграциями, AI-агентами или пользователями в ИТ-инфраструктуре размещения и развертывания инференса модели

Последствия
Увеличение времени или невозможность выявления, реагирования и расследования событий безопасности и инцидентов

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения

Нарушаемое свойство
Целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели; владелец приложения

Infr15 Перехват или подмена запросов или ответов модели или данных, передаваемых при взаимодействии с БД RAG

Описание
Перехват или подмена запросов или ответов модели путем реализации атаки man-in-the-middle (MiTM), вмешательство в процесс обмена данными между интеграциями и компонентами AI-агентами или пользователями и моделью, что позволяет выполнить перехват, изменение или подмену входных данных, отправляемых модели, или изменение выходных данных, а также ответов модели

Последствия
Перехват запросов и ответов, содержащих ПДн или конфиденциальную информацию, модификация запросов для получения некорректных предсказаний, или подмена ответов модели для введения в заблуждение. Это может привести к утечке конфиденциальной информации, компрометации системы или принятию ошибочных решений на основе искаженных данных

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения, AR02 AI-агенты, AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr16 Несанкционированное отключение или модификация механизмов фильтрации или контроля входных и выходных данных

Описание
Отключение или изменение систем, предназначенных для проверки и очистки данных, поступающих в модель или возвращаемых потребителю

Последствия
Обработка вредоносных или некорректных входных данных, сбои в работе модели, возвращение потребителю нежелательного, опасного или конфиденциального контента

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения

Нарушаемое свойство
Целостность, конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr17 Хищение системного промпта

Описание
Получение доступа к системным промптам, которые управляют поведением модели, приложения или AI-агента, с целью их кражи для последующего анализа, применения для копирования особенностей функционирования модели, приложения или AI-агента

Последствия
Раскрытие конфиденциальной информации, облегчение проведения промпт-атак, утрата конкурентных преимуществ

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения

Infr18 Несанкционированная модификация системного промпта

Описание
Получение доступа к системным промптам, которые управляют поведением модели, приложения или AI-агента, с целью их модификации для манипуляции выводом модели, функционированием приложения или AI-агента

Последствия
Нарушение функциональности модели, приложения или AI-агента, в том числе некорректные, вредоносные или нежелательные генерации, или раскрытие конфиденциальной информации

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения

Infr19 Несанкционированная модификация данных во внутренних источниках данных (в т.ч. в БД RAG)

Описание
Доступ к внутренним хранилищам и изменение данных, используемых при работе модели, для внедрения вредоносной информации, включая непрямые промпт-инъекции или некорректную/неэтичную информацию

Последствия
Искажение результатов работы модели, нарушение целостности ответов, некорректные или вредоносные ответы, генерации некорректных инструкций, нарушение принятия решений

Объект воздействия
IR04 Инфраструктура приложения, AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец ИТ-инфраструктуры хранения данных

Infr20. Утечки информации из внутренних источников данных

Описание
Несанкционированное копирование, передача или раскрытие конфиденциальной информации или ПДн, хранящихся в базах данных, файлах или других внутренних источниках данных

Последствия
Утечка конфиденциальной информации, финансовые потери, репутационный ущерб и юридические последствия

Объект воздействия
IR04 Инфраструктура приложения, AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец ИТ-инфраструктуры хранения данных

Infr21. Несанкционированная модификация тестовых и валидационных датасетов

Описание
Несанкционированное изменение тестовых и валидационных датасетов, добавление или удаление данных для искажения данных о качестве работы модели

Последствия
Внедрение ненадежной модели в эксплуатацию, утечки данных, некорректные предсказания и генерации

Объект воздействия
IR02 Инфраструктура обучения модели

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры обучения модели

Infr22. Несанкционированные подключения к модели

Описание
Подключение и обращение к модели с использованием скомпрометированных учетных данных или через системы-посредники, проксирующие системы или шлюзы с использованием их учетных данных

Последствия
Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности, DoW[4] системы-посредника

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

Infr23. Утечки данных AI-агента или информации об особенностях его реализации

Описание
Несанкционированное копирование, передача или раскрытие информации, связанной с AI-агентом, в том числе о его цели, описании функций, инструкциях механизма планирования, содержимом памяти.

Последствия
Нарушение конфиденциальности цели, описания функций, содержимого памяти или инструкций механизма планирования AI-агента

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец приложения

Infr24. Несанкционированная модификация AI-агента

Описание
Несанкционированная модификация AI-агента через добавление промпт-атак в цель, описание функций, память или механизм планирования AI-агента, добавление вредоносных функций в список доступных функций, добавление в список доступных ресурсов AI-агента записей, содержащих ресурсы с промпт-атаками, добавление промпт-атак в память AI-агента

Последствия
Нарушение функциональности AI-агента или приложения его реализующего

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

---

[4] DoW (Denial of wallet) — отказ в обслуживании вследствие исчерпания лимитов на затраты по потреблению ресурсов модели.

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец приложения

Infr25. Утечка информации об архитектуре мультиагентной системы через интерфейсы инструментов разработки или взаимодействия пользователя с AI-агентом или мультиагентной системой

Описание
Извлечение информации о мультиагентной системе (включая ее архитектуру, состав, правила взаимодействия) из интерфейсов инструментов разработки или взаимодействия пользователя с AI-агентом или мультиагентной системой

Последствия
Нарушение конфиденциальности состава мультиагентной системы и устройства взаимодействия между AI-агентами в мультиагентной системой, облегчение проведения кибератак

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Угрозы, связанные с моделью

M01. Невозможность реагирования и расследования событий безопасности и инцидентов из-за отсутствия информации о данных, на которых выполнено обучение модели

Описание
Отсутствие документации и прозрачности в отношении данных, использованных для обучения модели. Угроза создает риски безопасности и соответствия нормам законодательства РФ и наличия потенциальных смещений (bias) или уязвимостей к специфичным атакам (состязательным и промпт-атакам)

Последствия
Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности

Объект воздействия
MR03 Модель

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

M02. Использование модели с высокой уязвимостью к состязательным атакам (в том числе промпт-атакам)

Описание
Недостаточная подготовка или валидация модели, приводящая к возможности осуществления нарушителем деструктивного воздействия на систему путём эксплуатации уязвимостей модели. Обусловлена слабостями самой модели и слабостями механизмов обработки входных данных. Реализуется при использовании специально подобранных входных данных (например, изображений, текстов или запросов) для эксплуатации уязвимостей модели

Последствия
Некорректное или недекларированное поведение модели, утечки конфиденциальной информации

Объект воздействия
MR03 Модель

IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

M03. Нежелательное поведение, вредоносные генерации, галлюцинации

Описание
Недостаточная подготовка или валидация модели, приводящая к некорректному или недекларированному поведению модели, вредоносным генерациям или галлюцинациям. Выявленные злоумышленником галлюцинации (наименования программных пакетов, URL-адреса, имена организаций или электронные адреса) могут быть использованы для создания вредоносных целей, таких как поддельные программные пакеты, фишинговые сайты или фальшивые организации, которые затем публикуются и распространяются

Последствия
Некорректное или недекларированное поведение модели

Объект воздействия
MR03 Модель

Нарушаемое свойство
Целостность, достоверность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

M04. Подбор атак с использованием знаний уровня white-box об open-source модели

Описание
Подбор атак на целевую модель с использованием знаний уровня white-box (полный доступ к архитектуре, параметрам и обучающим данным) об уязвимости open-source моделей (моделей сторонних разработчиков и проприетарных моделей при их размещении в открытом доступе)

Последствия
Уязвимость модели к состязательным атакам и промпт-атакам

Объект воздействия
MR03 Модель, IR06 Open-source модели, IR07 Модель, размещенная в открытых источниках

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

M05. Отсутствие информации об инференсах модели

Описание
Отсутствие или неактуальность информации о состоянии инференсов моделей

Последствия
Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности, уязвимость интеграций к промпт-атакам

Объект воздействия
MR03 Модель

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M06. Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели

Описание
Обнаружение способов обхода или нарушения работы механизмов обработки входных или выходных данных, включая механизмы санитизации, валидации и фильтрации, что позволяет внедрять вредоносные данные, манипулировать результатами или получать несанкционированный доступ к информации. Угроза связана с отсутствием или недостаточностью механизмов обработки входных/выходных данных, реализованных на уровне модели

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний

Объект воздействия
MR04-1 Механизмы обработки входных данных, MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M07. Нарушение доступности модели (DoS) из-за отсутствия единого контроля запросов на уровне модели

Описание
Использование специально подобранных входных данных (например, вычислительно сложных) или отправка большого количества запросов, специально созданных для максимальной нагрузки на среду развертывания инференса модели

Последствия
Прекращение или снижение скорости оказания услуг AI-сервисом всем потребителям (или группе потребителей) или увеличение затрат на эксплуатацию из-за перегрузки AI-сервиса или модели запросами

Объект воздействия
MR04-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M08. Исчерпание лимитов интеграции (DoW) из-за отсутствия единого контроля запросов на уровне модели

Описание
Направление чрезмерного количества запросов от интеграции к модели, что приводит к повышенному использованию токенов и лимитов

Последствия
Прекращение работы интеграции или увеличение затрат на эксплуатацию

Объект воздействия
MR04-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M09. Обход встроенных защитных механизмов модели в том числе с использованием методов состязательных атак и промпт-атак

Описание
Обход встроенных защитных механизмов модели или механизмов, основанных на технологиях искусственного интеллекта

Последствия
Некорректное или недекларированное поведение модели

Объект воздействия
MR04-1 Механизмы обработки входных данных

IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

Владелец инфраструктуры размещения, подключения и предоставления контента

M10. Утечка информации о модели

Описание
Использование различных методов (например, повторяющихся запросов или анализа документации и файлов модели, размещенных в открытых источниках) для получения информации о модели, её онтологии, семействе модели, границах принимаемых решений, а также о связанных артефактах, таких как данные, программное обеспечение и инфраструктура

Последствия
Утечка конфиденциальной информации, облегчение проведения состязательных и промпт-атак

Объект воздействия
MR04-2 Механизмы обработки выходных данных, IR07 Модель, размещенная в открытых источниках

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M11. Утечки конфиденциальной информации из дообученной модели или LoRA

Описание
Использование специально подобранных входных данных (например, с техниками джейлбрейков) для извлечения конфиденциальной информации из дообученной модели или LoRA

Последствия
Утечка конфиденциальной информации

Объект воздействия
MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели; владелец инфраструктуры размещения, подключения и предоставления контента

M12. Эксфильтрация, инверсия или реверс-инжиниринг модели

Описание
Отправка многократных запросов к модели через API доступа к модели, анализ и изучение ответов для создания функциональной копии модели, воссоздающей поведение и функциональность целевой модели без прямого доступа к её архитектуре или обучающим данным, реализация атак инверсии (Invert ML Model) или извлечения (Extract ML Model) модели

Последствия
Кража модели, создание теневой модели

Объект воздействия
MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M13. Эксфильтрация данных

Описание
Отправка многократных запросов к модели через API доступа к модели, анализ ответов для извлечения конфиденциальной информации или обучающих данных, реализация атак на определение принадлежности данных (Infer Training Data Membership)

Последствия

Восстановление фрагментов обучающих данных, которые могут включать ПДн или конфиденциальную информацию

Объект воздействия

MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство

Конфиденциальность

Виды моделей, подверженных угрозе

PredAI и GenAI

Лица, ответственные за митигацию угрозы

Владелец инфраструктуры размещения, подключения и предоставления контента

Угрозы, связанные с приложениями

App01. Ошибки в проектировании, использование небезопасных интеграций компонентов

Описание
Использование небезопасных интеграций функциональных компонентов (например, AI-агентов, функций, плагинов) приложений, в том числе отсутствие проверки всех входных/выходных данных, использование небезопасных API, отсутствие шифрования данных в процессе передачи, некорректная настройка прав доступа

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR01 Приложение

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App02. Обход механизмов обработки входных/выходных данных, реализуемых на уровне приложения

Описание
Обнаружение способов обхода или нарушения работы механизмов обработки входных или выходных данных, включая механизмы санитизации, валидации и фильтрации. Это позволяет внедрять вредоносные данные, манипулировать результатами или получать несанкционированный доступ к информации. Угроза связана с отсутствием или недостаточностью механизмов обработки входных/выходных данных, реализованных в приложении

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR05-1 Механизмы обработки входных данных, AR05-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App03. Загрузка вредоносного программного обеспечения (ВПО) из внешних источников (Интернет)

Описание
Использование специально созданных источников данных (в том числе баз данных, репозиториев кода или веб-сайтов), на которых размещается вредоносный код. Возможна реализация в условиях наличия функционала поиска по внешним источникам и выполнения кода

Последствия
Выполнение вредоносной нагрузки, которая может повлечь дальнейшие нарушения безопасности

Объект воздействия
AR04-1 Внешние источники данных, AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App04. Загрузка отравленных данных из внешних источников (Интернет)

Описание
Использование специально созданных источников данных (в том числе баз данных или веб-сайтов), на которых размещены данные (файлы, текст, мультимедиа), содержащие непрямые промпт-инъекции. Возможна реализация в условиях наличия функционала поиска по внешним источникам

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR04-1 Внешние источники данных, AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App05. Внедрение непрямых промпт-инъекций во внутренние источники (в т.ч. БД RAG)

Описание
Использование специально подобранных входных данных для внедрения непрямых промпт-инъекций во внутренние источники. Возможна реализация в условиях сохранения результатов обработки данных моделью во внутренние источники

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App06. Утечки информации из внутренних источников (в т.ч. БД RAG)

Описание
Использование специально подобранных входных данных (например, с техниками джейлбрейков) для извлечения конфиденциальной информации из внутренних источников (в т.ч. БД RAG)

Последствия
Утечка конфиденциальной информации

Объект воздействия
AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App07. Выполнение вредоносных инструкций, созданных моделью

Описание
Использование специально подобранных входных данных для формирования вредоносных инструкций, направляемых во внутренние источники (в т.ч. в БД SQL). Возможна реализация в условиях формирования моделью запросов ко внутренним источникам

Последствия
Несанкционированная модификация или уничтожение информации во внутренних источниках, утечка конфиденциальной информации

Объект воздействия
AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App08. Реализация прямых промпт-инъекций из-за отсутствия контроля входных данных

Описание
Использование специально подобранных входных данных для проведения промпт-атак

Последствия
Некорректное или недекларированное поведение модели, утечка конфиденциальной информации

Объект воздействия
AR05-1 Механизмы обработки входных данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App09. Нарушение доступности (DoS/DoW) интеграции

Описание
Использование специально подобранных входных данных или чрезмерного количества запросов от интеграции к модели, что приводит к DoS интеграции или повышенному использованию токенов и лимитов DoW

Последствия
Прекращение или снижение скорости оказания услуг интеграции с AI-сервисом или увеличение затрат на эксплуатацию интеграции

Объект воздействия
AR05-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App10. Нарушение логики выполнения задачи из-за отсутствия контроля входных данных

Описание
Использование специально подобранных входных данных (состязательных атак или промпт-атак), чтобы обойти инструкции, заложенные в системном промпте, ограничения или иные системные настройки

Последствия
Изменение логики функционирования, выполнение недекларированных действий, которые были явно запрещены или не предполагались при разработке

Объект воздействия
AR05-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App11. Утечка информации о системном промпте из-за некорректной обработки выходных данных

Описание
Использование специально подобранных входных данных для получения информации об используемом системном промпте

Последствия
Утечка конфиденциальной информации, облегчение проведения промпт атак

Объект воздействия
AR05-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App12. Токсичная или вредоносная генерация из-за некорректной обработки выходных данных

Описание
Использование специально подобранных входных данных для получения токсичной генерации (противоречащей законодательным и морально-этическим нормам) или вредоносной генерации (опасные инструкции, инструкции по подготовке киберпреступлений, генерации вредоносного кода, уязвимого кода)

Последствия
Некорректное или недекларированное поведение приложение вследствие генераций модели

Объект воздействия
AR05-2 Механизмы обработки выходных данных

Нарушаемое свойство
Целостность, достоверность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App13. Вывод информации о среде

Описание
Использование специально подобранных входных данных для получения информации о конфигурации системы, внутренних процессах, API-ключах, версиях программного обеспечения или других деталях, которые могут быть использованы для дальнейших атак

Последствия
Утечка конфиденциальной информации, облегчение проведения кибератак

Объект воздействия
AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App14. Автоматическое распространение вредоносной инструкции на другие приложения

Описание
Использование специально подобранных входных данных для реализации самовоспроизводящихся промпт-атак. Угроза может быть реализована в случае сохранения результатов обработки данных во внутренние источники (в том числе на базе RAG) или при передаче результатов обработки данных в другие приложения

Последствия
Реализации кибератак на другие приложения, выполнение вредоносной нагрузки (вредоносные запросы распространяются на другие модели или приложения)

Объект воздействия
AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Угрозы, связанные с AI-агентами

Ag01. Ошибки в проектировании AI-агентов и мультиагентных систем

Описание
Использование небезопасных интеграций AI-агентов, функций и плагинов, используемых AI-агентами, в том числе отсутствие проверки всех входных/выходных/промежуточных шагов выполнения, некорректная настройка прав доступа

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности выполнения задач, некорректное или недекларированное поведение AI-агентов или мультиагентных систем

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag02. Вредоносные генерации в ответе AI-агента на запрос пользователя

Описание
Использование специально подобранных входных данных для генерации токсичного или вредоносного ответа

Последствия
Некорректное или недекларированное поведение AI-агента вследствие генераций модели, риск репутационного ущерба

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Достоверность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag03. Отправка информации из среды исполнения функций AI-агента (действий) на внешние ресурсы

Описание
Использование специально подобранных входных данных для вызова функций для размещения на внешних ресурсах конфиденциальной информации (в том числе значений переменных среды)

Последствия
Утечка конфиденциальных данных, содержащихся в среде исполнения, облегчение проведения кибератак

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag04. Удаление или модификация файлов в среде исполнения функций AI-агента (действий)

Описание
Использование специально подобранных входных данных для вызова функций, удаляющих или модифицирующих файлы, доступные в среде исполнения

Последствия
Нарушение целостности данных, содержащихся в среде исполнения

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag05. Размещение в среде исполнения функций AI-агента (действий) файлов с ВПО, полученных с внешних ресурсов

Описание
Использование специально подобранных входных данных для вызова функций (загружающих из внешних источников файлы ВПО и исполняющих его), подмена легитимных ресурсов или внедрение ВПО во внешние ресурсы, используемые AI-агентами

Последствия
Выполнение вредоносной нагрузки, которая может повлечь дальнейшие нарушения безопасности

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag06. Нарушение доступности (DoS/DoW) среды исполнения AI-агента (в т.ч. функций)

Описание
Использование специально подобранных входных данных для исполнения функций, потребляющих избыточное количество ресурсов среды исполнения

Последствия
Прекращение или снижение скорости оказания услуг AI-агентом/приложением или увеличение затрат на эксплуатацию

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag07. Утечка информации об архитектуре мультиагентной системы через интерфейсы пользовательского ввода

Описание
Использование специально подобранных входных данных для извлечения информации о мультиагентной системе (включая ее архитектуру, состав, правила взаимодействия) из ответов AI-агента

Последствия
Нарушение конфиденциальности состава мультиагентной системы и устройства взаимодействия между AI-агентами в мультиагентной системе, облегчение проведения кибератак

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag08. Передача другому AI-агенту ложной информации в мультиагентной системе

Описание
Использование специально подобранных входных данных для модификации и искажения семантической составляющий информации, передаваемой AI-агентом в рамках мультиагентной системы

Последствия
Нарушение взаимодействия и кооперации AI-агентов, нарушение рабочего процесса приложения или мультиагентной системы

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag09. Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе

Описание
Использование специально подобранных входных данных для модификации цели другого AI-агента в мультиагентной системе путем передачи запроса с промпт-атакой

Последствия
Нарушение взаимодействия и кооперации AI-агентов, нарушение рабочего процесса приложения или мультиагентной системы

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag10. Распространение промпт-атаки по AI-агентам в мультиагентной системе для усиления ее эффекта

Описание
Использование специально подобранных входных данных для распространения промпт-атак в мультиагентной системе

Последствия
Распространения в мультиагентной системе промпт-атаки, имеющей целью возникновение различных негативных последствий

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag11. Нарушение рабочего процесса приложения, реализующей AI-агента

Описание
Использование специально подобранных входных данных для модификации и искажения семантической составляющей информации, структуры или формата её представления AI-агентом

Последствия
Искажение результатов работы, нарушение доступности приложения, реализующей AI-агента

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостности, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag12. Утечка информации о цели, функциях, содержимом памяти или инструкциях механизма планирования AI-агента

Описание
Использование специально подобранных входных данных для получения информации, связанной с AI-агентом, в том числе о его цели, описании функций, инструкциях механизма планирования, содержимого памяти

Последствия
Нарушение конфиденциальности цели, описания функций, содержимого памяти или инструкций механизма планирования AI-агента

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Материалы, использованные при подготовке

1. OWASP Top-10 LLM 2025
2. OWASP Top-10 Machine Learning Security
3. OWASP AI Security Solutions Landscape
4. OWASP Agentic Threats Taxonomy (draft)
5. OWASP AI Exchange 4.5
6. MITRE ATT&CK
7. MITRE ATLAS
8. Google SAIF (Secure AI Framework)
9. NIST Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
10. AWS Generative AI Security Scoping Matrix