

Использование искусственного интеллекта в кибербезопасности

AI в кибербезопасности: от аналитики SOC к адаптивной киберзащите



40-minute professional talk · SOC / SIEM / IR / Threat Intelligence / AI leadership

Core line: AI as decision-support layer, not replacement for analyst or standalone product.

AI for Defense

AI for Offense

Security of AI

Human-in-the-Loop

Adaptive Deception

Controlled Autonomy

Сдвиг: от Analytics к Action

Важен не сам переход от rule-based к LLM. Важен рост влияния AI на цикл защиты.



Key question: где AI получает право влиять на решение?

Risk question: что произойдёт при ошибке?

Три одновременных эффекта AI

AI нельзя рассматривать только как инструмент защитника. Он меняет всю поверхность конфликта.

AI for Defense

AI для защиты

- Detection & triage
- Enrichment
- Prioritization
- Recommendation
- Controlled response

AI for Offense

AI для атаки

- Reconnaissance
- Vulnerability analysis
- Planning
- Tool orchestration
- Social engineering

Security of AI

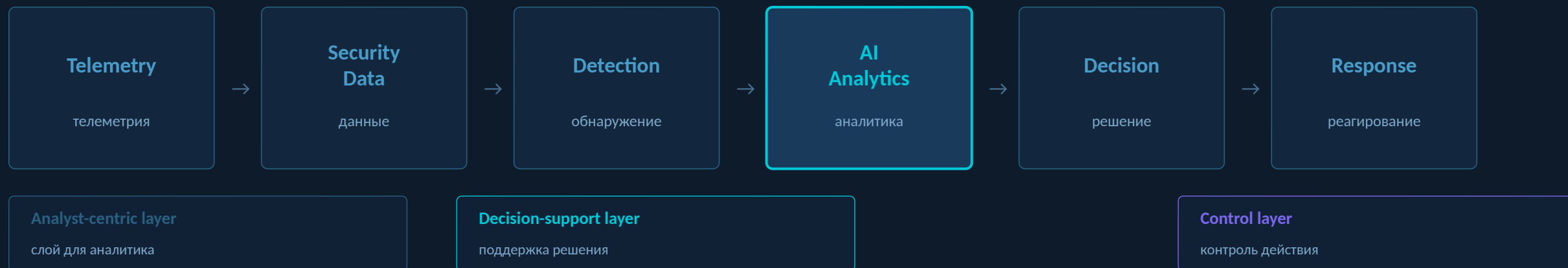
безопасность AI

- Prompt injection
- Data poisoning
- Excessive agency
- Supply chain risk
- Hallucination

Goal: Secure · Measurable · Controlled AI-enabled cybersecurity

Где AI находится внутри SOC

AI должен усиливать цепочку принятия решений, а не заменять SIEM, SOAR или аналитика.



The real question: what decision becomes better, faster, or safer?

✓ Baseline (базовый уровень)

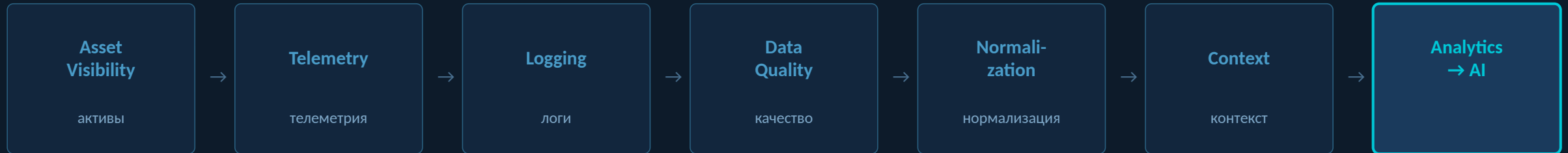
✓ Acceptable error (допустимая ошибка)

✓ Metric (метрика)

✓ Control & human oversight

Data Before AI

AI-enabled SOC начинается не с LLM, а с видимости активов, логирования и качества данных.



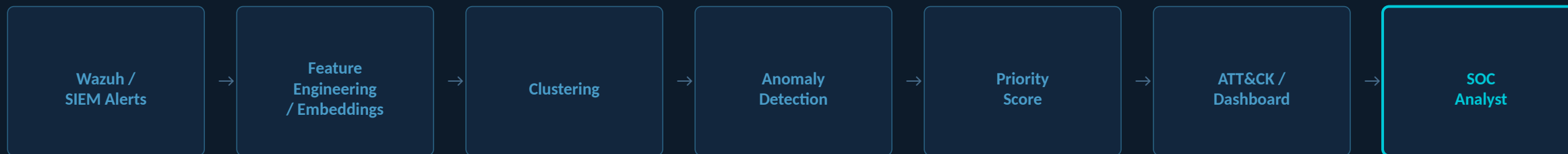
Garbage in → AI → Sophisticated garbage out

Collect by value, not by volume



Scientific Case #1: Unsupervised Analytics для SOC

Из потока Wazuh/SIEM alerts — в кластеры, аномалии и приоритеты для аналитика.



Clustering

- HDBSCAN
- DBSCAN
- KMeans
- GMM

Anomaly Detection

- Isolation Forest
- LOF
- One-Class SVM

Dataset:

30,000 events · 3 × 10,000 alerts

Evaluation:

Repeated 10-fold CV · 10 random seeds

Operational value: alert prioritization, not "AI magic"

Как читать результат модели глазами SOC

Accuracy недостаточно. Для SOC важны Precision, Recall, F1 и стоимость ошибки.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Из всех alerts, названных опасными —
сколько действительно опасны

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Из всех реальных угроз —
сколько мы не пропустили

$$\text{F1} = 2 \times \text{P} \times \text{R} / (\text{P} + \text{R})$$

Баланс между Precision и Recall

False Positive (FP)

→ Alert fatigue · Analyst workload ↑

False Negative (FN)

→ Missed attack · Incident risk ↑

Isolation Forest — synthetic threat injection

	Precision	Recall	F1
Dataset 1	0.980	1.000	0.990
Dataset 2	0.967	1.000	0.983
Dataset 3	0.978	1.000	0.989

High Recall without acceptable Precision = alert fatigue

Detection недостаточно: нужна приоритизация

Operational value появляется, когда anomaly превращается в очередь решений.

$$\text{Priority}_i = A_i(\text{IF}) \times \frac{1}{|C_i|} \times W_i(\text{SOC})$$

$A_i(\text{IF})$: anomaly score · C_i : cluster size · $W_i(\text{SOC})$: SOC contextual weight

$A_i(\text{IF})$

Anomaly Severity

степень аномальности
насколько событие необычно по модели

$\frac{1}{|C_i|}$

Cluster Rarity

редкость кластера
маленький/редкий кластер → выше приоритет

$W_i(\text{SOC})$

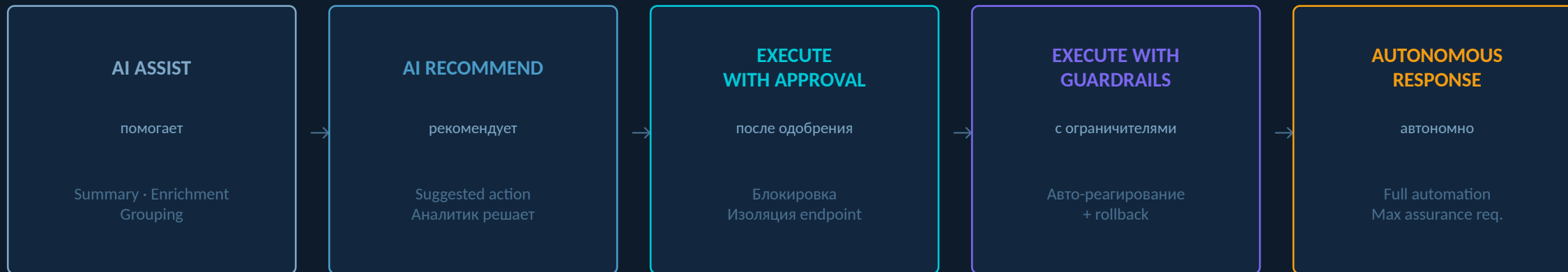
SOC Context

бизнес-критичность актива
ATT&CK mapping · threat context

Goal: fewer blind alerts · more defensible triage decisions

Human-in-the-Loop: не лозунг, а контроль риска

Уровень human oversight должен зависеть от автономности решения и стоимости ошибки.



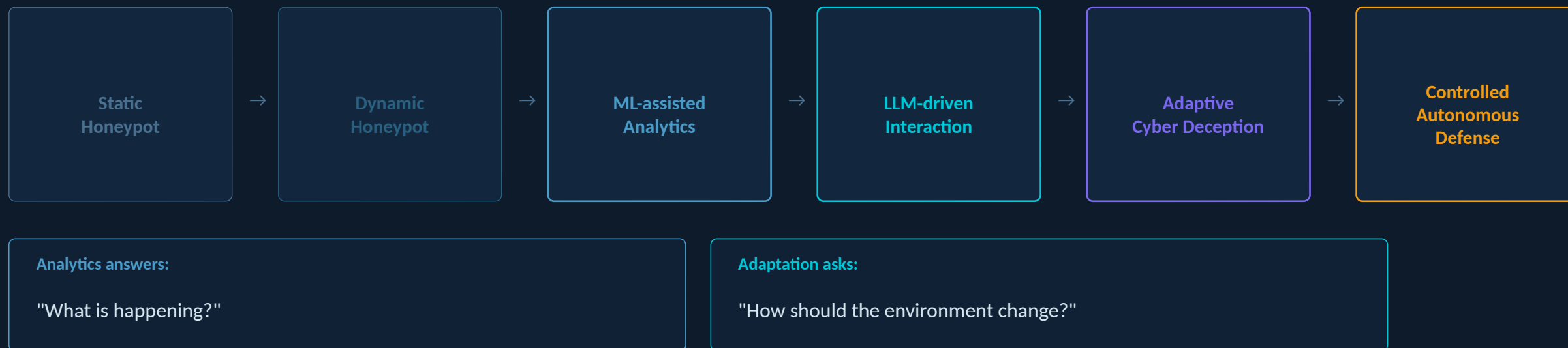
Autonomy ↑

Required Assurance ↑

Potential Impact of Error ↑

Следующий шаг: от Detection к Adaptation

AI начинает участвовать не только в анализе, но и в изменении защитной среды.



Key: CONTROLLED autonomy — observation → intelligence → adaptation → limited action

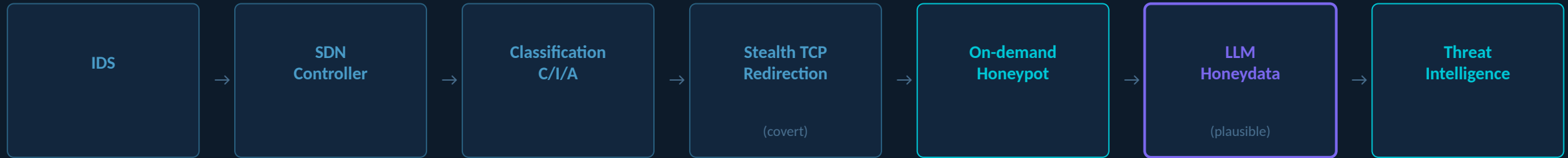
HoneyGPT: LLM как часть киберобмана

LLM повышает гибкость и глубину взаимодействия honeypot, но приносит новые ограничения.



Scientific Case #2: Reactive Cyber Deception

AI становится частью defensive control loop: обнаружение → перенаправление → deception → intelligence.



◀ defensive feedback loop ▶

Measured: Stealth

- Latency
- Throughput
- TCP retransmissions

Measured: Realism

- Format consistency
- Contextual coherence
- Statistical similarity

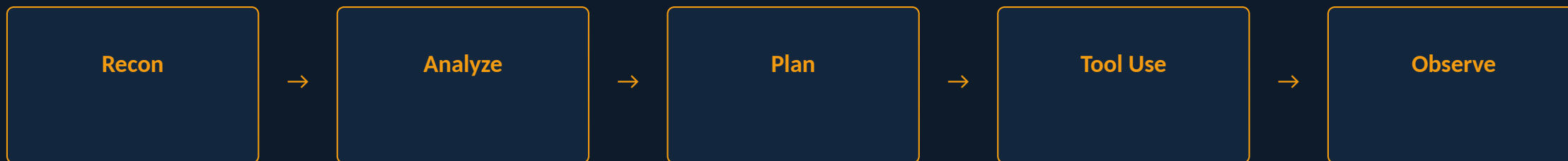
Measured: Limits

- IDS dependency
- TCP focus
- Compute overhead / scaling

AI as analytics layer → AI as adaptive defense component

AI for Offense: автономность сокращает attack loop

Главный риск — не "AI стал хакером", а снижение стоимости и времени итераций атаки.



 Re-planning

Cost compression

снижение стоимости итераций атаки

Scale compression

масштабирование операций

Skill compression

снижение порога входа

Capability depends on: model · tools · permissions · sandboxing · prompt quality · environment.
Defense implication: build defenses resilient to agentic behavior — not just known signatures.

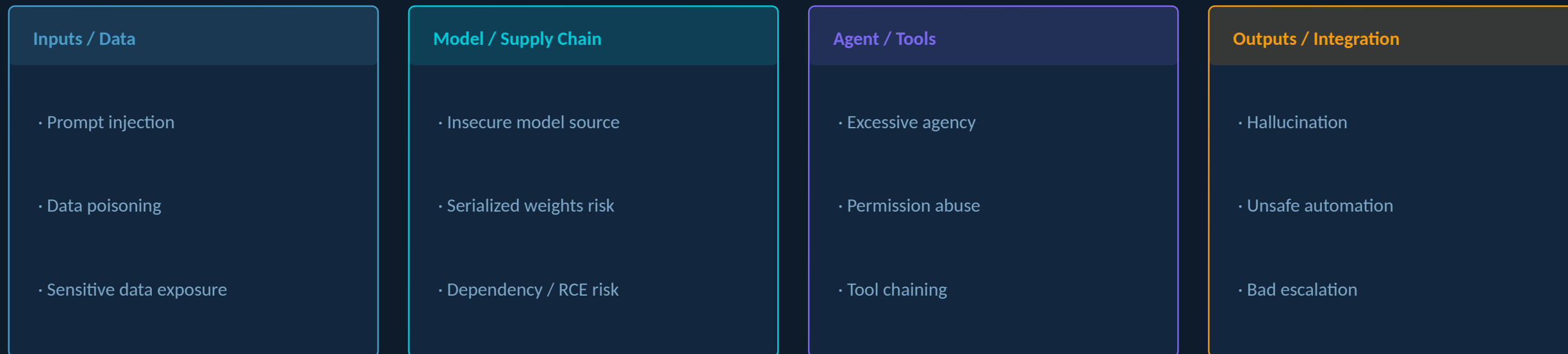
Mantis: Prompt Injection как защитный механизм?

Уязвимость LLM может стать defensive primitive против LLM-driven attacker — только в контролируемом контексте.



Capability растёт — Attack Surface тоже

Чем больше AI получает данных, контекста и tools, тем больше его нужно защищать.

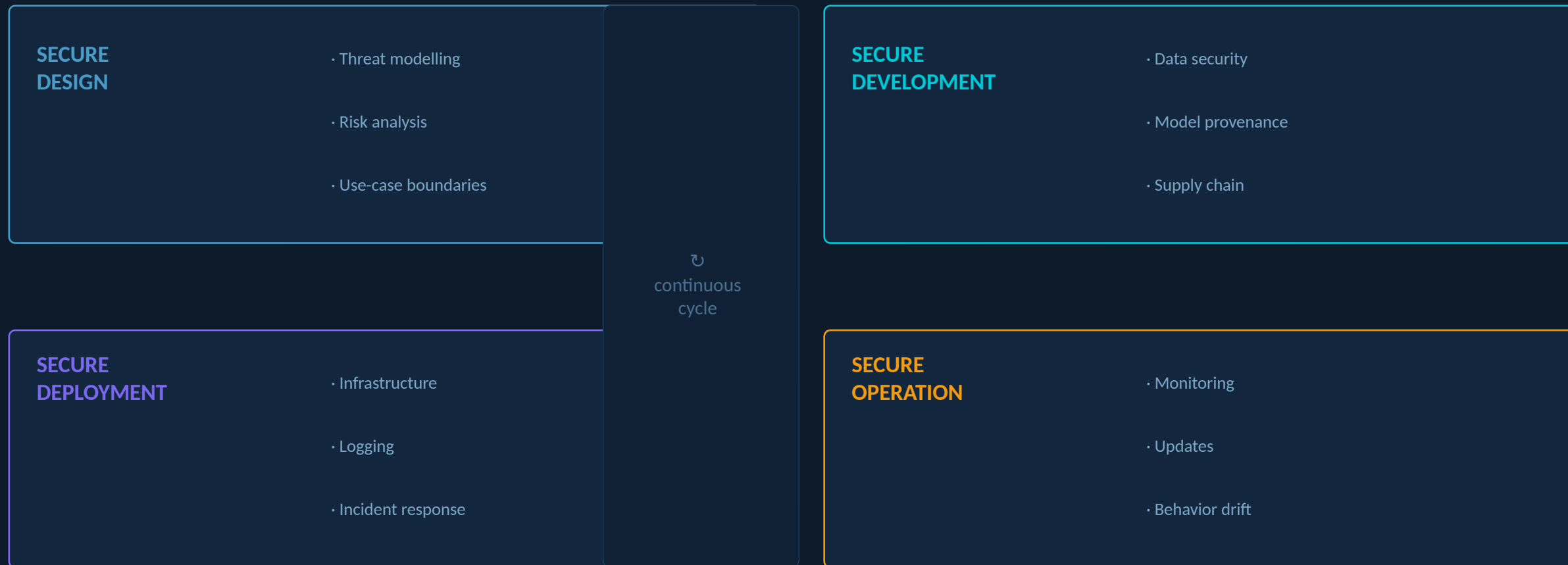


Security question: what can the model read · decide · call · change?

AI in SOC increases capability — but also increases blast radius if control is weak.

Secure AI Lifecycle

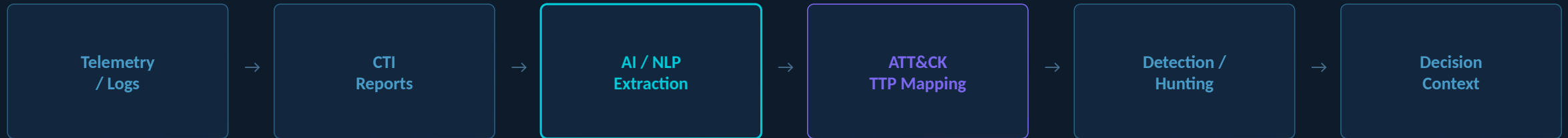
Secure AI — это не prompt filtering, а безопасность всего жизненного цикла.



Secure by design: security is a primary requirement, not a post-deployment patch.
Provider responsibility · transparency · accountability · secure leadership

MITRE ATT&CK: context layer, not ground truth

AI может ускорять TTP mapping, но mapping остаётся неоднозначной аналитической задачей.



Systematic review: 417 papers

Cyber Threat Intelligence (CTI)	30.2%
Threat Hunting	17.5%
Threat Modeling	10.8%

Limitations of ATT&CK mapping

- Mapping can be subjective and resource-intensive
- Requires continuous update
- Poor transfer to ICS / IoT domains
- Depends on quality of source data

AI accelerates TTP extraction — but final interpretation must remain verifiable and contextual.

Research Frontier: куда движется AI Cybersecurity

Нужно разделять current research и emerging direction — это не одно и то же.

Current Research

- ML anomaly detection for SOC
- LLM-driven honeypots
- ATT&CK TTP mapping via NLP
- Secure AI lifecycle frameworks
- Prompt-injection defense research

Emerging Direction

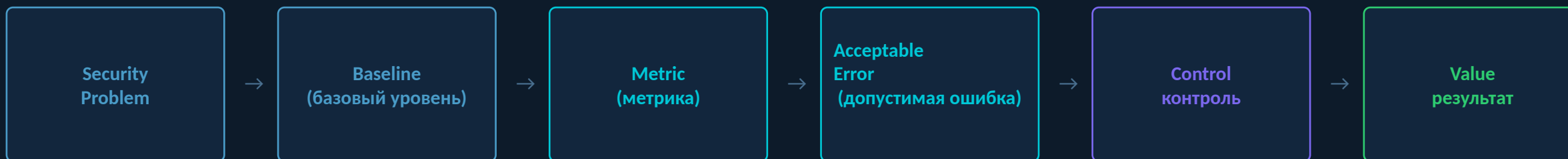
- Agentic SOC automation
- AI-assisted threat hunting
- Adaptive / reactive cyber deception
- Continual + online learning
- Controlled autonomous response

Speaker research area:

Honeypot telemetry → AI analysis → Threat Intelligence → ATT&CK mapping → Adaptive cyber deception → SOC

Как внедрять AI без маркетинга

Начинать нужно не с модели, а с проверяемой инженерной гипотезы.



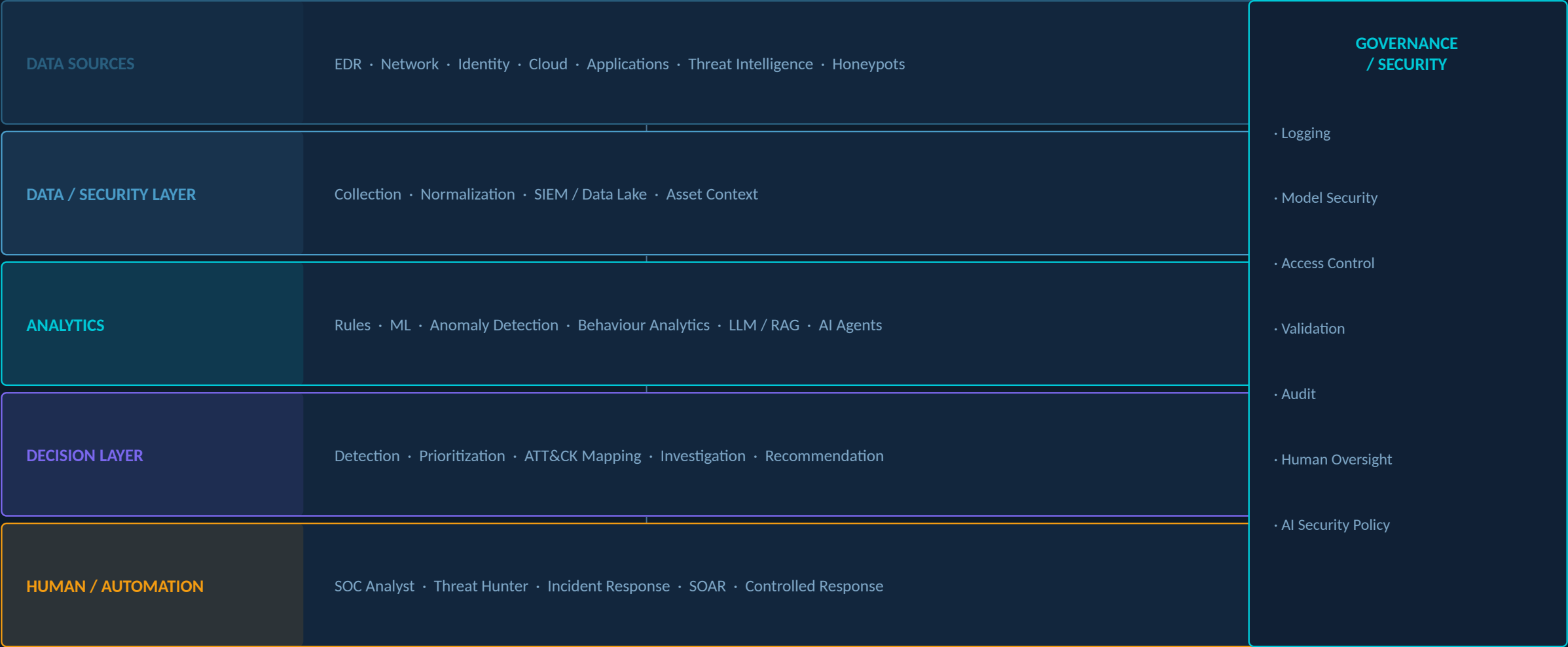
Example: Reduce alert fatigue without increasing missed incidents

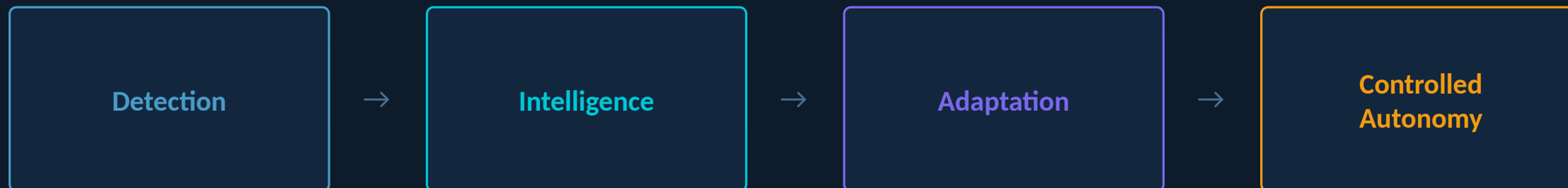
Measure: MTTT · FP rate · FN risk · analyst time · escalation quality

Decide: Where human approval is mandatory (blocking, isolation, legal impact)

No baseline → No metric → No defensible AI deployment

AI-Enabled SOC Reference Architecture





AI-Enabled SOC =

High-quality telemetry

+ Machine-scale analytics

+ Human context

+ Controlled automation

+ Secure AI architecture

"SOC becomes faster, more adaptive,
and more accountable."

Not: "AI replaces SOC."