Рой без интеллекта

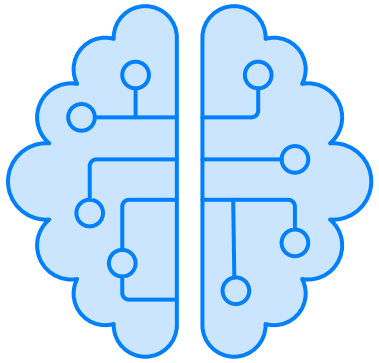
Реальные киберугрозы для мультиагентных систем

Виткова Лидия Андреевна

К.т.н., Начальник Аналитического центра кибербезопасности
ООО «Газинформсервис»

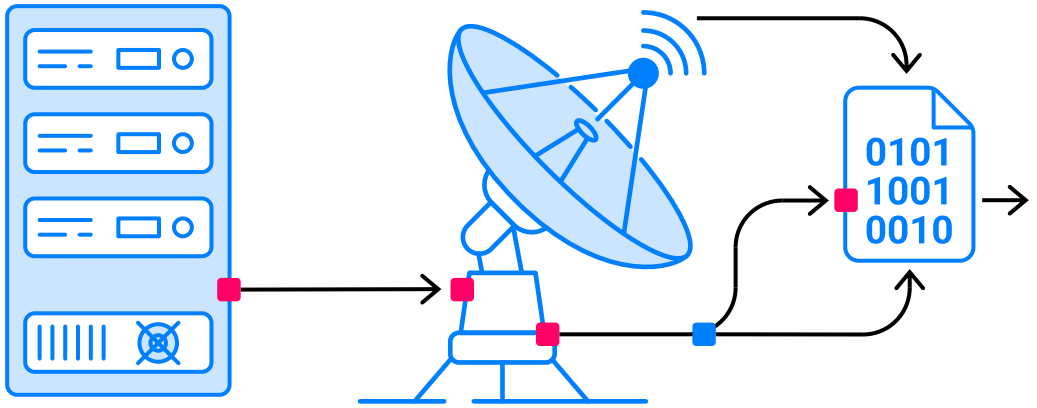
GIS ГАЗИНФОРМ СЕРВИС

Искаженная модель угроз



Маркетинговый миф

Статьи про роевые интеллект звучат тревожно и заставляют искать угрозы не там, где они возникают.



Инженерная реальность

В исследованиях атаке подвергается материальная, цифровая инфраструктура: сенсоры, телеметрия, каналы связи, хосты.

Поверхность атаки:
узлы, каналы связи, данные

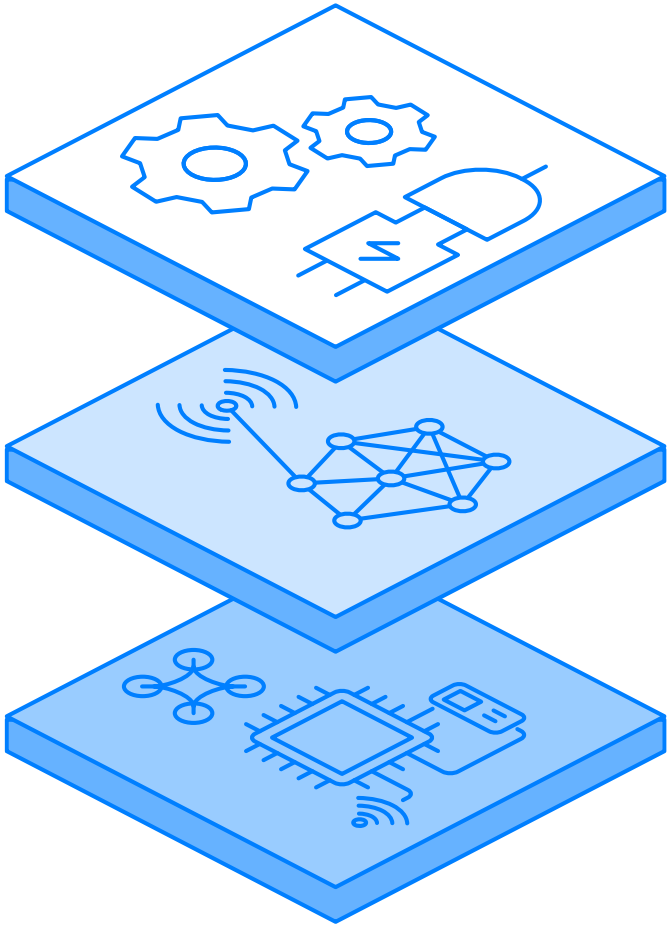
Разделение понятий: биология против инженерии

GIS ГАЗИНФОРМ СЕРВИС

В прикладных системах используется координация простых агентов

	Роевый интеллект	Роевый алгоритм
Происхождение	Биологическая метафора (эмерджентное поведение, стаи птиц, рои пчел, муравейники)	Распределенные вычисления (инженерное направление)
Базовый механизм	Единый коллективный субъект	Локальные правила и обмен сигналами между узлами
Типовые воплощения	Маркетинговое описание «Рой ИИ», «Рой контейнеров»	ASO, PSO, протоколы консенсуса, управление строем (роем)
Вектор атаки	Борьба с центром координации	Компрометация алгоритмов маршрутизации, перераспределение задач

Архитектура мультиагентной системы



Прикладной слой

Планирование задач, локальная логика, консенсус, фильтрация сообщений, общие цели

Коммуникационный слой

Mesh, радиоканалы, телеметрия, ретрансляторы, синхронизация времени, идентификация узлов

Физический слой

Устройства, сенсоры, исполнительные модули, системы навигации, цепи питания

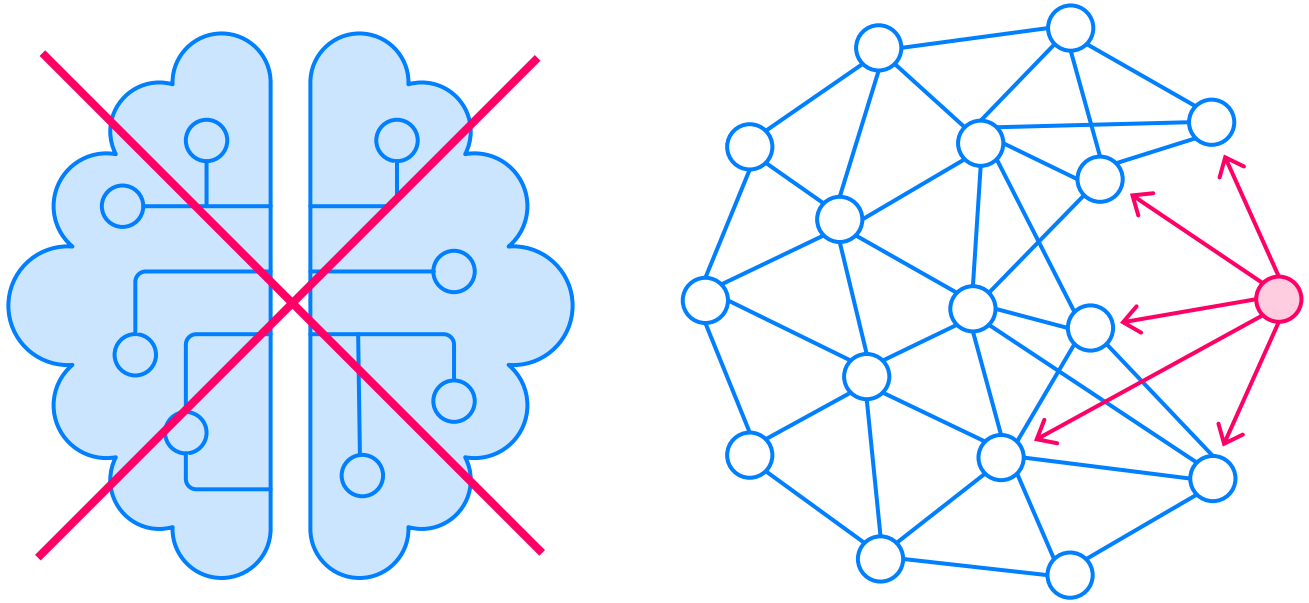
Даже при наличии выделенного координатора уязвимости отдельных узлов не исчезают, добавляется лишь еще одна критическая точка отказа

Ошибка в модели угроз

GIS ГАЗИНФОРМ СЕРВИС

Когда угроза сформулирована неправильно, защита начинает бороться не с тем объектом

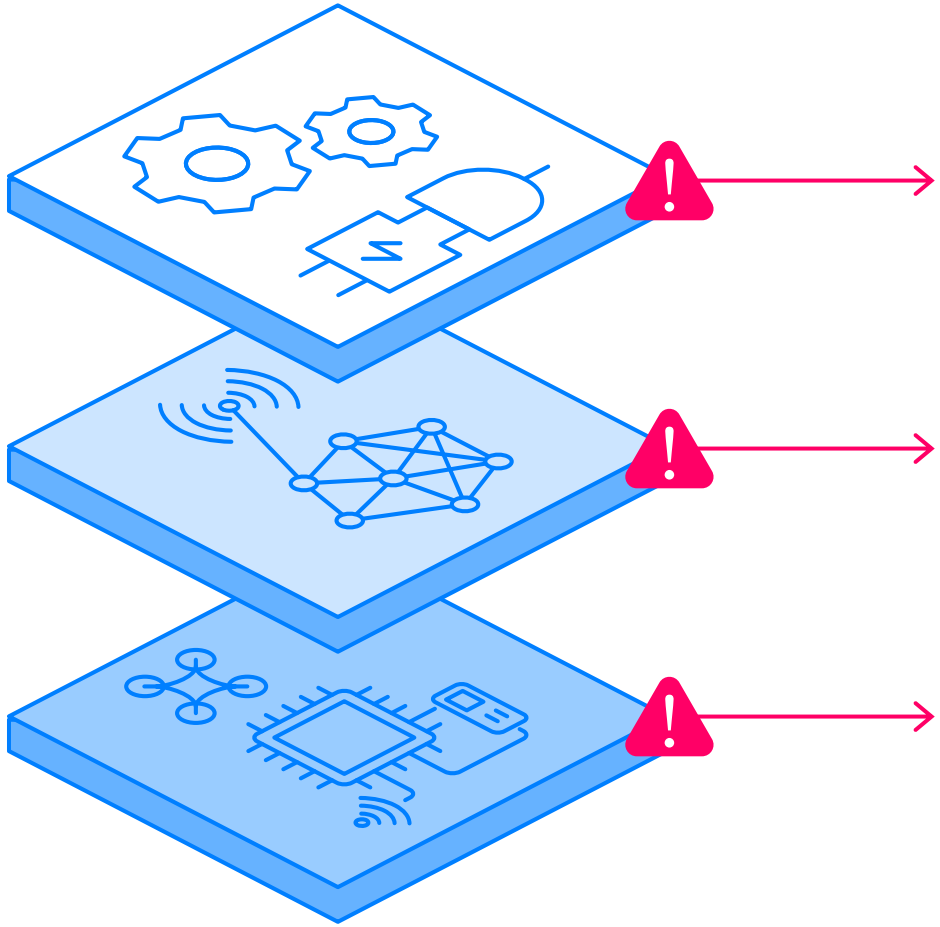
- Противник не пытается «перепрограммировать» коллективный разум
- Противник находит наименее защищенную точку доверия (один узел)
- Противник использует доверие других узлов к этому элементу для разрушения всей координации



Вывод: поверхность атаки локализована на уровне узлов, каналов и данных

Карта реальных киберугроз

GIS ГАЗИНФОРМ СЕРВИС



Прикладной слой

False Data Injection: неверная телеметрия искажает групповую картину мира
Host execution: в AI агентах опасна связка промт-инструмент-хост

Коммуникационный слой

DoS/Jamming: потеря видимости соседей, распад координации
Sysbil/Byzantine: один узел создаёт множество ложных идентификаторов с целью сместить консенсус в распределённой системе

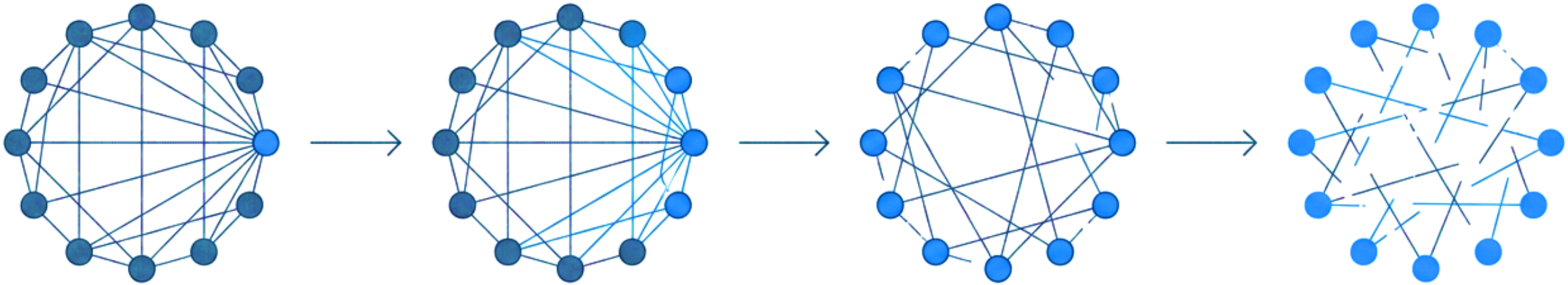
Физический слой

GPS/GNSS Spoofing: подмена координат для срыва маршрута

Механика каскадного сбоя

GIS ГАЗИНФОРМ СЕРВИС

Коллективный отказ всегда начинается с локальной ошибки, которая многократно усиливается алгоритмами кооперации



1 Компрометация части узлов

Достаточно захватить малую часть сети, чтобы запустить цепную реакцию

2 Ложная групповая картина

Соседние узлы принимают искаженные данные, подтверждают неверные сигналы и увеличивают доверие к ошибке

3 Срыв координации

Алгоритмы маршрутизации, распределение задач и формирование роя начинают дрейфовать от изначальной цели

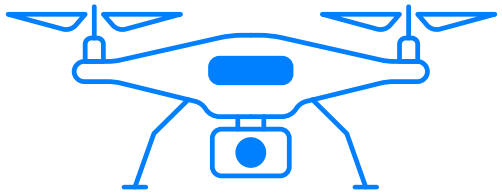
4 Деградация миссии

Система не «сходит с ума» мгновенно, но начинает последовательно действовать хуже и так до полного отказа

Примеры наследования атак

GIS ГАЗИНФОРМ СЕРВИС

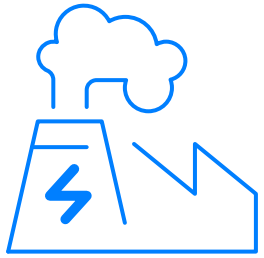
Одинаковый вектор компрометации повторяется как в физических устройствах, так и в AI-агентах



(GPS) Spoofing attacks

В статье дана пошаговая инструкция как создать недорогое устройство для атаки на дрон на примере Matrice 100 quadcopter. Логика переносима и на AI агента

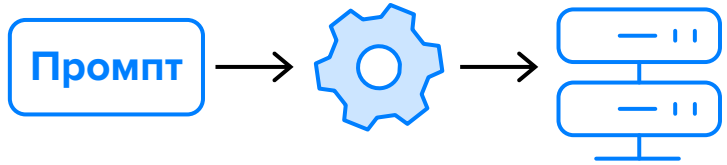
Horton, E., Ranganathan, P. Development of a GPS spoofing apparatus to attack a DJI Matrice 100 Quadcopter. J. Glob. Position. Syst. 16, 9 (2018). https://doi.org/10.1186/s41445-018-0018-3



Сеть агентов в энергетике

В статье предложен механизм, по которому агент присваивает сам метрики доверия своим соседям, информация от агентов с низким уровнем игнорируется

Matei, I. and Srinivasan, V. (2011), Trust Based Multi-Agent Filtering for Increased Smart Grid Security, Proceedings of the 2nd IEEE International Conference on Smart Grid Communications (SmartGridComm), Brussels, BE (Accessed May 10, 2026)



RCE-уязвимости в фреймворках AI-агентов

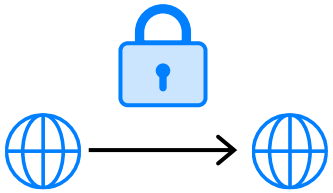
7 мая 2026 года Microsoft Defender researchers опубликовали статью, в которой написали: Мы обнаружили уязвимый путь в Microsoft Semantic Kernel, который может привести к удаленному выполнению кода на уровне хоста (RCE).

https://www.microsoft.com/en-us/security/blog/2026/05/07/prompts-become-shells-rce-vulnerabilities-ai-agent-frameworks/

Инженерный фокус защиты

Отказоустойчивость устройств

Резервирование систем, строгая самодиагностика, изоляция и переход в безопасный режим при локальном сбое узла.

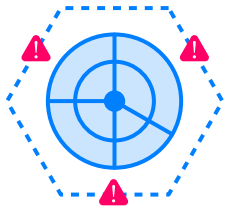


Резервирование связи

Криптографическая аутентификация узлов, альтернативные mesh-пути, резервирование маршрутизации, аппаратная устойчивость к помехам.

Целостность телеметрии

Криптографическая верификация происхождения данных и алгоритмическая фильтрация аномалий.



Изоляция скомпрометированных агентов

Локальная автономность принятия решений, протоколы карантина, автоматическое снижение уровня доверия к подозрительным соседям.

Архитектурные принципы безопасного роя

GIS ГАЗИНФОРМ СЕРВИС

Безопасность мультиагентных систем – это строгая дисциплина проектирования распределенных вычислений

Фундамент Zero Trust

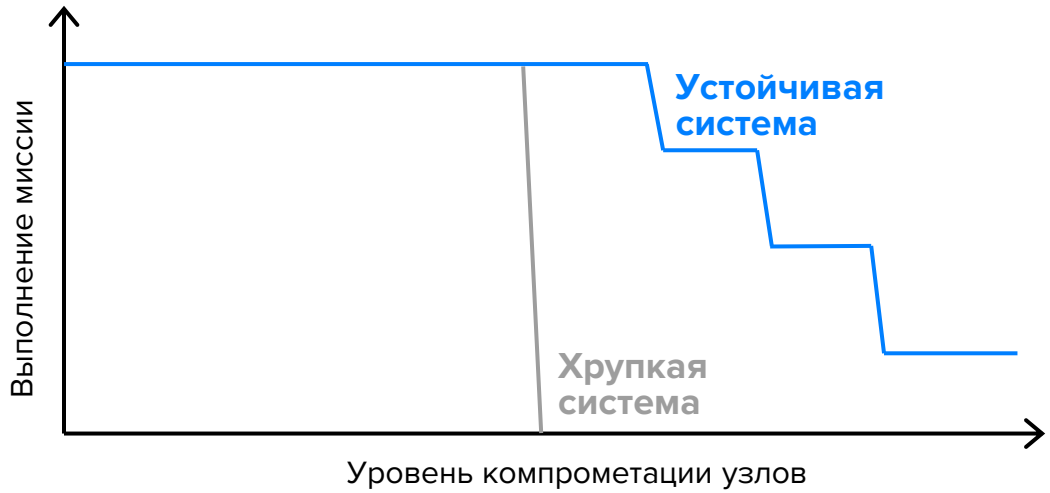
Zero Trust: ни один узел и ни одно сообщение не считаются безопасными по умолчанию.

Least Privilege: минимальные права для каждого агента и каждого вызываемого инструмента.

Cross-Validation: перекрестная проверка сигналов. Запрет на слепое доверие одному сенсору или агенту.

Observability: доступность логов, телеметрии решений и непрерывный аудит для восстановления причин сбоя.

Контролируемая деградация



При частичном отказе узлов или потере связи миссия должна ухудшаться контролируемо, а не превращаться в системных хаос

Корректная модель угроз, правильная защита

GIS ГАЗИНФОРМ СЕРВИС

1

Роевой алгоритм ≠ роевой интеллект

Большинство прикладных систем координируется математическими локальными правилами, но пока что технологии не достигли уровня роевого интеллекта и чаще всего есть центр управления.

2

Реальные атаки бьют по инфраструктуре

Навигация (GPS), каналы связи, консенсус, права доступа и хост — это и есть поверхность атаки в мультиагентной системе.

3

Защита должна быть структурной

Надежность физических узлов, криптографическая защита данных и контролируемая деградация.

GIS ГАЗИНФОРМ СЕРВИС

Спасибо за внимание

Виткова Лидия Андреевна

К.т.н., Начальник Аналитического центра кибербезопасности ООО «Газинформсервис»

Vitkova-l@gaz-is.ru

GIS ГАЗИНФОРМСЕРВИС