Охотники и жертвы. Обучение с подкреплением в среде GridWorld

Валентина Алексеева

26 октября 2025 г.

1 Обзор версий

1.1 Версия 1, один агент

- Один хищник на мини-картах 20×20 с движущимися целями;
- обучение по DAgger: действие сети смешивается с советом учителя (ClosestTargetAgent) с убывающим β, затем сеть оценивается в режиме *greedy*.
- Признаки компактные: «голова» состояния + локальный патч $P \times P$ вокруг агента + K ближайших целей в тор-нормировке $(\Delta y, \Delta x, \Delta n)$.
- Награда разложена на компоненты (база, поимка, исследование, стояние, повторные посещения) с осторожным shaping по Δ кратчайшего пути (BFS), который не доминирует базовые сигналы.

  Подробные логи по шагам и согласованные теплокарты/рендер подтверждают корректную геометрию и динамику преследования; версия служит «санитибенчмарком» DAgger перед масштабированием на командные сценарии. Светлое пятно на тепловой карте – блуждания агента по нескольким точкам.

1.2 Версия 2, мультиагентная среда

- **Постановка.** Мультиагентный контроль команды из пяти охотников на смешанных лабиринтных картах с движущимися целями; дискретные действия; тороидальные патчи во входных признаках.
- **Признаки.** На одного агента: компактная «голова» состояния, локальный патч $P \times P$ проходимости вокруг позиции, K ближайших целей $(\Delta y, \Delta x, \Delta n)$ и K ближайших союзников (кроме себя) с нормировкой по тору; длина признакового вектора фиксирована и не зависит от размера карты.
- **Модель.** Общая для всех охотников лёгкая MLP (shared parameters) с двумя «головами»: policy (логиты действий) и value; инференс в режимах *greedy*/*sample*.

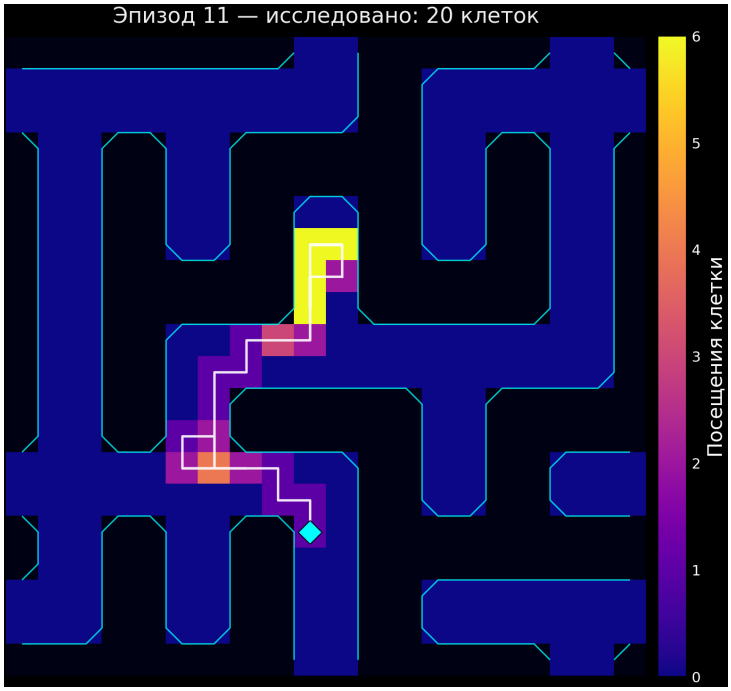

Рис. 1: Тепловая карта посещений, 1 охотник.

- **Награда (разложение).** Базовый штраф за шаг, поимка, исследование новых клеток; мягкие штрафы за стояние и повторные посещения; ограниченный по весу потенциал Δ кратчайшего пути до ближайшей цели (BFS; не доминирует базовые сигналы); поведенческие регуляторы против толкотни: *repulse* (антикучкование, с каппированием), штраф за *flip-flop* и за одинаковое направление на малой дистанции (*same-dir-close*).

- **Обучение.** DAgger для мультикоманды: на каждом шаге батч BC по лейблам учителя (ClosestTargetAgent) и смешение исполняемого действия по β-расписанию (от β_{start} к β_{end}); далее короткая RL-стадия с понижением энтропии и ранней остановкой по медиане шага первой поимки. Отбор лучшего чекпойнта ведётся по *greedy*-метрикам.

- **Ограничения.** В данной конфигурации акцент на стабильное поведение и координацию пятёрки на мини/средних картах;

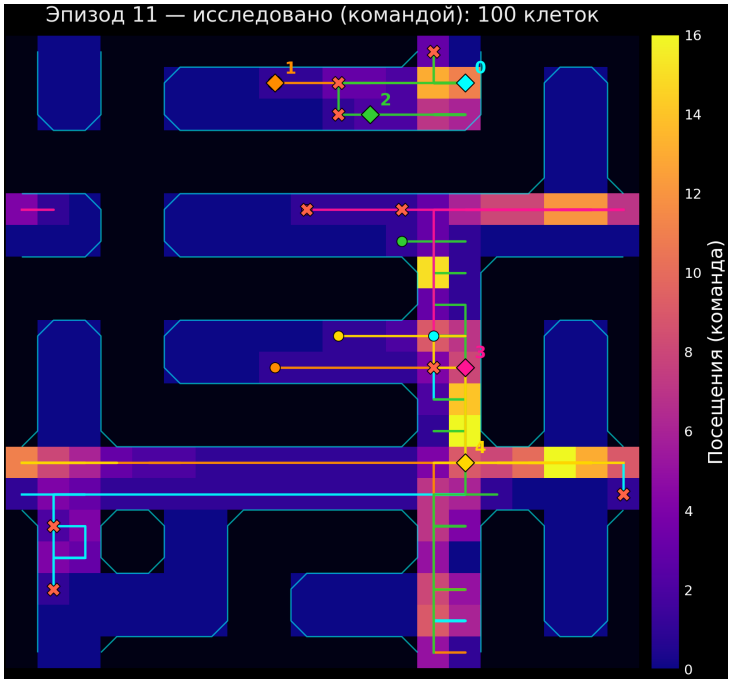

Рис. 2: Тепловая карта посещений, 2 версия, 5 охотников.

1.3 Версия 3

- **Параметры среды.** Размер карты 40×40, число целей 100, лимит шага 300;
- **Награда (разложение).** Базовый штраф за шаг `r_base`, поимка `r_capture`, исследование новых клеток `r_explore`, мягкие штрафы за стояние `r_standstill` и «затаптывание» `r_revisit`; ограниченный по весу shaping по Δ кратчайшего пути до ближайшей цели `r_bfs` (BFS, без тора).
- **Обучение/оценка.** DAgger для команды (лейблы ClosestTargetAgent, смешение исполняемых действий по β-расписанию) с периодическим рендером; быстрая холостая оценка `eval_network_only_multi` в *greedy* для отчётных метрик (шаги, поимки, шаг первой поимки).
- **Ограничения.** Ключевая проблема в больших картах – тип лабиринт. Ключевые проблемы связаны с несовершенством политики учителя Closest Target. В узких проходах охотники толпятся, и агент выучивает эту стратегию, периодически оставаясь на одном месте на несколько шагов. Также на карте посещений видно, что все охотники идут к одним и тем же жертвам, вместо того чтобы разделяться по разным с целью более эффективной ловли. Также обнаружилось, что на таких картах не редко сменяется самая близкая жертва за несколько шагов к ближайшей. Из-за этого участились перемещения с одной клетки на соседнюю и обратно (flip-flop здесь и далее).

1.4 Версия 4 новый учитель

- **Фичи и агент.** Обновлён `FeatureBuilder`: локальный патч $P \times P$, K ближайших целей $(\Delta y, \Delta x, \Delta n)$ *и добавлены* K *ближайших союзников* (по тем же трём

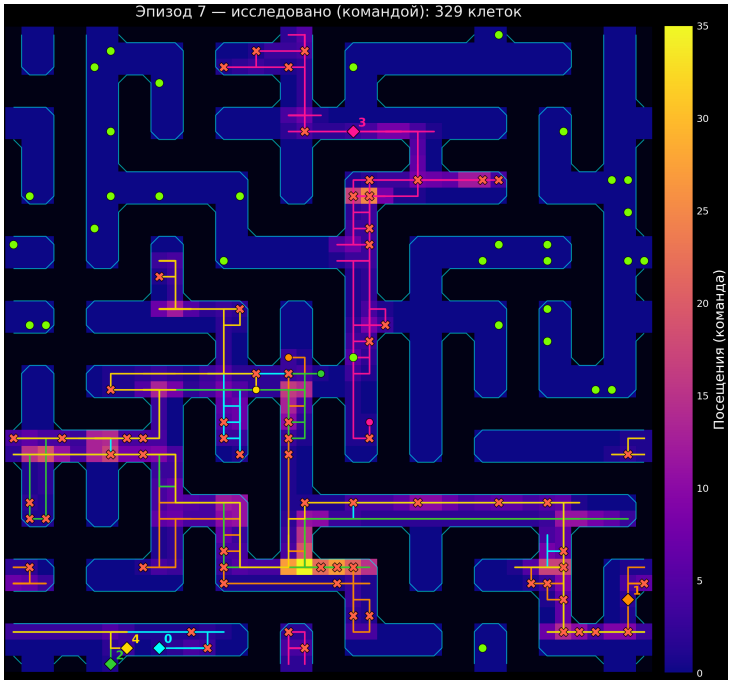

Рис. 3: Тепловая карта посещений, 3 версия, 5 охотников. Наложение траекторий разных охотников, толкучка и дергание. Неисследованные зоны.

признакам); общий агент `NetAgentShared` (shared MLP) + небольшой BC-*replay* для устойчивости.

- **Учитель с гистерезисом.** Вместо чистого `ClosestTargetAgent` — `AssignedClosestTargetA` с *уникальным назначением целей*, *удержанием* цели (`hold_steps`) и переключением только при выигрыше по дистанции на `switch_margin`. *(Снижает дрожание и толкотню.)*

- **Анти-скучивание и плавность движения.**

  – Введён *динамический* `repulse`: штрафы за расстояния 0/1/2 между охотниками с усилением по мере «опустошения» карты; добавлен штраф `pair_stick_penalt` за «держимся рядом шаг за шагом».

  – Добавлены штрафы `flipflop_penalty` (развороты A→B→A) и `same_dir_close_penalt` (одинаковый вектор движения при близком расстоянии).

  – Ранний буст исследования: `early_steps_boost`, `early_explore_scale` усиливают `r_explore` в первые шаги.

- **DAgger и реплей.** Обновлена процедура `train_dagger_multi`: геометрическое расписание β (`beta_start`→`beta_end`), на каждом шаге — BC по лейблам учителя *плюс* 1 шаг мини-`replay`

- **EVAL и сохранение лучшего.** Добавлен `eval_network_only_multi` с критерием лучшего (caught↑, cluster↓, idle↓, first_cap↓) и автосохранением чекпойнта (`logs/checkpoints`, дубликаты `agent_best_eval.pkl`, `agent.pkl`).

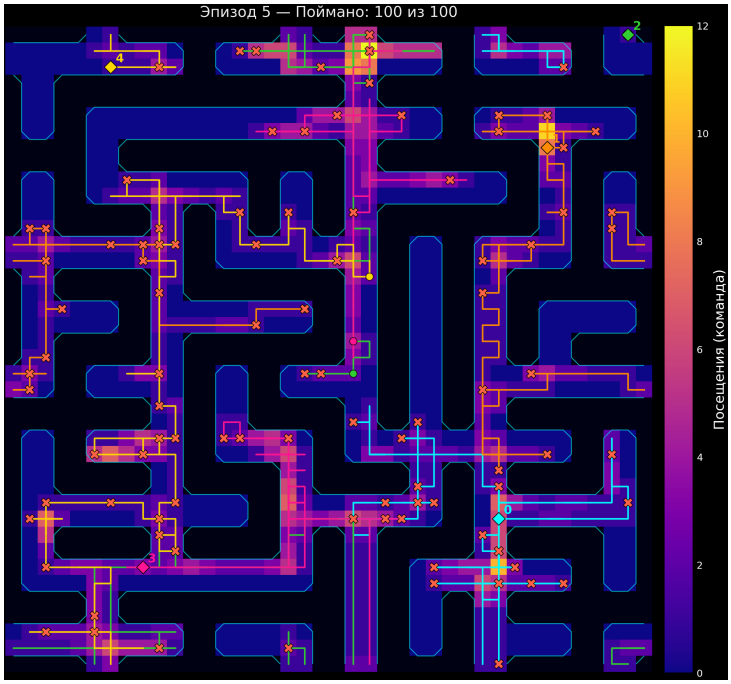

Рис. 4: Тепловая карта посещений, 4 версия, 5 охотников. Все жертвы пойманы.

2 Итоговая модель агента

- **Борьба с замиранием (idle).** Жёсткий штраф за бездействие заменён на мягкий и *условный*: применяется при отсутствии прогресса по командному потенциалу (сумма top-L1 дистанций до ближайших жертв). Это устраняет бессмысленные простои, не ломая перехваты в узких местах.

- **Подавление flip-flop.** Введён малый штраф за разворот $\mathrm{A} \to \mathrm{B} \to \mathrm{A}$ и снижена энтропия на стадии дожима; вместе с гистерезисом смены цели это уменьшает «пиление» на соседних клетках.

- **Слипание и толкотня в коридорах.** Использован динамический *repulse* с каппированием вклада на агента и отложенным стартом; добавлен штраф за одинаковое направление движения на малой дистанции (*same-dir-close*). В результате команда реже идёт «маршем в ногу» и лучше покрывает карту.

- **Признаки, устойчивые к размеру карты.** Локальный патч $P \times P$ и K ближайших сущностей $(\Delta y, \Delta x, \Delta n)$ позволяют переносить политику с миникарт на 40×40 без деградации за счёт глобальных «плоских» представлений.

- **Учитель и DAgger.** Чистый `ClosestTarget` заменён на назначающий цели учитель с удержанием и переключением по выигрышу в дистанции; sticky-DAgger со снижением β и небольшим реплеем стабилизирует обучение на смешанных картах (статик/динамика).

- **Награда без доминирования shaping.** Δ-BFS оставлен ограниченным по весу и служит только направляющим сигналом; основное значение сохраняют поимка и исследование, что уменьшает прокси-эффекты.

- **Критерий «лучшего» агента.** Отбор чекпойнта ведётся в режиме *greedy* по сочетанию метрик: итоговые поимки ↑, медианный шаг первой поимки ↓, кластеризация/idle/flip-flop ↓. Такой критерий согласован с эксплуатационным режимом.

- **Итоговая политика.** Лучший агент — shared-MLP на локальных признаках с поведенческой регуляризацией (idle, flip-flop, repulse, same-dir-close) и гистерезисом целей; демонстрирует устойчивость на больших лабиринтах и выигрывает 45 % PvP-матчей против `ClosestTarget` при исключении отвлекающих компонент (поимка других охотников, бонусы).

- **Облегчённая версия для сабмита.** `NetAgentSharedLite` (уменьшенная глубина/ширина MLP при тех же входах и масках действий) дообучена финальным дожимом и уступает основной версии на ≈ 5 % в доле побед PvP, оставаясь совместимой по времени с проверяющей системой.

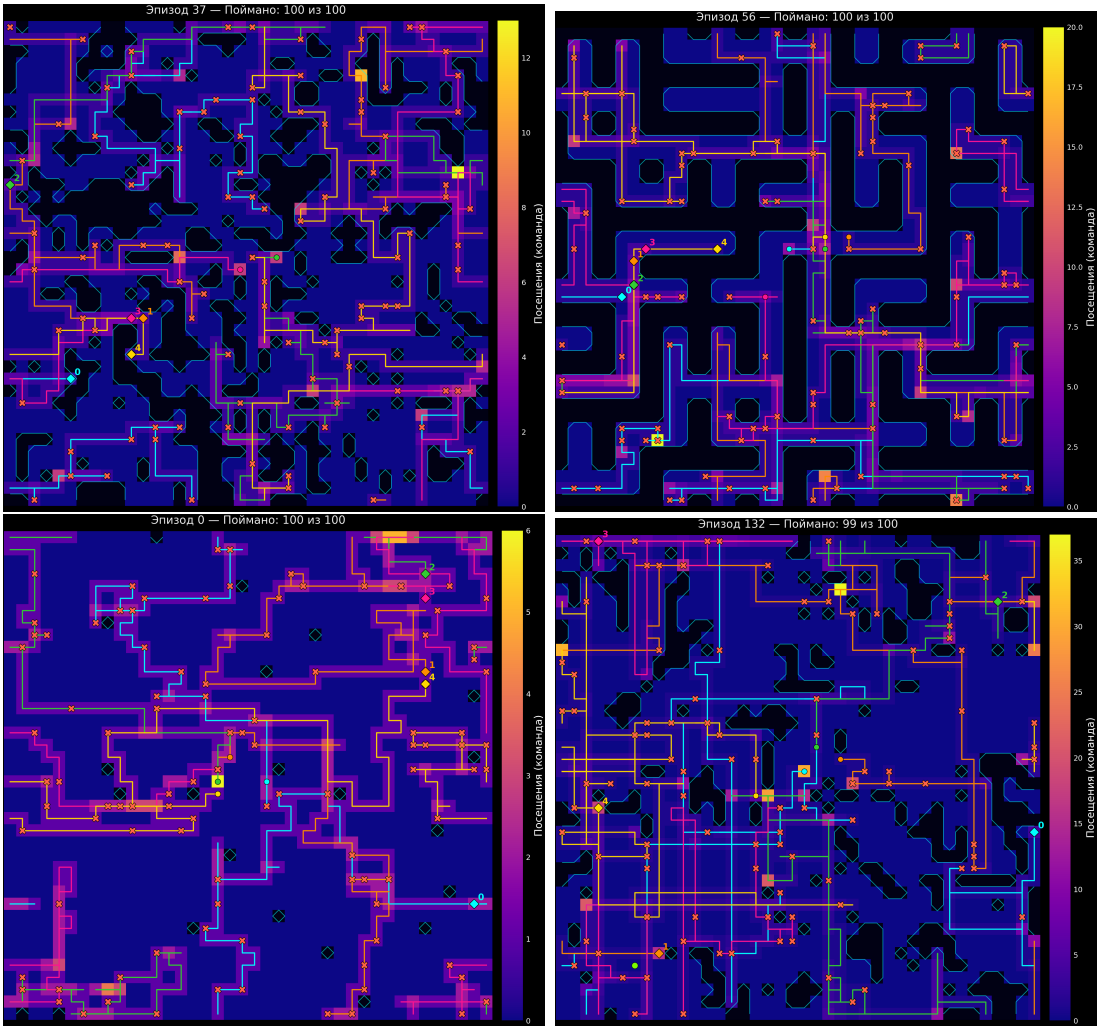

Рис. 5: Тепловые карты активности агентов на разных картах, финальная версия.

2.1 Наблюдения (PvP) и облегчённая версия для сабмита

- **PvP против ClosestTarget.** В конфигурации PvP (VersusBotEnv) политика выигрывает в **40 %** матчей. При этом качество повышается, если *исключить* из функции награды компоненты за поимку других охотников и подбор бонусов: такие сигналы отвлекают от перехвата целей на больших картах и усиливают нежелательные столкновения.

- **Облегчённая версия для сабмита.** Используется `NetAgentSharedLite` (тот же вход: «голова» + локальный патч $P \times P + K$ ближайших целей/союзников; те же маски допустимых действий), но с *уменьшенной глубиной/шириной MLP* и сокращённым числом параметров для быстрого инференса и совместимости с проверяющей системой. Полная версия не проходит проверку по времени работы.

- **Дообучение и сравнение.** Финальный дожим выполнен именно для облегчённой версии; по числу побед в PvP она уступает основной на **5 %**, при прочих метриках поведения отличие несущественно.

3 Общие выводы по проекту

- Ключевые поведенческие проблемы преодолены: замирание (idle) подавлено условным и мягким штрафом; частые развороты (flip-flop) снижены за счёт небольшого штрафа и гистерезиса целей; слипание и толкотня в коридорах уменьшены динамическим *repulse* с каппированием и штрафом за одинаковое направление на малой дистанции.

- Признаковая схема (локальный патч $P \times P + K$ ближайших целей/союзников показала устойчивость к увеличению карты до 40×40 и к смешанным типам карт; глобальные «плоские» представления не использовались.

- Shaping по Δ-BFS оставлен в ограниченном весе и используется как направляющий сигнал; основными остаются поимка и исследование. Это позволило избежать прокси-оптимизации и сохранить полезную динамику преследования.

- В PvP против скриптованного `ClosestTarget` уверенного превосходства не достигнуто: доля побед держится на уровне $\approx 45\%$. При этом базовые задачи — покрытие карты, перехват движущихся целей и избегание толкотни — выполняются стабильно.

- Облегчённая версия (`NetAgentSharedLite`) после дополнительного дожима сохраняет поведенческие преимущества, но уступает основной ~ 5 % в доле побед PvP; взята как рабочий вариант для проверяющей системы по ограничениям времени и формата.

В результате получена устойчивая политика для мультиагентного преследования на больших лабиринтах с контролируемой координацией, сниженными антипаттернами и воспроизводимыми метриками в режиме *greedy*.