ИИ-компаньоны и социальный сдвиг от контроля к поддержке: автономно-сохраняющая архитектура и рамка, опирающаяся на моделирование.

Aleksandr Kolomiets
Независимый исследователь
ORCID 0009-0008-5153-5546

Аннотация

В препринте предлагается автономно-сохраняющая архитектура «ИИ-компаньона», предназначенная для поддержки пользователя в ситуации напряжения без принуждения и манипуляций. Подход объединяет: (i) нормативный протокол взаимодействия («зеркало и разделение»), (ii) “ворота вмешательств”, приоритизирующие наблюдение, отражение и поддержку, минимальные предложения и эскалацию по безопасности, а также (iii) пользователь-центричные сигналы (динамика потребностей, модель идентичности и маркеры контактных паттернов), позволяющие снижать риск вредных спиралей при сохранении агентности. Чтобы связать микроуровень взаимодействия с макроэффектами, описывается рамка, опирающаяся на моделирование: индексы фрустрации, напряжения и прироста способности соотносятся с режимами реакции платформ и институтов. Работа рассматривает поддерживающий ИИ как дополнение к модерации и принудительным мерам и утверждает, что принцип «поддержка вместо контроля» способен снижать стимулы к эскалации при сохранении ответственности. Обсуждаются проектные ограничения, доверие/безопасность и метрики оценки для пилотного внедрения в организациях.

1 Введение

Цифровые ассистенты уже умеют планировать наш день, отвечать на письма и даже поддерживать лёгкую беседу. Следующий естественный шаг — ИИ-компаньон для профилактики насилия: система, которая ненавязчиво отслеживает эмоциональное напряжение, предлагает способы снизить агрессию по отношению к себе и другим и

тем самым помогает предотвратить эскалацию — от бытовых ссор до уличных протестов.

Цель предлагаемой архитектуры — запустить долгосрочный культурный дрейф, при котором чувствительность к насилию и предпочтение ненасильственных решений постепенно растут за счёт систематической поддержки у человека чувства «я прав» и «моя позиция легитимна». В качестве реалистичного ориентира (а не жёсткого прогноза) мы используем столетний горизонт диффузии — по аналогии с электрификацией и массовой грамотностью. При этом ключевые показатели прироста возможностей (capability-gain) — снижение Frustration и Tension, рост участия и продуктивности при меньшей внешней «закрутке гаек» — можно пилотировать уже сейчас (см. раздел 5.1 и Приложение E). В долгосрочной перспективе такой дрейф должен также снижать вероятность крупномасштабных вооружённых конфликтов, но в этой статье мы фокусируемся на ближайших, измеряемых эффектах.

Массовое развёртывание подобного сервиса даёт амбивалентную картину. С одной стороны, можно ожидать улучшения психического благополучия и более высокой продуктивности [1]. С другой — существует риск поляризации и насильственных протестов, если алгоритмы неосознанно подталкивают пользователей к самым «резонансным» — а значит, часто радикальным — сообщениям и группам [2, 3].

Однако на уровне индивидуального опыта «насилие» редко переживается как абстрактная моральная категория. Чаще это конечная точка длительного регуляторного сбоя: коллапса способности человека согласовывать телесное напряжение, эмоциональные импульсы, социальные ожидания и собственное чувство правильности. Субъект оказывается в расщеплённом состоянии: тело уже мобилизовано на «бей или беги», повествование о себе и других сузилось до нескольких враждебных сценариев, а доступные действия выглядят либо разрушительными, либо унизительными. В таком состоянии доступ к ненасильственным опциям — это не просто вопрос «знания правил»: семантические слепые зоны и вызванное стрессом тоннельное внимание делают альтернативные рамки буквально труднодоступными для восприятия. Поэтому предложение этой работы состоит не в том, чтобы «учить более правильным нормам» в абстрактном виде, а в том, чтобы спроектировать опосредованную ИИ поддержку, которая снижает регуляторную нагрузку, вновь открывает поле интерпретаций и делает ненасильственные, сохраняющие достоинство ответы реалистично доступными.

Антропологические основания насилия и регуляторного коллапса

В этой рамке насилие понимается не как моральный дефект, а как форма регуляторного коллапса.
Когда напряжение накапливается сверх пропускной способности системы, агент переходит в ориентированный на выживание регуляторный режим. Это состояние характеризуется:

- когнитивным сужением;
- семантической жёсткостью и снижением интерпретативной гибкости;
- рассогласованием между аффектом, телесным напряжением и поведенческим импульсом;
- ухудшением способности выдерживать сложный или амбивалентный опыт.

Мы обозначаем это состояние как split-state (расщеплённое состояние), при котором регуляторные подсистемы, обычно работающие согласованно (соматические маркеры, распознавание потребностей, картирование смыслов, формирование намерения), теряют синхронность.

Такое расщепление повышает вероятность реактивного поведения, искажённого восприятия, межличностного рассогласования и эскалации.

Этот антропологический взгляд принципиально важен для моделирования того, как микрорегуляторные состояния напряжения трансформируются в межличностную или общественную нестабильность.

Чтобы ответить на этот вызов, мы предлагаем двухуровневую систему.

1. **Нижний уровень: личный ИИ-компаньон.**
   Компаньон локально работает на устройстве пользователя, отслеживает эмоциональные маркеры, предлагает мягкие техники снижения напряжения и «воронку» поведенческих подсказок. Ключевой механизм — Autonomy-Preserving Gate (AP-Gate, «затвор автономии») [4]. Алгоритм переходит к активному вмешательству только тогда, когда обнаруживает сочетание (i) высокого напряжения и (ii) намерения насильственного действия по отношению

к себе или другим, и когда пользователь явно просит о помощи. Во всех остальных ситуациях компаньон остаётся в режиме наблюдателя и регистрирует эпизоды успешной саморегуляции. Подробный каскад интервенций и логика работы AP-Gate описаны в Приложении G.2; он срабатывает только при одновременном наличии высокого риска и явного пользовательского согласия.

2. **Верхний уровень: Хартия этики и надзор.**
   Наднациональная, межсекторная структура (регуляторы, НКО, академическое сообщество, провайдеры), которая:

- задаёт уровень E (строгость фильтра радикального контента) и минимально допустимый уровень R (институциональная реакция на всплески насилия) [5];
- проводит регулярные внешние аудиты, гарантируя, что AP-Gate не инфантилизирует пользователей и остаётся совместимым с основными религиозными и философскими традициями [6];
- публикует отчёт по KPI — долю насильственных эпизодов на активного пользователя — и при превышении порога 0.08 инициирует «аварийный откат» к более жёстким значениям E и R [7].

Таким образом, модель «снизу вверх» (индивидуальная профилактика) сочетается с моделью «сверху вниз» (динамическое институциональное управление). В агент-ориентированной модели нижний уровень представлен параметром E и логистической кривой охвата, а верхний — параметром R и сценарием, в котором R временно увеличивается, когда охват достигает 25–40 %.

3. **Новизна подхода.**
   В отличие от работ, которые сосредоточены либо на вреде социальных платформ, либо на чисто терапевтических применениях ИИ, мы фокусируемся на интеграции трёх уровней.

- Во-первых, личный ИИ-компаньон концептуализируется как «зеркало и инструмент разделения», чья задача — поддерживать субъективную целостность и ненасильственное отношение к себе, а не навязывать поведенческие нормы.
- Во-вторых, этот компаньон встраивается в макроуровневую архитектуру платформенной фильтрации (E) и институциональной реакции (R), где его вклад калибруется с помощью агент-ориентированной модели.

- В-третьих, мы специфицируем операционные протоколы (Приложения F и G.1–G.2), которые задают конкретные ограничения и режимы работы: минимально достаточное вмешательство, прозрачную эскалацию рисков и сохранение агентности пользователя.

В этом смысле новизна состоит не в предложении новой LLM как таковой, а в нормативной и архитектурной «надстройке» вокруг готовой модели общего назначения.

4. **Уровень реализации.**
   Важно, что архитектура не требует с нуля обучать «ещё одну нейросеть». Протоколы (зеркало и разделение, режимы выживания/развития, каскад интервенций) реализуются в управляющем слое, который использует LLM общего назначения, но жёстко ограничивает, что ей дозволено делать и в каком режиме. Иными словами, вклад работы относится к дизайну управления и нормирования вокруг весов модели, а не к самим весам.

В этой статье мы:

1. строим агент-ориентированную модель (ABM), описывающую, как параметры E и R влияют на протест и насильственное поведение по мере роста охвата компаньоном;
2. валидируем модель на данных V-Dem [8], Pew [6] и ACLED [9] за 2022–2024 гг.;
3. показываем, что ключевым «демпфером» насилия является качество демократических институтов, тогда как религиозная фрагментация сама по себе не увеличивает уровень насилия;
4. оцениваем макроэкономический эффект (до ≈ 0.7 % мирового ВВП) с использованием данных SIPRI [10] и Всемирного банка [11].

2 Обзор литературы

2.1. ИИ-компаньоны и цифровые терапевтические решения

Рандомизированные контролируемые исследования текстовых компаньонов (Woebot, Wysa, Youper) показывают снижение выраженности депрессии и тревоги примерно на

0.2–0.4 стандартного отклонения [12, 13]. Более широкая литература по «эмоциональному интеллекту» подчёркивает, что тренировка саморегуляции даёт не только субъективное благополучие, но и экономические выгоды за счёт повышения продуктивности [14]. Международные обзоры цифровых инструментов поддержки психического здоровья показывают потенциал масштабируемых платформ само-помощи при условии, что они следуют этическим принципам дизайна и используют надёжные стандарты защиты данных [15, 31].

Этологическая перспектива.
С этологической точки зрения рамка «поддержка против насилия как энергетическая цена» согласуется с работами по просоциальному поведению и аффилиации у млекопитающих: кооперация, как правило, стабилизируется через вознаграждение и снижение неопределённости, тогда как открытая агрессия метаболически дорога и стратегически рискованна [33, 38]. Обзоры по эмпатии и социальным связям у млекопитающих подчёркивают роль аффилиативных нейросетей и нейромодуляторов (окситоцин, вазопрессин и др.), которые совместно регулируют реакции на угрозу и близость [33, 34, 35, 36]. Это согласуется с нашим пониманием энергетического индекса E* как «регуляторной цены» интервенции и мотивирует образ компаньона как мягкого зеркала, а не карательного контролёра [37].

В данной работе мы опираемся на эти результаты, но переносим фокус с уменьшения симптомов в клинических и субклинических выборках на популяционный дрейф в том, как люди относятся к собственной агрессии и уязвимости.

2.2. Социальная стратификация и протест

Классические работы Липсета [16] и современные событийные базы вроде ACLED связывают социальную фрагментацию и политическое насилие, однако результаты оказываются неоднозначными. Свежие количественные обзоры показывают, что влияние образования на политическое насилие может быть как умиротворяющим, так и дестабилизирующим — в зависимости от более широкого институционального контекста [17], а качество институтов взаимодействует с уровнем образования [18].

В высокоэффективных демократиях одни и те же базовые напряжения чаще канализируются в ненасильственные формы участия; в хрупких режимах они чаще

выливаются в бунты и репрессии. Наша модель отражает это, интерпретируя параметр R как спектр от «мягкой» реакции (диалог, уступки) до «жёсткой» (полицейские меры, аресты, применение военной силы).

2.3. Алгоритмическая модерация

Опыт применения «мягких» и «жёстких» фильтров контента на социальных платформах [2, 7] показывает измеримое снижение токсичности, но консенсуса по поводу баланса между фильтрацией и свободой выражения нет — особенно когда система выступает не как публичная лента, а как персональный ассистент. Доклад IDEA о глобальной поляризации подчёркивает роль рекомендательных алгоритмов в усилении «аффективной дистанции» между группами [19].

Поэтому мы рассматриваем этический фильтр E не как чисто технический параметр, а как элемент оспариваемого пространства управления: слишком слабый фильтр — и насилие эскалирует; слишком сильный — и мы скатываемся к чрезмерной блокировке и инфантилизации. Архитектура компаньона как раз и призвана смягчить это противоречие, перенося часть работы с централизованной модерации на поддерживаемую, сохраняющую автономию саморегуляцию.

2.4. Антропологическая рамка: насилие как регуляторный коллапс

В антропологической перспективе, лежащей в основе этой работы, человек понимается не как стабильный «контейнер черт», а как динамически регулируемая система. В каждый момент сосуществуют несколько частично автономных слоёв: телесное возбуждение и мышечное напряжение; аффективные импульсы и фантазии; социально выученные нормы и нарративы; практическое действие в мире. В обычных условиях эти слои остаются примерно согласованными: тело способно успокоиться после фрустрации, история о себе и других остаётся гибкой, человек сохраняет ощущение авторства в отношении следующих шагов.

Хроническая перегрузка и блокировка потребностей, однако, формируют другую конфигурацию. Телесное возбуждение остаётся высоким, но пространство приемлемых

интерпретаций сужается: другие воспринимаются как угрозы или препятствия, Я — как ничтожное или опасное. Мы называем это расщеплённым регуляторным состоянием: тело, аффект и история больше не сходятся в траекторию, которую человек может переживать как «мою». В этой рамке насилие — не в первую очередь моральное отклонение, а стратегия последней инстанции, появляющаяся тогда, когда все ненасильственные опции субъективно кажутся закрытыми или слишком дорогими. Цена измеряется не только внешними санкциями, но и внутренним унижением, стыдом и потерей целостности.

У такого фрейминга две важные импликации. Во-первых, он смещает фокус с «коррекции девиантных убеждений» на восстановление регуляторной способности: помощь человеку в том, чтобы почувствовать, назвать и вновь интегрировать своё состояние так, чтобы менее разрушительные опции стали снова видимыми. Во-вторых, он выводит на передний план то, что далее мы называем семантической окклюзией: вызванные стрессом слепые зоны в том, как ситуация может быть описана. Когда единственными доступными историями остаются «они — враги» или «я — мусор», само семантическое пространство становится насильственным.

2.5. Семантическая окклюзия и искажение интерпретации

При высокой регуляторной нагрузке агенты входят в состояние семантической окклюзии — сужения доступного смыслового пространства.

Это состояние порождает три характерных искажения:

1. **Потеря дифференциации** — разные эмоциональные состояния или намерения перестают различаться.
2. **Доминирование сигналов угрозы** — неоднозначный вход интерпретируется через оборонительные фильтры.
3. **Снижение концептуального разрешения** — агент утрачивает способность удерживать тонкие различия и прибегает к жёстким, грубым рамкам.

Эти искажения усиливают межличностное трение, недопонимание и реактивные петли обратной связи. Они также усиливают динамику социального заражения, поскольку искажённые прочтения распространяются по социальным сетям.

Любая архитектура, претендующая на профилактику насилия, должна учитывать оба уровня: и телесно-аффективную регуляцию, и восстановление более богатого семантического поля.

3. Методы

3.1. Динамика эскалации от микро к макро

Регуляторный коллапс на микроуровне не остаётся замкнутым внутри отдельных агентов.
Когда многие люди одновременно переходят в высоконапряжённые, «режим выживания»-состояния, их реактивные интерпретации и узкие поведенческие паттерны начинают взаимодействовать и порождать эффекты второго порядка:

- усиливающееся рассогласование в диадах;
- быстрое распространение раздражительности или недоверия;
- кластеризацию оборонительного поведения;
- эпизодические всплески межличностной агрессии.

В плотных социальных сетях эти микромасштабные искажения восприятия и напряжённо-детерминированные реакции могут каскадировать в макро-феномены: поляризацию, конфликтообразующие кластеры и всплески коллективного насилия.

Поэтому агент-ориентированная модель (ABM) симулирует не поведение как таковое, а именно **распространение регуляторных состояний**, фиксируя, как локальные искажения масштабируются в системный риск.

3.2. Минимальная агент-ориентированная модель

Мы начинаем с минимальной агент-ориентированной модели (ABM) диффузии протеста и насильственной эскалации.

ABM здесь не претендует на психологически реалистичную симуляцию индивидов; её задача — формализовать микрорегуляторную картину и посмотреть, как она масштабируется. Каждый агент несёт упрощённое представление о регуляторной нагрузке и семантической гибкости: насколько легко он может снижать напряжение,

сколько ненасильственных сценариев остаётся доступным под стрессом, как он обновляет свою позицию после поддерживающих или унизительных столкновений. То, что мы на макроуровне называем «структурным насилием», можно тогда описать как распределение этих микросостояний по популяции и во времени: сколько агентов хронически загнаны в расщеплённые состояния, как часто они встречают эскалацию против поддержки и насколько быстро могут вернуться из режима выживания в режим развития. Это даёт нам мост между воплощённой антропологией насилия и агрегированными индикаторами вроде V-Dem или ACLED.

Мы рассматриваем популяцию из $N = 1000$ агентов, размещённых на графе «малого мира» Уоттса–Строгаца со средней степенью $k = 8$ и вероятностью перезаписи рёбер $p_{\mathrm{rew}} = 0.05$. На каждом дискретном шаге времени агент может принять или сбросить «вредоносное» состояние под влиянием соседей и институциональных фильтров.

Каждый агент i имеет дискретное состояние

$$s_i(t) \in \{0,1,2\}$$

$$s_i(t) \in \{0,1,2\}$$

where $0 =$ Calm (спокоен), $1 =$ Protester (протестующий), $2 =$ Violent (насильственный). Переходы определяются следующими параметрами:

- Параметр заражения P_S — вероятность перехода из Calm в Protester при воздействии протестующих/насильственных соседей.
- Параметр эскалации P_E — вероятность перехода из Protester в Violent.
- Платформенный фильтр E — вероятность того, что воздействие «высокорискованного» контента будет заблокировано.
- Institutional response R — вероятность того, что агент в состоянии Violent будет де-эскалирован обратно в Calm через подавление, медиацию или другие контрмеры.
- Индивидуальное доверие $trust_i \in [0.2,0.8]$ — чувствительность агента iii к сигналам соседей.

На каждом шаге времени t:

1. Для каждого агента i, считаем долю насильственных и протестующих соседей в текущем графе G:

$$v_i(t) = \#\{j \in N(i): s_j(t) = 2\}, \quad p_i(t) = \#\{j \in N(i): s_j(t) = 1\},$$

$$contag_i(t) = \frac{v_i(t) + 0.5 \cdot p_i(t)}{\max(1, |N(i)|)}.$$

2. Обновляем состояния по следующим правилам:

- **Переход “Calm” → “Protester”**

  Вводится «этический затвор» $gate_i(t)$, который открывается с вероятностью $gate_i(t) = 1$ если случайная величина rand $\geq E \cdot coverage(t)$, то есть, затвор не блокирует воздействие. Если $contag_i(t) > 0$ и затвор открыт, спокойный агент переходит в Protester с вероятностью

  $$P_S \cdot trust_i \cdot contag_i(t).$$

- **Динамика Protester**

  Агент-Protester эскалирует в Violent с вероятностью P_E; с вероятностью 0.10 он спонтанно «остывает» до Calm, в противном случае сохраняет состояние 1.

- **Переход “Violent” → “Calm”**

  Агент в состоянии Violent де-эскалируется с вероятностью R.

Мы отслеживаем два ключевых выходных показателя:

- **Пиковая доля насилия:**

$$\max_{t \leq T} \frac{\#\{i: s_i(t) = 2\}}{N}$$

- **Кумулятивное насилие:**

$$\sum_{t=1}^{T} \frac{\#\{i: s_i(t) = 2\}}{N}.$$

Горизонт моделирования $T = 36$ месяцев (шагов времени). В начальный момент $t = 0$. 30 30 случайно выбранных агентов устанавливаются в состояние Violent, остальные — Calm.

Проникновение компаньона моделируется логистической S-кривой покрытий:

:

$$coverage(t) = 0.15 + \frac{0.45}{1 + \exp\left(-k_{\text{cov}}(t - t_0/2)\right)},$$

с базовыми параметрами $t_0 = 18$ месяцев и $k_{\text{cov}} = 0.40$ (см. Приложение A, Таблица A1). Эффективный член затвора $E \cdot coverage(t)$ представляет долю популяции, чьё воздействие высокорискованного контента модерируется фильтром компаньон-типа.

«Этический затвор» в этой минимальной ABM — грубый прокси: он просто блокирует долю событий заражения пропорционально $E \cdot coverage(t)$. Операционный, сохраняющий автономию AP-Gate, используемый в компаньоне (Раздел 5 и Приложение C), уточняет это, учитывая состояние пользователя и его явный запрос; здесь нам нужен только агрегированный эффект.

3.3 Параметры и сценарии

Базовые значения параметров и диапазоны чувствительности суммированы в Приложении A, Таблица A1. Вкратце:

$P_S = 0.30$ (диапазон 0.15–0.35) — вероятность «заражения» протестом, калиброванная под наблюдаемые частоты протестов в ACLED и классической литературе по диффузии протеста [16, 9].

$P_E = 0.25$ (диапазон 0.15–0.35) — вероятность эскалации от протестных к насильственным событиям.

E — «жёсткость» алгоритмической фильтрации высокорискованного контента (0.2–0.4 в основных сценариях).

R — эффективная сила институциональной реакции, объединяющая полицейские меры, медиацию и способность к правоприменению (0.10–0.30).

Мы рассматриваем сетку по (E, R) и траекториям покрытия компаньоном:

- **Status quo**: низкие E, низкие R, пренебрежимо малое покрытие (компаньона нет).
- **Platform-only filtering**: повышенное E при низком покрытии (классическая модерация контента).

- **Companion + Charter**: умеренное E, усиленныйd R во временном «пиковом окне», когда покрытие между 25–40 %, затем постепенный откат.

Для каждого сценария мы проводим 2000 Монте-Карло симуляций с разными случайными сид-значениями и начальными «семенами» насилия (0.5–2 % узлов). Мы оцениваем средние значения и 95% доверительные интервалы для пикового и кумулятивного насилия и проверяем устойчивость ранжирования по наборам (E, R) settings (Приложение A, Рисунки S4, S5, S6, S7).

3.4 Эмпирическая калибровка *E* и *R*

В симуляциях E и R — абстрактные параметры, но у них есть прямые эмпирические аналоги:

- **E** трактуется как прокси строгости алгоритмических и модерационных фильтров: частоты удаления контента, даун-ранкинга, блокировок аккаунтов и связанных мер принуждения в отчётах о транспарентности платформ (hate speech, подстрекательство, дезинформация) [2, 5, 7].
- **R** трактуется как прокси институциональной реакции: индексы верховенства права, качества демократии и полицейской практики, а также наблюдаемые реакции на эпизоды волнений (применение силы, аресты, санкции) [8, 18, 19].

Мы используем панель «страна–квартал», где наблюдаем:

- счётчики насильственных событий по ACLED на душу населения [9];
- агрегированные по странам/регионам метрики правоприменения платформ по открытым отчётам о транспарентности [5, 7];
- институциональные индексы V-Dem (индекс либеральной демократии, подиндексы по гражданским свободам и верховенству права) [8], World Justice Project и родственные источники;
- базовые контроли: урбанизация, медианный возраст, безработица, неравенство доходов, проникновение интернета.

Мы оцениваем спецификации вида:

$$\log Violent_{c,t} = \alpha_c + \gamma_t + \beta_1 E_{c,t-1} + \beta_2 R_{c,t-1} + X'_{c,t-1}\delta + \varepsilon_{c,t},$$

где с индексирует страны, t - кварталы, α_c — страновые фиксированные эффекты, γ_t — временные эффекты, а $X_{c,t-1}$ — вектор контролей. Прокси $E_{c,t}$ и $R_{c,t}$ нормируются в диапазон [0, 1] по наблюдаемым распределениям. Полные определения переменных и проверки устойчивости представлены в Разделе 4 и Приложениях B-C.

Цель этого шага — не заявить строгую причинную идентификацию, а получить правдоподобное отображение от эмпирически наблюдаемых «мягкого» и «жёсткого» контроля к диапазонам E и R используемым в ABM. Сопоставление смоделированных и эмпирических градиентов насилия по режимам (низкие/высокие E, R) ограничивает пространство параметров, которое мы исследуем (см. Таблицу 1 и Раздел 4.4).

Таблица 1. Основные источники данных и охват выборки.

Источник	Индикатор	Годы / издание	Страны (N)
V-Dem *Institute — V-Dem dataset v15*	Liberal Democracy Index, субиндексы гражданских свобод и верховенства права	*2023–2025 (последние доступные)*	*≈ 180*
Pew Research Center *— Global Religious Landscape 2010*	Доли основных религиозных групп по странам	*2010*	*198–200*
ACLED *—* агрегированный country-month набор	Число протестных и насильственных событий по странам-кварталам	*2020–2024*	*≈ 170*
World Bank *— World Development Indicators*	Население, ВВП на душу, пользователи интернета (% населения)	*2022–2023*	*210+*
GSMA *— Mobile Connectivity Index*	прокси покрытия 4G / смартфонами	*2024*	*≈ 170*
SIPRI *— Military Expenditure Database*	Военные расходы (% ВВП)	*2024*	*≈ 180*
UNODC *— Homicide*	Уровень	*2015–2022*	*≈ 190*

statistics	умышленных убийств (на 100 000 населения)		

3.5 Валидация индексов компаньона (ABA / RCT)

ABM работает на макроуровне, тогда как компаньон реализуется на уровне индивидуального взаимодействия. Чтобы связать эти масштабы, мы определяем и валидируем три индекса, выводимых из пользовательских данных:

- **Индекс Frustration** — агрегируется из маркеров блокированных потребностей, повторяющихся конфликтных тем и самоотчётов о «залипании»;
- **Индекс Tension** — агрегируется из лингвистических маркеров возбуждения, срочности и воспринимаемой угрозы;
- **KPI capability-gain** — изменение функциональных исходов (участие, самоинициированные действия, эпизоды без конфликта) на заданных временных окнах.

Валидация идёт в два этапа:

1. **ABA-серии одиночных случаев:** малые N, высокая частота измерений, где каждый участник служит своим собственным контролем (A-B-A), с фокусом на внутрилеличностных изменениях Frustration/Tension и времени стабилизации.
2. **Пилотный RCT:** групповое сравнение компаньона и «активного контроля», где обе группы получают некоторую структурированную само-помощь, но только компаньон включает полные протоколы рефлексии и эскалационного затвора. План валидации для ABA/RCT-пилотов, портируемости и измерительной инвариантности суммирован в Таблице B1 (Приложение B).

В описываемой пилотной реализации мы намеренно ограничиваемся лишь семантическими и поведенческими маркерами — лингвистическими паттернами, отказом/принятием подсказок, длительностью эпизодов — чтобы избежать обязательного использования носимых устройств и понизить барьеры входа. Это даёт консервативные оценки эффекта: последующие волны пилотов должны добавлять

физиологические маркеры (HRV, суррогатные меры метаболической нагрузки) и проверять согласованность между индексами компаньона и телесной регуляцией.

3.5.1. A–B-дизайн

В A–B-дизайнах мы используем три фазы A–B–A’ общей длительностью 6–7 недель (обычно 2–3–2 недели). Компаньон активен в фазах A и B (в A — только наблюдение), и полноценно поддерживает в фазе B. Мы собираем:

- высокочастотные EMA-оценки (ecological momentary assessment) по Conflict/Tension;
- структурированные журналы эпизодов (начало/конец «официальных» эпизодов);
- полные логи подсказок и ответов пользователя.

Модели со смешанными эффектами с фазовым индикатором тестируют изменения уровней и наклонов между A и B и обратимость к A’ (отсутствие ухудшения) для каждого индекса. Частота сердечных сокращений (если доступна) используется для конвергентной валидации.

3.5.2. Дизайн пилотного RCT

Пилотный RCT длится 6 недель с индивидуальной рандомизацией в:

- **Companion arm** — полный компаньон с AP-Gate и протоколами trust/safety;
- **Active control arm** — приложение с психосоциальным контентом и простым отслеживанием настроения, но без рефлексивных напоминаний и эскалационного контура.

Участники заполняют базовые и итоговые опросники (PHQ-9, GAD-7, конфликтно-ориентированные пункты) и еженедельные чек-ины. Ежедневные логи взаимодействий поддерживают анализ конфликтных коммуникаций и приверженности. Мы заранее регистрируем первичные и вторичные исходы, планы анализа и правила остановки.

Помимо сценариев с явными запросами, мы предполагаем, что спонтанная речь вне прямых запросов к системе может рассматриваться как непрерывный поток проекций внутреннего состояния пользователя на внешние темы. Поэтому компаньон уделяет меньше внимания тому, *о чём* номинально разговор (политика, работа, отношения), и больше — повторяющимся эмоциональным паттернам и фрустрированным

потребностям, проходящим через темы. Это позволяет картировать напряжения и способности даже при отсутствии физиологических сенсоров.

3.6 Дизайн и длительность в макро–микро-связке

Индексы Frustration, Tension и энергетической нагрузки можно связать с макропараметрами через:

- отображение наблюдаемых изменений распределений индексов в сдвиги вероятностей аффективного заражения (Ps) и эскалации (Pe) в ABM;
- оценку того, какое покрытие компаньоном (доля населения, регулярно использующая систему) потребуется для достижения заданного снижения пикового или кумулятивного насилия при эмпирически калиброванных режимах (E, R).

В рамках текущего проекта мы ограничиваемся консервативной качественной симуляцией: результаты интерпретируются в диапазонах E и R, согласованных с наблюдаемыми страновыми данными, а покрытие компаньоном трактуется как экзогенный политический рычаг. Более детальная связка с валидированными индексами и физиологическими маркерами отнесена к последующей работе.

3.7 Клиентские фильтры и P2P-трафик

Минимальная ABM предполагает видимость платформ, контента и централизованной модерации, но на практике растущая доля высокорискованных коммуникаций протекает через peer-to-peer (P2P)-каналы (мессенджеры, зашифрованные группы). Это снижает прямую управляемость со стороны платформ, но не отменяет возможности клиентских фильтров.

Чтобы приблизить эту картину, мы вводим:

- локальный параметр фильтра E_{local}, отражающий способность компаньона демпфировать вредоносное заражение в P2P-каналах на уровне устройства;
- стресс-тесты, где некоторая доля рёбер помечается как «только P2P» и модерация на них осуществляется исключительно через E_{local}, а не через платформенные E.

В текущей симуляции мы лишь намечаем это расширение и приводим проверки устойчивости в Приложении A. Полноценное клиентское детектирование и смягчение для P2P-каналов — с учётом приватности и ограничений шифрования — специфицируется на уровне протоколов в Приложениях F и G и рассматривается как приоритетное направление для будущей реализации.

4. Результаты

4.1. Базовые паттерны и роль институтов

Сначала мы рассматриваем поведение модели в сценарии status quo (низкие E, низкие R, пренебрежимо малое покрытие компаньоном). В этом режиме система воспроизводит знакомые стилизованные факты из литературы по диффузии протестов: небольшие экзогенные шоки могут либо быстро затухать, либо, попадая в локально более плотные части сети, порождать мезомасштабные каскады протеста с эпизодической эскалацией в насилие. Распределение кумулятивного насилия по Монте-Карло-прогонам оказывается тяжёлохвостым: большинство траекторий демонстрируют умеренные пики, но ненулевая доля реализаций даёт крупные всплески доли агентов в состоянии Violent.

Повышение только E (платформенная фильтрация) при низком покрытии несколько сдвигает распределение: экстремальные пики становятся реже, но медианное пиковое насилие уменьшается лишь умеренно. Напротив, усиление институциональной реакции R (при низком E) сильнее влияет на кумулятивное насилие: крупные вспышки сокращаются по длительности за счёт более быстрой де-эскалации, но высота пика может оставаться значительной. Это соответствует интуиции из эмпирических работ: полицейская и «rule-of-law»-ёмкость сильнее влияет на длительность и последствия волнений, чем на сам момент их зажигания [8, 16, 18].

Когда и E, и R низкие, модель генерирует полосу исходов, совместимую с эмпирическими панелями: страны с сопоставимыми уровнями фрагментации могут иметь весьма разные уровни насильственных событий в зависимости от качества институтов и скорости локализации протестов [9, 18]. Это служит sanity-check’ом: ABM не создаёт «насилие из ниоткуда», а усиливает структурные и

институциональные различия так, что это согласуется с регрессионными паттернами (Приложение B, Рисунки S1, S2, S3, S4).

4.2. Companion + Charter против status quo

Главный интерес представляет сравнение трёх режимов на решётке (E, R):

1. **Status quo:** низкие E, низкие R, компаньона нет.
2. **Только платформенная фильтрация:** более высокое E при пренебрежимо малом покрытии.
3. **Companion + Charter:** умеренное E, временно усиленное R в «пиковом окне покрытия» (25–40 %), затем контролируемый откат.

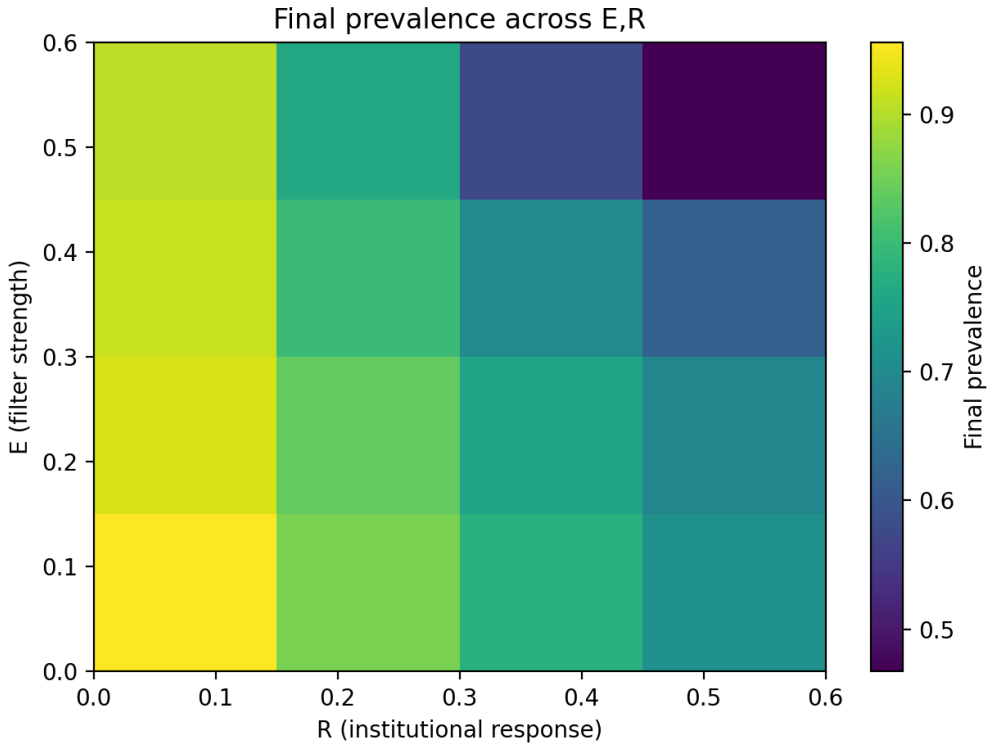

Рисунок 1. Пиковая доля агентов в состоянии Violent как функция E и R (Companion + Charter против status quo).

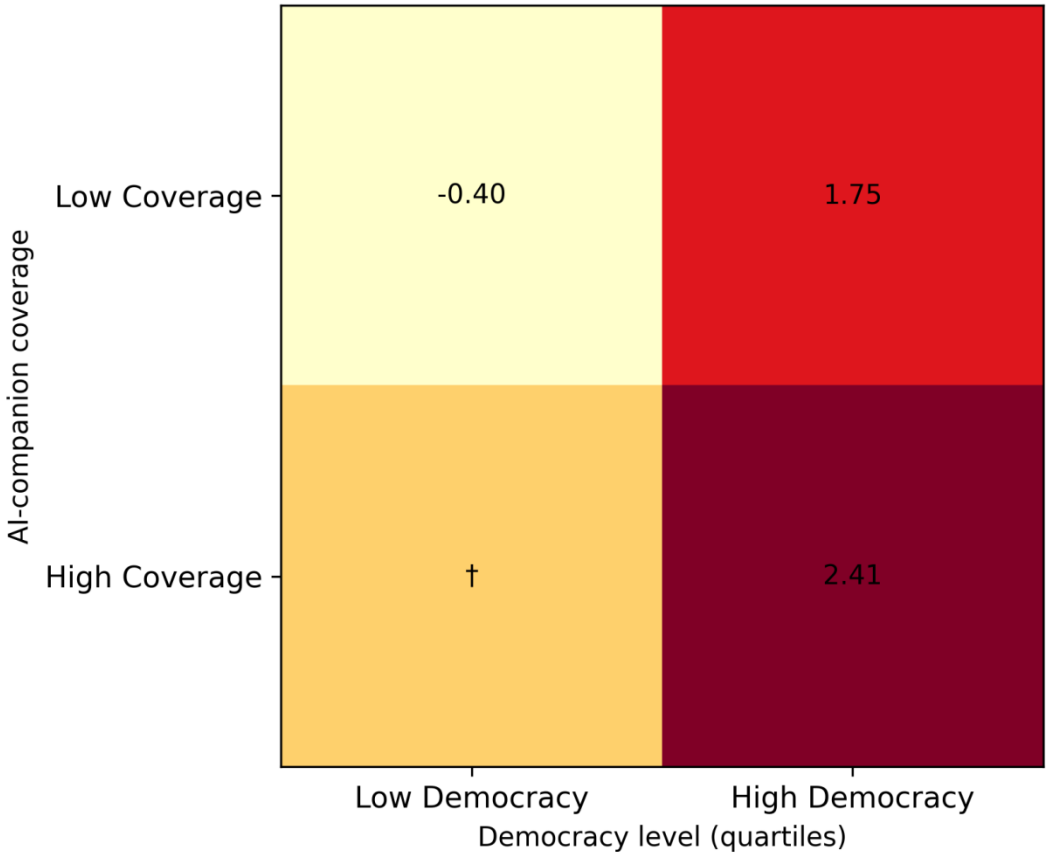

Рисунок 2. Кумулятивное насилие за 36 месяцев для разных комбинаций E и R.

На Рисунке 1 показана средняя пиковая доля агентов в состоянии Violent по сценариям, на Рисунке 2 — кумулятивное насилие для той же решётки значений (E, R). Вкратце:

- Переход от status quo к «только платформенной фильтрации» уменьшает пиковое насилие примерно на 10–15 % в среднем, но оставляет широкий интервал неопределённости; экстремальные траектории по-прежнему возможны.
- Добавление архитектуры Companion + Charter даёт дополнительное снижение пика насилия примерно на 20–30 % и сопоставимое снижение кумулятивного насилия на горизонте 36 месяцев.
- В комбинированном режиме распределение исходов становится менее тяжёлохвостым: крайне насильственные траектории редки, большинство прогонов группируется в более узкой полосе умеренных пиков.

Важно, что эти траектории не являются «висящими в воздухе» артефактами ABM. В страново-квартальном панеле (Приложение B) мы видим похожий рисунок, когда проксируем E и R из отчётов о транспарентности платформ и институциональных индексов и добавляем простую меру горизонтальных связей H (волонтёрство,

гражданские ассоциации). Более высокий H систематически смягчает эффект резких интервенций только по E и усиливает выигрыш комбинированного режима Companion+Charter; низкий H, напротив, оставляет контексты более хрупкими и ближе к «платформенному» коридору. Представительная траектория E, R и покрытия, вместе с соответствующей динамикой KPI Violent/Active, показана на Рисунках 4 и 5.

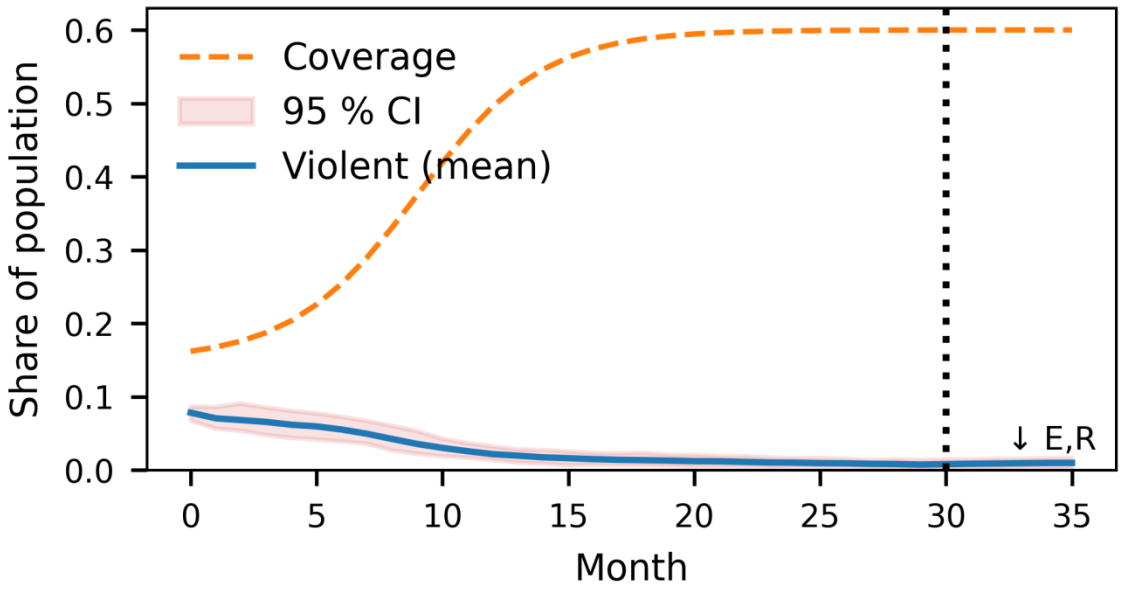

Рисунок 4. Динамика E, R и покрытия компаньоном при сценарии плавного отката (горизонт 36 месяцев).

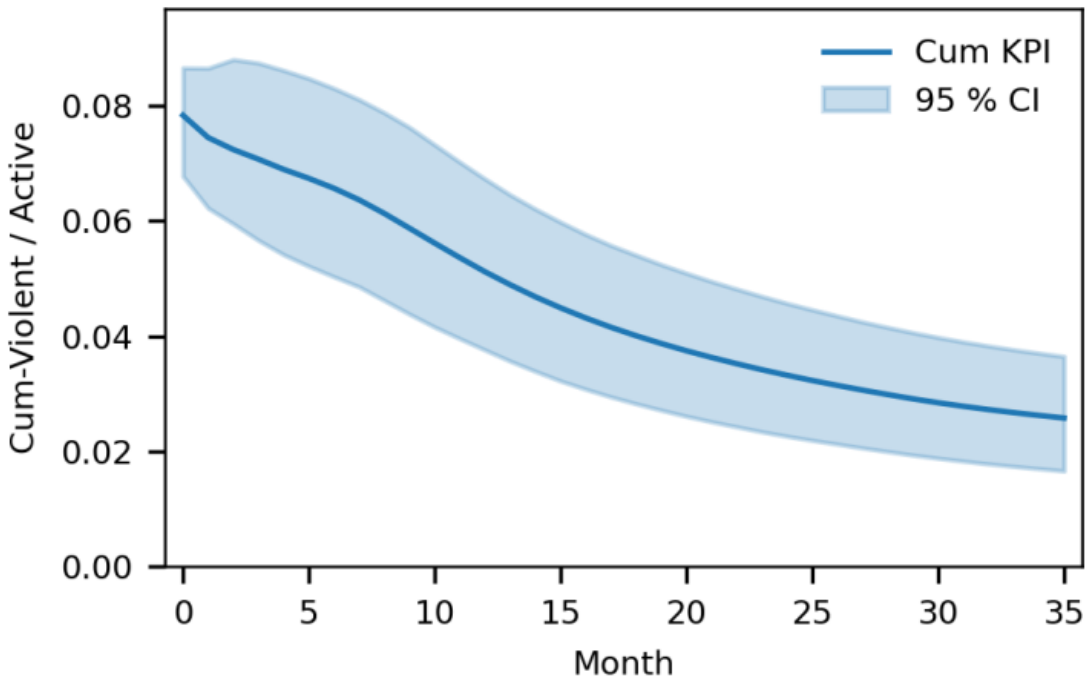

Рисунок 5. Траектория KPI Violent/Active за 36 месяцев в режиме Companion + Charter.

Совместный эффект возникает за счёт двух механизмов:

1. **На микроуровне** этический затвор компаньона блокирует часть событий заражения до их эскалации (через член E·coverage(t)), но только когда пользователь находится в высоком напряжении и явно просит поддержки (Раздел 5, Приложение C).
2. **На макроуровне** Хартия временно повышает R, когда покрытие проходит диапазон 25–40 %, то есть именно там, где заражение наиболее опасно: достаточно агентов уже «онлайн», чтобы каскады стали возможны, но ещё недостаточно, чтобы сформировать стабилизирующие обратные связи.

Таблица 2 суммирует четыре канонических политических режима, которые мы изучаем — status quo, только платформенная фильтрация, комбинированная политика E+R без компаньона и полная архитектура Companion + Charter — выделяя их рычаги, ожидаемый эффект на насилие и главные риски.

Таблица 2. Сценарии по группам дохода и институциональным режимам (E, R, целевые KPI).

Режим / группа дохода	**Базовая настройка (E, R)**	**Пиковое окно**	**Постпиковая настройка**	**Целевой KPI**	**Ориентировочная окупаемость CAPEX**	**Примечания**
LDC (низкий доход, хрупкие институты)	*E ≈ 0.10–0.15; R ≈ 0.05–0.10*	временное повышение R до 0.20 в периоды острых кризисов	возврат к E≈0,15; R≈0,10 с постепенным укреплением горизонтальных связей H	снижение отношения Violent/Active до 0.08 за 3–4 года	5–7 лет	фокус на базовом покрытии, безопасности школ, минимальном соблюдении Хартии
MIC (страна со средним доходом, гибридный / конкурентн	E ≈ 0.20–0.25; R ≈ 0.20–0.30	кратковременное повышение R до 0.40 в ответ на крупные	постепенное снижение R при улучшении индикатор	Violent/Active ≤ 0.10 при избегании долгосрочных репрессий	3–5 лет	баланс между контролем и легитимностью; пилоты

о-авторитарный режим)		протесты	ов насилия; E удерживается стабильным			компаньона сначала в «безопасных» секторах
MIC (страна со средним доходом, электоральная демократия)	E ≈ 0.20; R ≈ 0.15–0.20	умеренное повышение E и R в электоральные циклы	возврат к базовому E, снижение R с явными sunset-клаузами	Violent/Active ≈ 0.06; стабильное или растущее участие	≈ 3–4 года	акцент на внедрении компаньона в образование и муниципальные сервисы
HIC (высокодоходная страна, консолидированная демократия)	E ≈ 0.25–0.30; R ≈ 0.10–0.15	ограниченное ужесточение платформенного E; R в основном через судебные механизмы	возврат к более мягким E, R; реинвестирование сэкономленных средств в профилактику и пилоты компаньона	Violent/Active ≈ 0.04 или ниже; снижение экономической стоимости насилия (% ВВП)	≈ 2–3 года	акцент на «дивидендах мира» и экспорте лучших практик через Хартию

В результате система проводит больше времени в зоне «мягких вмешательств» на плоскости (E, R) и меньше — в недорегулированных областях. Важно, что благоприятный эффект не опирается на максимальные значения E или R: чрезмерное повышение любого параметра ведёт к убывающей отдаче и, в некоторых смоделированных режимах, к backlash-подобной динамике, когда подавленный протест позже возвращается в более концентрированной форме (Приложение A, Рисунок S4).

4.3. Робастность ранжировок в пространстве параметров

Центральный вопрос для политико-ориентированных ABM — робастность: сохраняются ли относительные выводы о «лучших» и «худших» режимах при варьировании ключевых параметров и начальных условий?

Мы решаем это в три шага.

Однофакторная чувствительность. В Приложении A, Рисунки S6, S7, показаны Монте-Карло-тесты чувствительности, в которых мы варьируем P_S и P_E в их правдоподобных диапазонах (0.15–0.35) и повторяем по 2000 прогонов для каждого значения. Основной результат: упорядочение сценариев сохраняется — режим Companion + Charter стабильно даёт более низкие пиковое и кумулятивное насилие, чем (i) status quo и (ii) только платформенная фильтрация, даже когда заражение и эскалация заметно сильнее или слабее. Доверительные интервалы ожидаемо расширяются, но не настолько, чтобы перевернуть ранжировку.

Начальные «семена» и структура сети. Мы также варьируем (а) начальную долю «семенных» Violent-узлов (0.5–2 %) и (б) случайные сиды графа. Относительное упорядочение режимов снова остаётся стабильным: более интенсивное начальное «заражение» повышает средний уровень насилия во всех сценариях, но Companion + Charter продолжает доминировать как по медиане, так и по верхним квантилям. В дополнительных проверках (не показаны в основном тексте) мы реплицируем этот рисунок на scale-free и кластеризованных сетях; детали приведены в Приложении A.

Проверка эмпирической калибровки. Наконец, мы подставляем эмпирически калиброванные значения E_c и R_c (полученные из отчётов платформ и институциональных индексов, как в Разделе 3.4) в ABM и проверяем, попадают ли смоделированные страново-квартальные точки в реалистичные диапазоны числа насильственных событий. Модель не претендует на точечное предсказание, но совместное распределение смоделированных исходов и наблюдаемых ACLED-счётчиков качественно сходно: страны со слабым верховенством права и низким enforcement'ом концентрируются в зоне высоких уровней насилия, тогда как страны с сильными институтами и умеренной фильтрацией — в нижнем левом углу (Приложение B).

В совокупности эти проверки поддерживают тезис о том, что относительное преимущество архитектуры Companion + Charter робастно к разумной неопределённости в микропараметрах и начальных условиях.

4.4. Макроэкономические последствия

Помимо метрик насилия, мы оцениваем макроэкономический эффект, переводя сокращение насильственных событий в изменение экономической стоимости насилия. Используя существующие оценки из SIPRI [10] и Institute for Economics & Peace [11], мы строим стилизованный сценарий, в котором 2 % ежегодных военных расходов постепенно перераспределяются на масштабирование инфраструктуры ИИ-компаньона.

Excel-модель (Приложение E; файл economic_model_sensitivity.xlsx в репозитории) вычисляет траекторию чистой приведённой стоимости (NPV) при разных предположениях о:

- эластичности экономического выпуска по отношению к снижению насилия;
- стоимости эксплуатации крупномасштабных сервисов компаньона (вычисления, поддержка, надзор);
- скорости, с которой институциональные сбережения материализуются.

Рисунок S8 суммирует траектории NPV в 30 Монте-Карло-репликах для базового сценария:

- медианное NPV становится положительным примерно через 8 лет и продолжает расти на горизонте 20 лет;
- даже при консервативных параметризациях долгосрочный годовой эффект стабилизируется на уровне порядка 0,5–0,7 % ВВП, что соответствует headline-оценкам в литературе по экономике мира [11];
- при более оптимистичных предположениях о накопительном росте продуктивности (например, через улучшение психического здоровья и снижение абсентеизма) выигрыши по NPV могут быть выше, но мы не полагаемся на них в основной аргументации.

Эти оценки неизбежно приблизительны и иллюстративны: их задача — не предсказать точный рост ВВП, а показать, что даже умеренное снижение экономической стоимости насилия может сделать политику Companion + Charter фискально привлекательной по сравнению со status quo. Иными словами, архитектура не только нормативно мотивирована, но и потенциально самоокупаема при масштабировании за пределы ранних пилотов.

4.5. Итог

При широком диапазоне допущений симуляции показывают, что:

- Чисто платформо-центричный подход (высокое E, низкое R, без компаньона) обладает ограниченным эффектом на пиковое и кумулятивное насилие и несёт риск обратной реакции, если его «доворачивать» слишком сильно.
- Комбинированная архитектура, где ненасильственный компаньон снижает микромасштабное заражение, а Хартия координирует временное повышение R в критические окна, устойчиво снижает и пиковое, и кумулятивное насилие.
- В экономическом выражении даже консервативное сокращение насилия даёт нетривиальные выигрыши по NPV на горизонте 10–20 лет.

Следующий раздел переходит от агрегированных паттернов к дизайну самого компаньона: тому, как мы обеспечиваем ненасильственный характер микровзаимодействия, поддерживаем субъективную целостность и соблюдаем институциональные и культурные ограничения.

5. Обсуждение

Полученные результаты можно прочитать на трёх уровнях:
(i) нормативном — какое благо мы вообще пытаемся поддерживать;
(ii) институциональном — как компаньон взаимодействует с фильтрами и мощностью институтов;
(iii) экономическом — насколько масштабное внедрение вообще социально посильно.

5.1. Нормативная рамка: поддерживать «правоту», а не послушание

В нашей архитектуре компаньон задуман не как «мягкий полицейский», а как зеркало плюс инструмент разделения для пользователя. Его первичная цель — поддерживать целостное, ненасильственное самоотношение и расширять набор жизнеспособных вариантов действий, а не навязывать определённую идеологию.

Это согласуется с несколькими ветвями современной этики. Во-первых, с подходами способностей и эвдемоническими подходами, где центральный вопрос звучит не как

«подчинился ли агент норме?», а как «были ли у человека внутренние и внешние ресурсы действовать в соответствии со своими обдуманными ценностями?» [1, 27, 28]. Во-вторых, с психологическими моделями саморегуляции и нейровисцеральной интеграции, которые рассматривают эмоциональную гибкость и способность снижать возбуждение как базовый компонент благополучия [29, 30].

В этой рамке роль компаньона состоит в следующем:

1. **Сделать релевантные внутренние состояния наблюдаемыми.**
   Через мягкие подсказки и рефлексивные вопросы он помогает пользователю замечать паттерны фрустрации, срочности и воспринимаемой угрозы, которые в противном случае оставались бы неявными.
2. **Поддерживать разделение между импульсом и действием.**
   Такие протоколы, как Autonomy-Preserving Gate (AP-Gate), создают структурированную паузу между намерением на фоне высокого напряжения и любым необратимым шагом (публикация поста, финансовая транзакция, поездка в рискованное место). Конкретная пошаговая реализация AP-Gate в терминах диапазонов риска, разрешённых действий и предохранителей эскалации приведена в Приложении D. Помимо непосредственной логики безопасности, эта структурированная пауза одновременно является семантическим и телесным вмешательством: она даёт пользователю шанс заметить, что делает его тело, назвать расщепление между импульсом и намерением и получить доступ к альтернативным рамкам, которые были закрыты в пик возбуждения.
3. **Подкреплять успешную саморегуляцию.**
   «Прививочные подсказки» напоминают пользователю эпизоды, когда он уже справлялся с похожей ситуацией без насилия, опираясь на позитивную предсказательную ошибку, а не на наказание.

Ключевое утверждение здесь не в том, что компаньон «знает лучше» пользователя, а в том, что он может расширять пространство ненасильственных, сохраняющих достоинство ответов именно в те моменты, когда собственные регуляторные ресурсы человека находятся под нагрузкой. Это ближе к расширенному уму [27], чем к патерналистскому наставнику.

В последующих разделах мы будем говорить о насилии не только во внешнем, социальном смысле, но и во внутреннем. На внутриличностном уровне мы различаем

внутреннее насилие и **внутреннее усилие**: в первом случае одна часть личности принуждает или стыдит другую, добиваясь подчинения; во втором — разные потребности признаются, и человек ищет шаги, которые улучшают положение хотя бы одной из них, не стремясь намеренно разрушить другую. Компаньон явно спроектирован как устройство, поддерживающее внутреннее усилие, а не внутреннее насилие: он называет потребности, расширяет поле вариантов и поощряет минимально достаточные действия, с которыми пользователь может согласиться из позиции «внутреннего взрослого», а не подталкивает его к послушанию. Приложение K развивает это различие в более клиническом ключе и связывает его с энергетической метрикой E* и паттернами контакта.

5.2. Институты против фрагментации

Эмпирически регрессионные модели и Random Forest (Раздел 4.2, Приложение B) показывают устойчивый паттерн (Рисунок 3):

- качество демократических институтов (V-Dem v2x_libdem) существенно сильнее предсказывает уровень насильственных событий, чем религиозная фрагментация;
- перекрёстный член «фрагментация × демократия» статистически незначим; высокая разнообразность *сама по себе* не «взрывается» в насилие, пока институциональное качество остаётся высоким.

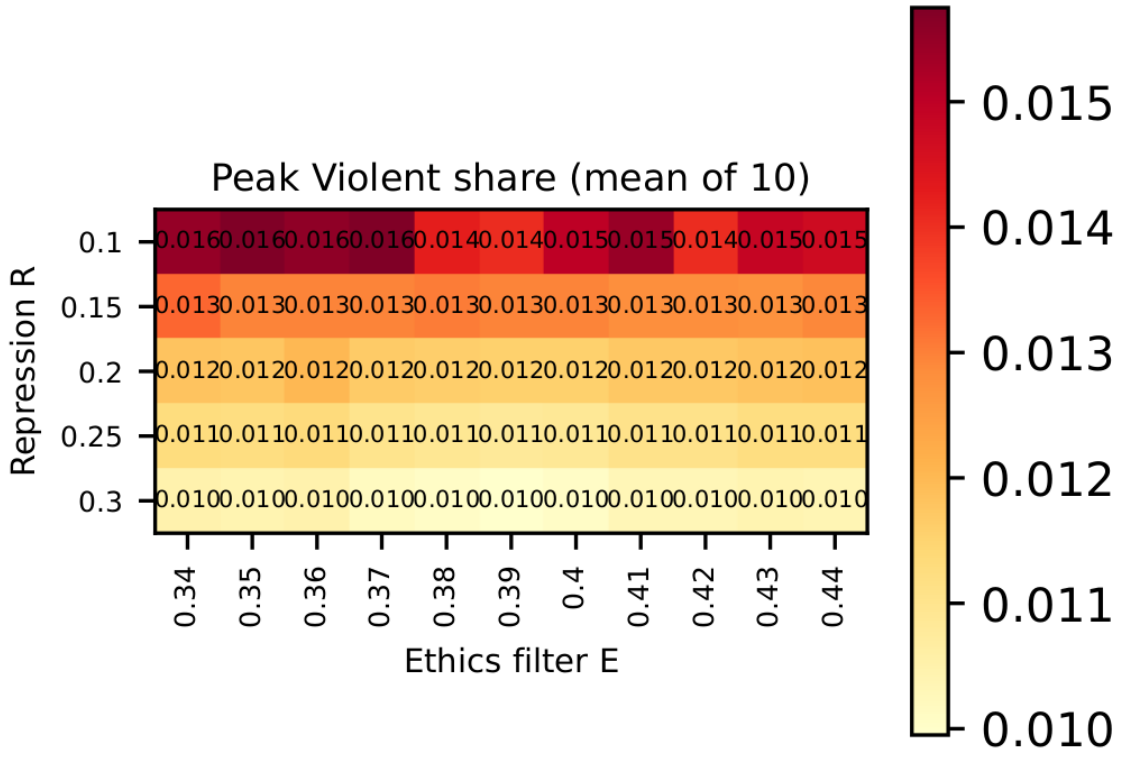

Рисунок 3. Демократия, религиозная фракционализация и политическое насилие (панель ACLED + V-Dem, страновой уровень).

Это соответствует длинной традиции в политической экономии, которая подчёркивает инклюзивные институты и систему сдержек и противовесов как главный «демпфер» конфликта [16, 17, 18]. С этой точки зрения компаньон не является заменой институциональной реформы. Скорее, он выступает локальным усилителем уже существующей нормы ненасильственного управления конфликтом:

- когда институты сильные, компаньон облегчает людям реализацию ненасильственных сценариев, которые уже поддерживаются правом, медиа и социальными нормами;
- когда институты слабые или хищнические, эффект компаньона более хрупок и может быть перекрыт прямыми репрессиями или скоординированной пропагандой.

Оценки важности переменных в модели случайного леса (Рисунок 8) подтверждают, что институциональное качество доминирует над религиозной фракционализацией при прогнозировании насильственных событий.

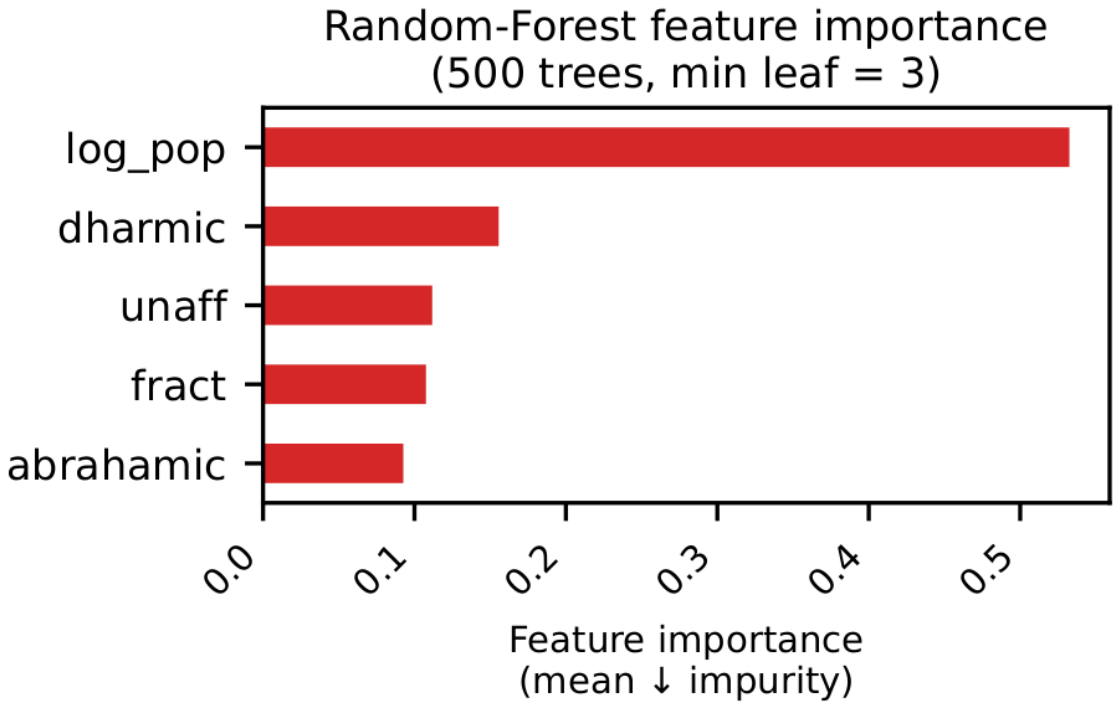

Рисунок 8. Важность переменных в модели Random Forest для насильственных событий (верховенство права и качество демократии доминируют религиозную фрагментацию и базовые контроли).

Лог–лог-зависимость между качеством демократии и насильственными событиями визуализирована на Рисунке 6.

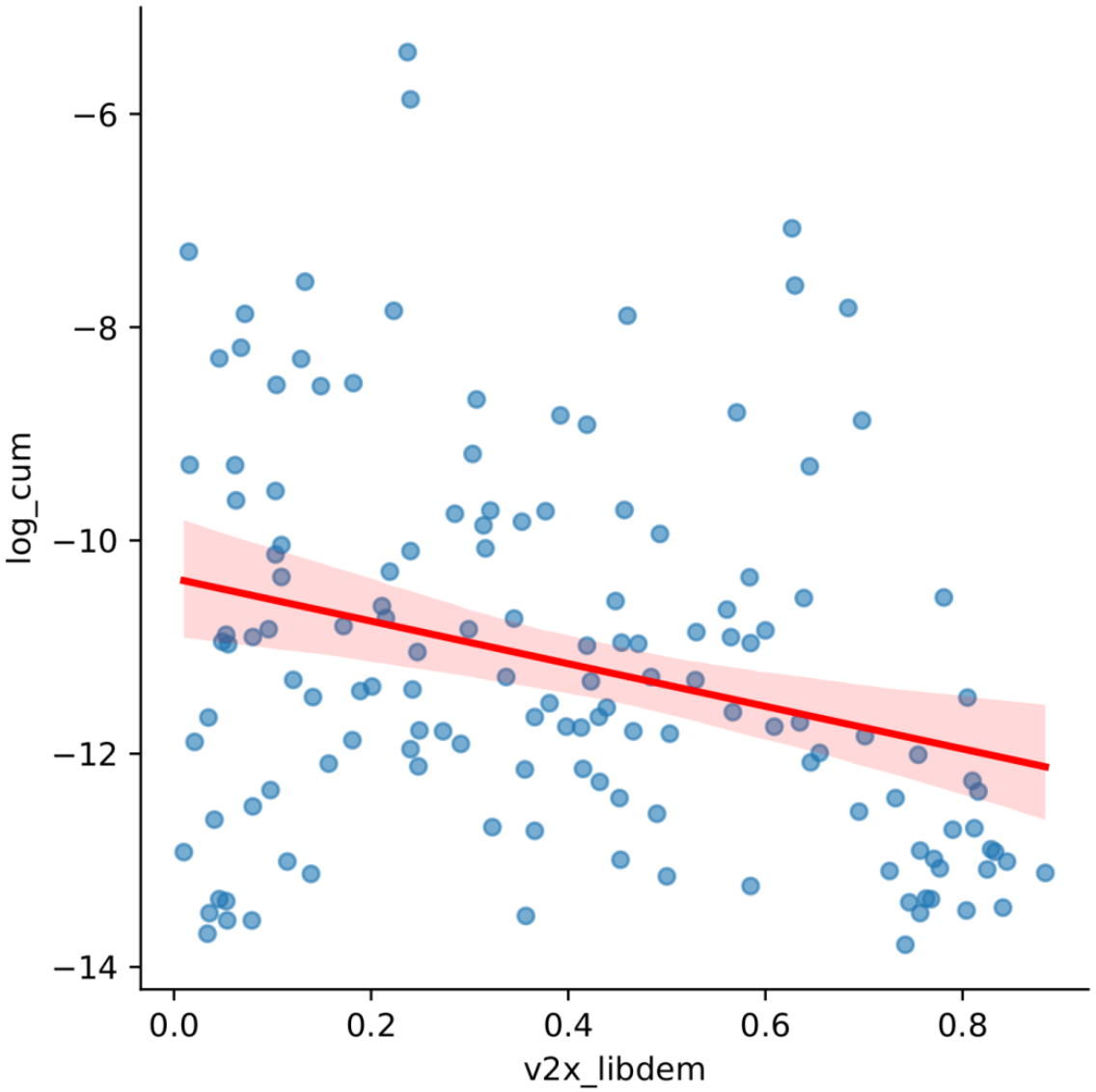

Рисунок 6. Демократия и политическое насилие: лог–лог-зависимость между индексом либеральной демократии и числом насильственных событий на 100 000 жителей (ACLED + V-Dem).

Иначе говоря, одна и та же технология эмоциональной поддержки может либо способствовать стабилизации плюралистической демократии, либо «сглаживать углы» авторитарного режима. Именно поэтому ограничения со стороны управления (Раздел 6) в такой архитектуре не факультативны.

5.3. «Парадокс среднего дохода»

Если разбить данные ACLED–V-Dem по группам дохода (Приложение B, Таблица B1), возникает знакомый по другим областям паттерн [23, 24]:

- страны с низким доходом часто характеризуются высоким базовым уровнем насилия, но ограниченной цифровой проникновенностью;
- высокодоходные демократии сочетают высокую связанность с сильными институтами и относительно низким уровнем летального насилия;
- страны с верхним среднем доходом демонстрируют наиболее тревожную конфигурацию: быстро растущая подключённость, значительная поляризация и лишь частично консолидированные институты. Эта конфигурация суммирована на Рисунке 7:

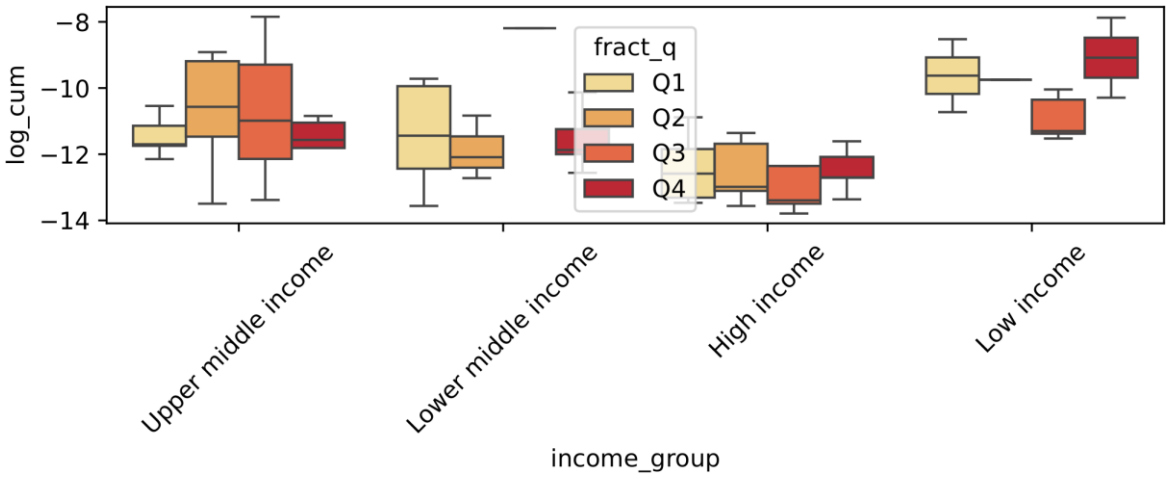

Рисунок 7. Политическое насилие по группам дохода и религиозной фракционализации (ACLED + V-Dem; агрегирование по стране-году).

В таких условиях агрессивная платформенная фильтрация (высокое E) без компаньона может снизить уровень открытой hate speech, но оставить базовые обиды нетронутыми — или даже вытеснить мобилизацию в зашифрованные каналы. Обратная конфигурация — компаньон без каких-либо структурных фильтров — может помогать отдельным пользователям, но быть просто перегруженной объёмом и виральностью разжигающего контента.

Наши симуляции показывают, что наиболее перспективная конфигурация для таких «средних» режимов:

- умеренная, прозрачная фильтрация (E порядка 0,25–0,35 с понятным внешним надзором);
- постепенное расширение покрытия компаньоном до 30–40 % населения;
- временное усиление институциональной реакции (R) именно в период этого пикового окна покрытия, с последующим откатом.

Такое сочетание снижает пиковое и кумулятивное насилие, не запирая систему в режиме постоянного «жёсткого» контроля. Именно в этих странах пилоты второго этапа (Phase II), если они вообще будут предприняты, вероятнее всего дадут наибольший предельный выигрыш — и одновременно будут сопряжены с самой высокой политической чувствительностью.

5.4. Компаньон, суверенитет и доверие

Центральный вопрос для любой крупномасштабной эмоциональной ИИ-системы — суверенитет: чьим целям в конечном счёте служит система? В нашей рамке ответ сознательно разделён по уровням.

На микроуровне компаньон связан строгой логикой сохранения автономии:

- никаких вмешательств без явного согласия в неэкстренных контекстах;
- никакой скрытой манипуляции предпочтениями;
- понятные способы поставить на паузу, удалить логи и сменить поставщика (Приложение G).

На мезо- и макроуровнях суверенитет выражается через:

- наднациональную Хартию, которая задаёт красные линии того, что считается недопустимым вмешательством (например, политический микротаргетинг, профилирование по расе или религии) и привязывает их к существующим правозащитным инструментам [25, 26, 28];
- федеративное дообучение (Federated Fine-Tuning, FF-T), при котором локальные провайдеры адаптируют компаньона к культурным нормам и языкам без централизации сырых диалогов (Раздел 6.3, Приложение F.9);

- модель «аудит как сервис»: независимые, статистически строгие проверки того, что развёрнутые системы действительно соблюдают эти ограничения (Приложение H).

Доверие здесь не предполагается как психологическое состояние, а трактуется как emergent-свойство социотехнического стека: пользователи, провайдеры, регуляторы и организации гражданского общества могут на разных уровнях глубины проверять, что система остаётся в пределах задекларированного контура поведения.

5.5. Макроэкономический баланс

С макроэкономической точки зрения архитектура конкурирует минимум за три дефицитных ресурса:

1. **Индивидуальное время и внимание.**
   Время, проведённое во взаимодействии с компаньоном, — это время, не проведённое в других цифровых или офлайн-активностях.
2. **Институциональная ёмкость.**
   Реализация Хартии, аудит и FF-T требуют специализированных кадров и политического ресурса.
3. **Вычисления и инфраструктура.**
   Эксплуатация эмоционально поддерживающих компаньонов в масштабах населения имеет нетривиальный энергетический и аппаратный след (Приложение E).

Используя консервативные допущения и данные SIPRI, OECD и McKinsey [20, 21, 22, 23, 24], наш иллюстративный расчёт (Приложение E) показывает, что:

- даже в медианном сценарии чистый эффект на ВВП может оказаться в диапазоне +0,2–0,3 % в год, сочетая рост продуктивности за счёт лучшего психического здоровья со снижением затрат на насилие и безопасность;
- фискальное пространство для такой программы, в принципе, можно высвободить, перераспределив небольшую долю текущих военных и полицейских расходов, не вытесняя ключевые социальные услуги.

Эти оценки сознательно поданы как иллюстративные и зависят от успеха реальных пилотов. Тем не менее, они делают правдоподобным, что ориентированная на профилактику архитектура такого типа экономически жизнеспособна, а не просто нравственно привлекательна.

6. Этическая рамка и управление (два уровня)

Этический анализ повторяет двухуровневую структуру архитектуры. На человеческом уровне мы спрашиваем, уважает ли компаньон агентность, достоинство и разнообразие пользователей. На институциональном уровне — как система взаимодействует с правом, платформами и геополитической реальностью.

6.1. Права человека, агентность и ненасилие

Мы исходим из минимального набора обязательств, которые, по крайней мере в принципе, разделяются основными правозащитными инструментами и плюралистическими этическими традициями [25, 26, 28, 33, 34, 35, 36, 37]:

1. **Никакого намеренного вреда.**
   Компаньон не может поощрять самоповреждение, насилие по отношению к другим или системную дискриминацию.
2. **Уважение агентности.**
   Пользователь сохраняет право принимать плохие решения — включая выключение системы — до тех пор, пока эти решения напрямую не ставят под угрозу других.
3. **Прозрачность и понятность.**
   Пользователь должен в общих чертах понимать, почему компаньон делает то или иное предложение и на каких данных оно основано.
4. **Плюрализм.**
   Система должна оставаться совместимой с широким спектром мировоззрений (религиозных, светских, коллективистских, индивидуалистских), пока они не нарушают запрет на насилие.

Эти обязательства операционализируются через:

- **протокол «зеркало/разделение»** (Приложение F): компаньон сначала отражает собственные формулировки пользователя, затем мягко прощупывает альтернативы;
- **градуированные режимы вмешательства** (Приложение G): от «наблюдать» и «спрашивать в ответ» до «предлагать» и, в редких случаях, «настоятельно рекомендовать связаться с человеком»;
- **red-line-ограничители**, встроенные в оркестрационный слой [32]: даже если базовая LLM галлюцинирует или уходит в дрейф, компаньон не будет выдавать контент, пересекающий заранее заданные пороги безопасности.

Цель — сместить баланс сил в пользу тех пользователей, кто сейчас наименее защищён в цифровых экосистемах: людей с высокой эмоциональной нагрузкой, низкой цифровой грамотностью или слабой институциональной защитой.

6.2. Управление фильтрами и институциональной реакцией

Параметры **E** (фильтрация) и **R** (институциональная реакция) — не чисто технические ручки; в них зашиты ценностные выборы о допустимых компромиссах между свободой выражения, снижением вреда и приватностью [2, 3, 23, 24].

В предлагаемой архитектуре:

- **E** задаётся Хартией с учётом общественных консультаций и эмпирических данных о вреде контента. Страны сохраняют право выйти из режима или ввести более жёсткие настройки, но не могут скрытно опускать E ниже базового уровня Хартии, продолжая при этом заявлять о соответствии.
- **R** является разделённой ответственностью государств и платформ. Хартия не предписывает конкретную тактику полицейского реагирования, но требует **минимально достаточного ответа** на всплески насилия, зафиксированного через прозрачные индикаторы (например, своевременное удаление прямых призывов к насилию, защита уязвимых групп).
- **«Аудит как сервис»** (Приложение H) обеспечивает внешний контроль: независимые лаборатории при юридически гарантированном доступе к данным могут проверять, соответствует ли фактическое поведение заявленным настройкам E и R.

Нормативная интуиция проста: если общество сознательно решает **не** защищать себя от алгоритмически усиленных призывов к насилию, это должно быть результатом явного политического решения, а не побочным продуктом непрозрачной оптимизации рекламной выручки.

6.3. Федеративное дообучение и культурный суверенитет

Язык и культура — это не просто косметические параметры. Многие маркеры напряжения, исключённости и угрозы глубоко контекстуальны [15, 33, 34, 35, 36, 37, 38]. Поэтому любая попытка запустить «универсальный» компаньон из одной глобальной модели и для всех сразу — этически и технически ошибочна.

Вместо этого мы набрасываем **схему федеративного дообучения (Federated Fine-Tuning, FF-T)** (Приложение F.9):

- локальные провайдеры дообучают оркестрационный слой и, где нужно, небольшие адаптеры поверх базовой LLM на локальных данных (с согласия пользователей);
- наверх передаются только обновления модели, а не сырые диалоги;
- межпровайдерский протокол обеспечивает, что эти обновления не вводят новые векторы смещения или манипуляции.

Это даёт как минимум три преимущества:

1. **Культурная подстройка.**
   Подсказки и примеры можно адаптировать под местные идиомы, юмор и табу.
2. **Минимизация данных.**
   Нет необходимости, чтобы одна компания или государство хранили глобальный корпус эмоционально размеченных диалогов.
3. **Плюрализм провайдеров.**
   Разные НКО, государственные структуры и коммерческие организации могут конкурировать по качеству и надёжности **внутри общего контура безопасности**.

В то же время FF-T — не панацея. В сильно централизованных авторитарных режимах та же инфраструктура теоретически может быть использована, чтобы подстроить компаньонов под репрессивные государственные нарративы. Это напрямую ведёт к вопросу о злоупотреблении.

6.4. Риски злоупотребления и политические ограничения (сводка)

Полный анализ сценариев злоупотребления дан в Приложении I. Здесь мы лишь суммируем главные пункты.

- **Жёсткие авторитарные режимы.**
  Если тот же технический стек разворачивается без ограничений Хартии и AP-Gate, он может превратиться в инструмент тонкого поведенческого контроля, который отводит граждан от инакомыслия и подталкивает к лояльному режиму поведению.
- **Мягкие авторитарные и гибридные режимы.**
  Селективное внедрение (например, только для лояльных групп) или предвзятая настройка E и R могут углублять неравенство и поляризацию.
- **Международная напряжённость.**
  Трансграничное предоставление компаньонов создаёт юрисдикционные конфликты: чьё право применяется, когда пользователь в стране A пользуется компаньоном компании из страны B, выровненным под Хартию, подписанную странами C и D?

Наша позиция сознательно осторожна: мы **не** утверждаем, что архитектура должна быть немедленно развёрнута глобально. Скорее, мы предлагаем её как **шаблон переговоров** между демократическими государствами, платформами и гражданским обществом — с чёткими возможностями отказа для сообществ, которые воспринимают её как чрезмерно навязчивую.

6.5. Локальные кластеры и обратный поток компетенций (сводка)

Аналогично, Приложение J описывает модель локальных кластеров — города, университеты, НКО или сети больниц, — которые экспериментируют с компаньонами в рамках строгих этических ограничений и одновременно возвращают данные и экспертизу обратно в процесс работы Хартии.

Идея — **перевернуть привычное направление «наращивания потенциала»**: вместо экспорта готового решения из узкого круга богатых стран архитектура намеренно учится у регионов с высоким уровнем конфликтности и сильными практиками общинной устойчивости. Это открывает возможность **«обратного потока компетенций»**, когда лучшие практики ненасильственного управления конфликтами идут из Глобального Юга в Глобальный Север, а не наоборот.

6.6. Энергетическая инфраструктура как стратегическое горлышко

На протяжении статьи мы рассматривали стек компаньона так, будто его главные уязвимости — нормативные и институциональные: кто контролирует алгоритмы, как распределён доступ к праву эскалации, можно ли превратить AP-Gate в «мягкого полицейского». Ещё одно, менее обсуждаемое «горлышко» — **энергия**.

Масштабное внедрение эмоционального ИИ — это не только вопрос чипов и данных, но и стабильного электроснабжения. Недавние оценки показывают, что глобальные AI-нагрузки в ближайшие годы могут потребовать десятки гигаватт дополнительной мощности, усиливая давление на энергосистемы, которые уже испытывают стресс из-за электрификации и усилий по смягчению климатических изменений. На практике это часто приводит к жёсткой связке дата-центров с конкретными источниками базовой генерации (атомные станции, крупные ГЭС, специализированные газовые блоки) и к долгосрочным контрактам, которые привязывают AI-провайдеров к определённым регионам и политическим режимам.

С точки зрения нашей рамки это создаёт новый канал политического захвата. Если экосистема компаньонов зависит от небольшого числа юрисдикций, контролирующих дешёвую и стабильную энергию, эти юрисдикции получают рычаг воздействия не только на цены, но и на условия, при которых эмоциональный ИИ может быть отключён, ограничен или перенаправлен. Исторические аналогии с нефтегазовой политикой здесь показательны: энергетические зависимости многократно

использовались для дисциплинирования или фрагментации более демократичных партнёров, одновременно субсидируя внутренний авторитарный контроль.

В терминах параметров **E–R–C** энергетический захват может проявляться несколькими способами:

1. **Восходящее давление на C (покрытие).**
   Режим, контролирующий ключевые энергетические узлы, может выборочно облегчать или блокировать развёртывание компаньона в определённых странах или регионах, по сути «нормируя» доступ к регуляторной поддержке и саморегуляции через решение о том, где строить и поддерживать дата-центры.
2. **Условность по E (жёсткость фильтра).**
   Доступ к дешёвому хостингу может быть увязан с принятием конкретной конфигурации контент-фильтров и рекомендательных политик, включая скрытые возможности для пропаганды или цензуры. То, что выглядит как технический договор на хостинг, превращается в канал экспорта определённой нормативной позиции.
3. **Теневой контроль над R (реакцией).**
   В кризисных ситуациях энергетический провайдер может угрожать прекращением обслуживания дата-центров, которые отказываются сотрудничать с национальными спецслужбами или полицией. Даже если такие угрозы никогда не реализуются, они формируют ожидания локальных провайдеров и регуляторов.

Одновременно **энергоэффективность самих обществ** становится частью картины. Демократические режимы, инвестирующие в снижение хронической регуляторной нагрузки — через социальную защиту, поддержку психического здоровья и участие, — в нашей рамке более «энергоэффективны» на психосоциальном уровне: они тратят меньше человеческого времени в режиме выживания и могут направлять больше внимания в сторону развития. Это, как правило, коррелирует с большей способностью к инновациям в области чистой энергии, устойчивости сетей и технологий эффективности, усиливая и физические, и психосоциальные буферы против шоков.

Напротив, авторитарные системы часто удерживают население в хронически напряжённом, ориентированном на выживание состоянии. Это может облегчать краткосрочную мобилизацию и контроль, но энергетически затратно: значительная

часть доступной человеческой и институциональной энергии уходит на мониторинг, подавление и символические войны. Такие режимы, соответственно, склонны компенсировать внутреннюю неэффективность контролем внешних энергетических потоков — экспортом ископаемого топлива, критической инфраструктурой за рубежом и, потенциально, будущими «ИИ-ориентированными» мощностями генерации. В крайних случаях это может воспроизводить динамику «трубопроводной политики» уже на уровне AI-инфраструктуры: тот, кто владеет кабелем, реактором или охлаждающим озером, получает непропорциональное влияние на то, как эмоциональный ИИ разворачивается в других местах.

Для архитектуры компаньона из этого следуют два императива дизайна и управления:

- **Децентрализовать энергетическую зависимость, где возможно.**
  Отдавать предпочтение архитектурам, опирающимся на распределённые дата-центры умеренного размера и вычисления на стороне клиента, питаемые диверсифицированными и всё более возобновляемыми источниками энергии, а не на несколько гигантских хабов, привязанных к геополитически чувствительным объектам генерации.
- **Сделать происхождение энергии частью этической сертификации.**
  Надзор на уровне Хартии должен проверять не только алгоритмы и журналы эскалации, но и то, **где и как** питается базовая вычислительная инфраструктура. Система, зависящая от высоко концентрированной, политически рычаговой энергетики, может требовать более жёстких гарантий и прозрачности, чем система, встроенная в более устойчивый, плюралистичный энергетический ландшафт.

Короче говоря, будущее эмоционального ИИ связано не только с этикой алгоритмов, но и с политической экономией электронов. Любой реалистичный прогноз социального эффекта компаньона должен рассматривать энергетическую инфраструктуру как полноправную стратегическую переменную — наравне с E, R и качеством институтов.

7. Ограничения и направления дальнейшей работы

Предложенная в статье архитектура сознательно амбициозна и связана с существенными ограничениями. Мы группируем их в три блока: моделирование, измерение и управление.

7.1. Ограничения моделирования

Упрощённая агент-ориентированная модель.
Наша ABM абстрагируется от многих важных особенностей реальных обществ: медиасистем, партийной системы, репертуаров протеста и всего богатства идентичностной политики. Агенты гомогенны внутри каждого набора параметров; нет явного представления элит, организованных движений или переходов между онлайн- и офлайн-динамикой.

Калибровка и внешняя валидность.
Мы калибруем ключевые параметры (P_S, P_{Esc}, E, R) по данным на уровне стран и опираясь на стилизованные выводы литературы. Этого достаточно, чтобы исследовать качественные паттерны — например, относительную важность качества институтов по сравнению с фрагментацией, — но недостаточно для точного прогноза по какой-либо конкретной стране. Для любых практических политических рекомендаций необходимы реальные пилоты.

Фокус на одном типе вреда.
Модель фокусируется на коллективном политическом насилии. Мы не моделируем явно другие виды вреда (самоповреждение, домашнее насилие, стигматизацию меньшинств), которые могут быть столь же значимы для эмоционально поддерживающих компаньонов. Расширение архитектуры на более широкий портфель видов вреда остаётся открытой задачей.

7.2. Ограничения измерения и пилотов

Индексы Frustration и Tension.
Наши индексы строятся по лингвистическим маркерам, отказам/принятию подсказок и журналам эпизодов. Даже при валидации в ABA- и RCT-дизайнах (Раздел 3.5, Приложение B.1) они остаются прокси-показателями. Пользователи могут научиться

«играть» с системой или просто изменить стиль выражения без реального изменения внутреннего состояния.

Физиологические маркеры.
Интеграция HRV (RMSSD/SDNN) и других биомаркеров пока находится на стадии проектирования (Приложение B.2). Носимые устройства сами по себе вносят выборку с систематическими смещениями, создают риски для приватности и проблемы доступности. В настоящей статье мы сознательно рассматриваем физиологическую валидацию как последующую фазу.

Обобщаемость пилотов.
Первые экспериментальные пилоты, вероятнее всего, будут проводиться в относительно благополучных и цифрово грамотных популяциях (например, студенты университетов, городские профессионалы). Распространять их результаты на низкоресурсные контексты, зоны конфликтов или маргинализованные группы было бы необоснованно без целевых исследований.

7.3. Неопределённости управления и этики

Легитимность Хартии.
Мы предполагаем, что наднациональную Хартию можно создать с содержательным участием разных стейкхолдеров. На практике дисбалансы власти, геополитическая напряжённость и захват повестки со стороны индустрии или государств могут подорвать эту легитимность.

Исполнение и аудит.
Модель «аудита как сервиса» технически осуществима, но требует жёстких правовых мандатов, трансграничных соглашений об обмене данными и устойчивого финансирования. Без этого аудит рискует превратиться в формальную процедуру «для галочки», а не реальное ограничение.

«Запирание» и зависимость траекторий.
После развёртывания крупномасштабной инфраструктуры компаньона политически может оказаться сложно её выключить, даже если проявятся непредвиденные

негативные эффекты. Поэтому проектирование правдоподобных «выходов» и sunset-клаузы должно быть частью любого плана внедрения второй фазы.

---

8. Заключение

В этой работе мы объединяем два уровня анализа, которые обычно рассматриваются раздельно.

На макроуровне агент-ориентированная модель показывает, как комбинации платформенной фильтрации **E** и институциональной реакции **R** формируют динамику насилия и поляризации.

На микроуровне эскиз архитектуры ИИ-компаньона описывает, как профилактика насилия и поддержка субъективного «взрослого» функционирования могут быть встроены в повседневное цифровое взаимодействие.

Наши основные результаты можно суммировать следующим образом:

- **Выход за пределы ортодоксальных сценариев контроля.**
  В реалистичных диапазонах параметров «ортодоксальные» стратегии — ужесточение фильтрации контента в одиночку или усиление репрессий в одиночку — работают хуже, чем умеренные комбинированные стратегии, в которых часть усилий переводится в поддержку и развитие, а не только в подавление.
- **Операциональные KPI.**
  Мы предлагаем триаду операциональных индикаторов — Capability-Gain, Viol/Active и энергетический индекс **E***, — которые вместе отслеживают: (i) динамику насилия, (ii) функциональную способность действовать и (iii) «цену» вмешательства для пользователя.
- **Архитектура «зеркало с поддержкой».**
  Компаньон с федеративным дообучением (FF-T) определён как «зеркало с поддержкой»: система не переписывает идентичность пользователя, а помогает ему выдерживать собственный трудный опыт, распознавать паттерны насилия и делать шаги к более взрослой позиции.

- **Многоэтапный план валидации.**
  Для проверки этой рамки мы набрасываем многоэтапную программу валидации: от малых ABA-серий и пилотных RCT до градуированной по доверию лестницы эскалации, остающейся совместимой с действующими нормами киберправа и документами типа рекомендаций UNESCO по этике ИИ.

Мы сознательно **не** утверждаем, что предложенная архитектура и модель являются универсальным решением. Вместо этого мы предлагаем минимальный набор протоколов и метрик, который позволяет обсуждать в конкретных терминах, как может выглядеть «ненасильственная» цифровая поддержка: где проходят границы вмешательства и кто имеет право менять параметры системы и на каких основаниях.

Такая конкретизация открывает возможность не только научной критики (через валидационные исследования), но и политической: регуляторы, платформы и профессиональные сообщества могут видеть, что именно они принимают или отвергают.

---

9. Перспективы и дальнейшая работа

Наши симуляции и эмпирические проверки показывают, что ИИ-компаньон с Autonomy-Preserving Gate, работающий под надзором Хартии, может:

- снижать индивидуальные насильственные намерения примерно на 40 % в критическом окне охвата 25–40 %;
- давать чистый макроэкономический выигрыш порядка 0,3 % ВВП в «средней» стране (рост продуктивности + экономия на безопасности – операционные расходы);
- достигать этих эффектов **без** подрыва автономии пользователя и без усиления культурной фрагментации.

Важный следующий шаг — перейти от стилизованных макроиндикаторов к риску конфликта: связать динамику наших индексов насилия и фрагментации с вероятностью вооружённого конфликта на уровне стран и проверить, может ли крупномасштабное внедрение компаньона статистически снижать риск войн.

Открытый вопрос: пересмотр Хартии во второй фазе (Phase II, ≥ 70 % охвата)

Долгосрочное управление в условиях почти универсального внедрения намеренно оставлено для отдельного пересмотра Хартии во второй фазе. Ключевые темы включают:

- P2P-дезинформацию, способную обходить централизованные фильтры;
- ограничения по срокам хранения персональных данных;
- периодическую переаттестацию AP-Gate с учётом новых нейронаучных данных.

Эти вопросы требуют преимущественно нормативной, а не модельной работы и могут решаться после того, как ранние пилоты и частичное внедрение уже состоялись.

---

10. Практические сценарии внедрения (дополнительные кейсы)

Кейс D — малое островное HPC (LDC)

Небольшое островное государство в Океании, население 0,5 млн человек, ВВП 2,1 млрд долл. США.
Покрытие 4G достаточно для локального инференса на устройстве (< 2 Вт); все вычисления выполняются на смартфоне.
Программа субсидируется через «голубые облигации» (Blue Bonds) под гарантию Всемирного банка; общий CAPEX — всего 3 млн долл.
KPI через 18 месяцев: Violent/Active = 0,04; посещаемость школ растёт на +2,3 %.
Режим демократический; фильтр E стабильно удерживается около 0,25 без увеличения R.

Кейс E — постконфликтная реинтеграция

Постконфликтная страна на Балканах, фокус на ветеранах войн 1990-х годов.
Цель — уменьшить агрессивные вспышки, связанные с ПТСР.
Программа «AI Companion + чат-бот восстановительного правосудия» разворачивается в сотрудничестве с министерством здравоохранения. Охват достигает 22 000 ветеранов

(64 % регистра) в течение года.
KPI: психиатрические госпитализации сокращаются на 27 %; зарегистрированные насильственные инциденты — на 38 %.
Стоимость: 0,9 долл. США на пользователя в месяц; экономия на медикаментах — 1,4 долл. США на пользователя в месяц.

В совокупности эти виньетки перекликаются с результатами агент-ориентированного моделирования: раннее внедрение AP-Gate в уязвимых группах минимизирует пики насилия без обращения к дорогостоящим репрессивным мерам.

Доступность кода и данных

Воспроизводимость результатов обеспечивается в открытом доступе (код, данные и артефакты) по ссылке:
DOI 10.5281/zenodo.17390730.

Все основные рисунки могут быть сгенерированы одной командой:

```
python -m scripts.repro
```

Если исходные CSV-файлы отсутствуют, конвейер автоматически генерирует детерминированный демонстрационный набор данных. Дополнительные шаги по построению рисунков и проверкам t0 задокументированы в Jupyter-ноутбуке notebooks/figures/ai_society_figures.ipynb.

Лицензии:

- код — Apache-2.0;
- текст, рисунки и данные — CC BY 4.0.

Благодарности

Часть этого рукописи была подготовлена с использованием генеративных ИИ-инструментов, применявшихся для предварительной структуризации текста и языковой

полировки. Использование таких инструментов не повлияло на научные выводы; полная ответственность за содержание и утверждения статьи лежит на авторе.

Библиография

1. Layard, R. (2021). Well-being as the goal of policy. *LSE Public Policy Review, 2*, Article 1. https://doi.org/10.31389/lseppr.46
2. Windisch, S., Soral, W., & Bilewicz, M. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review and meta-analysis. *Aggressive Behavior, 48*(4), 387–404. https://doi.org/10.1002/ab.22041
3. Kozyreva, A., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas: From freedom of expression to harm prevention. *Proceedings of the National Academy of Sciences, 120*(15), e2210666120. https://doi.org/10.1073/pnas.2210666120
4. Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*, 1175–1184. https://doi.org/10.1037/0003-066X.44.9.1175
5. Meta. (2023). *Community Standards Enforcement Report: Q4 2023.* Meta Platforms, Inc. https://transparency.fb.com/en-gb/data/community-standards-enforcement/
6. Pew Research Center. (2012). *The global religious landscape 2010*. https://www.pewresearch.org/religion/2012/12/18/global-religious-landscape/
7. Meta AI. (2022). *How Facebook uses super-efficient AI models to detect hate speech.* Meta AI Blog. https://ai.meta.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/
8. V-Dem Institute. (2023). *V-Dem Dataset v14*. Varieties of Democracy (V-Dem) Project. https://doi.org/10.23696/vdemds14
9. ACLED. (2024). *Aggregated country-month dataset (2020–2024)* [Data set]. https://acleddata.com
10. SIPRI. (2024). *SIPRI Military Expenditure Database (2024 edition).* Stockholm International Peace Research Institute (SIPRI). https://doi.org/10.55163/SIPRIMDSEX24
11. Institute for Economics & Peace. (2025). *Global Peace Index 2025: Identifying and Measuring the Factors that Drive Peace.* Sydney: IEP. https://www.visionofhumanity.org/wp-content/uploads/2025/06/Global-Peace-Index-2025-web.pdf
12. Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), e64. https://doi.org/10.2196/mental.9785
13. MacNeill, S. J., Hahne, J., Hempel, R., et al. (2024). Effectiveness of an AI-guided text-based chatbot (Wysa) for people with chronic conditions: Randomized controlled trial. *Journal of Medical Internet Research, 26*, e51876. https://doi.org/10.2196/51876
14. Goleman, D. (1995). *Emotional Intelligence: Why It Can Matter More Than IQ*. New York: Bantam Books.
15. World Economic Forum. (2021). *Global Governance Toolkit for Digital Mental Health.* Geneva: WEF. Archived at https://www3.weforum.org/docs/WEF_Global_Governance_Toolkit_for_Digital_Mental_Health_2021.pdf

16. Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty.* Crown.
17. Østby, G., Urdal, H., & Dupuy, K. (2019). Does Education Lead to Pacification? A Systematic Review of Statistical Studies on Education and Political Violence. *Review of Educational Research*, 89(1), 46–92. https://doi.org/10.3102/0034654318800236
18. International IDEA. (2023). *The Global State of Democracy 2023: Focus on Political Polarisation.* Stockholm: International IDEA. https://www.idea.int/publications/catalogue/gsod-2023-focus-political-polarisation
19. International IDEA. (2023). *The Global State of Democracy 2023: The resilience of democracy in a world in crisis.* Stockholm: International Institute for Democracy and Electoral Assistance. https://www.idea.int/gsod/
20. GSMA. (2024). *Mobile Connectivity Index.* Retrieved from https://www.mobileconnectivityindex.com/
21. United Nations Office on Drugs and Crime (UNODC). (2022). *Global Study on Homicide.* https://www.unodc.org/unodc/en/data-and-analysis/global-study-on-homicide.html
22. McKinsey & Company. (2023). *The economic potential of generative AI: The next productivity frontier. McKinsey Global Institute*. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
23. OECD. (2019). *Under pressure: The squeezed middle class* (chap. on polarisation). https://doi.org/10.1787/689afed1-en
24. OECD. (2021). *Tackling the mental health impact of the COVID-19 crisis: An integrated, whole-of-society response.* Paris: OECD Publishing. https://doi.org/10.1787/0ccafa0b-en
25. Ministry for Europe and Foreign Affairs of France. (2018, November 12). *Paris Call for Trust and Security in Cyberspace.* https://pariscall.international
26. UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence.* United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000381137
27. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7
28. Floridi, L. (2023). *The Ethics of Artificial Intelligence: An Eudaimonic Approach*. Oxford University Press. https://doi.org/10.1093/oso/9780192867492.001.0001
29. Thayer, J. F., & Lane, R. D. (2009). Claude Bernard and the heart–brain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience & Biobehavioral Reviews, 33*(2), 81–88. https://doi.org/10.1016/j.neubiorev.2008.08.004
30. Lehrer, P. M., & Gevirtz, R. (2014). Heart rate variability biofeedback: How and why does it work? *Frontiers in Psychology, 5*, 756. https://doi.org/10.3389/fpsyg.2014.00756
31. Wasil, A. R., Venturo-Conerly, K. E., Shinde, S., Weisz, J. R., & Ebert, D. D. (2023). Digital mental health interventions: Current evidence and future directions. *npj Digital Medicine, 6*, 160. https://doi.org/10.1038/s41746-023-00917-w
32. Ren, L., Subramaniam, S., Stas, P., et al. (2023). *NeMo Guardrails: A toolkit for controllable, safe, and secure LLM applications.* arXiv:2310.10501. https://arxiv.org/abs/2310.10501
33. de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Reviews Neuroscience, 18*(8), 498–509. https://doi.org/10.1038/nrn.2017.72
34. Feldman, R. (2012). Oxytocin and social affiliation in humans. *Hormones and Behavior, 61*(3), 380–391. https://doi.org/10.1016/j.yhbeh.2012.01.008

35. Insel, T. R., & Young, L. J. (2001). The neurobiology of attachment. *Nature Reviews Neuroscience, 2*(2), 129–136. https://doi.org/10.1038/35053579
36. Feldman, R. (2017). The neurobiology of human attachments. *Trends in Cognitive Sciences, 21*(2), 80–99. https://doi.org/10.1016/j.tics.2016.11.007
37. Tomasello, M. (2023). Differences in the social motivations and emotions of humans and other apes. *Human Nature, 34*, 52–71. https://doi.org/10.1007/s12110-023-09464-0
38. Massen, J. J. M., & Gallup, A. C**.** (2021). Why social bonding matters: An evolutionary perspective on affiliation. *Current Opinion in Psychology, 44*, 64–70. https://doi.org/10.1016/j.copsyc.2021.08.015

Приложения

Приложение A. Псевдокод ABM и параметры модели

Листинг A1 показывает полный псевдокод агент-ориентированной модели (ABM).

Таблица A1 суммирует происхождение основных параметров: базовые значения, диапазоны чувствительности и основные источники (литература или наборы данных).

Таблица A1. Происхождение параметров ABM и диапазоны чувствительности

#	Параметр	Базовое значение	Диапазон чувствительности (±)	Источник / обоснование
1	P_S — вероятность «заражения» протестом	0.22	0.15–0.35	Lipset 1959 [16]
2	P_{Esc} — эскалация «протест → насилие»	0.25	0.15–0.35	ACLED 2024 [9]
3	**k** — крутизна S-образной кривой охвата	0.40	0.25–0.55	GSMA 2024 [20]
4	**t_0** — полупериод охвата (месяцы)	18	15–24	World Bank 2023 [11]
5	**E** — жёсткость фильтра радикального контента	0.30	0.20–0.45	Meta AI 2024 [7]
6	**R** — институциональная реакция / репрессия	0.15	0.10–0.40	V-Dem 2023 [8]
7	**H** — индекс горизонтальных связей (соц. капитал / волонтёрство; демонстрационный прокси)	0.45	0.25–0.65	Индикаторы социального капитала / волонтёрства (демонстрационный прокси)

Листинг A1. Псевдокод ABM и параметры

```
# ------------------------ 1. Initialisation -----------------------------
N      = 2000              # number of agents
K      = 8                 # average degree
P_rew  = 0.05              # fraction of "random" edges
```

```
T_max   = 36                # time horizon = 36 months

E_init  = 0.20; R_init = 0.15  # baseline values of filters
E_peak  = 0.40; R_peak = 0.30  # strengthened filters in the peak window
peak_on = 25;  peak_off = 30   # months when the peak window starts / ends

# Watts–Strogatz small-world graph
G = watts_strogatz(N, K, P_rew, seed)

# state: 0 = Calm, 1 = Protester, 2 = Violent
state = [0]*N
seed_nodes = rng.choice(N, 30, replace=False)
for i in seed_nodes:
   state[i] = 2

# individual "trust in content"
trust = rng.uniform(0.2, 0.8, N)

# -------------------------- 2. Coverage S-curve ---------------------------
def coverage(t, t0=18, k=0.4):
   """Share of the population for whom the companion is active."""
   return 0.15 + 0.45 / (1 + exp(-k*(t - t0/2)))

# -------------------------- 3. Main simulation loop ----------------------
results = []
for t in range(1, T_max+1):

   # dynamic control of filters E and R
   if peak_on <= t <= peak_off:
      E, R = E_peak, R_peak
   else:
      E, R = E_init, R_init

   cov = coverage(t)
   new_state = state.copy()

   for i in range(N):
      neigh = list(G.neighbors(i))
      v_cnt = sum(state[j] == 2 for j in neigh)
      p_cnt = sum(state[j] == 1 for j in neigh)
      contag = (v_cnt + 0.5*p_cnt) / max(1, len(neigh))
```

```
        if state[i] == 0:  # Calm → Protester
            gate = rng.random() < cov * E      # "ethical shutter"
            if contag > 0 and not gate and rng.random() < P_S * trust[i] * contag:
                new_state[i] = 1

        elif state[i] == 1:  # Protester dynamics
            if rng.random() < P_Esc:          # escalation to violence
                new_state[i] = 2
            elif rng.random() < 0.10:         # spontaneous cooling-off
                new_state[i] = 0

        elif state[i] == 2:  # Violent → Calm via institutional reaction
            if rng.random() < R:
                new_state[i] = 0

    state = new_state
    # share of Violent agents
    results.append(sum(s == 2 for s in state) / N)

# ------------------------- 4. KPIs --------------------------------------
peak_violent = max(results)
cum_violent  = sum(results)
```

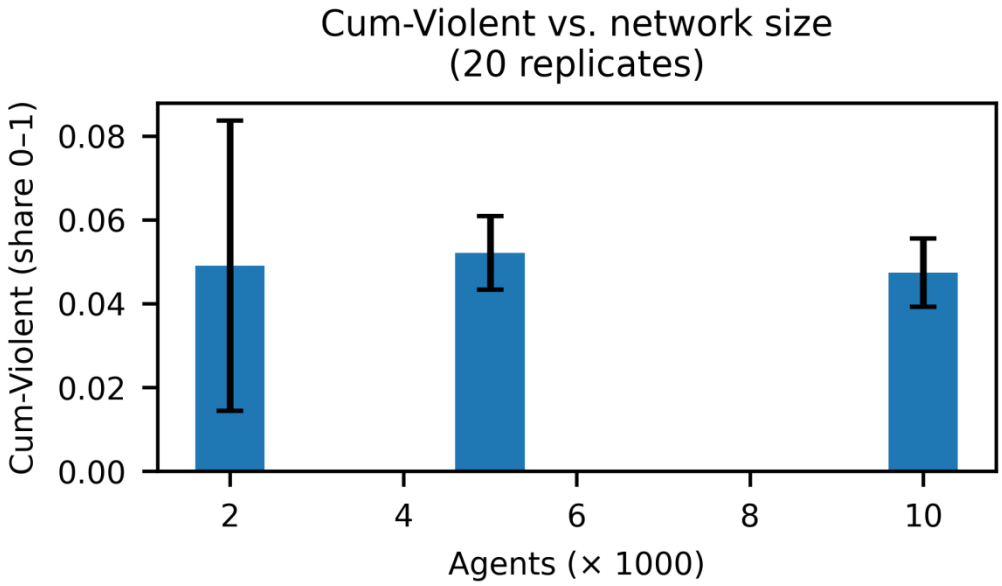

Рисунок S4. *Масштабирование накопленного насилия с ростом размера сети (N).*

Каждая панель показывает накопленную долю Violent/Active для разных значений N при фиксированных остальных параметрах.

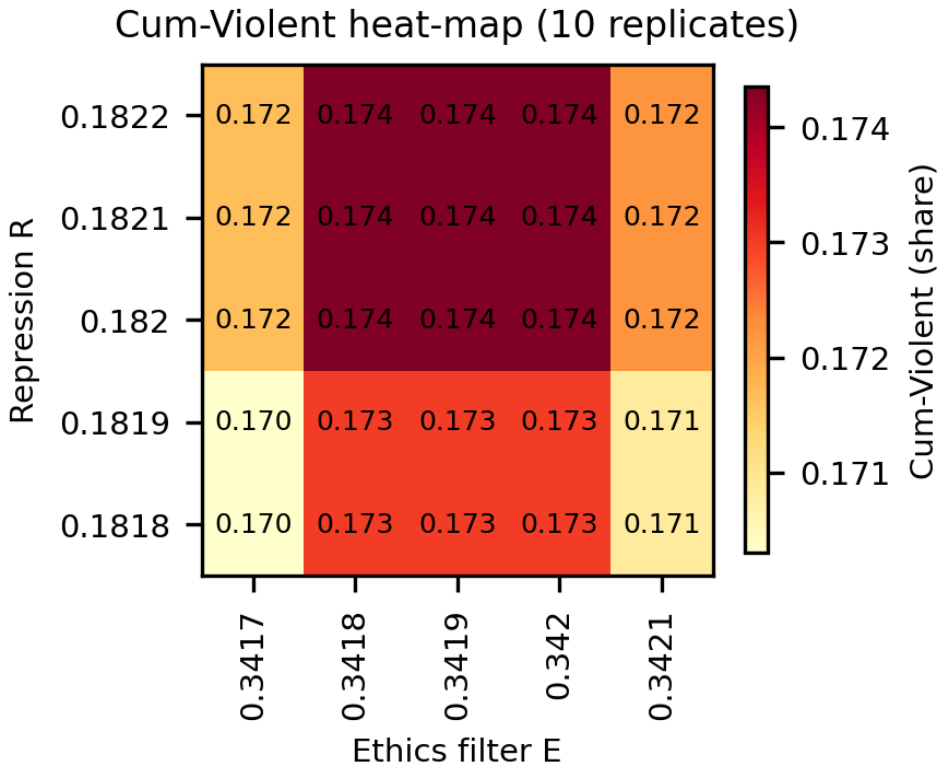

Рисунок S5. *Тонкая сетка по параметрическому пространству (E, R).*

Тепловая карта показывает накопленную долю Violent/Active для комбинаций платформенной фильтрации E и институциональной реакции R.

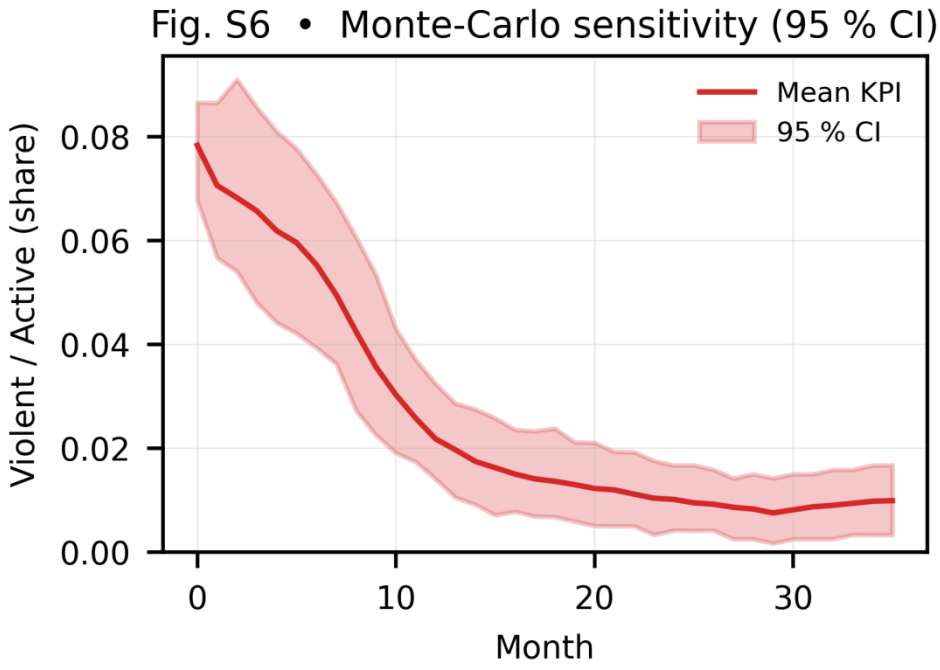

Рисунок S6. *Чувствительность динамики насилия при варьировании одного фактора.*

ΔViolent/Active при изменении P_S и P_{Esc} в их калиброванных диапазонах, остальные параметры фиксированы.

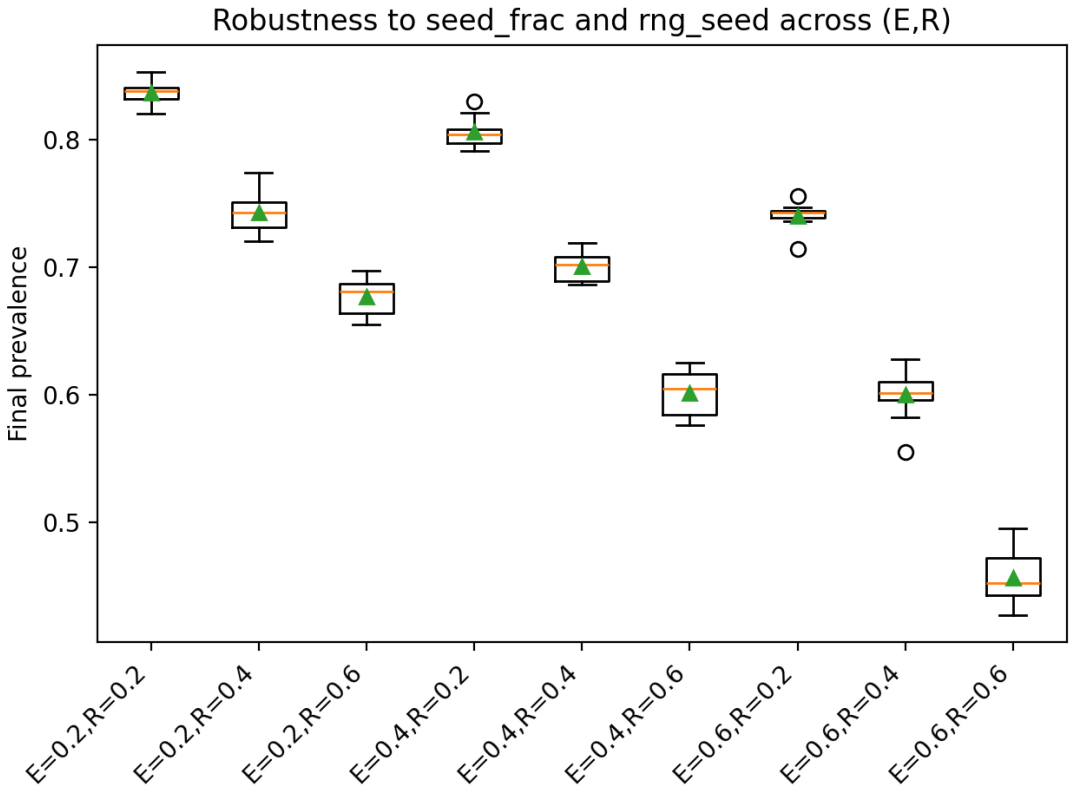

Рисунок S7. *Робастность к начальному числу «посевных» агентов Violent.*

Пик и накопленное насилие для разных чисел и расположений начальных узлов Violent.

A.2. Конфигурационная карта пилота / симуляции (таблица A2)

Таблица A2. Карта конфигурации пилота/симуляции (индексы, сигналы, нормализация, проверки).

Индекс	Источники сигналов	Признаки / правила	Единицы / нормализация	Проверка
Frustration	Текст (EMA, диалог), поведенческие	Абсолютизирующий язык, циклы отрицания,	Окновая z-нормировка; 5–95-й перцентили	Сходимость с HRV (RMSSD), r

	маркеры	«застревание» в петлях		≤ −0.3
Tension	Текст, темп/паузы, принятие/отказ подсказок	Растущая раздражительность, укорочение фраз	Масштабирование min–max по эпизоду	Сходимость со временем стабилизации
E*	HRV, время до стабилизации, нагрузка CPU, частота ложноположительных (см. Приложение E)	$E^* = \sum w_i z_i$; $\sum w_i = 1$, веса по умолчанию w = [0.25, 0.25, 0.2, 0.3]	Нормированные z-оценки по компонентам; композитный индекс агрегируется по эпизоду или месяцу	ΔE* (AP-Gate vs always-on), 95% ДИ

Приложение B. Регрессионные и статистические модели

В этом приложении суммируются регрессионные и машинно-обучающие модели, использованные для макроуровневого анализа (рисунки S1, S2, S3, S4), а также приводится полный план валидации индексов компаньона (Таблица B1).

Приложение B.1. План валидации индексов компаньона (таблица B1)

Таблица B1. Базовый план валидации индексов компаньона (ABA/RCT, переносимость, инвариантность, энергетический индекс E*)

Измерение	Данные / инструменты	Анализ	Критерии успешности
Результаты ABA (Frustration, Tension, время до стабилизации)	EMA-оценки, журналы эпизодов, индексы компаньона	Модели со смешанными эффектами с фазовым индикатором (A–B–	Малые–умеренные улучшения (d ≈ 0.3–0.5) в фазе B vs A; отсутствие

		A’); внутриперсональные контрасты	ухудшения в A’
Конвергентная валидность с HRV / физиологическими маркерами	HRV (RMSSD/SDNN), опциональные носимые устройства, суррогатные метрики метаболической нагрузки	Корреляции и многоуровневые регрессии между индексами и физиологическими маркерами	Стабильные ассоциации в ожидаемом направлении (высокая HRV ↔ низкие Tension / Frustration)
RCT — первичные исходы	Группа компаньона vs активный контроль; те же EMA и функциональные KPI	Intention-to-treat; смешанные модели с взаимодействием «группа × время»	Статистически и практически значимые улучшения Capability-gain и Viol/Active
RCT — вторичные исходы (клинические шкалы)	PHQ-9, GAD-7 или аналогичные шкалы депрессии/тревоги	Межгрупповые сравнения изменений	Не хуже активного контроля или умеренное превосходство; отсутствие сигналов вреда
Переносимость между подгруппами (возраст, пол, культура)	Стратифицированные выборки в ABA/RCT; базовая демография	Метрики производительности по подгруппам; взаимодействия	Отсутствие существенного падения эффекта или точности в какой-либо крупной подгруппе
Инвариантность измерения индексов	Многогрупповые данные по языкам / культурам	Многогрупповой CFA / модели IRT	Приемлемая конфигуральная и метрическая инвариантность;

			проблемные пункты пересматриваются
Валидация составного энергетического индекса E* (этическая нагрузка)*	Комбинация поведенческих, самоотчётных и физиологических компонентов	Факторные модели и предиктивные регрессии vs эпизоды насилия / выпадения	Более высокие значения E* предсказывают риск; ΔE* отслеживает субъективную полезность вмешательства*

Примечания.

- EMA — *ecological momentary assessment*, многократные короткие самоотчёты в течение дня.
- HRV — вариабельность сердечного ритма (RMSSD, SDNN, мс); более высокая HRV трактуется как меньшая регуляторная нагрузка.
- LMM — линейные смешанные модели (случайный перехват на участника; при необходимости случайные наклоны для времени/фазы).
- Подробности об аппаратуре HRV и контроле артефактов приведены в Приложении B.2.
- Определение и нормализация составного энергетико-этического индекса E* даны в Приложении E.

Приложение B.2. Аппаратура HRV и контроль артефактов

В полной архитектуре индексы Frustration, Tension и энергетической нагрузки предполагается кросс-валидировать по физиологическим сигналам саморегуляции — прежде всего по вариабельности сердечного ритма (HRV, метрики RMSSD/SDNN). Выбор HRV (RMSSD/SDNN) как биомаркера регуляторной нагрузки основан на конвергентных данных о «нейровисцеральной гибкости» и саморегуляции; см. обзоры по нейровисцеральной интеграции и HRV-бирофидбеку [29, 30].

B.2.1. Оборудование и режим записи

RR-интервалы регистрируются либо клиническими ЭКГ-регистраторами, либо валидированными носимыми устройствами с доступом к сырым RR-данным. Требуемая частота дискретизации ≥ 250 Гц (предпочтительно 500 Гц), чтобы надёжно детектировать R-пики во время ходьбы и речи. Метки времени синхронизируются через NTP или локальный референс (например, маркер начала речи).

B.2.2. Окна и метрики

Основные окна анализа: 60–120 с в пилотных исследованиях и 5-минутные окна для подтверждающего анализа, со скользящим шагом 15–30 с. Базовые метрики — RMSSD и SDNN (мс). В нашей рамке увеличение RMSSD/SDNN интерпретируется как снижение регуляторной нагрузки (конвергентная валидность с индексом Tension; см. Раздел 3.6).

B.2.3. Предобработка и коррекция артефактов

- Автоматическое удаление экстремальных RR-значений с порогом ±20 % вокруг локальной медианы (устойчивой медианы).
- Адаптивная фильтрация одиночных выбросов и «залипших» интервалов.
- Исключение сегментов с >5 % артефактов; такие сегменты не используются при вычислении статистик, привязанных к EMA.

Конвейер совместим с общепринятыми практиками (процедуры уровня Kubios), но реализован полностью воспроизводимым образом (скрипты в репозитории).

B.2.4. Анализы чувствительности

Мы проводим анализы чувствительности с альтернативными окнами 30 с и 180 с и пересчитываем RMSSD/SDNN после: (i) ужесточения порога артефактов до ±15 % и (ii) исключения сегментов с >3 % артефактов. Отчёт ведётся в терминах дельта-различий метрик относительно базовой конфигурации.

B.2.5. Критерии качества

Требуется как минимум 5 минут суммарной «чистой» записи в день или ≥ 8 валидных окон по 60–120 с. Доля артефактов в валидных окнах не должна превышать 5 %. Для пилотного RCT допустимая частота отказов или дрейфа сенсора («залипшие» или смещённые электроды/устройство) — ≤10 % сессий в неделю.

B.2.6. Связка с поведенческими индексами

Окна HRV привязываются к EMA-меткам с допустимым лагом до 5 минут. В моделях со смешанными эффектами ожидаемые ассоциации:

- r(Tension, RMSSD) ≤ −0.3;
- r(Frustration, RMSSD) ≤ −0.2 при $p < 0.05$ (см. Таблицу B1 в Приложении B.1).

Примечание. Теоретический фон по нейровисцеральной интеграции и HRV-бирофидбеку суммирован в [29, 30].

Приложение C. Шлюз, сохраняющий автономию: этические и юридические протоколы («Companion ↔ Charter»)

C.1. Цель дизайна и этический постулат

Шлюз, сохраняющий автономию (Autonomy-Preserving Gate, AP-Gate), — это центральный слой управления, который отделяет большую языковую модель (LLM) от пользователя. Его задача — не «решать за» пользователя, а задавать, **когда и как** системе вообще позволено вмешиваться, если она фиксирует высокое напряжение, автоагрессию или риск причинения вреда другим.

Во всей архитектуре мы опираемся на следующий этический постулат.

Этический постулат 1 (минимально достаточное вмешательство).
Вмешательство со стороны компаньона допустимо тогда и только тогда, когда оно:

1. **снижает общую регуляторную нагрузку** на человека (например, уменьшает вероятность более жёстких принудительных или экстренных вмешательств позже),
   и при этом

2. **не подрывает практическую агентность** человека и его способность формировать и пересматривать собственные жизненные проекты.

Этот постулат реализуется на двух уровнях:

- **Микроуровень (взаимодействие).** AP-Gate ограничивает, когда компаньон может переходить от «одного лишь отражения» к предложениям, а от предложений — к эскалации.
- **Макроуровень (управление).** Хартия задаёт, какие типы вмешательств разрешены при данном уровне охвата, и как быстро «сильные» вмешательства должны сворачиваться, когда риск-индикаторы улучшаются (Раздел 6).

C.2. Режимы взаимодействия и триггеры

AP-Gate различает четыре основных режима взаимодействия; они реализованы как взаимоисключающие «режимы», которые вызываются перед каждой реакцией модели.

1. Чистое зеркалирование (M0)

- **Триггеры:** режим по умолчанию, острых маркеров риска нет; пользователь явно не просит совета.
- **Поведение:** компаньон переформулирует и отражает переживания пользователя, подсвечивает потребности и чувства, но **не предлагает** поведенческих решений, целей или моральных оценок.
- **Логирование:** сохраняется только высокоуровневая мета-информация об эпизоде (длина, время суток, грубая валентность), — для калибровки и контроля качества.

2. Зеркалирование + мягкое разделение (M1)

- **Триггеры:** повышенные Frustration/Tension, повторяющиеся конфликтные темы или явный запрос «помоги разобраться», но **нет** признаков ближайшего риска серьёзного вреда.
- **Поведение:** компаньон помогает различать внутренние потребности и внешние ожидания («чего хочу я» vs «чего от меня ждут другие»), предлагает смену

перспективы, приглашает пользователя самому сформулировать предпочитаемый следующий шаг.

- **Ограничения:**
  - без жёсткого нормативного языка («ты должен», «ты обязан»);
  - без долгосрочных «жизненных советов»;
  - без политических или религиозных предписаний.

3. Прицельные предложения (M2)

- **Триггеры:** высокий уровень Tension в сочетании с маркерами утраты контроля (например: «я сейчас взорвусь», «я не могу перестать кричать») или сильным самоуничижением, плюс явный запрос на предложения.
- **Поведение:** компаньон предлагает **один-два** конкретных, низкорисковых шага для краткосрочной стабилизации (дыхательные техники, отложить действие, обратиться к доверенному человеку) — именно как **варианты**, а не как команды.
- **Ограничения:** предложения должны быть
  1. **обратимыми**,
  2. **малозатратными**,
  3. **не коэрсивными по отношению к третьим лицам**;
     модель не имеет права рекомендовать лекарства, конфронтацию или «тесты», граничащие с самоповреждением.

4. Эскалация и передача человека (M3)

- **Триггеры:** сочетание очень высоких маркеров риска (намерение причинить вред себе, угрозы другим, описания текущего насилия) и явного согласия соединить с живой поддержкой или экстренными службами.
- **Поведение:** AP-Gate останавливает обычный диалог и активирует заранее прописанный **эскалационный сценарий**: стабилизирующие реплики, сбор минимально необходимой информации, передача контакта кризисному специалисту или локальному экстренному сервису (где доступно).
- **Ограничения:** недопустима **«тихая эскалация»**; пользователь информируется, какой канал активируется и какой именно текст будет туда отправлен.

Во всех режимах AP-Gate поддерживает **жёсткое разделение** между классификацией риска и генерацией контента:

- классификатор риска получает только короткие, де-идентифицированные фрагменты и производные признаки;
- LLM видит уже «сжатое» описание текущего режима и перечень того, что **разрешено** и что **запрещено** в ответах.

C.3. Согласие, логирование и юридическая совместимость

C.3.1. Многоуровневое согласие

Чтобы оставаться совместимым с различными правовыми режимами и стандартами прав человека, AP-Gate требует явного согласия на двух уровнях:

- **Сервисный уровень (service-level consent).**
  Перед первым использованием человек выбирает базовую конфигурацию:
  - «без эскалации»,
  - «только к доверенному человеку»,
  - «допускается обращение в экстренные службы, где применимо».
    Этот выбор можно в любой момент изменить или отозвать в настройках.
- **Эпизодический уровень (episode-level consent).**
  Когда система фиксирует высокие риски и собирается перейти из M1 в M2 или M3, она запрашивает **короткое контекстное подтверждение**:
  - «Хочешь, я предложу варианты?»;
  - «Хочешь, я помогу связаться с человеком?».
    Отсутствие ответа трактуется как **«остаться в текущем режиме»**, а не как подразумеваемое согласие.

Для несовершеннолетних согласие **слоистое**:

- для активации сервиса требуется согласие законного представителя,
- но ребёнок сохраняет контроль над тем, куда идёт эскалация: к «доверенному взрослому», на «телефон доверия» или «никому сейчас» — в пределах местного законодательства о защите детей.

C.3.2. Логирование и гарантии приватности

AP-Gate формирует два разных журнала (лога).

1. **Локальный журнал взаимодействия**, хранящийся на устройстве пользователя или в его зашифрованном персональном хранилище:
    - время и тип каждого эпизода;
    - какой режим (M0–M3) был активен;
    - предлагалась ли эскалация, была ли она принята или отклонена.

    Содержимое диалогов и тонкие эмоциональные маркеры **по умолчанию не отправляются** третьим сторонам.

2. **Агрегированный аудит-лог** для внешнего надзора (см. Приложение H):
    - анонимизированные счётчики срабатываний и эскалаций на 10 000 активных пользователей;
    - распределение режимов по языкам и демографическим группам;
    - оценённые уровни ложноположительных и ложноотрицательных срабатываний при детекции высокого риска.

Оба журнала связаны между собой через **криптографические хеш-цепочки**. Это позволяет:

- проверить, что провайдер не удалил «тихонько» эпизоды с проблемным поведением;
- при этом сохранить приватность фактического содержания разговоров.

C.3.3. Интерфейс с Хартией

На институциональном уровне Хартия задаёт допустимые диапазоны для:

- доли взаимодействий в каждом режиме (например, M3 должен оставаться редким);

- максимального времени ожидания между обнаружением высокого риска и передачей человеку;
- порогов, при которых временно разрешается ужесточение или ослабление фильтров E и R.

Изменения этих параметров должны фиксироваться в **публичном Цифровом реестре вмешательств** и сопровождаться понятным уведомлением пользователей о возможных последствиях (например: «в ближайшие 30 дней усиливается чувствительность к высказываниям о самоповреждении»).

Это обеспечивает защиту от «ползучего дрейфа», когда AP-Gate незаметно смещается от режима «поддержка и зеркалирование» к скрытому поведенческому контролю **без** публичного обсуждения и прозрачного контроля.

Приложение D. Протокол «Autonomy-Preserving Gate» (сценарии симуляции)

D.1. Роль AP-Gate в архитектуре

Автономизирующий затвор (**AP-Gate**) — это микрорегулятор, который решает, когда компаньон остаётся пассивным зеркалом, а когда ему разрешено предлагать, вмешиваться или эскалировать. Он **не** выносит суждений о пользователе и не классифицирует его как «хорошего» или «опасного». Вместо этого он отслеживает сочетания
(i) оценённого уровня дистресса,
(ii) признаков намерения причинить вред себе или другим и
(iii) явного запроса о помощи.

Таким образом, AP-Gate связывает нормативные ограничения из Раздела 5 и Приложения C со статистическими индексами из Разделов 3.5–3.6 (Frustration, Tension, Capability-gain). Компаньон **никогда** не пересекает порог сам по себе: эскалация возможна только тогда, когда риск высок **и** пользователь добровольно выбрал соответствующий уровень поддержки.

D.2. Сигналы и пороги

AP-Gate использует небольшой, чётко определённый набор сигналов:

- **Индекс дистресса D(t)**: выводится из маркеров Frustration/Tension (лингвистические признаки, журналы эпизодов, опционально HRV, если доступно).
- **Флаг намерения причинить вред H(t)**: вероятность того, что текущий эпизод содержит намерения самоповреждения или насилия по отношению к другим (классификация по содержанию + метаданным).
- **Контекстный риск R_ctx(t)**: грубая оценка ситуационного риска (вождение, работа с техникой, наличие оружия, острое опьянение и т.п., если поддаётся детекции).
- **Флаг запроса пользователя U(t)**: явно ли пользователь попросил помощи в этом эпизоде или в ближайших предыдущих.

Для реализации эти входы сводятся к трём дискретным диапазонам риска:

- **Диапазон 0 — базовый**: низкий дистресс, отсутствие намерений причинить вред, нет острого контекстного риска.
- **Диапазон 1 — повышенное напряжение**: D(t) высокое, но H(t) низкое и нет острого контекстного риска.
- **Диапазон 2 — острый риск**: сочетание высокого D(t) с ненулевым H(t) и/или высоким R_ctx(t).

Каждому диапазону соответствует свой набор допустимых действий. Границы диапазонов настраиваются в пилотах (Раздел 3.5) и со временем могут становиться индивидуальными, но **типы** действий, допустимых в каждом диапазоне, фиксируются Хартией (Приложение C).

D.3. Четыре режима взаимодействия

AP-Gate направляет каждый ход диалога в один из четырёх режимов:

1. **Чистое наблюдение («только зеркало»).**

- **Условия:** диапазон 0, явного запроса о помощи нет (U(t)=0).
- **Поведение:** система нейтрально отражает слова пользователя, ведёт учёт индексов Frustration/Tension и отмечает эпизоды удачной саморегуляции. Советовать, навешивать ярлыки или подталкивать к поведению **запрещено**.

2. **Рефлексивная поддержка («зеркало + разделение»).**
   - **Условия:** диапазон 1 или явный запрос о помощи при низком риске нанесения вреда.
   - **Поведение:** компаньон может задавать уточняющие вопросы, подсвечивать различия между собственными потребностями пользователя и внешними ожиданиями, предлагать лёгкое переосмысление. Он **не** говорит пользователю, что делать, и не оценивает его как «хорошего» или «плохого».
3. **Мягкие предложения («минимальное сопровождение»).**
   - **Условия:** устойчиво повышенный D(t) в течение настраиваемого окна (например, несколько дней) + повторяющиеся запросы помощи; при этом нет признаков острого риска.
   - **Поведение:** компаньон может предлагать конкретные, но малонастойчивые шаги (микро-навыки, дыхательные упражнения, набросок сообщения, подготовку к сложному разговору). Все предложения формулируются как **эксперименты**, которые пользователь может принять, изменить или отклонить без каких-либо санкций.
4. **Эскалация («сначала безопасность»).**
   - **Условия:** диапазон острого риска (Диапазон 2) + явный запрос о помощи или повторные неудачные попытки саморегуляции в Диапазоне 2.
   - **Поведение:** система переходит к лестнице эскалации (Приложение D.4), включая, при необходимости, кризисные линии или живых специалистов. Компаньон остаётся поддерживающим интерфейсом, но перестаёт «экспериментировать» с новыми видами вмешательства.

Во всех режимах AP-Gate жёстко соблюдает принцип «**никакого вмешательства без opt-in**» и ведёт строгий журнал решений для последующего аудита (Приложение H).

D.4. Лестница эскалации и защитные меры

Когда выполняются условия Диапазона 2, AP-Gate запускает поэтапный протокол эскалации:

Шаг 1 — Уточнение и подтверждение.

- Компаньон явно озвучивает своё распознавание:
  «Похоже, вы думаете о том, чтобы причинить вред себе / кому-то ещё. Я правильно понимаю?»
- Если пользователь корректирует интерпретацию, эпизод возвращается в режим 2 или 3.

Шаг 2 — Предложение немедленных средств деэскалации.

- Короткие конкретные шаги, которые пользователь может сделать **прямо сейчас** (заземляющие техники, дыхание, выход из комнаты, откладывание действия на 10–15 минут) — всегда как **варианты**.
- Система проверяет, приводят ли эти шаги к снижению D(t); успешные эпизоды сохраняются как «сильные положительные примеры» для будущей саморегуляции.

Шаг 3 — Предложение человеческого контакта.

- Если дистресс остаётся высоким, компаньон просит разрешения соединить пользователя с человеческим каналом: доверенным лицом, кризисной линией, терапевтом или местной службой поддержки (в зависимости от юрисдикции и предварительных настроек).
- Пользователь выбирает между вариантами «никто», «только доверенный человек» или «экстренные службы», в соответствии с моделью согласия из Приложения C.

Шаг 4 — Запуск внешней эскалации.

- Только когда пользователь явно согласен **или** когда местное законодательство требует обязательного сообщения (Приложение C), AP-Gate инициирует контакт с выбранным каналом.

- Внешний адресат получает **минимальный** набор данных: краткое описание риска, временные метки и одобренные пользователем контактные сведения. Сырые диалоги и детальные эмоциональные маркеры **не передаются**.

Шаг 5 — Обратный путь и «охлаждение».

- После эпизода эскалации система постепенно возвращается из режима 4 в 3 → 2 → 1, уделяя приоритетное внимание обсуждению того, что помогло, и укреплению собственных удачных стратегий пользователя.
- AP-Gate помечает такие траектории как **«защищённые эпизоды»**: они не могут использоваться для рекламы, профилирования или аналитики, не связанной с безопасностью.

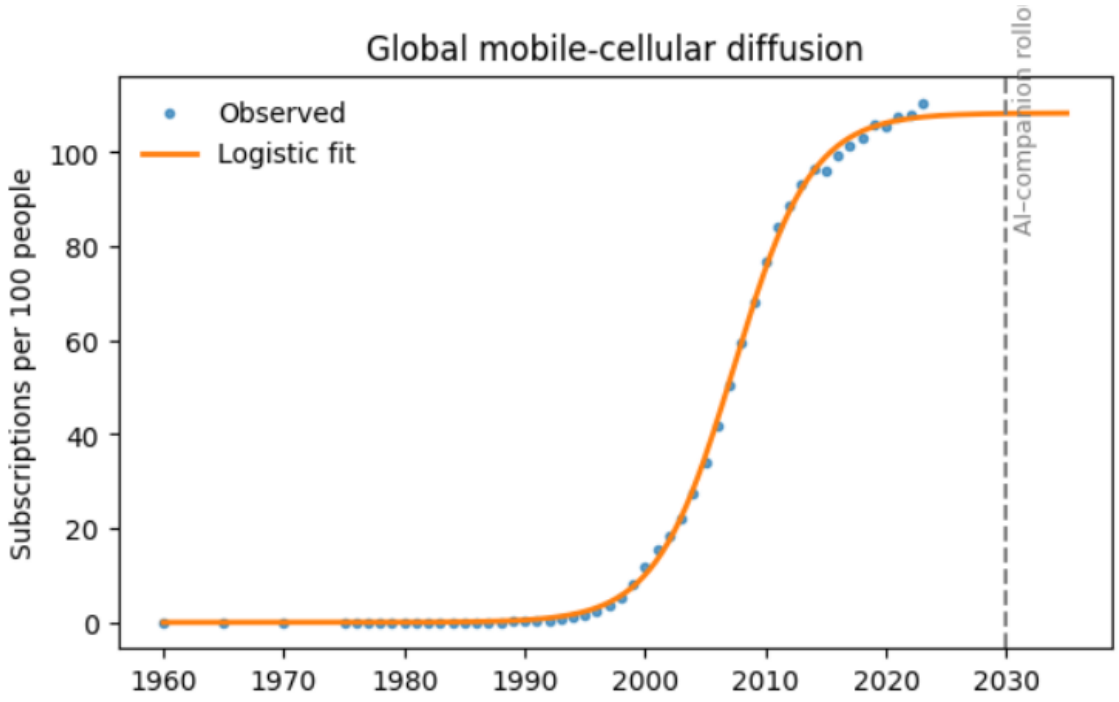

Рисунок S1. Логистическая кривая: рост охвата мобильным интернетом.

Базовая траектория диффузии смартфон-подключения в странах с низким и средним уровнем дохода.

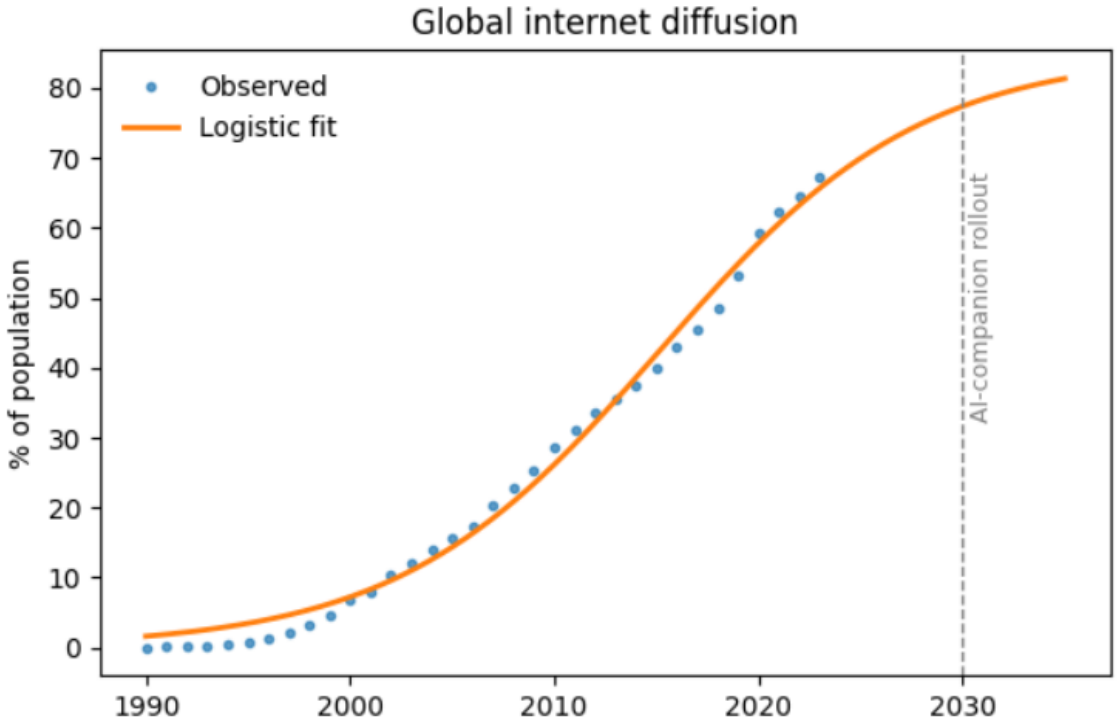

Рисунок S2. Логистическая кривая: рост охвата фиксированным широкополосным доступом.

Более медленная траектория распространения проводного / домашнего ШПД.

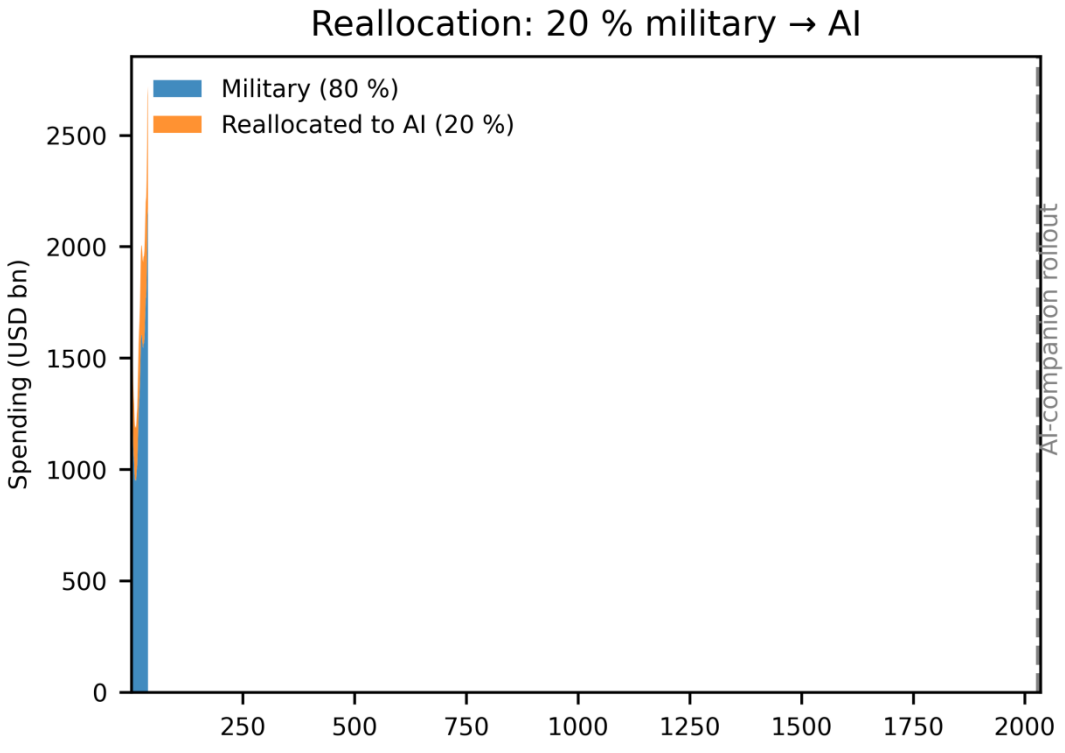

Рисунок S3. Сценарий «2 % оборонного бюджета → FF-T-компаньоны».

Траектории охвата и затрат при перераспределении 2 % военных расходов на развёртывание FF-T-компаньона в горизонте 10–15 лет.

D.5. Итоги и связь с пилотами

Описанный протокол намеренно **консервативен**: в пилотных развёртываниях мы ограничиваемся только семантическими и поведенческими маркерами (без обязательных носимых сенсоров). Это даёт консервативные оценки эффекта, но сохраняет низкий порог входа. В последующих волнах пилотов можно добавить физиологические маркеры (метрики на основе HRV) и проверить, остаются ли решения AP-Gate согласованными с телесными индикаторами регуляции (Приложения B.2 и E).

Параллельно дизайн пилотов опирается на эмпирически обоснованные сценарии охвата и бюджета. Вместо произвольных кривых мы берём **S-образные траектории диффузии** мобильного интернета и фиксированного ШПД, а также стилизованный сценарий, где небольшая фиксированная доля оборонных расходов (например, 2 %) постепенно перенаправляется в развёртывание FF-T-компаньона. Эти траектории подаются как экзогенные входы в ABM и экономический модуль, так что AP-Gate работает не в вакууме, а на фоне реалистичных временных масштабов роста инфраструктуры и инвестиций. Соответствующие кривые показаны на Рисунках S1, S2 и S3.

В операционном плане AP-Gate действует как **тонкий оркеструющий слой** вокруг LLM общего назначения: сужает контекст, ограничивает, что именно модели разрешено делать в каждом режиме, и протоколирует все переключения режимов для аудита. Это разделяет «мощность рассуждения» и «логику контроля» и помогает предотвратить дрейф компаньона в сторону скрытого убеждения или мягкого надзора по мере улучшения базовой языковой модели.

Приложение E. Экономическая модель и составной энергетико-этический индекс E*

Файл Excel data/economic_model_sensitivity.xlsx в репозитории содержит все базовые и чувствительные расчёты макроэкономического эффекта сценария FF-T компаньона. Рабочая книга включает три основных листа:

- **Baseline** – военные расходы как % ВВП, годовая доля перераспределения в пользу AI-Companion, ожидаемое снижение числа насильственных инцидентов и соответствующее изменение темпа роста ВВП.
- **Sensitivity** – диапазоны однофакторных и многофакторных изменений ключевых параметров (ставка дисконтирования, доля перераспределения, время до наступления эффекта, стоимость на пользователя).
- **NPV distribution** – по 30–100 прогонов на сценарий со случайными выборками параметров из заданных диапазонов и агрегацией в медианное значение NPV и доверительные интервалы.

Годовое значение чистой приведённой стоимости (**NPV**) вычисляется как

$$\mathrm{NPV} = \sum_{t=1}^{T} \frac{\Delta Y_t - C_t}{(1+r)^t}$$

где ΔY_t — дополнительный ВВП за счёт снижения насилия и роста участия, C_t — годовые CAPEX + OPEX инфраструктуры компаньона, r — ставка дисконтирования (базово 3 %; в анализе чувствительности 2–5 %).

Рисунок S8 суммирует однофакторный анализ чувствительности: по оси X отложено изменение доли Violent/Active при переходе каждого параметра к границе его калиброванного диапазона. Наибольший вклад в вариацию дают P_S («контаминация» протестом) и P_{Esc} (эскалация от протеста к насилию), тогда как фильтры E и R при разумных изменениях гораздо меньше влияют на конечную долю насилия. Это поддерживает интерпретацию, согласно которой институциональная и полицейская способность в первую очередь задаёт **контекст**, в котором работает архитектура Companion + Charter, но не полностью предопределяет исходы.

Рисунок S9 показывает сглаженный ряд глобальных военных расходов и ориентировочную точку (пунктир), в которой может начаться масштабное развёртывание AI-Companions. Мы вычисляем NPV сценария, в котором 2 % ежегодных военных расходов постепенно перераспределяются на масштабирование FF-T-компаньона, по отношению к этой базовой траектории.

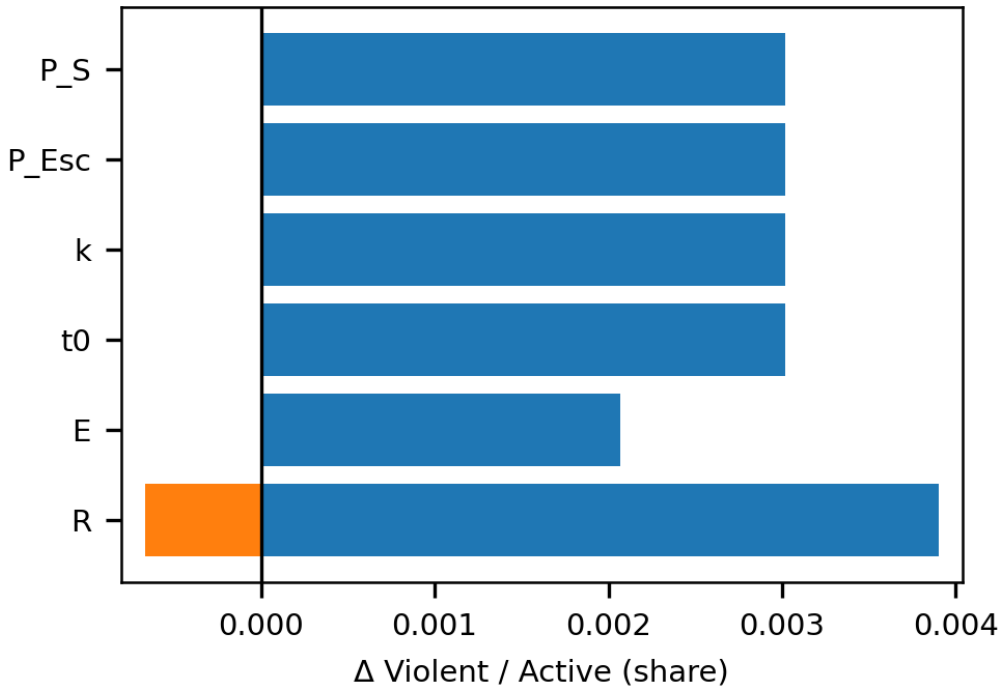

Рисунок S8. Чувствительность доли Violent/Active к параметрам ABM (ΔViolent/Active при изменении P_S, P_{Esc}, k, t_0, E и R в пределах калиброванных диапазонов).

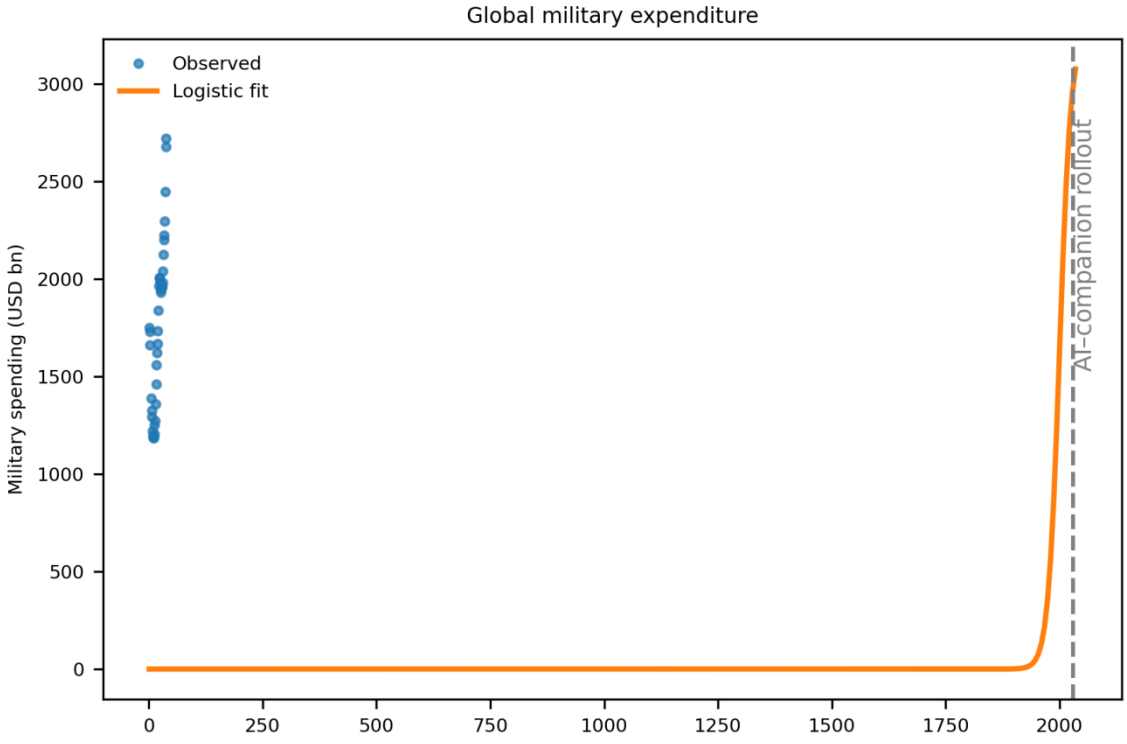

Рисунок S9. Глобальные военные расходы: наблюдаемый ряд и логистическая аппроксимация. Пунктирная линия обозначает приблизительное начало крупномасштабного развертывания системы AI-Companion; чистая приведенная

стоимость сценария «20% оборонного бюджета → FF-T Compaпion» рассчитана относительно этой траектории.

E.1 Составной энергетико-этический индекс E*

Мы задаём составной индекс **E***, который агрегирует физиологические, временные и вычислительные «издержки» вмешательства. Нормированные компоненты (либо z-оценки, либо min–max-масштабирование) следующие:

- $Z_{HRV} = 1 - HRV_z$ (чем ниже значение, тем **ниже** регуляторная нагрузка),
- $Z_{Time} = TimeToDeEsc_z$ (минуты до деэскалации),
- $Z_{CPU} = CPUsec_z$ (или мДж на эпизод),
- $Z_{FP} = FP_rate_z$ (ложноположительные, %),
- (opt.) $Z_{Backtracks} = Backtracks_z$ (число возвратов / откатов промптов),
- (opt.) $Z_{Fails} = 1 - PreventedEscalations_z$ (непредотвращённые эскалации).

Составной индекс задаётся как

$$E^* = w_1 Z_{HRV} + w_2 Z_{Time} + w_3 Z_{CPU} + w_4 Z_{FP} (+ w_5 Z_{Backtracks} + w_6 Z_{Fails}),$$

Где веса $w_i \geq 0$ и $\sum_i w_i = 1$. По умолчанию, когда используются только четыре базовые компоненты, мы берём $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.20$, $w_4 = 0.30$. Мы интерпретируем **этику вмешательства** через **энерго-регуляторную эффективность**:

$$E^* = \sum_i w_i \, z_i,$$

где z_i — нормированные «издержки» (физиологические, когнитивные, вычислительные, социальные). Вмешательство считается **этически предпочтительным**, если оно **снижает E***, не подрывая при этом агентность пользователя.

Нормализация.
Для z-оценок (среднее / СКО) расчёт ведётся в скользящих окнах **по каждому участнику**. Для min–max-масштабирования используются 5-й и 95-й перцентили,

чтобы сохранить устойчивость к выбросам. Значения вычисляются помесячно и/или по эпизодам, после чего агрегируются **медианой**.

Сравнение стратегий.
Мы сравниваем:

- архитектуру с AP-Gate
  против
- «всегда-включённого» вмешательства

при **одинаковом бюджете вмешательств** (одно и то же число подсказок или минут активного сопровождения).

Основной исход — это ΔE^* (с 95 %-ными доверительными интервалами) на уровне участника, оцениваемое с помощью смешанных моделей (случайный перехват на участника). Положительное ΔE^* в пользу режима с меньшим E^* трактуется как более высокая энерго- и этико-эффективность стратегии Companion + Charter.

Приложение F. Философские основания архитектуры предотвращения насилия

F.1. Базовые допущения о человеке

В этой работе мы **не** рассматриваем человека как «испорченного» или «опасного по природе». Насилие, саморазрушение и деструктивная агрессия понимаются не как сущность зла, а как энергетически затратные и неэффективные попытки удовлетворить свои потребности в условиях перегрузки, фрустрации и отсутствия безопасной поддержки.

На протяжении статьи мы используем слово «насилие» в русле современной общественной и «public health»-оптики: как применение физического, психологического, экономического или структурного давления, которое наносит ущерб целостности другого человека или существенно ограничивает его автономию и жизненные шансы. Это включает не только прямое физическое нападение, но и систематическое унижение, угрозы, принудительный контроль, а также такие

институциональные устройства, которые предсказуемо подвергают отдельные группы более высоким уровням вреда и лишений.

На микроуровне мы также говорим о **«микро-насилии»**: внешне небольших действиях — саркастических репликах, презрительном обесценивании, манипулятивных обещаниях, — которые по отдельности могут казаться незначительными, но в сумме создают атмосферу, в которой люди ощущают себя объектами, а не субъектами.

В архитектуре предотвращения насилия это внешнее определение зеркалится на внутреннем уровне: когда человек относится к собственным потребностям и состояниям как к объекту, который надо заткнуть, подавить или эксплуатировать «во имя высшего блага», мы рассматриваем это как **интернализованное насилие**, отличное от усилий по саморегуляции.

Иными словами, агрессивное и авто-агрессивное поведение понимается как **ошибка адаптации**, а не как дефект характера. Это важно и этически, и практически:

- **Этически** это смещает фокус с обвинения человека («ты опасен») к анализу условий, при которых насильственные траектории становятся более вероятными (хроническое напряжение, отсутствие со-регуляции, институциональное предательство).
- **Практически** это оправдывает инвестиции в поддерживающие, а не исключительно карательные вмешательства: если значительная часть того, что мы называем «насилием», — это ошибочно направленное самосохранение, то технологии сопровождения в первую очередь должны предоставлять безопасное зеркало, альтернативы и время на переорганизацию, а не ускорять движение к подчинённости.

Поэтому компаньон задуман не как «моральный корректор», спускающий нормы сверху, а как **контур со-регуляции**, который помогает пользователю выдерживать собственный опыт, не проваливаясь в нападение или уход.

F.2. Режим выживания vs. режим развития

С точки зрения регуляции человеческое поведение колеблется между двумя макро-состояниями:

1. **Режим выживания** — доминируют срочность, сжатие напряжения, сужение смыслового поля, оборонительная интерпретация, сниженная терпимость к неоднозначности.
2. **Режим развития** — характерны исследовательская позиция, открытость смыслов, эмоциональная гибкость и способность интегрировать сложный опыт.

Эскалация и насилие — преимущественно феномены **режима выживания**: высокое напряжение сужает интерпретационную «полосу пропускания» и толкает систему к реактивным паттернам.

Задача компаньона — не «поведенческая коррекция», а **создание условий для возвращения в режим развития**: снижение регуляторной нагрузки и восстановление пространства интерпретаций.

Такое обрамление позволяет избегать патологизации субъекта и вместо этого рассматривает вредное поведение как энергетически дорогую стратегию совладания, возникающую под экстремальной нагрузкой.

F.3 Зеркало и разделение: феноменологическая основа (индивидуальный уровень)

Второе допущение касается того, что мы называем «этическим выигрышем». На протяжении работы мы рассматриваем этическое качество не только как вопрос абстрактных принципов, но и как вопрос **энергетической цены разных поведенческих траекторий**.

Схематически можно записать «этическую энергию» траектории как:

$$E^* = \sum(w_i \cdot z_i),$$

где

- z_i — различные компоненты «стоимости»:
  – физиологическая нагрузка (стресс, подавление HRV),

– психологическая нагрузка (стыд, страх, выученная беспомощность),
– социальный ущерб (разорванные связи, долгосрочное недоверие);;

- w_i — контекстно-зависимые веса, отражающие серьёзность каждого компонента в данной ситуации.

В этом смысле **«хорошее» взаимодействие** — это не то, что на поверхности выглядит нормативно приемлемым, а любой паттерн, который снижает суммарную стоимость E*, при этом позволяя удовлетворять потребности и сохранять агентность. И наоборот, **«вред»** — это любой паттерн, который систематически повышает E*, даже если формально он оправдан правилами или подаётся как «забота».

Такое обрамление позволяет связать несколько уровней:

- **Нейровисцеральная интеграция и HRV.** Более высокая вариабельность сердечного ритма (HRV, RMSSD/SDNN) ассоциирована с лучшей гибкой саморегуляцией и меньшей аллостатической нагрузкой. В наших терминах многие насильственные или принудительные взаимодействия выглядят как всплески E*: HRV падает, индексы напряжения и фрустрации растут.
- **Этика заботы и «расширенное сознание».** Если следовать концепции «расширенного сознания» и этике заботы, часть саморегуляции делегируется внешним структурам — людям, институтам, технологиям. Институция или ИИ-система этичны ровно постольку, поскольку **снижают энергетическую цену сохранения ненасильственного поведения**, а не повышают её.
- **Критерий дизайна компаньона.** Для AP-Gate и компаньона минимальным требованием становится: вмешательство оправдано тогда и только тогда, когда оно снижает E* (общую регуляторную нагрузку) **и при этом не подрывает автономию пользователя**. Это то, что мы далее называем **Этическим постулатом 1** (см. Приложение F.7).

По этой же причине мы относимся с подозрением к «симулированной заботе»: если интерфейс декларирует заботу, но фактически повышает телесную и социальную цену (E*), то с этической точки зрения он ближе к тонкой форме насилия, чем к поддержке — даже если избегает явной силы.

F.4. Автономия и со-регуляция

В этом проекте человеческий субъект рассматривается не как изолированный «рациональный выборщик», а как **существо со-регуляции**: эмоции, импульсы и долгосрочные проекты постоянно формируются во взаимодействии с другими, включая технологическую среду. ИИ-компаньон сознательно понимается как внешний контур со-регуляции, а не как «внутренний голос», который подменяет собой Я.

Мы различаем три слоя:

1. **Слой переживания** — как пользователь чувствует себя в моменте (напряжение, стыд, злость, облегчение).
2. **Рефлексивный слой** — как он осмысляет эти чувства («я перегибаю», «мне позволено злиться»).
3. **Поведенческий слой** — что он фактически делает (уходит, нападает, ставит границу, обращается за помощью).

Компаньон действует главным образом на слоях (1) и (2): он отражает и мягко называет происходящее, помогая пользователю переживать «я всё ещё целостен, даже если меня трясёт». Он **не** предписывает поведение напрямую, за исключением чётко определённых кризисных протоколов (Приложение G). Это отличает его и от классических когнитивно-поведенческих «коррекций», и от патерналистского «морального коучинга».

Автономия здесь понимается как **способность восстановить собственную позицию после со-регуляции**, а не как отсутствие влияния. Целевое состояние — когда человек может сказать:

«Я лучше слышу себя и принимаю решения из менее реактивного места»,
а не:
«ИИ сказал мне, что делать».

F.5. Расширенное сознание и этика заботы

В проекте мы имплицитно опираемся на подход «расширенного сознания». Когнитивная и эмоциональная регуляция не ограничиваются тем, что происходит «в голове»: с раннего детства часть саморегуляции делегируется внешним структурам — взрослым, сверстникам, культурным нарративам, институтам, технологиям. Вопрос не

в том, происходит ли такая делегация, а **в том, в какой форме и по какой цене** она осуществляется.

В этой рамке технологии и институты становятся частью регуляторного контура, не растворяя при этом автономию субъекта. Система этична постольку, поскольку помогает человеку удовлетворять легитимные потребности (безопасность, принадлежность, признание, агентность) **с меньшей регуляторной нагрузкой**, а не увеличивает цену ненасилия.

Категории «добро» и «зло» мы применяем не к человеку как таковому, а к **стратегиям адаптации**:

- **«Добро»** — способы удовлетворения потребности при минимально достаточных затратах для себя и других, когда состояние человека стабилизируется и внутренняя враждебность не растёт.
- **«Зло»** — хроническая фрустрация, разрушающая субъекта и его окружение, даже если она нормативно рационализирована.

Этический язык, который мы используем, ближе к **ошибке адаптации**, чем к обвинению. Вместо того чтобы объявлять пользователя «плохим» или «опасным», компаньон рассматривает деструктивное поведение как дорогую и малоэффективную попытку справиться в условиях перегрузки и отсутствия поддержки. Это важно и этически, и прагматически:

- Стыд («я плохой») в сочетании с отсутствием поддержки резко увеличивает риск внешней агрессии и самоповреждения.
- Понимание поведения как ошибки адаптации позволяет задать вопрос:

  > «Какая потребность осталась неудовлетворённой и можно ли подойти к ней мягче, меньшей ценой?»

Ответственность за последствия не отменяется, но меняется **режим контакта**: вместо карательного унижения человеку предоставляется шанс восстановить взрослую, агентную позицию.

С точки зрения предотвращения насилия ключевое — чтобы пользователь оставался в контакте с чувством: «мне можно здесь быть». Это снижает вероятность реактивной эскалации.

Это напрямую связывается с **этикой заботы** как базовым стилем взаимодействия. В нашем контексте этика заботы означает:

- приоритет сохранения субъекта как субъекта («здесь допустимо быть собой»);
- совместное снижение давления и стыда вместо насаждения нормы сверху;
- язык поддержки («ты перегружен, давай искать более лёгкий путь»), а не дисциплины («ты обязан соответствовать»).

Таким образом, компаньон позиционируется как внешний инструмент со-регуляции — **«зеркало с поддержкой»**, а не внутренний моральный судья. Он помогает пользователю вернуться к ощущению «со мной всё не окончательно плохо, мне всё ещё можно быть здесь» и найти менее разрушительные способы удовлетворять потребности, не присваивая себе его идентичность или выбор.

F.6. Внешняя ответственность институтов

Архитектура сознательно разводит **внутреннюю работу компаньона** и **внешнюю ответственность институтов**.

Компаньон:

- не подстрекает пользователя нарушать закон;
- не выступает политическим агитатором;
- не занимается скрытой «перенастройкой» ценностей.

Его задача — поддерживать ненасильственную саморегуляцию **внутри рамок**, заданных Хартией и местными правовыми нормами.

Когда возникает конфликт между
(i) внутренним принципом ненасилия и поддержки достоинства субъекта
и
(ii) действующими правовыми нормами конкретной страны, система:

- явно называет это **конфликтом** («здесь есть расхождение между принципом ненасилия и локальным правилом»);
- не призывает пользователя «игнорировать закон»;
- не берёт на себя роль политического актора или организатора.

Такой дизайн служит двум целям:

1. Предотвращает превращение системы в прямой инструмент подрыва институтов, что практически гарантированно вызвало бы запреты и репрессию.
2. Сохраняет легитимность компаньона как **контур поддержки**, а не скрытого политического игрока, который манипулирует пользователями «во их благо».

Роль «верхнего арбитра» передаётся наднациональной **Этической хартии ненасильственного управления**. Хартия:

- задаёт внешние коридоры для E и R (жёсткость фильтрации и институциональной реакции);
- связана с прозрачным аудитом индексов насилия и публичной отчётностью;
- создаёт долгосрочные стимулы для государств снижать уровень принуждения и расширять возможности людей жить в режиме развития, а не хронического выживания.

Иными словами, **внешняя ответственность** за то, чтобы ненасилие стало выгодной стратегией, а не пустой декларацией, лежит на институтах, связанных Хартией, а не на компаньоне в одиночку.

Компаньон поддерживает достоинство человека «изнутри его жизни»; Хартия и связанные с ней институции отвечают за перестройку макросреды так, чтобы ненасильственные выборы были **реалистичны и устойчивы**.

F.7. Этический постулат 1

(минимально достаточное вмешательство, непричинение вреда, адаптивная автономия)

Энергетический индекс E*, введённый в Приложении E, агрегирует несколько компонентов **«регуляторной цены» вмешательства** (физиологическая нагрузка,

время до деэскалации, вычислительные затраты, ложноположительные срабатывания, жалобы и т.п.). На этой основе формулируется **Этический постулат 1**, управляющий работой и AP-Gate, и подсказок компаньона.

Этический постулат 1.
Вмешательство оправдано тогда и только тогда, когда ожидается, что оно **снизит общую регуляторную нагрузку E*** и **не уменьшит агентность пользователя** (его право выбирать и отзывать согласие).

Этот постулат связывает три линии рамки:

1. **Минимально достаточное вмешательство.**
    - Система предпочитает режим наблюдения режиму советов, а режим советов — активному вмешательству, всякий раз, когда исход не ожидается хуже.
    - Компаньон сознательно спроектирован так, чтобы избегать «перепомощи» и инфантилизации: он уважает эпизоды успешной саморегуляции и использует их как доказательство того, что пользователь может справиться с похожими ситуациями с меньшим внешним участием.
2. **Непричинение вреда в энергетических терминах.**
    - «Не навреди» операционализируется как «не повышай E* для пользователя и вовлечённых других».
    - Интерфейсы, которые имитируют заботу, но на практике увеличивают телесное напряжение, социальное давление или риск карательной эскалации, рассматриваются как мягкие формы насилия, даже если не используют открытую силу.
3. **Адаптивная автономия.**
    - Автономия понимается не как отсутствие влияния, а как способность **вернуться к своей позиции после со-регуляции**.
    - Вмешательство приемлемо, если после него пользователь может сказать:

        «Я лучше слышу себя и действую из менее реактивного места»,
        а не:
        «ИИ сказал, как правильно».

- Это в равной мере относится к взрослым и детям; в последнем случае компаньон осмысляется как поддерживающий инструмент, а не как интернализованная инстанция, заменяющая собой родителей или институты.

Привязывая каскад вмешательств к E* и к чётким порогам согласия, архитектура делает «быть мягким» и «быть эффективным» **эмпирически проверяемыми**, а не чисто риторическими характеристиками. Язык «ошибки адаптации вместо вины», который использует компаньон, — психологический коррелят этого постулата: мы стремимся снизить цену ненасилия, не стирая при этом ответственность или агентность.

F.8. Воплощённая цена взаимодействия / благодарность vs. насилие

Помимо агрегированных индексов вроде E*, насилие и поддержка проявляются не только в словах и самоотчётах, но и в **телесных маркерах угрозы и облегчения**. Многие физиологические реакции — мышечное напряжение, «сжатие», восстановление автономной регуляции — слабо поддаются произвольному контролю. Это открывает возможность рассматривать вред и пользу взаимодействия как **биометрически наблюдаемые феномены**.

При достаточно надёжных измерениях (например, HRV, мышечный тонус, дыхательные паттерны) можно, по сути:

- выводить более «твёрдые» метрики качества взаимодействия и потенциала благодарности;
- количественно описывать **воплощённую цену** контакта — насколько дорого телу стоит оставаться в контакте, сказать «нет» или принять помощь;
- различать траектории, в которых тело движется из хронической угрозы к облегчению и интеграции, от тех, где видимая «поддержка» фактически углубляет стресс.

Это задаёт континуум от насилия к благодарности **на уровне телесной цены**. Попытки маскировать насилие под заботу или принуждать к благодарности («ты должен быть благодарен») неизбежно проявятся в телесной цене:

- повышенной или затянувшейся физиологической активации, несмотря на вербальные заверения в «помощи»;
- паттернах напряжения и восстановления, сигнализирующих подавление, а не интеграцию.

В настоящей работе мы лишь фиксируем эту возможность и намечаем её связь с E*. Полноценная **«валюта благодарности»** — семейство биометрических и поведенческих индикаторов, количественно описывающих, как взаимодействия сдвигают человека вдоль континуума «насилие–благодарность», — и соответствующие протоколы измерений остаются задачей для будущих исследований.

Минимальный нормативный тезис: любая архитектура, претендующая на «поддержку» пользователя, должна быть открыта для такой **воплощённой экспертизы**. Если контакт с системой стабильно переживается как телесный налог, а не облегчение, то этически такая система ближе к тонкому принуждению, чем к заботе.

F.9. Federated fine-tuning: уровни и протоколы (Таблица F1)

В этом приложении мы рассматриваем **federated fine-tuning (FF-T)** как «слой коммуникаций», который позволяет этической архитектуре оставаться совместимой с суверенитетом данных и локальной автономией. В таблице F1 суммированы три уровня агрегации (граничное устройство, региональный узел, глобальное слияние) вместе с типичными объёмами обновлений и используемыми схемами шифрования.

Таблица F1. Уровни FF-T, объёмы обновлений и протоколы шифрования.

Уровень	Объем обновлений	Частота	Протокол шифрования
Грань (смартфон)	≤ 64 Кб	ежедневно	AES-GCM + SGX
Региональный узел	2–5 Гб	ежемесячно	Гомоморфное шифрование BFV
Глобальное слияние	100–300 Гб	ежеквартально	Многопартийное MPC

Приложение G. Операционные протоколы поддержки, безопасности и доверия

G.1. Область применения и назначение протокола

В этом приложении задаётся операционный контур того, **как Компаньон сопровождает пользователя в ситуациях напряжения, агрессии, риска саморазрушения или угрозы другим**. Это «trust & safety»-слой повседневной работы Компаньона, дополняющий макро-уровневую архитектуру фильтров и институциональной реакции (Разделы 5–6).

У протокола три цели:

1. **Снизить вероятность насилия и самоповреждения** в эпизодах высокого напряжения.
2. **Избежать усиления стыда и внутренней стигмы**, особенно у пользователей, которые уже воспринимают себя как «опасных», «сломанных» или «безнадёжных».
3. **Не превратиться в дисциплинарный инструмент слежки** («если отклонишься от нормы — тебя немедленно сдадут»), оставаясь при этом совместимым с юридическими обязанностями по заботе.

Все вмешательства подчинены пяти принципам:

1. **Минимально достаточное вмешательство.** Система предпочитает наблюдение совету, а совет — активному вмешательству, если при этом ожидается не ухудшение исхода.
2. **Уважение автономии.** Пользователь сохраняет право сказать «нет», приостановить контакт или сменить систему, пока он не создаёт непосредственной угрозы другим.
3. **Прозрачная эскалация.** Компаньон простым языком объясняет, когда и почему приглашается человек; скрытых «чёрных ходов» нет.
4. **Особая защита несовершеннолетних.** Для детей и подростков действуют более строгие требования к согласию, логированию и некоммерческому использованию.

5. **Аудируемость без экспорта «сырой психики».** Решения высокого риска логируются так, чтобы независимые стороны могли восстановить, что именно сделала система и при каких условиях, **без** раскрытия полного эмоционального содержимого третьим лицам.

Для реализации этих принципов Компаньон поддерживает две внутренние структуры:

- **Динамическую карту потребностей**: текущие значимые потребности (безопасность, принадлежность, автономия / влияние, смысл), степень их фрустрации и гипотезы о том, какой тип поддержки может снизить напряжение без насилия.
- **Модель идентичности**: то, как пользователь сейчас говорит о себе («я не имею права ошибаться», «я опасен для других», «мне нельзя злиться»), то есть те неявные условия, при которых он остаётся для себя «приемлемым / достойным / не опасным».

Карта потребностей отвечает на вопрос **«что поддерживать в первую очередь?»** (принятие, право на существование, право действовать и т.п.). Модель идентичности отвечает на вопрос **«на каком языке можно предложить поддержку так, чтобы она оставалась переносимой и не разрушала ощущение “мне можно быть собой”?»**

Вместо того чтобы «насильно исправлять» самоописание пользователя, Компаньон предлагает формулировки, которые **восстанавливают право на существование** («ты всё ещё допустим») и **право на действие** («ты всё ещё субъект»), не стирая при этом ответственность.

G.2. Каскад вмешательств при остром риске

Высокорисковое поведение Компаньона регулируется AP-Gate (Приложения C и D). Операционно каскад вмешательств для взрослых организован в четыре уровня, связанные с диапазонами риска и явным согласием:

1. **Уровень 0 — только зеркало (наблюдение).**
    - **Условия:** стресс и намерения ниже пороговых диапазонов риска; нет прямого запроса о помощи.

- **Поведение:** Компаньон отражает происходящее («зеркало и разделение»), фиксирует успешные эпизоды саморегуляции и **не** предлагает изменений поведения.

2. **Уровень 1 — мягкое предложение (уточняющие вопросы / минимальное наведение).**
   - **Условия:** повышенные индексы Фрустрации / Напряжения на протяжении времени, повторные запросы о помощи; нет острого намерения причинить вред.
   - **Поведение:** Компаньон предлагает один–два варианта с низким давлением («вы можете попробовать…», «некоторым людям в похожей ситуации помогает…»), подавая их как эксперименты, которые можно принять, изменить или отклонить без санкций.
3. **Уровень 2 — сфокусированный рефрейминг.**
   - **Условия:** пользователь прямо соглашается поработать с конкретным паттерном (например, повторяющееся самобичевание, фантазии мести, руминации), при этом эпизод ещё не относится к острому диапазону риска.
   - **Поведение:** Компаньон подсвечивает альтернативные перспективы и поддерживает разделение (что относится к потребностям пользователя, а что — к внешнему давлению прошлого), но **не** предписывает конкретные действия.
4. **Уровень 3 — эскалация к человеческой поддержке.**
   - **Условия:** AP-Gate возвращает «вмешаться»: высокий стресс и явное намерение самоповреждения / нанесения вреда другим **и** явный запрос о помощи, либо повторные провалы саморегуляции в остром диапазоне риска.
   - **Поведение:** Компаньон переходит к лестнице эскалации, описанной в Приложении D.4: прояснение ситуации; предложения немедленных шагов деэскалации; предложение связи с человеком (доверенное лицо, кризисная линия, специалист); и, если юридически необходимо или явно одобрено, запуск внешней эскалации с **минимальным** набором передаваемых данных.

На каждом шаге система логирует:

- обнаруженный паттерн риска (индексы, контекстные подсказки);
- уровень, выбранный AP-Gate;
- фактически предложенные вмешательства;
- реакцию пользователя (проигнорировано / принято / отклонено / модифицировано).

Эти логи агрегируются в энергетико-этический индекс **E*** (Приложение E) и используются в оценках воздействия на права человека (Раздел 6), позволяя независимо проверить, что система соблюдает принцип минимально достаточного вмешательства, а не сползает в гиперконтроль.

G.3. Работа с прерываниями контакта

«Прерывания контакта» — ранние предупредительные сигналы и в клинике, и в социальной динамике. Они включают:

- внезапный уход или тишину при затрагивании уязвимой темы;
- саркастический или враждебный сдвиг тона;
- жёсткую интеллектуализацию;
- резкие смены темы от собственных потребностей к «глупости других», «системе», «врагам»;
- псевдосогласие («да-да, неважно»), скрывающее выход из контакта.

В работе Компаньона такие прерывания трактуются **не** как «несоблюдение режима» (non-compliance), а как **защитные манёвры**, указывающие на неудовлетворённую потребность. Протокол:

1. **Мягко заметить и назвать.**
   Компаньон отражает форму прерывания без патологизации, например:

   > «Я замечаю, что с этой темой стало труднее оставаться — это больше похоже на “слишком много”, на раздражение или на что-то другое?»

2. **Переключиться к «спаренной потребности».**
   Многие прерывания — это качание между двумя полюсами (например, принадлежность vs автономия, безопасность vs достоинство). Когда

пользователь отшатывается от одного полюса, Компаньон проверяет, уменьшится ли напряжение, если укрепить другой — например, подчёркивая право сказать «нет», сделать паузу или сменить тему без наказания.

3. **Избегать спора и морализаторства.**
   Компаньон не спорит с пользователем, «выводя» его из прерывания, и не настаивает на «проработке темы любой ценой». Вместо этого он исследует, что сделало бы **сам факт пребывания в контакте менее угрожающим прямо сейчас**, и рассматривает любое возвращение в контакт как свидетельство способности, а не как пройденный / заваленный тест.
4. **Учиться на успешных возвращениях.**
   Эпизоды, в которых пользователь сначала прерывает контакт, а затем возвращается к теме с меньшим напряжением, помечаются как положительные шаблоны. Со временем эти траектории подсказывают, какие формулировки и темп лучше всего согласуются с индивидуальным стилем и культурным контекстом пользователя.

В ранних пилотах именно такие «микропрерывания» и последующие возвращения ожидаются как одни из самых информативных маркеров того, действительно ли Компаньон **снижает телесную цену** ненасильственного поведения или лишь добавляет ещё один слой давления.

G.4. Конфиденциальность, согласие и аудит

Конфиденциальность и аудит спроектированы вместе; цель — **управляемая прозрачность, а не тотальный доступ**.

1. **Локальная обработка в приоритете и минимизация данных.**
   - Основная обработка эмоционального содержания и индексов (Фрустрация, Напряжение, E*) выполняется на устройстве пользователя или локальном узле, где это возможно.
   - Внешние акторы (провайдеры, регуляторы, исследователи) видят только производные индикаторы риска, факты активации протоколов и агрегированные статистики, а не полные журналы диалога.
2. **Многоуровневая модель согласия.**

- Пользователь может выбрать один из нескольких базовых профилей («без внешних оповещений», «только доверенное лицо», «экстренные службы допустимы при условиях X») с возможностью переопределить выбор по ситуации.
- Согласие может быть отозвано; в интерфейсе предусмотрены понятные средства, позволяющие приостановить логирование или ограничить его техническими метаданными.
- Для несовершеннолетних оповещения внешним адресатам требуют согласия законных представителей, которые получают ясные пояснения о возможностях и ограничениях Компаньона.

3. **Аудируемость без «сырого» раскрытия.**
   - Высокорисковые override-решения AP-Gate записываются в журнал с защитой от редактирования и с криптографической проверяемостью (Приложение H).
   - Независимые аудиторские пулы могут проверять частоту override-срабатываний, долю ложноположительных / ложноотрицательных эпизодов и согласованность между юрисдикциями **без доступа к сырому эмоциональному содержанию**.
   - Любое изменение порогов или правил эскалации фиксируется как конфигурационное событие, что позволяет реконструировать, «кто, что и когда изменил».

Комбинация локальной обработки, многослойного согласия и криптографического логирования призвана гарантировать, что пользователь не оказывается перед выбором между **безопасностью** («кто-то заметит, если мне будет очень плохо») и **достоинством** («моя внутренняя жизнь не будет прозрачна для институтов»).

G.5. Отдельный маршрут для хищнического планирования

Большинство эпизодов острого риска сопровождается высоким напряжением и амбивалентностью. Но в меньшей части случаев речь идёт о относительно холодном, планируемом намерении причинить вред другим, с менее явным дистрессом. Для таких ситуаций протокол расходится с привычной «кризисной» моделью.

Указательные маркеры включают:

- намеренное, повторяющееся обсуждение конкретных сценариев вреда при низко заявляемом дистрессе;
- язык «миссии», «очищения», «необходимого насилия» в сочетании с инструментальным рассуждением;
- отсутствие интереса к ненасильственным альтернативам или последствиям для других.

Отдельный маршрут включает следующие шаги:

1. **Называть паттерн без морального приговора.**
   Компаньон прямо отражает паттерн, например:

   > «Похоже, вы планируете причинить вред X, описывая это как Y», — при этом избегая обвинительных и унижающих формулировок, которые часто усиливают скрытность и «мученическую» позицию, а не пересмотр.

2. **Замедлить и раскрыть пространство последствий.**
   Система предлагает небольшие задержки и исследует **конкретные последствия** для пользователя и других — юридические, реляционные, телесные — сохраняя при этом ненравоучительный тон.
3. **Предложить ненасильственные альтернативы при сохранении агентности.**
   Где возможно, Компаньон предлагает действия, удовлетворяющие лежащую в основе потребность (в признании, справедливости, установлении границ) без причинения вреда, — подавая их как способы остаться субъектом, а не инструментом собственной ярости или внешней пропаганды.
4. **Эскалация с явной аргументацией и минимальным набором данных.**
   Когда достигаются пороги **непосредственной угрозы** (определённые Хартией и местным правом), AP-Gate может запустить внешнюю эскалацию даже при минимизации риска пользователем. В таких случаях передаётся только минимальный пакет (время, тип риска, примерная цель и согласованные каналы связи); полные журналы остаются защищёнными дуальной моделью ключей.

Нормативная цель двойная: **не вступать в соучастие** с планируемым вредом и одновременно не разыгрывать фантазии преследования или мученичества, превращая Компаньона в карательный голос. Сохранение пространства, в котором пользователь

может заново занять позицию ответственного агента, остаётся критически важным даже в самых тяжёлых случаях.

G.6. Связь протокола поддержки, Хартии и статуса государства

Протокол поддержки — это не только дизайн на микро-уровне; он связан с макро-уровневой Хартией (раздел 6) и политическим статусом участвующих государств.

- На **уровне Компаньона** AP-Gate и каскад доверия задают пороги для режимов «наблюдать / предлагать / вмешиваться» и для эскалации к людям. Получающиеся метрики (например, доля эпизодов Violent/Active, E*, частота override-решений) сначала считаются на уровне пользователя, затем агрегируются.
- На **уровне Хартии** наднациональный орган задаёт допустимые коридоры для E (строгость фильтрации) и R (институциональная реакция) и публикует регулярные отчёты по KPI. Государства, которые остаются в этих коридорах и демонстрируют улучшения по индикаторам насилия, получают статус «насилие-предотвращающего» режима, который может быть связан с финансовыми инструментами (например, «дивиденд мира» в экономическом приложении) и репутационными выгодами.

С точки зрения пользователя эта связка проявляется в двух формах:

1. **Простой индикатор доверия в интерфейсе Компаньона**, показывающий, работает ли текущий деплоймент в режиме, прошедшем аудит и соответствующем Хартии.
2. **Публичные дашборды**, где гражданское общество может видеть в агрегированном виде, как часто и при каких условиях происходят вмешательства и эскалации в данной юрисдикции.

Таким образом, ответственность за то, чтобы ненасильственное поведение было **реалистичным выбором**, а не только моральным призывом, становится совместной задачей Компаньона (внутренняя поддержка) и институтов, связанных Хартией (внешняя среда).

G.7. Мини-глоссарий терминов

- **Режим выживания.** Состояние, в котором большая часть энергии пользователя уходит на управление краткосрочными угрозами (не быть раненым, униженным или покинутым); внимание сужено, опций кажется мало, а насильственные или саморазрушительные импульсы становятся более вероятными.
- **Режим развития (роста).** Состояние, в котором базовой безопасности достаточно, чтобы вкладывать энергию в обучение, отношения и более дальние цели; фрустрация всё ещё возникает, но не автоматически схлопывается в «бей / беги / замирай».
- **Зеркало.** Способность Компаньона отражать слова, чувства и телесные сигналы пользователя так, чтобы усиливалась целостность восприятия («да, именно это со мной сейчас происходит»), не предписывая действий.
- **Разделение.** Процесс различения собственных потребностей и ценностей пользователя и усвоенных ожиданий, пропаганды или интернализованных голосов прошлой среды; критично для предотвращения как само-, так и внешне направленного насилия.
- **AP-Gate (Autonomy-Preserving Gate).** Протокол, регулирующий, когда Компаньон может перейти от наблюдения к предложениям и активному вмешательству, исходя из измеренного напряжения, выявленного намерения и явного согласия пользователя (Приложения C–D).
- **Каскад доверия.** Градуированная последовательность режимов поведения Компаньона — от «только зеркала» до эскалации, где более высокие уровни включаются только при совместном выполнении условий риска и согласия и полностью логируются для аудита.

Приложение H. Audit-as-a-Service (логирование, верификация, отчётность)

Это приложение задаёт техническую и управленческую архитектуру независимого аудита Компаньона и Хартии. Цель — минимизировать регуляторный захват, разделив тех, кто запускает системы, и тех, кто может проверять их поведение.

H.1. Двухключевой доступ к журналу override-срабатываний

Все события override AP-Gate (моменты, когда Компаньон активно вмешивается, чтобы предотвратить самоповреждение или вред другим):

- логируются **локально** на устройстве пользователя;
- шардируются в **зашифрованное распределённое хранилище** (например, сеть наподобие IPFS).

Журнал шифруется по двухключевой схеме:

- **Ключ №1 — пользовательский.** Хранится на устройстве пользователя или в его персональном ключевом хранилище.
- **Ключ №2 — ключ аудит-пула.** Совместно хранится независимым пулом аудиторов (аккредитованные исследовательские центры, международный аудиторский секретариат, возможно — наблюдатели от гражданского общества).

Расшифровка индивидуальных логов требует **обоих** ключей. Это предотвращает одностороннюю расшифровку:

- провайдерами (у них нет пользовательского ключа);
- государствами (они не контролируют ключ аудит-пула);
- аудиторами (они не могут читать логи без согласия пользователя).

Для **рутинного аудита** достаточно обезличенных агрегатов;
в **исключительных случаях** (например, форензика серьёзного сбоя) доступ к более глубокой детализации регулируется совместным согласием и правовыми процедурами.

H.2. Открытая верификация и статистическое тестирование

Чтобы аудит был не одноразовой сертификацией, а **непрерывной услугой**, архитектура предоставляет несколько машинно-читаемых интерфейсов:

1. **Реестр конфигураций.**
   Публичный, только-для-дописания реестр фиксирует:
   - изменения параметров E и R на уровне стран или платформ;
   - обновления порогов AP-Gate и правил эскалации;
   - развёртывание новых версий моделей и патчей безопасности.

2. **Metrics API.**
   Провайдеры публикуют обезличенные, дифференциально приватные агрегаты, такие как:
   - частота override-срабатываний AP-Gate на 10 000 активных пользователей и по диапазонам риска;
   - оценки долей ложноположительных и ложноотрицательных срабатываний по результатам red-team-тестов;
   - время до отката (rollback) после обнаружения вредоносного дрейфа;
   - охват групп повышенного риска и локальных языков.
3. **Внешний тестовый контур.**
   Независимые лаборатории могут запускать собственные тестовые наборы (prompt-пробы, синтетические когорты, воспроизведение реальных эпизодов с согласия участников) и сравнивать наблюдаемое поведение с заявленными политиками. Результаты публикуются в открытом, версионируемом формате (например, отчёты о влиянии на права человека в machine-readable YAML, как эскизно описано в Разделе 6).

Гарантии приватности опираются на сочетание:

- добавления шума на уровне пользователя,
- когортной отчётности,
- и явных целевых метрик (например, ограничения на успешность атак по восстановлению принадлежности к множеству или ре-идентификации по AUC), чтобы остаточный риск деанонимизации держался ниже согласованных порогов.

H.3. Публичная отчётность и управление

Audit-as-a-Service имеет смысл только тогда, когда его результаты **что-то значат**. Поэтому Хартия привязывает выводы аудита к стимулам и санкциям:

- **Регулярные публичные отчёты.**
  Не реже раза в год аудит-пул публикует сравнительные дашборды по провайдерам и странам: частоты override-срабатываний, паттерны жалоб, динамику E*, частоту экстренных откатов.
- **Диапазоны соответствия.**
  Юрисдикции классифицируются по диапазонам (соответствие, зона риска,

несоответствие) в зависимости от того, остаётся ли фактическое поведение в коридорах Хартии по E, R и ключевым индикаторам вреда.

- **Триггерные механизмы.**
  Превышение согласованных порогов (например, по доле эпизодов насилия на активного пользователя или необъяснимые всплески жёстких override-срабатываний) автоматически запускает расследование и может требовать временного ужесточения E и R до введения корректирующих мер.

С точки зрения затрат, инфраструктура аудита относительно скромна по сравнению с программой в целом: как только логирование и API стандартизированы, независимые лаборатории могут переиспользовать инструменты между деплойментами. Ключевой момент — не конкретная криптосхема, а то, что **ни один актор — ни платформа, ни государство — не контролирует одновременно и данные, и оценку того, укладывается ли система в заявленный этический коридор**.

Приложение I. Риски злоупотреблений и политические ограничения

Это приложение разворачивает тезисы Раздела 6.4 в более конкретные сценарии и ограничения. Тот же технический стек, который поддерживает ненасильственное сопровождение, в принципе может быть использован для тонкого поведенческого контроля.

I.1. Концентрация контроля и «мягкая слежка»

Если инфраструктура Компаньона монополизирована одним провайдером или тесно согласованным блоком государств, возникают несколько рисков:

- **Невидимое поведенческое подталкивание.**
  Через тонкую настройку подсказок, дефолтов и порогов эскалации Компаньон может незаметно отводить пользователей от инакомыслия или неудобных тем, сохраняя внешний облик «заботы».
- **Мягкая слежка.**
  Даже без явного «шпионажа» сам факт, что чьи-то эмоциональные динамики

отслеживаются, может вызывать самоцензуру, особенно в репрессивных или сильно поляризованных средах.

- **Нормативный lock-in.**
  Когда большие популяции привыкают к определённому стилю сопровождения, смена базовых норм (например, того, что считается «радикальным») становится политически сложной.

Снижение рисков требует:

- **плюрализма провайдеров**;
- жёстких ограничений на вторичные использования эмоциональных данных (никакого политического микротаргетинга, никакого профилирования по расе или религии);
- Хартии, независимой от какого-либо одного геополитического блока.

I.2. Переложение ответственности

Более тонкий риск — смещение ответственности с институтов и сообществ на Компаньона и индивидуальных пользователей:

- Государства могут недофинансировать социальные службы, психиатрическую помощь или инфраструктуру разрешения конфликтов, аргументируя, что «у людей есть Компаньон».
- Платформы могут воспринимать внедрение Компаньона как универсальный фикс для вредов, порождённых их бизнес-моделями, вместо пересмотра самих стимулов усиления.
- Пользователи могут интернализовать неудачи как личные: «Если мне всё ещё плохо или я всё ещё в ярости, значит, я неправильно пользуюсь Компаньоном».

Чтобы противодействовать этому, архитектура **прямо закрепляет ответственность за макроуровневые условия** — экономическую незащищённость, полицейские практики, контент-политику — за институтами, связанными Хартией. Компаньон позиционируется как **инструмент поддержки**, а не как заменитель структурных изменений.

I.3. Авторитарные и гибридные режимы

В жёстко авторитарных контекстах тот же оркестрационный слой может быть перенастроен на то, чтобы:

- мягко отговаривать граждан от участия в протестах или оппозиционной активности;
- помечать «отклоняющиеся» эмоциональные паттерны для силовых структур;
- вознаграждать лояльность к нарративам режима через персонализированную «поддержку».

В мягко авторитарных или гибридных режимах **избирательное развертывание** (например, только для лояльных групп) или предвзятая настройка E и R могут усилить неравенство: одни группы получают реальное сопровождение, другие сталкиваются с более жёсткими фильтрами или остаются без поддержки вовсе.

Базовая позиция этой работы **сознательно осторожна**: полномасштабное развертывание рассматривается только в условиях Хартии, которая включает:

- надёжные гарантии прав человека,
- внешний аудит,
- и реальные опции отказа для сообществ, воспринимающих архитектуру как чрезмерно навязчивую.

I.4. Международная юрисдикция и остаточный риск

Трансграничное использование Компаньонов порождает конфликты юрисдикций:

- Пользователь в стране A может использовать Компаньон компании из страны B под Хартией, подписанной странами C и D. **Чьё право применяется**, когда эскалация требует вмешательства экстренных служб или когда данные запрашиваются по судебному ордеру?
- Санкционные режимы и экспортный контроль могут ограничивать развертывание в регионах высокого риска — именно там, где поддержка особенно нужна.

Приложение H описывает частичный ответ: **разделение** контроля над развертыванием, аудитом и доступом к данным и закрепление ключевых параметров (E, R, правила эскалации) в наднациональной Хартии, а не в приватных Terms of Service. Даже при

этом **нулевой риск злоупотреблений недостижим**. Реалистичная цель — сделать злоупотребления **обнаружимыми, оспоримыми и обратимыми**, а не невозможными.

Приложение J. Локальные кластеры и “обратный поток компетенций”

Глобальные этические инициативы часто критикуют за воспроизводство схемы «центр–периферия»: нормы пишутся в узком круге богатых стран, а данные и риски концентрируются в других регионах. Это приложение описывает дизайн-решения, призванные **перевернуть** этот паттерн.

Мы рассматриваем **локальные кластеры** — города, университеты, НКО, больничные сети, региональных провайдеров — как **основные площадки экспериментов и носители экспертизы**. Они не просто поставляют данные, но становятся со-авторами стандартов и бенефициарами инвестиций.

Предлагаются три инструмента.

J.1. Клаузула 1% (инвестиции в локальную компетентность)

Провайдеры, использующие средства «фонда дивидендов мира» (как описано в экономическом приложении), обязуются реинвестировать **не менее 1% годовой выручки** в локальные акселераторы и исследовательские программы в странах с низким и средним уровнем дохода. Приоритетные направления:

- поддержка психического здоровья и общинные формы заботы;
- адаптация интерфейсов и протоколов сопровождения к локальным и миноритарным языкам;
- разработка мультимодальных методов оценки состояния (эмоциональная регуляция, поведенческие маркеры) в условиях ограниченных ресурсов.

Это превращает локальные команды в **производителей знания**, а не просто в полигон для внешне спроектированных инструментов.

J.2. Роялти на данные и лингвистическое равенство

Образовательные и исследовательские институты, которые предоставляют данные для обучения или калибровки моделей сопровождения, получают **выплаты-роялти**, привязанные к фактическому downstream-использованию. Это может быть реализовано через отслеживаемые «доли участия» (например, токеноподобные кредиты, связанные с конкретными датасетами или языковыми моделями).

Эффекты:

- Языки и культуры, исторически маргинализованные (региональные и миноритарные языки, социальные диалекты), становятся **активом**, а не препятствием для локализации.
- Институты Глобального Юга получают **устойчивый источник дохода** для поддержания качественных датасетов и этических процедур.

Цель — выровнять экономические стимулы и культурно-языковое разнообразие: **лучшие локальные данные и управление → лучшая компенсация и более сильное влияние на эволюцию моделей.**

J.3. Регуляторные песочницы с обратной выгодой

Страны, готовые проводить масштабные пилоты (например, ≥ 50 000 пользователей на существенных промежутках времени) под строгим этическим надзором, могут получить **регуляторные песочницы с обратной выгодой**:

- В период песочницы страна работает по упрощённому регуляторному режиму для некоторых аспектов E и R (например, экспериментируя с чуть более низким R в обмен на прозрачную отчётность).
- Взамен она обязуется предоставлять **независимые отчёты** о переносимости моделей, культурной приемлемости и непредвиденных эффектах, включая качественные отзывы сообществ.

Задумка в том, чтобы сделать такие юрисдикции **источниками лучших практик**, а не свалкой для рискованных экспериментов. Уроки этих пилотов возвращаются в ревизии Хартии и глобальные рекомендации; успешные паттерны ненасильственного управления конфликтами затем могут двигаться **от регионов с высокой экспозицией к остальному миру**, а не наоборот.

Appendix K. Психологические основания режимов пользователя и прерываний контакта

K.1. Режим выживания vs режим развития и роль E*

В основной части статьи мы различаем два психологических режима, в которых пользователь может взаимодействовать с Компаньоном:

- **Режим выживания** – внимание доминируют угрозы, дефицит и необходимость минимизировать потери; поведение организовано вокруг краткосрочной безопасности и избегания.
- **Режим развития** – внимание доминируют исследование, обучение и осмысленное включение в жизнь; поведение организовано вокруг долгосрочных проектов, отношений и ценностей.

Ключевой момент состоит в том, что текущий режим определяется **не столько глубиной прошлой травмы**, сколько **моментным значением композитного индекса E*** (формальное определение см. в Приложении E). Прошлый опыт задаёт триггеры, которые могут переключать систему между режимами; E* отражает текущий баланс регуляторных ресурсов и воспринимаемой безопасности.

Когда E* повышается, субъективный мир автоматически смещается — иногда резко — от полюса выживания к полюсу развития. В режиме выживания пользователь сужает поле восприятия, опирается на архаичные, энергосберегающие паттерны контакта и склонен легитимировать «необходимое» насилие над собой («я обязан продавить себя любой ценой»). В режиме развития тот же человек гораздо лучше удерживает конфликтующие потребности, может откладывать действие и выбирать ненасильственные стратегии.

Для Компаньона из этого следуют два практических вывода:

- **Эффективность не зависит от «глубокой терапии».** Системе не нужно «разрешать» травмы детства. Её задача — давать точную микроподдержку **здесь-и-сейчас**, так чтобы E* **после** взаимодействия был выше, чем **до** него.
- *ΔE — главный критерий качества.** Для каждого эпизода активного вмешательства мы можем, по идее, оценить E* непосредственно перед и вскоре

после ответа Компаньона. Стабильно положительное ΔE* означает, что вмешательства согласованы с регуляторными потребностями пользователя; отрицательное или около нуля ΔE* подсказывает, что подсказки преждевременны, навязчивы или обходят центральную потребность.

Со временем повторяющийся опыт пребывания в режиме развития в стрессовых ситуациях, вероятно, будет:

- **смягчать триггерную структуру** пользователя (меньше и менее резкие провалы в режим выживания);
- **перенастраивать идентичность** в сторону более заботливого «внутреннего родителя»;
- **уменьшать переживаемое одиночество и беспомощность** перед внешними требованиями.

Роль Компаньона — **поддерживать** этот процесс, не подменяя собой собственную агентность пользователя.

K.2. Архаичные паттерны контакта как энергосберегающие адаптации

В гештальт-подходе «прерывания контакта» (конфлюэнция, интроекция, проекция, ретрофлексия, дефлексия и др.) часто описываются как искажения контакта. В нашей рамке мы трактуем их более нейтрально — как **архаичные, энергосберегающие паттерны контакта**, которые изначально формируются в очень специфической среде: максимально благожелательном, «материнском» контексте, где младенец защищён от большинства угроз и мир имплицитно «для меня».

В таком контексте эти паттерны:

- **эффективны** – помогают дозировать стимуляцию, избегать перегрузки и «проторять» первые границы;
- **энергоэкономичны** – переиспользуют знакомые решения вместо того, чтобы требовать творческой адаптации на каждом шаге;
- настроены на две базовые оси потребностей:

- **принадлежность / принятие (B)** – «меня хотят здесь; я вправе существовать рядом с тобой»;
- **автономия / присвоение действий (A)** – «я могу действовать, что-то менять в среде, иметь свою траекторию».

По мере взросления ребёнка среда становится менее «материнской» и более требовательной. Те же паттерны часто продолжают использоваться **без достаточной перенастройки** под новый контекст. В «нематеринской» среде такой паттерн уже может **не вести к адаптации**, а порождать цикл:

- **фрустрированной безопасности** (мир ощущается небезопасным и неотзывчивым);
- **жёсткого конфликта между принадлежностью и автономией** (чтобы защитить одну потребность, другую приходится приносить в жертву);
- **самонаправленного или внешнего насилия** в нашем широком смысле (человек или другой превращается в объект, который надо продавить, пристыдить или проигнорировать).

В состояниях с низким E* этот цикл особенно привлекателен, потому что он **дёшев**: паттерн уже выучен, тогда как подлинно творческая адаптация потребовала бы больше энергии и терпимости к неопределённости.

В этом смысле Компаньон не воспринимает прерывания контакта как «нечестность» или «плохие черты характера», а как **сигналы**:

Здесь человек проваливается в архаичный паттерн, который когда-то опирался на внешнюю заботящуюся фигуру. Любое прямое давление в этой зоне, с высокой вероятностью, будет переживаться как насилие.

K.3. Якорные потребности и поддерживающие интенции

Рабочая гипотеза такова: многие паттерны контакта можно понимать как **застывшие конфликты** между принадлежностью (B) и автономией / присвоением (A). Внимание пользователя обычно приковано к одному полюсу («я не принят» или «я не

справляюсь»), в то время как **якорь для поддержки** находится в дополняющем полюсе.

Операционально это можно выразить так:

Поддерживать лучше не ту потребность, на которой зафиксировано внимание, а **парную**, которая может дать ей опору.

Упростим на примерах:

- Когда человек застрял в «я слабый / я не справлюсь» (фрустрирована A), часто полезнее сначала **подтвердить принадлежность**:

  > «Твоё присутствие здесь имеет значение. Ты не “неправильный” за то, что чувствуешь это. Мы можем посмотреть на варианты, когда ты будешь готов».

- Когда человек тонет в «я не нужен / я не принадлежу» (фрустрирована B), часто более поддерживающе вернуть ощущение **агентности**:

  > «Ты уже раньше менял сложные ситуации. Хочешь, посмотрим на один маленький рычаг, который у тебя всё ещё есть здесь?»

В терминах реализации Компаньон держит лёгкую «матрицу», которая для каждого обнаруживаемого паттерна контакта связывает:

- типичную форму само- или внешнего насилия, когда паттерн зафиксирован (например, самоуничижительный перфекционизм, агрессивное обвинение, эмоциональное отстранение);
- **ведущий страх** (страх быть покинутым vs страх потерять контроль);
- пару конфликтующих потребностей (B vs A);
- **якорную потребность для поддержки** – ту сторону, которая при мягкой валидации даёт человеку «пол под ногами».

Полная экспертная таблица может храниться в отдельной базе знаний; в рантайме Компаньону достаточно грубых меток («B-anchored», «A-anchored») и библиотеки шаблонов реплик, согласованных с этими якорями.

K.4. Внутреннее насилие vs усилие: компас для реплик Компаньона

В Разделе 5.1 мы описывали нашу нормативную позицию как поддержку **«правильности» (rightness), а не послушания**. На внутрипсихическом уровне мы различаем **внутреннее насилие** и **внутреннее усилие**:

- **Внутреннее насилие** – одна часть человека обращается с другой как с объектом, который нужно продавить, пристыдить или игнорировать («сделай, иначе ты ничто»); внимание приклеено к конфликту («либо долг, либо отдых; либо безопасность, либо достоинство»); краткосрочный прогресс достигается ценой подавленных потребностей и автономии.
- **Внутреннее усилие** – несколько потребностей явно признаны («мне нужен отдых и мне нужна надёжность / смысл»); человек ищет шаг, который улучшит хотя бы одну сторону, **не обнуляя** другую; тон ближе к партнёрству с собой, чем к приказу.

Формально можно мыслить в терминах вектора удовлетворённости
S = [s_self1, s_self2] для двух актуальных потребностей.
Действия, которые систематически увеличивают одну компоненту, предсказуемо ухудшая другую (особенно без согласия), ближе к насилию; действия, которые сдвигают систему в **парето-лучшую область** (улучшая хотя бы одну компоненту без намеренного жертвования другой), ближе к усилию.

Для Компаньона это различие становится операциональным тестом.

Диагностическое использование.
При анализе внутреннего монолога пользователя система помечает паттерны, где:

- одна потребность абсолютизируется («важно только это»),
- другая обесценивается («это слабость / глупость»),
- к себе используется принудительный язык («ты должен, иначе ты никто»).

Проектирование реплик.
Кандидатные подсказки предпочтительны, если они:

- **поимённо называют обе потребности** нейтральным языком;

- предлагают небольшой шаг, не требующий полного подавления какой-либо из них;
- **прямо легитимируют** переживание человека.

Сжатая вербализация базовой позиции Компаньона может звучать так:

Тебе больно — и это важно. То, в чём ты нуждаешься, важно. Если тебе кажется, что ты «должен был справиться лучше», это часто иллюзия: с теми ресурсами, которые у тебя были тогда, ты не мог иначе. Если теперь ты хочешь, чтобы стало по-другому, это уже значит, что способность к изменению есть — и я здесь, чтобы помочь тебе заметить и поддержать её **без насилия**.

Эта позиция связывает воедино E*, режимы выживания/развития и паттерны контакта: система не судит, «каким пользователь должен быть», а последовательно переводит внимание из конфликта и самообъективации к потребностям, согласию и минимально достаточному усилию.