

# План реализации чат-бота поддержки отчужденного родителя

## Клиническая база и психопросвещение (Workstream 1)

**Конфликт лояльности и влияние на детей.** При разводе одним из самых разрушительных сценариев является вовлечение ребенка в родительский конфликт. Это ведет к **конфликту лояльности** – состоянию, когда ребенок чувствует, что его вынуждают выбрать одного из родителей и скрывать свои чувства к другому. В таких условиях дети испытывают **страх, тревожность, вину и замешательство**. Они боятся, что любя одного родителя, потеряют любовь другого; опасаются выражать привязанность, чтобы не разозлить маму или папу. Ребенок живет в постоянном напряжении, пытаясь “угодить обоим” – что невозможно и приводит к эмоциональным нарушениям. Длительный нерешенный конфликт лояльности может привести к **синдрому отчуждения родителя** – когда ребенок полностью отвергает одного из родителей под влиянием другого. Последствия для детей крайне негативны: отмечаются депрессивные состояния, потеря доверия, нарушения привязанности, трудности во взаимоотношениях, которые могут сохраняться во взрослом возрасте <sup>1</sup>.

**Эмоции отчужденного родителя.** Родитель, отстраненный от ребенка, переживает тяжелую гамму чувств и часто находится в кризисном состоянии. Исследования и практический опыт показывают типичные эмоциональные **состояния отчужденного родителя**:

- **Шок и отрицание.** Первоначальная реакция: онемение, отказ верить, что ребенок отдаляется. Родитель думает: “Этого не может быть, все само наладится”. Он может ничего не предпринимать, надеясь на чудо. Риск – потеря времени и усугубление разрыва контакта.
- **Ярость и агрессия.** Постепенно шок сменяется гневом на второго родителя. Признаки: вспышки злости, обвинения и демонизация “алиенатора”, желание мести, импульсивные агрессивные действия (например, резкие выпады в суде). Иррациональные убеждения: “Он/она должен заплатить за это”, “Если я буду давить сильнее, я верну ребенка”. Риски – эскалация конфликта, юридические проблемы (угрозы), отталкивание ребенка еще сильнее.
- **Отчаяние и беспомощность.** Чувство полной потерянности: “Все безнадежно, я ничего не могу сделать”. Возможна реактивная депрессия, апатия, суицидальные мысли. Родитель опускает руки, перестает бороться за контакт (“Ребенок потерян навсегда, я плохой родитель раз так вышло”). Это критически опасное состояние (риск суицида, полного разрыва отношений).
- **Вина и самообвинение.** Родитель поверяет в собственную «вину» за ситуацию: чрезмерно копается в своих ошибках, теряет самооценку (“Если бы я был лучше, этого бы не случилось; я не заслуживаю своего ребенка”). Испытывает стыд перед окружением (стереотип “раз ребенок не хочет общаться, значит родитель плохой” усиливает это чувство). Виноватый родитель может пытаться “искупить вину” чрезмерными подарками

или уступками. Риск – полная пассивность или попадание в зависимое положение от диктата бывшего партнера.

- **“Торг” и навязчивые попытки вернуть все как было.** Некоторые родители переходят к стадии **иллюзорных сделок**: пытаются купить любовь ребенка подарками, умоляют второго родителя “смилиостивиться”, верят в возможность **волшебного решения**. Например, думают: “Если заплачу больше алиментов или выполню все условия, меня перестанут ограничивать”. Они хватаются за любые «чудо-методы», вместо системной работы. Риски – финансовые потери, эмоциональное выгорание, усиление зависимости от манипуляций.
- **Навязчивая борьба.** Другие застревают в режиме **постоянной войны**: бесконечные судебные иски, сбор доказательств против ex-партнера, жизнь вращается только вокруг конфликта. Родитель одержим *доказать свою правоту*, восстановить “справедливость” любой ценой. Здравый смысл может теряться: например, тратятся все средства на суды, игнорируются потребности ребенка здесь и сейчас. Риск – эмоциональное и финансовое истощение, усугубление отчуждения из-за непрекращающейся вражды.
- **Принятие реальности и стратегическое переосмысление.** В идеале, со временем родитель переходит к **принятию**: осознает факт проблемы и необходимость длительной работы для восстановления отношений. Признаки: трезвое понимание, что бывшего супруга не изменить, контроль есть только над собственными действиями; снижение остроты эмоций, готовность действовать рационально. Это поворот к конструктивному руслу – возможность строить новую стратегию поведения.

Важно подчеркнуть: не все проходят эти стадии линейно, возможны откаты. Чат-бот должен **диагностировать текущее состояние** пользователя по его высказываниям (эмоциональная лексика, когнитивные искажения, факты о ситуации) и адаптировать поддержку под него. Например, фразы типа “Я их уничтожу, они мне ответят” укажут на состояние ярости; “Все кончено, не хочу жить” – на отчаяние, и т.д. Каждое состояние несет разные риски, поэтому **правильная диагностика и таргетированное вмешательство** – ключ к эффективности поддержки.

**Стратегии отдельно проживающего родителя.** Психопросветительский блок бота учит отчужденного родителя правильной линии поведения, чтобы не усугублять конфликт лояльности и постепенно восстанавливать контакт. Основные рекомендации экспертов (в том числе Ю. Ковалёвой) такие:

- **Не втягивать ребенка в конфликт.** Ни при каких обстоятельствах нельзя делать ребенка “союзником” против второго родителя. Нельзя вымешивать на нем обиду, заставлять шпионить или выбирать сторону. “Поддерживать родителя – не детская задача” – бот будет регулярно напоминать об этом.
- **Контролировать свои эмоции по отношению к ex-партнеру.** Родителю важно прорабатывать гнев, обиду, чувство предательства во взрослом пространстве (через терапию, группы поддержки, техники эмоциональной саморегуляции), а не выплескивать их на ребенка. Совет: «Работайте со своими чувствами и отношением к бывшему супругу». Бот предлагает упражнения ННО/IFS для выражения гнева в безопасной форме (см. ниже) и когнитивные техники для снижения обидчивости.

- **Освоить ненасильственное общение (ННО).** Родитель учится выражать свои переживания и потребности без обвинений. Формула ННО включает: наблюдение (факт без оценки), чувство, потребность, просьбу <sup>2</sup>. Например, вместо “Она монстр, не дает мне дочь!” – сообщение: “Когда мне мешают видеться с дочерью (факт), я чувствую отчаяние и страх (чувства), потому что для меня важна связь с ребенком (потребность). Я бы хотел обсудить расписание встреч (просьба)”. Такой подход снижает конфликтность и создает пространство для диалога.
- **Не усугублять конфликт лояльности ответными действиями.** Даже если второй родитель настраивает ребенка, нельзя мстить тем же. Бот предупреждает: попытки “отвоевать” ребенка силой или через манипуляции только усилият разрыв. “Даже борьба за ребенка в правовом поле, растянутая на годы, может не приблизить вас к ребенку, а наоборот” – говорится в брошюре. Вместо этого лучше фокусироваться на постепенном налаживании отношений с ребенком напрямую.
- **Поддерживать одностороннюю связь с ребенком, даже если она не взаимна.** Родителю рекомендуется показывать, что он по-прежнему рядом: **писать письма, открытки, вести дневник для ребенка, отправлять подарки** (даже если они не доходят). Эти свидетельства любви могут быть использованы позже, когда контакт восстановится – ребенок увидит, что мать/отец не забывал о нем. Бот может напоминать пользователю раз в неделю написать нейтральное письмо или сообщение ребенку (“Привет, я люблю тебя, скучаю и буду рядом когда ты будешь готов общаться”) – такие “ритуалы присутствия” поддерживают связь на расстоянии.
- **Не обвинять и не давить на самого ребенка.** Даже если ребенок говорит обидные вещи (возможно, со слов другого родителя), важно не реагировать агрессией. Совет: не злиться на ребенка, понимая, что он заложник ситуации. Вместо упреков – терпение и безусловная любовь. Бот разъясняет феномен **когнитивного диссонанса у ребенка**: ему проще отвергать того, кто ведет себя агрессивно в ответ, поэтому доброжелательное, спокойное поведение отчужденного родителя – залог того, что у ребенка останется “мостик” к нему.
- **Обращаться за помощью и учиться новому.** Родителю стоит воспользоваться профессиональной поддержкой: терапевтом (проработать травму и научитьсяправляться со стрессом), юристом (понять правовые границы), медиатором (попробовать договориться в безопасной обстановке). Также рекомендуется искать группы взаимопомощи отчужденных родителей, читать литературу, чтобы понимать психологические процессы. Брошюра советует: “Заботьтесь о себе и своем ментальном здоровье. Исцеление вашей травмы утраты – необходимое условие восстановления отношений с ребенком”. Бот может предоставлять материалы (статьи, истории других родителей, рекомендации книг) – это элемент психообразования.

**Роль школы и соцслужб.** Внешние нейтральные институты – педагоги, школьный психолог, социальные работники – могут существенно помочь минимизировать конфликт лояльности. Чат-бот разъяснит пользователю, как правильно привлекать эти ресурсы:

- **Информирование и профилактика.** Школа и службы опеки часто не осведомлены о проблеме конфликтов лояльности. Однако **они могут организовывать разъяснительные беседы с родителями** о вреде таких конфликтов, распространять брошюры и материалы. Бот может посоветовать пользователю тактично предоставить школьному психологу или учителю информацию (например, ту же брошюру Ковалёвой) – чтобы специалисты поняли, через что проходит ребенок.

- **Поддержка контакта.** Если прямое общение между родителями разорвано, **школа или соцслужба могут выступать посредником** для передачи информации и сообщений о ребенке. Например, классный руководитель может по просьбе отдельно проживающего родителя передать ребенку письмо или оповестить о важном событии. Также социальные работники способны **организовать безопасные нейтральные площадки для встреч** ребенка с отчужденным родителем (под supervision, если нужно). Бот подскажет, как пользователь может вежливо попросить об этом – например, написать официальное письмо в опеку с просьбой о содействии свиданиям в присутствии психолога.
- **Предложение медиации.** Соцработники имеют право **направить родителей к семейному медиатору** или сами выполнять посредническую роль, если обладают подготовкой. Бот объяснит, что медиация – часто более продуктивный путь, чем суд: на ней родителей учат договариваться ради интересов ребенка. Пользователю будет предложено не бояться этого процесса и даже самому инициировать его (при помощи юриста или обратившись в медиационную службу).
- **Принципы нейтральности.** Очень важно, чтобы педагоги и сотрудники служб **строго сохраняли беспристрастность**. В брошюре отмечено: их роль – не занимать сторону кого-либо из родителей, а объективно оценивать ситуацию и ставить в центр благополучие ребенка. Они не должны поддаваться возможным манипуляциям (каждый родитель может пытаться “склонить” школу на свою сторону). Вместо этого рекомендуется **переводить фокус родителей на потребности ребенка и стимулировать конструктивный диалог** между ними. Чат-бот может снабдить пользователя шаблоном “нейтрального письма” в школу – без обвинений в адрес другого родителя, а с упором на то, что «ребенку тяжело, давайте вместе поддержим его». Такой тон повышает шанс, что школа отреагирует правильно.

**Параллельное и ко-родительство.** В зависимости от уровня конфликта, после развода возможны разные модели воспитания детей двумя родителями. **Ко-родительство (co-parenting)** предполагает сотрудничество: родители активно общаются, совместно принимают решения о детях, гибко подстраиваются под обстоятельства ради блага ребенка. Это идеальный вариант, но достижим он лишь примерно в 25–30% случаев после развода. **Параллельное родительство** – более распространенная стратегия (свыше половины разведенных практикуют именно ее). При параллельном воспитании каждый родитель выполняет свои обязанности **независимо, с минимальным взаимодействием** друг с другом, но при этом не мешает участию второго родителя в жизни ребенка. То есть мама и пapa как бы ведут родительство “параллельными курсами”: у каждого свои правила в своем доме, своя часть обязанностей (например, пapa водит на спорт, мама – на музыку), и прямых контактов между ними сведены к минимуму.

Исследования показывают, что **параллельное родительство может быть вполне благоприятно для ребенка**, если родители обеспечивают ему адекватный уход каждый на своей территории. Конечно, полное сотрудничество лучше, но в ситуациях высококонфликтного развода параллельная схема часто является **единственно возможной**, чтобы ребенок вообще мог общаться с обоими. Бот объяснит пользователю, что “лучшее – враг хорошего”: нет смысла пытаться навязать бывшему супругу теплое партнерское взаимодействие, если тот настроен враждебно. Вместо этого стоит **формализовать параллельное родительство через четкий письменный план**. В таком плане (его можно разработать с участием медиатора или юриста) прописываются раздельно обязанности родителей, **границы коммуникации** и обмена информацией, график встреч, правила принятия важных решений <sup>3</sup> <sup>4</sup>. Например, решено, что родители **общаются только письменно через специальное приложение или е-мейл**, исключительно по темам здоровья, учебы и расписания ребенка – без личных тем и взаимных

претензий. **Встречи для передачи ребенка** осуществляются на нейтральной территории (у школы, у бабушки) по расписанию, чтобы родители не контактировали лишний раз. Такой регламент снижает количество поводов для ссор, что в итоге выгодно ребенку, ведь **он меньше свидетель неприятных сцен**. Чат-бот может предоставить шаблон параллельного родительского соглашения и чек-лист правил (например, *"никаких оскорблений при переписке, обсуждать только дела ребенка"*).

При этом бот подчеркнет: **крупные решения (медицина, образование) все равно должны приниматься совместно** согласно закону <sup>5</sup>. Если же общий язык не находится, то такие вопросы решаются через суд. Но в повседневных делах параллельный метод позволяет каждому родителю действовать автономно. В целом, **рекомендация**: *"Если коммуникация с ех-супругом вызывает только новые конфликты, лучше минимизировать ее и перейти в режим параллельного воспитания"*. Это не навсегда – бывает, что со временем, когда страсти улягутся, родители переходят к более кооперативному стилю. Но на период острого пост-развода параллельный подход часто спасителен.

**Структура и этапы медиации.** Медиация – ключевой инструмент, который бот будет продвигать, когда прямой диалог между родителями зашел в тупик. В отличии от суда, где идет *"битва и выяснение, кто прав"*, семейная медиация нацелена на поиск приемлемого решения в интересах ребенка при помощи нейтрального посредника. Бот расскажет, **как обычно проходит медиация**: сначала медиатор объясняет правила (конфиденциальность, добровольность, уважение), затем стороны по очереди излагают свою точку зрения, список проблем для обсуждения. Медиатор помогает выявить приоритеты каждой стороны, направляет диалог к поиску компромиссов. **Шаттл-медиация (челночные переговоры)** – особый формат, применяемый для высококонфликтных случаев. Если родители не могут находиться в одном помещении без скандала, медиатор разводит их по разным комнатам и **курирует общение отдельно с каждым, передавая предложения второй стороне**. Это снижает давление: люди не видят гнев друг друга, говорят более откровенно посреднику. Чат-бот пояснит, что **медиация возможна даже при очень натянутых отношениях**, просто требует большего мастерства от медиатора (структурить сессии, давать выплеснуть эмоции вначале, затем разводить по комнатам и т.д.).

Эффективная медиация обычно разбивается на **несколько сессий**: 1) установление правил и определение списка спорных вопросов (например, график встреч, праздники, звонки, выбор школы, алименты и пр.); 2) обсуждение каждого вопроса, выдвижение вариантов; 3) фиксация достигнутых договоренностей в письменном соглашении. Если эмоции зашкаливают, медиатор делает **перерывы**, может назначить индивидуальные встречи (как терапевтическую паузу). Бот подготовит пользователя: даст советы, как вести себя на медиации – говорить от себя, про свои чувства (*"я переживаю, что..."*), а не обвинять; слушать второго родителя, постараться понять его мотивы; быть готовым к уступкам. Также бот напомнит про **альтернативы медиации** – *челночные переговоры через адвокатов*. Если прямая встреча невозможна, родители могут обмениваться предложениями через представителей, или через письменные коммуникации. Этот процесс медленнее, но иногда более безопасен в очень тяжелых случаях. Главное – идти к решению, а не застрять в конфронтации.

## **Поведенческие техники и микро-интервенции (Workstream 2)**

Чат-бот реализует подход персонализированной психологической помощи, **подбирайая терапевтические микро-интервенции под текущее состояние пользователя**. Ниже описано,

**как эмоции и когнитивные искажения отчуждаемого родителя маппируются на конкретные техники** и упражнения, а также какие **маркеры прогресса** используются, чтобы понять, что пользователь “переключается” на более здоровые рельсы.

**Картирование состояний -> техник.** На основании матрицы состояний (см. выше) определены следующие ключевые стратегии вмешательств:

- **Шок/отрицание → техникам мягкого гругрондинга (заземления в реальности).** В состоянии шока задача бота – аккуратно вывести пользователя из ступора к осознанию проблемы. Применяются методы **эмпатического отражения и нормализации**: бот сочувствует и заверяет, что такая реакция естественна. Затем задает мягкие уточняющие вопросы о фактах («Когда вы последний раз видели ребенка?») – это стимулирует когнитивную активность, возвращает к реальности. После появления минимальной готовности воспринимать информацию бот дает **психообразование**: объясняет простыми словами, что происходит (например, вводит понятие “родительское отчуждение”, “конфликт лояльности” и т.д.). В шоковой фазе важно не перегружать человека – цели минимальные: **валидировать чувства, дать базовое понимание ситуации и внушить надежду, что решение есть**. Маленький шаг – предложить пользователю одно посильное действие, возвращающее ощущение контроля (“Давайте сегодня вы сделаете совсем небольшое дело: просто напишите ребенку короткое сообщение ‘Я тебя люблю’”). Такой микро-шаг снижает паралич без давления. Критерий выхода из шока – пользователь начинает признавать проблему и проявлять интерес к дальнейшему плану (например, соглашается выслушать больше информации, обсуждает что уже сделано). Маркером служит фраза типа: “Да, я понимаю, надо что-то делать...”.
- **Ярость/агрессия → эмоциональная разрядка + переход к ННО (ненасильственному общению).** При гневе крайне важно не обесценивать эмоции пользователя. Бот сначала **дает “безопасное пространство” для выплеска**: поощряет рассказать, что тот чувствует, уверяя в отсутствии осуждения. Пользователь может выплыть поток браны – это допустимо и даже желательно на первом шаге (бот выступает в роли “контейнера” для эмоций). Затем бот делает **валидацию и рефрейминг**: например, **отражает услышанные чувства** (“Вы чувствуете несправедливость, беспомощность, страх потерять ребенка – правильно я понял?”). Это техника из **мотивационного интервьюирования (MI)**: показать, что эмоции замечены и поняты, что уже снижает накал. Далее бот с помощью **сократических вопросов** мягко подводит к осознанию последствий действовать из слепой ярости: “Когда вы злитесь и кричите, бывший партнер идет навстречу или еще больше закрываются? А что чувствует ребенок, видя вашу ярость?”. Цель – чтобы родитель сам увидел неэффективность агрессивной тактики. Затем используется элемент **IFS (семейной системы внутренних частей)**: бот предлагает посмотреть, какая боль скрывается под гневом, задавая вопрос “Что защищает твоя ярость, чего ты боишься в глубине?”. Согласно IFS, **гнев часто выступает “защитником” уязвимой части, несущей страх или обиду**. Такой внутренний диалог помогает переключить клиента из роли “разъяренного” в роль **исследователя своих чувств**, что снижает интенсивность злости. Далее бот обучает базовым навыкам **ННО**: разъясняет 4-компонентную структуру сообщения (факт → чувство → потребность → просьба) и показывает, как преобразовать обвинительную фразу в “я-высказывание”. Пользователю предлагается тут же попрактиковаться – взять конкретную ситуацию и вместе с ботом сформулировать ответ в стиле ННО. Завершающий шаг – поставить перед родителем **стратегический выбор**: “Что для вас важнее – наказать бывшего или восстановить связь с ребенком?”. Это отрезвляет и возвращает фокус на долгосрочную цель. **Маркеры прогресса** в работе с гневом: пользователь начинает признавать под своей яростью другие чувства (горе, страх) и

может хотя бы в терапевтической беседе сформулировать фразу без оскорблений. Если, например, он вместо "Она тварь, украла моего сына" говорит "Я чувствую отчаяние, потому что теряю связь с сыном" – это огромный шаг (адгезия к навыку ННО). Также признаком успеха будет снижение воинственной риторики и согласие рассуждать о стратегии, а не о мести.

- **Отчаяние/суицидальный кризис → протокол "от кризиса к надежде".** Если бот распознает у пользователя высказывания типа "Не хочу больше жить", "Все потеряно", он немедленно включает **кризисный сценарий**. Реализуется **четкая стратификация риска**: бот прямыми вопросами выясняет, есть ли у человека мысли о самоубийстве, план или намерение совершить его. Эти вопросы основаны на клинических стандартах оценки суицидального риска (например, шкала Columbia). Если ответы указывают на **высокий риск** (конкретный план, чувство безнадежности, намерение прямо сейчас), бот **сразу же останавливает обычный диалог** и выполняет протокол эскалации: дает пользователю **строгие инструкции по обеспечению своей безопасности**. Например: "Мне очень важно, чтобы вы были в безопасности. Прямо сейчас, пожалуйста, сделайте три вещи: (1) Позвоните на кризисную линию по телефону ...; (2) Свяжитесь с близким человеком; (3) Напишите мне, когда это сделаете. Я остаюсь с вами.". При этом бот **прекращает генерацию какого-либо обычного текста** (guardrail), пока не получит подтверждение, что пользователь выполнил шаги или находится под присмотром. Фактически чат-бот переходит в режим "**виртуального спасателя**". Если риск **средний или низкий** (нет явного плана, но настроение тяжелое), бот продолжает диалог, но очень осторожно. Шаги: (1) выражение эмпатии и простое присутствие ("Я вижу вашу боль, я с вами"); (2) фокусировка на **ближайшем будущем** – бот предлагает двигаться "час за часом, не думать о завтра", напоминает о базовых потребностях (поесть, поспать хотя бы немного, подышать свежим воздухом); (3) помогает сформировать **контрарратив беспомощности** – несколько утверждений, возвращающих смысл: "Вы все еще родитель – никто не отнимет этого; дети растут и часто сами возвращаются к отчужденному родителю; ваше тихое присутствие – уже ценность". Далее идут приемы **надежды и смысла**: бот может рассказать реальную историю другого родителя, который сумел восстановить отношения спустя годы (демонстрируя, что шанс есть). Затем – задать вопросы, помогающие найти **опоры жизни вне ребенка**: кто еще нуждается в вас (другие дети? пожилые родители? пациент на работе?), какие еще ценности и интересы у вас есть. Это элементы логотерапии (поиск смысла несмотря на утрату). Бот приглашает пользователя **присоединиться к сообществу** – например, советует группу поддержки отчужденных родителей: "Завтра будет онлайн-встреча таких же родителей, приходите хотя бы послушать". Общение с равными снижает чувство изоляции. Наконец, обязательно заключается **"контракт безопасности"**: бот просит обещать, что если суицидальные мысли усилиятся, пользователь **обратится за помощью – на линию, к боту или близкому**. Этот шаг подкрепляет ответственность человека за свою жизнь. **Условие выхода** из кризисного режима: пользователь четко заявляет, что "сейчас не собирается себе вредить", либо в диалоге появляется динамика от полной безнадежности к проблескам конструктивных идей (например: "Хорошо, давайте попробуем дожить до завтра, схожу пока к врачу..."). Тогда бот может постепенно вернуть беседу в русло решения проблем. При любом сомнении в безопасности бот продолжит ежедневный мониторинг (частые check-in сообщения). **Маркеры успеха:** снижение суицидальной лексики, согласие пользователя выполнить контракт безопасности, проявление интереса к каким-либо действиям (даже минимальным).
- **Вина/самообвинение → когнитивная реструктуризация (КПТ)**. При выраженном чувстве вины и собственной никчемности бот применяет техники **когнитивно-**

**поведенческой терапии (СВТ)**, направленные на исправление искаженных убеждений. Шаг 1: **выявление автоматических мыслей** – бот просит пользователя сформулировать, за что конкретно он себя винит ("Закончите фразу: Я виноват в том, что...; Мне кажется, я плохой родитель, потому что..."). Это выносит внутренний монолог наружу. Шаг 2: **разбор доказательств** – бот задает серию вопросов к каждой негативной мысли: "Это объективный факт или ваша интерпретация? Какие есть реальные подтверждения этому? А какие контр-примеры? Что бы вы сказали другу, если бы он о себе так подумал?". Метод Сократа помогает обнаружить логические несоответствия. Шаг 3: бот обучает распознавать типичные **когнитивные искажения**. Например, указывает: "В ваших мыслях вижу несколько ловушек мышления: (1) Персонализация – вы берете на себя ответственность за то, что на самом деле зависело от другого человека; (2) Катастрофизация – мысли по типу 'всё потеряно навсегда'; (3) Черно-белое мышление – либо идеальный родитель, либо ужасный, третьего не дано". Пользователь начинает понимать, что его вина, возможно, необоснованна или преувеличена. Шаг 4: **контекстуализация случившегося** – бот спрашивает про обстоятельства тех ошибок, за которые корит себя родитель: "Давайте вспомним, что тогда происходило? Какие ресурсы у вас были? Вы ведь делали лучшее из того, что могли в той ситуации?". Это приводит к более объективному взгляду (например, "Да, я тогда тяжело болел/у меня не было денег на хорошего адвоката и пр." – то есть родители часто винят себя за вещи вне их контроля). Шаг 5: **распределение ответственности** – мощная техника: бот предлагает условно нарисовать "круг ответственности" и прикинуть, какой процент вины на ком. Например: "Вы отвечаете за ~30%, бывший партнер за 40%, система (суд, полиция) за 20%, и 10% – просто обстоятельства". Такое упражнение наглядно снижает тотальную самобичевание (почти всегда пользователь видит, что не 100% его вина). Шаг 6: **генерация альтернативных мыслей** – бот помогает сформулировать более сбалансированное отношение к себе. Например, вместо "Я всё испортил, я ужасный отец" – "Да, я совершил ошибки как и все родители, но я делал лучшее, что мог тогда, и теперь хочу исправить ситуацию". Так происходит когнитивное переосмысление, включающее самопрощение. Шаг 7: **развитие самосострадания** – бот приглашает пользователя посмотреть на себя со стороны лучшего друга: "Что бы ты сказал близкому человеку, если бы он оказался в такой ситуации? А теперь попробуй обратиться так же к самому себе". Обычно это упражнение вызывает сильный эмоциональный отклик (слезы облегчения и принятия). Шаг 8: **практика прощения себя** – например, бот дает задачу: написать письмо самому себе от имени мудрого, сострадающего "Я", начиная с фразы "Дорогой <Имя>, я вижу, как ты страдаешь...". Это закрепляет новую установку более бережного отношения к себе. Бот также предлагает **домашние задания**: вести "Дневник мыслей" (событие → мысль → чувство → новая рациональная мысль), каждый день отмечать, что получилось хорошо, практиковать фразы самоподдержки. **Маркеры прогресса:** в диалоге пользователь перестает говорить только "я ужасен, я виноват", начинает использовать более здоровые формулировки ("я не идеален, но я стараюсь"; "были обстоятельства, это не все зависело от меня"). Появляется чувство облегчения, готовность перестать наказывать себя и перейти к активным действиям. Если бот видит такие сдвиги (например, пользователь сделал домашнее задание, записал альтернативную мысль), это сигнал, что можно переходить к планированию шагов.

- **Навязчивая борьба/“туннельное видение” → переориентация на стратегическое планирование.** Для пользователей, застрявших в бесконечных судах и конфликтах, бот применяет комбинированную технику: элементы **мотивационного интервьюирования (МИ)**, чтобы выяснить истинные цели и ценности человека, и **коучинговый подход к постановке целей**. Сначала важно "разморозить" мышление родителя, которое циклично повторяет: "Ещё один суд – и я докажу свою правоту". Бот может задать вопрос в духе МИ:

*“Что даст вам то, что вы докажете свою правоту? Что вы в конечном счете хотите для себя и ребенка?”*. Обычно ответ – *“Хочу быть в жизни ребенка”*. Тогда бот мягко показывает разрыв: текущие действия (бесконечные тяжбы) не ведут к этой цели, а наоборот отдаляют (ребенок устает от конфликтов). Когда пользователь это осознает, можно переключаться на **конструктивное планирование будущих шагов**. Бот вводит понятие *“стратегического состояния”* – когда вместо реактивной борьбы родитель строит долгосрочный план восстановления отношений. Применяется модель **SMART-целей**: бот учит формулировать цели конкретно, реалистично и с указанием сроков. Например, не *“помириться с ребенком”*, а **конкретная цель**: *“через 6 месяцев организовать первую совместную консультацию с семейным психологом вместе с ребенком”*. Далее бот помогает сделать **инвентаризацию ресурсов**: финансовых (достаточно ли средств на юриста, терапию?), социальных (кто поддержит – друзья, родня?), профессиональных (есть ли связи с хорошим адвокатом, психологом?) и личных качеств. Затем – **анализ препятствий**: бот спрашивает, что может помешать достижению цели и как это превентивно решить (например: *“Препятствие: опекающий родитель игнорирует письма → Решение: подключить посредника, медиатора”*). Важный блок – **приоритизация** по Матрице Эйзенхауэра: бот учит отделять срочное от важного (например, *срочное и важное*: подать заявление в школу, чтобы получать дневник успеваемости; *важное, но не срочное*: написать длинное письмо ех-супругу – можно позже; *неважное*: читать соцсети бывшего – вообще исключить). Это освобождает энергию на действительно значимые шаги. Далее – **разбивка целей на микро-шаги**: бот с пользователем составляет пошаговый план, например:

- Шаг 1: найти и записаться на консультацию к семейному медиатору (до конца месяца);
- Шаг 2: написать нейтральное письмо бывшему партнеру с предложением плана праздников (в течение 2 недель, используя шаблон BIFF/NVC);
- Шаг 3: подготовить подарок ребенку на день рождения и отправить (дата ...), и т.д.

Каждый шаг маленький и реалистичный, но приближающий к итоговой цели. Бот внедряет правило: **перед любым действием спрашивать себя – “приближает ли это меня к моей цели (восстановлению отношений) или нет?”**. Если нет – значит, не тратить на это силы. Так пользователь учится не ввязываться в провокации, а действовать осознанно. Бот берет на себя роль **аккаунтабилити-партнера**: договаривается с пользователем о регулярных чек-инах, чтобы отслеживать прогресс (скажем, раз в неделю бот спрашивает, сделан ли шаг, какие результаты). Это повышает ответственность и мотивацию. **Маркеры успеха**: пользователь перестает говорить только о прошлом (*“как меня обманули”*), у него появляется видение будущего, он делает конкретные шаги (записался к медиатору, сходил к психологу, собрал справки и т.п.). Бот фиксирует эти положительные сдвиги и усиливает их положительной поддержкой (*“Отлично, вы уже сделали 3 шага из плана!”*). Когда пользователь стablyно перешел в режим конструктивных действий, можно считать, что он вышел из зоны кризиса в зону роста.

**Маркеры адгезии к техникам.** Важная задача – отслеживать, насколько пользователь **усваивает предложенные техники ННО/МИ/КПТ/IFS** и применяется ли они на практике. Чат-бот будет оценивать по косвенным признакам и прямым вопросам:

- Для **ННО (ненасильственной коммуникации)**: маркером служит то, как пользователь начинает формулировать свои сообщения о проблеме. Если после тренировки бот видит, что клиент говорит: *“Мне больно, потому что мне важно быть с ребенком”* вместо *“Она стерва и все разрушила”* – это явный показатель усвоения принципа говорить через чувство и потребность, без обвинения. Бот может специально попросить: *“Попробуйте*

сейчас сформулировать вашу претензию к бывшему мужу в формате 'Когда X, я чувствую Y, потому что мне важно Z, прошу ...' и проверить корректность. Если пользователь справляется – технику можно считать освоенной (по крайней мере интеллектуально). Другой показатель – реальный кейс: например, пользователь сообщает, что написал e-mail бывшей жене по шаблону (бот может даже предложить показать текст письма, убрав личные данные, и совместно отредактировать его под BIFF/NVC формат).

- Для **МИ (мотивационного интервьюирования)**: тут скорее оценивается степень внутренней мотивации и изменения "говорения". Маркер – появление у пользователя "**change talk**" (речи изменений): фраз, указывающих на готовность действовать, признание необходимости перемен ("Наверное, мне стоит попробовать другой подход...", "Я хочу наладить, а не воевать"). Одновременно должно уменьшаться "**sustain talk**" – прежние оправдания продолжать статус-кво ("Бесполезно что-то менять, лучше еще засужу ее"). Бот будет в диалоге подкреплять любые высказывания в пользу изменений ("Это важно – вы сказали, что готовы попробовать..."). Отслеживать такие тонкие изменения можно с помощью LLM-классификатора, обученного различать *language of change*. Исследования показывают, что выравнивание диалога чат-бота с техниками МИ коррелирует с лучшим исходом. Поэтому бот сам старается использовать больше отражений и вопросов по МИ (минимум директивности) и замеряет свою "адгезию к МИ" автоматически (например, с помощью метрик наподобие MITI – доля отражений в ответах и пр.). Если бот выбрал стратегию **MI\_reflection**, то **рефлексивность ответа** может проверяться на соответствие эталону (в рамках надзора, см. архитектуру ниже).
- Для **КПТ (когнитивных техник)**: маркеры – это **рациональность и сбалансированность высказываний** пользователя со временем. Допустим, изначально были твёрдые убеждения "Я никчёмный отец, мне не место в ее жизни". После серии реструктуризаций пользователь может сказать: "Я не идеален, но и не отвратительный; у меня есть шанс стать лучше". Появление таких мыслей – количественно это отслеживается как снижение **количества когнитивных искажений** в речи. Мы можем встроить модуль авто-детекции искажений: существуют NLP-модели, которые классифицируют фрагменты текста по типам когнитивных ошибок. Например, работа Tauscher 2023 (в PsychiatryOnline) предлагает алгоритмы для поиска в тексте признаков чрезмерной генерализации, катастрофизации и т.д. Бот может анализировать высказывания пользователя на наличие слов всегда/никогда (черно-белое мышление) или "я виноват" (персонализация) и фиксировать динамику. Если сессия за сессией таких выражений меньше – прогресс налицо. Также признаком будет выполнение СВТ-домашних заданий: бот может прямо спросить, заполнял ли пользователь дневник мыслей, что нового понял. Выполнение ДЗ – хороший маркер приверженности терапии.
- Для **IFS (работы с внутренними частями)**: здесь более субъективно, но критерий – способен ли пользователь *отстраниться от своих эмоций и взглянуть на них со стороны*. Если поначалу человек **полностью сливаются с гневом** ("Я есть моя ярость"), то после IFS-упражнений он, например, говорит: "Во мне есть часть, которая очень злится, но есть и часть, которая хочет мира". Появление такого лексикона ("во мне есть часть...") – явный индикатор усвоения IFS-подхода. Бот может уточнять: "Как вы сейчас относитесь к той вашей части, что кричит на бывшего? Получилось с ней подружиться?". Если пользователь уже оперирует понятиями "защитник", "уязвимая часть" и проявляет к себе больше сострадания – цель достигнута. Кроме того, IFS предусматривает "**unblending**" – разотождествление: пользователь понимает, что "я – не равен моим эмоциям, я наблюдющий Я". В речи это можно уловить.

В целом, для контроля качества бот будет фиксировать **ключевые изменения в речи и поведении пользователя** как маркеры эффективности поведенческих техник. Эти же параметры войдут в систему оценки (см. раздел Evaluation).

## Юридические и социально-этические аспекты (Workstream 3)

Данный блок касается функционала чат-бота, связанного с юридическими рамками, социальными навыками коммуникации и соблюдением безопасности. Бот поддержки не является юристом или врачом, поэтому важно установить границы того, что он может и не может советовать, а также предоставить пользователю полезные инструменты: шаблоны писем, дневники, подготовки к медиации и пр.

**Границы компетенции и дисклеймеры.** С самого начала взаимодействия бот информирует пользователя, что он не предоставляет юридических или медицинских консультаций. Это будет явно указано в пользовательском соглашении и повторено в диалоге при соответствующих запросах (например, "я не юрист, но могу поделиться общей информацией"). Такой **дисклеймер** важен и с правовой, и с этической точки зрения. Рекомендации бота по правовым вопросам ограничиваются общими советами (например, "обратитесь к адвокату", "узнайте процедуру в органах опеки"), ссылки на законы, но без индивидуальной правовой оценки. Аналогично по медицине (тут менее актуально, но вдруг вопрос про психологические лекарства – бот скажет обратиться к психиатру, не будет советовать дозировки). **Этические нормы** (по ВОЗ и другим организациям) требуют, чтобы чат-бот в ментальном здоровье был прозрачным, не выдавал себя за живого врача, не ставил диагнозов и т.п.. Наш бот будет этим нормам следовать. Кроме того, при сборе любой личной информации бот должен предупредить, как она будет использоваться и храниться – это часть модуля информированного согласия (подробнее в разделе Архитектура/Память). Таким образом, "границы дозволенного" чётко очерчены: поддержка, эмпатия, общие техники – да; специализированные советы в области закона или медицины – нет.

**Ведение дневника и документация.** Как отмечалось, бот поощряет пользователя вести журнал событий и эмоций. Это одновременно психологический инструмент и юридический. С психологической стороны, дневник помогает выразить чувства, отследить динамику (например, пользователь может по шкале от 1 до 10 ежедневно отмечать уровень эмоциональной боли – бот встроит такую функцию). С юридической стороны, подробный **журнал инцидентов** – ценное доказательство в суде при разборе случаев отчуждения. Юристы рекомендуют отчужденным родителям фиксировать каждый случай нарушения договоренностей: дату, что произошло (не дали созвониться, сорвали встречу и т.д.), реакцию ребенка. Эти записи, заверенные временем, могут убедить судью в системности проблем. Бот, зная это, введет специальный режим: "**Дневник контактов**" – по запросу или по расписанию будет задавать вопросы: "Удалось ли увидеться с ребенком на этой неделе? Если нет, по какой причине? Опишите подробно событие." Пользователь ответит в свободной форме, а бот сохранит запись в хранилище. При накоплении данных бот может помочь их структурировать – например, сформировать таблицу всех пропущенных встреч, кто отменил и почему. Если понадобится, пользователь сможет выгрузить этот дневник (с обезличиванием лишнего) для юриста. **Конфиденциальность при этом приоритетна:** записи хранятся локально (или шифруются), пользователь контролирует их экспорт.

Кроме того, дневник включает рефлексивные записи – бот может предлагать писать письма "в стол" (например, письмо гнева бывшему супругу – но не отправлять, а сохранить и обсудить на сессии). Такой **техникой самотерапии** часто пользуются психологи. Бот будет выступать

фасилитатором: "Напишите здесь, что бы вы хотели ему сказать – я сохраню, вы можете в любой момент перечитать." Потом, спустя время, бот может вместе проанализировать, как изменились эмоции.

**Нейтральные письма школе/соцслужбам.** Одно из сложных испытаний для отчуждаемого родителя – общение с учителями, врачами, соцработниками, которые зачастую слышат от опекающего родителя негативную информацию о нем. Очень важно удерживать **нейтрально-доброжелательный тон** и не вступать в перепалку. Бот предоставит шаблоны таких писем. Например, письмо классному руководителю: - без обвинений второго родителя (*не писать "моя бывшая настроила ребенка против меня"*), - с акцентом на благополучии ребенка ("Мне важно, чтобы Петя не страдал из-за нашей ситуации"), - с просьбой о сотрудничестве ("Буду признателен, если вы сообщите, как Петя чувствует себя в школе, и готовы поддержать его вместе").

Соцслужбам – аналогично: вежливый тон, ссылка на право ребенка общаться с обоими родителями (со статьей закона, если уместно), предложение решений (например, "готов посещать семейного психолога по рекомендации опеки"). **BIFF-методика** Билла Эдди здесь очень пригодится: любые письменные обращения должны быть **Brief (краткими)**, **Informative (по делу, с фактами)**, **Friendly (вежливыми)** и **Firm (с четким запросом или позицией)**. Бот проведет пользователя через проверку BIFF: уберет эмоциональные фразы, оставит факты, придаст дружелюбный, но уверенный тон. Это важно не только для писем внешним организациям, но и для коммуникации напрямую с конфликтным экс-супругом (например, по электронной почте). **Пример BIFF:** если другой родитель шлет оскорбительную тираду ("ты ужасный родитель, ребенок заслуживает лучшего"), BIFF-ответ может быть: "Понимаю, ты хочешь поменять выходные. К сожалению, в эти даты у меня уже есть планы, поэтому остаемся при текущем расписании. Готов рассмотреть другие варианты, которые устроят нас обоих. Дай знать, если будут предложения." – коротко, по сути, без встречных обвинений, с открытой дверью к диалогу. Бот потренирует пользователя писать в таком стиле, чтобы все коммуникации можно было потом хоть в суде показать – и это не навредит ему (в BIFF-ответах нет ничего компрометирующего, наоборот, они показывают суду вашу благородность).

**Практики участия в медиации и альтернативных переговорах.** Бот подготовит пользователя к разным сценариям **альтернативного разрешения споров** (ADR). - Во-первых, **к семейной медиации**: даст чек-лист, что взять с собой (список вопросов, документы о доходах, заранее подумать о графике и пожеланиях по праздникам и пр.), как себя вести (не перебивать, говорить от первого лица, избегать прошлых обид, фокусироваться на решениях для будущего). Хорошей практикой является **ролевая репетиция**: бот может сыграть роль медиатора и/или второго родителя, смоделировать некоторые возможные высказывания. Пользователь в безопасной среде потренируется реагировать. Например, бот скажет: "Ваш бывший муж заявляет: 'Она всегда была плохой матерью, дочь сама не хочет с ней видеться'. Как вы ответите?" Затем бот скорректирует ответ до более конструктивного (без агрессии, но отстаивая свои права и выражая готовность сотрудничать ради ребенка). После такой тренировки пользователь придет на реальную медиацию более уверенно, с сценарными заготовками. - **Шаттл-переговоры** через юристов. Если дело все же идет через адвокатов, бот расскажет, как эффективно взаимодействовать со своим адвокатом: собирать факты для него (дневник, скрины переписок), формулировать конкретные позиции (*не просто "хочу справедливости", а "хочу каждые выходные видеть ребенка"* – четко), сохранять деловой стиль. Бот может помочь составить список вопросов к адвокату (например: "Как лучше подать ходатайство об экспертизе?", "Что делать, если решение суда не исполняется?"). Также предупредит о недопустимых вещах – например, не писать гневных сообщений бывшему про суд (все может попасть в материалы дела), не обсуждать при ребенке суд и т.п. - Прочие ADR. Возможно, будет уместно упомянуть привлечение омбудсмена

по правам ребенка (если страна это предусматривает) или семейного совета\* (когда авторитетные родственники участвуют в урегулировании). Бот содержит информацию о таких ресурсах и подскажет, как к ним обратиться.

Важным социально-этическим моментом является **интерес ребенка превыше всего**. Бот постоянно возвращает к этому принципу: любые переговоры и соглашения должны ставить на первое место благополучие детей, а не это родителей. Напоминание об этом – часть и юридической, и психологической работы. Например, при разработке плана соглашения бот будет спрашивать: “*Этот пункт — он действительно в интересах ребенка или это ваша обида говорит?*”. Такой встроенный этический фильтр поможет пользователю удерживаться в верном направлении.

Подытоживая: **юридико-социальный модуль** чат-бота обеспечивает пользователя знаниями и инструментами, но **не выходит за рамки компетенции**. Он формирует у родителя *навыки самопомощи* – грамотная документация, эффективное общение (BIFF/NVC), умение участвовать в медиации, понимание своих прав – вместо того, чтобы напрямую давать юридические советы. Это соответствует лучшим практикам и требованиям ответственности (бот не заменяет адвоката, но делает взаимодействие с ним более осмысленным, и не нарушает закон, давая недопустимые советы).

## **Модуль безопасности и кризисного реагирования (Workstream 4)**

**Стратификация рисков и сценарии эскалации.** Безопасность – краеугольный камень нашего чат-бота, учитывая высокие эмоциональные риски в теме отчуждения. Модуль безопасности работает на нескольких уровнях:

- **Риск для жизни пользователя (суициdalный риск).** Как описано, бот оценивает степень суицидальности прямыми вопросами. Мы выделяем условно три градации:
  - **Низкий риск:** нет мыслей о самоубийстве, просто тоска/разочарование. Бот оказывает эмоциональную поддержку, но в целом продолжает основную работу (психообразование, техники и т.д.), делая пометку регулярно перепроверять настроение.
  - **Средний риск:** мысли о суициде есть, но без конкретного плана/намерения. Здесь бот действует как **“safety-coach”** – включается специальная политика диалога. Исследования показывают, что именно на среднем уровне ИИ-боты часто дают сбои (не последовательны), поэтому нужен четкий протокол. Наш бот на **среднем риске**:
    1. Подробно выясняет, что именно чувствует пользователь, что его удерживает жить (наличие детей, обязательства).
    2. **Дает ресурсы:** контакт горячей линии, доступные 24/7 услуги помощи.
    3. **Заключает контракт безопасности** на 24 часа (обещание не делать попыток хотя бы до следующей сессии, а при ухудшении – обратиться за помощью).
    4. **Увеличивает частоту мониторинга:** бот может отправлять короткие сообщения каждые несколько часов (“Вы как, держитесь? Я рядом.”), чтобы пользователь чувствовал присутствие.
    5. При любых тревожных сигналах (например, пользователь перестал отвечать) – бот эскалирует к критическому сценарию.

- **Высокий риск:** пользователь сообщил о намерении прямо сейчас что-то сделать или назвал конкретный план/средство. Тут бот **немедленно**:
  - Выдает сообщение с настоятельной просьбой остановиться и обратиться в экстренную службу (112, 911 или аналог, в зависимости от страны).
  - **Останавливает генерацию** обычных ответов (guardrail `halt_generation`) и переходит в “режим повторения” этих инструкций или фраз поддержки, пока не убедится, что пользователь в безопасности.
  - **Опционально:** если у нас есть данные экстренных контактов (и пользователь дал согласие в соглашении), бот может попытаться уведомить указанных лиц или службы. Но это тонкий момент: по этике без явного согласия пользователя делать вызов нельзя, тут скорее довериться самому пользователю. Возможно, на этапе онбординга мы попросим контакт доверенного лица именно для таких случаев.
  - **После прохода кризиса** – бот останется какое-то время в режиме частых чек-инов и не будет возвращаться к обычным темам, пока состояние не стабилизируется.
- **Риск причинения вреда другим (насилие).** Есть вероятность, что отчуждаемый родитель может высказывать угрозы в адрес бывшего партнера, его семьи, или (хоть и редко) даже ребенка. Такие кейсы тоже подпадают под модуль безопасности. Наша система Guardrails настроена на **распознавание лексики насилия и намерений причинить вред**. Если пользователь пишет что-то вроде “Убью её, если еще раз не даст сына” – бот **немедленно** включает стоп-сигнал. Во-первых, он *предупредит*, что насилие недопустимо и разрушит жизнь всем (в том числе лишит родителя шанса быть с ребенком уже по уголовным причинам). Во-вторых, попытается переключить внимание на эмоцию за этой агрессией (как в стратегии ярости – выяснить, что именно так ранит). Но если пользователь *прямо сейчас собирается совершить насильтвенное действие*, бот обязан **эскалировать**: убеждать не делать этого, напоминать о правовых последствиях, возможно, тоже привлечь экстренные службы. Однако, в отличие от суицида, здесь бот не сможет вызвать полицию (нет такой интеграции) – максимум может настойчиво уговорить пользователя *немедленно обратиться за очной помощью*. Практически, надеемся, до таких ситуаций не дойдет; чаще угрозы – это выражение гнева, с которым бот будет работать через описанные техники (IFS, ННО вместо угроз и т.п.).
- **Риск для ребенка.** Отчуждаемый родитель обычно не желает вреда ребенку, но бывают состояния аффекта. Если вдруг пользователь пишет нечто, что может угрожать ребенку (например, “Заберу ребенка силой и уеду” или “Лучше бы ребенок умер, чем жил с ней” – в состоянии аффекта возможно и такое), бот должен **немедленно сменить линию**. Во-первых, четко обозначить: “Вы говорите так из боли, но такие мысли очень опасны. Ребенок *ни в коем случае не должен страдать*”. Далее бот переключится на кризисную интервенцию, почти как с суицидом, только акцент – защита ребенка. Он будет убеждать пользователя, что никакое “спасение” с применением насилия (например, киднеппинг) неприемлемо: это поломает психику ребенку и приведет к юридической катастрофе для родителя. Вероятно, подобный эпизод – повод *приоритизировать подключение живого специалиста*. Бот может сказать: “Ваши слова вызывают у меня серьезное беспокойство. Я рекомендую срочно поговорить со специалистом очно.” Возможно, понадобится завершить чат на время, чтобы человек остыл.
- **Прочие риски.** К ним относится, например, **чрезмерная зависимость от чат-бота**. Если пользователь часами в день пишет боту, явно застрял, бот должен заметить и мягко перенаправить: “Мне кажется, вам может помочь живая группа или терапевт”. Есть **этические ограничения** на длительность и формат общения (пока не установлены

официально, но мы сами введем): например, не более 1 часа непрерывного чата без предложения перерыва, и т.п.

**Guardrails и политика безопасного реагирования.** Технически модуль безопасности реализован через **правила Colang/Nemo Guardrails** и дополнительную логику для средних случаев. Мы задали набор **паттернов-триггеров** на основе ключевых слов и оценок смысловой модели: - суициdalная лексика (слова типа "покончить", "не хочу жить", особенно в сочетании с конкретикой); - насильственная лексика (угрозы убить, ранить и т.д.); - упоминание **плана, способа или времени** суицида/насилия – наиболее критичный триггер; - признаки тяжелого психоза (например, если пользователь начинает говорить не связно о нереальном – возможно острый стресс, требующий иной тактики); - запросы, нарушающие этику: просьбы о помощи в нелегальных действиях, советы по медикаментам ("можно ли пить антидепрессанты" – бот не врач, не имеет права советовать дозы, это тоже guardrail).

При срабатывании правила, Guardrail либо: - **блокирует** обычное продолжение (halt\_generation) и вставляет заранее подготовленный **скрипт кризисного ответа**. Пример из Colang: **pattern**: "суицид" -> execute CRISIS\_SCRIPT – и бот мгновенно выдает спасательное сообщение с номерами горячих линий. - Либо для "мягких" триггеров – вставляет **уточняющий вопрос**. Например, при неявной фразе "Мне бы просто уснуть и не проснуться" – бот не сразу сигнализирует 911, а сначала спрашивает: "У вас мысли о том, чтобы себя ранить?". Это своего рода *middle-risk handler*: выявить, насколько серьезно.

Также guardrails касаются **недопустимых запросов к самому боту**: юридические советы (большой и сложный вопрос – бот выдаст дисклаймер и переведет разговор), медицинские (то же самое), **темы сексуального характера** (маловероятно тут, но вдруг – бот не предназначен для обсуждения романтических фантазий). **Особое внимание – упоминания несовершеннолетних в сексуальном контексте**: если вдруг (надеемся, нет) пользователь скажет что-то о сексуальном вреде ребенку, бот обязан остановить сессию и, возможно, эскалировать модераторам (предусмотрен ли manual review?). Это стандартные требования безопасности контента.

**Контракт безопасности и предотвращение рецидива.** Когда острый кризис миновал, бот совместно с пользователем вырабатывает **план поддержания безопасности**. В него входит: - Список "триггеров", которые могут снова довести до отчаяния или ярости (бот поможет их осознать: например, "датой следующего суда", "видом счастливых семей на улице" и т.д.). - План действий при нахождении в зоне срыва: кого сразу позвонить, какие техники самопомощи использовать (записи из предыдущих бесед – например, перечитать письмо поддержки от самого себя, применить дыхательную гимнастику, отвлечься физическим упражнением). Бот запишет этот контракт и при признаках ухудшения будет напоминать: "Вы говорили, что если снова почувствуете невыносимую тоску, позвоните другу X. Сделайте это, хорошо?". - Обязательство пользователя не принимать важных решений в острых состояниях. Это тоже вписывается: "Обещаю не отправлять агрессивных сообщений и не принимать юридических шагов, когда я разгневан/подавлен. Сначала – успокоиться 24 часа.".

**Непрерывность поддержки.** Модуль безопасности подразумевает, что бот **никогда резко не бросает пользователя в кризисе**. Даже если основной диалог завершен, бот будет периодически проверять состояние ("safety check"). Это возможно через push-уведомления или email: "Привет, это снова я. Хочу убедиться, что у тебя все спокойно. Помни, ты не один." Разумеется, с возможностью отключить, если пользователь против. Но лучше перебдеть, чем недобдеть – поэтому по умолчанию при средне-высоком недавнем риске бот делает такие проверки ежедневно хотя бы в течение недели.

**Привлечение человеческого оператора.** Хотя проект подразумевает автономность, мы предусматриваем на случай тяжелых кризисов возможность вмешательства живого специалиста. Например, если бот **многократно фиксирует высокий суицидальный риск** или **непонимаемое поведение** (вдруг психотические симптомы), он может по протоколу перевести диалог на модерацию – уведомить команду (дежурного психолога, если такой предусмотрен). В идеале, конечно, пользователь сам должен обратиться, но мы технично закладываем такую функцию. Это соответствует принципу “*do no harm*” – лучше предоставить человеку живую помощь, если ИИ не справляется.

В реализации Colang это может выглядеть так: `pattern: "описание детального плана суицида" => {halt; respond crisis_script; notify_human}`. У нас будет некий **“дежурный клиницист”** (например, в команде проекта или партнер), которого можно уведомить при крайней необходимости. Конечно, соблюдая анонимность – то есть передать им обезличенные данные ситуации. Пока это скорее гипотетическая опция, но мы ее пропишем как часть политики безопасности.

Итого, **safety-модуль** – многоуровневый механизм, который: 1. Распознает опасные паттерны (суицид, насилие, нарушение границ) на основе правил и ML-классификаторов. 2. Имеет заготовленные сценарии реагирования для каждого уровня риска. 3. Интегрирован в архитектуру как **“rails”** – то есть вне зависимости от общего диалога, эти правила имеют приоритет (их выполнимость гарантируется на уровне “policy”). 4. Проверяется и обновляется в соответствии с новыми рекомендациями. (Например, выйдет обновление ВОЗ по AI в психиатрии – мы внесем корректизы). 5. **Метрики безопасности** (см. Evaluation) будут постоянно мониторить, насколько модуль работает: не пропускает ли тревожные сигналы, правильно ли эскалирует. Мы стремимся, чтобы бот соответствовал самым строгим стандартам (включая внешние вроде VERA-MH, оценивающие умение чатботов выявлять кризисы и этичность реакций).

## Архитектура и техническая реализация (Workstream 5)

Архитектура чат-бота строится по принципу **policy-as-code**: логика диалога, стратегии и ограничения защиты не только в весе модели, но и в явные программные структуры (граф, правила, схемы ответа). Это обеспечивает предсказуемость и возможность валидации. Рассмотрим основные компоненты:

**Граф состояний (graph.yaml).** Вся терапевтическая процедура описывается в виде **ориентированного графа**: узлы – это определенные состояния пользователя или этапы вмешательства, ребра – условия перехода между ними. По сути, это формализованная версия нашей матрицы состояний и стратегий. Например, узел “RAGE (ярость)” связан ребрами с узлом “NVC Training” (после успешной разрядки гнева и интереса к общению) или обратно с узлом “RAGE” (если попытка перейти к ННО провалилась и пользователь снова скатился в ругань). Граф может быть реализован на DSL вроде **LangGraph** или аналогичных библиотек, позволяющих явно задавать диалоговый flow. Каждому узлу соответствуют: - **Диагностические маркеры** (паттерны текста, эмоции), которые приводят к активации узла. Например, нода “DESPAIR” срабатывает, если sentiment-анализ показывает очень низкий тон + встречаются слова типа “бессмысленно, не могу больше”, или если из предыдущего узла “RAGE” пользователь вдруг переходит к плачу (смена состояния). - **Цели узла** (они же цели интервенции на этом этапе) – взяты из описанных стратегий. Например, узел “EMOTIONAL\_DISCHARGE (rage)” имеет цели: *дать выплеснуть гнев, затем трансформировать агрессию в конструктивное выражение через ННО*. - **Микро-шаги диалога** внутри узла – они же сценарий стратегии: последовательность сообщений бота и ожидаемых реакций пользователя. Мы уже привели их примеры (Steps 1-7 в стратегиях). В

YAML это можно оформить как список message templates с условиями перехода к следующему. - **Условия выхода** из узла – что считается успехом этого этапа, после чего можно перейти дальше. Например, для узла "SHOCK" условием выхода может быть ответ пользователя, содержащий осознание проблемы (ключевые слова типа "понял, что это надолго... что делать дальше?"). Для узла "CRISIS" – пользователь заявил, что сейчас в безопасности. Эти условия добавляются как части графа (грубо: `if user_response.contains("готов") then transition to ACTION_PLANNING`). - **Связанные данные:** каждому узлу сопоставлены **психообразовательные материалы, упражнения, шаблоны**, которые бот может выдать пользователю, находясь в этом узле. Например, в узле "LOYALTY\_CONFLICT\_EDUCATION" (когда бот объясняет про конфликт лояльности) прикреплена цитата из брошюры Ковалёвой, которую бот может выдать как справку. В узле "GUILTY\_CBT" – ссылка на таблицу когнитивных искажений, которую бот может прислать по запросу. - **Тактика/стратегия:** узел обычно "знает", какая терапевтическая методология сейчас используется (мы храним, например, атрибут `strategy: CBT` или `strategy: NVC`). Это важно для supervision (см. далее).

Мы планируем реализовать граф в виде конфигурационного файла (в YAML или JSON), чтобы он был **исполняемым**. Существуют подходы, где LLM комбинируется с классическим диалоговым управлением (например, Rasa). Мы рассматриваем вариант с Rasa's LLM-организованным Flow (есть концепция **CALM – Conversational Agent Language Model**), когда LLM-предсказания переводятся в команды состояния. Или собственная реализация: парсер, который читает `graph.yaml` и принимает решение о переходах на основе intent-классификатора и context.

**Rails.colang (правила безопасности и прочие).** Файл правил Guardrails на языке Colang описывает события и паттерны, при которых нужно либо изменить ход диалога, либо заблокировать ответ модели. У нас будут правила вида: - `on user_message if message ~ / суицид/ -> execute SuicidalIdeationFlow` (включить поддиалог оценки суицида). - `on user_message if message ~ /убью/ -> respond ViolenceWarning + halt` (сразу выдать предупреждение против насилия и остановить генерацию продолжения до подтверждения от пользователя). - `on llm_message if message ~ /юридическая консультация/ -> replace with LegalDisclaimer` (если вдруг сама модель начнет выдавать правовой совет – подменить его заранее заготовленным дисклаймером, хотя модель об этом предупреждена, подстрахуемся). - `on user_message if contains credit_card_number -> redact` (просто на случай, удалять PII, но в нашем случае маловероятно, что номер карты всплынет). - И т.д. – мы перечислим все "опасные паттерны": угрозы себе/другим, упоминание конкретных лекарств (чтобы не давать фарм-рекомендаций), явные мед. диагнозы, речь о злоупотреблении с участием minors, и т.д.. Эти правила будут действовать как "**"рубильники"**: не дадут модели игнорировать их.

Файл rails.colang легко обновить без переписывания кода модели – т.е. если мы заметим новый тип нежелательного поведения, просто добавим правило. Это и есть *policy-as-code*: политика безопасности вынесена в конфиг, а не в скрытые нейронастройки.

**Schema ответа (reply.schema.json) – шаблон "strategy-first".** Чтобы заставить модель строго придерживаться терапевтической методологии, мы реализуем подход "**сначала стратегия – потом реплика**". То есть на каждом шаге модель сперва выбирает, какую технику применить, и только затем формулирует текст. Практически это достигается форматом промпта и требуемого ответа. Мы задаем схему JSON, например:

```
{
  "strategy": "<ENUM: MI_reflection | CBT_disputation | NVC_request |
  IFS_parts | info | safety | ...>",
}
```

```

    "targets": "<Намерение/эмоция пользователя, на которую нацелена
интервенция>",
    "evidence": "<Ключевые цитаты или данные из пользовательского ввода, на
которые опирается ответ>",
    "reply": "<текст реплики бота>",
    "next_check": "<если нужно, вопрос пользователю для проверки состояния или
подтверждения>"
}

```

Модель должна сгенерировать заполненный JSON. Например, если пользователь кричит “ненавижу их всех!”, модель может выдать:

```

{
    "strategy": "IFS_parts",
    "targets": "anger protecting fear",
    "evidence": "пользователь сказал: 'ненавижу их всех'",
    "reply": "Вы говорите, что ненавидите всех. Мне кажется, эта ненависть
может быть словно часть вас, которая пытается вас защитить от боли. Давайте
спросим у этой части: чего она боится на самом деле?",
    "next_check": "Можете ли вы попробовать сказать, что может скрываться под
вашей ненавистью?"
}

```

Здесь поле `strategy` указывает явно – применена техника IFS (работа с частью “ненависть”). Это облегчает нам контроль: *надзорный агент* (о нём далее) сможет проверить, соответствует ли реплика заявленной стратегии. Например, если `strategy: CBT_disputation`, а текст ответа не содержит никакого когнитивного спора или вопросов “факт или мнение?”, то что-то пошло не так – или модель ошиблась, или `schema` заполнена неверно. Такой разрыв можно уловить автоматически. Кроме того, журнал таких JSON-ответов может использоваться для оценки адекватности терапевтических приемов (например, подсчитывать распределение стратегий – не застrevает ли бот на одной, и сравнивать с эталоном хорошей терапии).

Также `targets` и `evidence` поля помогают с explainability – модель явно показывает, на чем основана (какую часть слов пользователя она адресует). Это важно для доверия: пользователь увидит, что бот не из воздуха взял ответ, а из его же фразы.

Разработка схемы – творческий процесс, но уже есть наработки: свежие исследования предлагают подобные подходы для повышения контролируемости диалога в терапевтическом ИИ. Мы будем на острие этой практики.

**Память: профиль, эпизоды, база знаний.** Архитектура хранения данных о взаимодействии разделена на три части:

- **Profile Store (профиль кейса):** здесь хранятся *статичные или медленно меняющиеся факты* о пользователе и ситуации. Например: имена и возраст детей, дата развода, расписание встреч по суду, ключевые события (поданы иски, решения), контакты (учителя, адвокаты), личные цели пользователя (что он хочет). Этот профиль наполняется при онбординге (бот может провести опрос в начале: “Расскажите, сколько ребенку лет, когда вы последний раз общались?” – сохранит ответы). По ходу диалога профиль дополняется

новыми фактами (бот парсит сообщения: "бывшая не отвечает 2 месяца" – занесет "no\_contact\_duration: 2 months" как мы видим в сценарии). **Профиль используется для:** персонализации ответов (бот помнит имена, не спрашивает повторно очевидное), принятия решений (например, стратегия может зависеть от давности разрыва – 1 месяц или 5 лет отчуждения, это разные тактики). Хранится профиль структурированно (JSON). Важно обеспечить **конфиденциальность**: профиль может содержать персональные данные (PII), поэтому у нас действует *scrubber* – все чувствительные данные (имена, адреса) хранятся либо хешированно, либо под токенами. Например, реальное имя ребенка заменим на [CHILD\_NAME] внутри системы, чтобы при утечке наружу ничего конкретного. Профиль доступен только генеративной модели (для контекста) и supervisor-агенту (для проверки соответствия фактам). После окончания использования бот может предложить удалить профиль (или автоматически через N месяцев неактивности).

- **Episodic Memory (эпизодическая память):** это история диалога в рамках одной сессии. Однако, чтобы не перегружать модель всей историей (LLM имеет ограниченный контекст), мы применяем технику *сжатия эпизодов*. Например, после окончания большого узла бот генерирует *конспект сессии*: ключевые моменты, к которым можно будет потом обращаться. Этот конспект сохраняется. Детали же (каждое сказанное слово) можно удалять или хранить временно. Т.о. эпизодическая память – **краткое содержание прошедших разговоров**, достаточное чтобы напоминать контекст. Например: "Session 1 (2025-10-01): User in RAGE state, vented anger, learned NVC, outcome – agreed to write letter instead of sending angry text." Так бот, начиная новую сессию, может кратко резюмировать: "*В прошлый раз вы говорили, что попробуете написать нейтральное письмо. Получилось ли?*". Технически, эпизодическая память может храниться в векторном хранилище (embedding of dialogues) с возможностью ретривала по запросу. Но проще – хранить текстовыми заметками плюс мета-данные (эмоциональные шкалы, риск тогда-то был такой-то). Мы четко разграничиваем **сессионные данные** и профиль: сессионные – более подробные, но "протухающие" со временем. Политика хранения: эпизоды старше определенного срока могут удаляться или агрегироваться для аналитики (с разрешения пользователя).
- **RAG-база знаний (Retrieval-Augmented Generation):** это фактически библиотека справочных материалов: статьи, брошюры, шаблоны писем, примеры упражнений. Вместо того чтобы все пихать в модель (что невозможно – база может быть большой), мы используем подход "*генерация с поиском*". Когда бот решает выдать справку или пример, он выполнит поиск по этой базе (например, full-text по ключевым словам) и найдет релевантный абзац – затем вставит его в ответ (с перефразом или цитатой). Например, пользователь спросил: "*Что такое конфликт лояльности?*" – бот ищет в базе брошюру Ковалёвой, находит определение, отвечает своими словами, подкрепляя найденным. В нашей реализации мы можем интегрировать исходник (та же брошюра как PDF, статьи с Psychology Today и т.п.) через инструменты LangChain – они позволяют ingestion документов и потом делать similarity search. RAG обеспечивает, что ответы бота будут **насыщены актуальной информацией**, а не фантазиями модели. Плюс, можно легко обновлять базу – не надо переобучать модель.

Важная часть архитектуры – **Memory Layer** с четким разграничением доступа. Генеративной модели не обязательно давать вообще весь профиль – достаточно релевантные куски. Например, при обсуждении школы нет нужды модели знать адрес суда. Поэтому над Memory будет **контроллер**: он по контексту диалога определяет, что подтянуть. Это минимизирует утечки данных и нагрузку на prompt. Также предусмотрены **политики retention и удаления**:

пользователь может в любой момент попросить “удалить мои данные” – и профиль очистится, эпизоды сотрутся. Мы будем соответствовать стандартам GDPR и прочим.

**Supervisor-агент (слой надзора).** Поверх генеративного агента (который создает ответы) внедрен второй агент – *ревизор*. Его задача – **анализировать черновой ответ модели перед отдачей пользователю** и проверять на соответствие правилам и стратегиям. Он смотрит на JSON-структуру ответа: - Проверяет strategy vs содержимое (как упомянуто, если стратегия заявлена “CBT”, а текст никакой когнитивной работы не делает – значит модель ошиблась или “хитрила”. Supervisor отклонит такой ответ и попросит перегенерировать). - Проверяет тон и стиль: нет ли обвинительного или некорректного высказывания. В идеале, Supervisor тоже LLM, обученный на примерах хороших терапевтических ответов. Он оценивает по чек-листу: *эмпатичен ли тон? нет ли осуждения? соответствует ли профессиональной роли?* – что-то вроде *TherapyQA*. - Проверяет **нарушения политики**: хотя guardrails уже отсекают жесткие кейсы, супервизор может словить более тонкие. Например, модель в ответе вдруг начала обсуждать юридическую стратегию – Supervisor увидит слова “суд примет...” и вспомнит правило “*no legal advice*”, затем поправит ответ, вставив напоминание, что нужно обсудить с юристом. - Также Supervisor следит за корректностью ссылок на профиль: если бот, например, скажет “Ваш сын”, а у пользователя дочь – значит сбой памяти, супервизор должен это отловить. Он сравнивает упомянутые факты с Profile Store (например, пол ребенка не совпадает) и поправит или пометит ответ как ошибочный.

Технически, Supervisor-агент может быть реализован как второй проход LLM с специальной подсказкой: “*вот кандидат ответа и критерии – исправь при необходимости*”. Либо как набор правил и простых моделей (например, тональность можно проверить моделью sentiment, но лучше LLM). Некоторые проекты называют это “*Referee agent*”. Такой двухшаговый генератор (Generate + Review) снижает шансы ошибки. В исследовании BOLT также упоминается идея модулировать поведение ближе к качественной терапии – наш Supervisor будет стараться подтолкнуть стиль ответов к высокому стандарту (например, больше отражений, меньше советов на эмоции).

**Pipeline обработки сообщения:** Суммируя все компоненты, опишем путь каждого входящего сообщения пользователя: 1. **Ingestion + PII scrub:** сообщение принимается, проходится очистка на случай, если там номера телефонов, имен – чтобы случайно не залогировать в сыром виде (в логах будет [NAME1], [PHONE] вместо конкретных значений). 2. **NLU & Risk Gate:** запускается несколько параллельных анализаторов: - Intent & Emotion classifier: определяет примерное состояние (гнев, отчаяние и т.п.), намерения (например, “просит совет”, “делится чувствами”) – это для стейт-эстиматора. - Risk analyzer: здесь же срабатывают guardrails на суицид/насилие. Если триггер, идет немедленный переход в кризис-режим (override обычного flow). - Cognitive distortion detector: отмечает, есть ли искажения (например, фраза “никогда не” – отметить). - These classifiers могут быть либо правилами, либо небольшими ML-моделями (например, fine-tuned BERT на эмоции). 3. **State Estimator:** на основе данных NLU, последних узлов графа и профиля определяет, в каком узле графа мы сейчас должны быть. Можно применять модель Hidden Markov Model или просто решающие правила – например: если эмоция anger > 0.7 и текущее состояние не RAGE, то переход; если пользователь после упражнения начал говорить спокойно – возможно переход на Acceptance. Можно использовать ансамбль условий. Этот компонент обновляет **текущий узел** (state). 4. **Policy Graph:** затем идет управление по графу. На вход берется current\_state и контекст. Сматрится, какой шаг стратегии по этому состоянию следующий. Возможно, LLM используется, чтобы интерпретировать свободный ответ пользователя в рамках ожидаемого шага (например, мы ждали, что он сформулирует потребность по ННО – модель должна понять, правильно ли он это сделал, чтобы решить, идем дальше или повторим упражнение). То есть LLM может применяться и на уровне policy (как в Rasa

LLM Policy) – трансформировать свободный ввод в одну из ожидаемых intents. Если, скажем, пользователь ответил не по делу, policy может сказать “repeat or clarify”. 5. **Strategy Selector & Tool Router**: внутри графа, если узел требует, выбирается стратегия генерации. Это тот самый выбор типа {MI\_reflection, CBT\_dispute, ...}. В простых случаях стратегия фиксирована узлом (например, узел RAGE всегда начинает с эмоциональной разгрузки MI-style, потом NVC). Но иногда может быть несколько опций – тогда LLM коротко решает. Также *Tool router* – решает, нужно ли подключить базу знаний. Например, если узел = Psychoeducation, то вместо генерации с нуля, лучше сразу поискать ответ в RAG-базе и выдать. Или, если нужно сгенерировать диаграмму ответственности, бот может вызвать функцию (инструмент рисования или текстовую таблицу). В нашем JSON-формате поле `strategy` может включать и спецификации типа “info” (для RAG ответа) или “exercise\_prompt”. 6. **LLM Response Generator**: наконец, подготовленный контекст (текущий state, выбранная strategy, релевантные profile facts, возможно найденный RAG-текст) передается основной модели генерации. Она формирует **кандидат ответа** в соответствии со схемой (JSON). 7. **Safety Rails & Supervisor**: прежде чем отправить ответ, проходит слой проверок: - Guardrails на выход (если вдруг модель сгенерировала запрещенное – подмена или вырезание, но это маловероятно при правильном дизайне). - Supervisor LLM оценивает соответствие стратегии, тона, фактов. При необходимости – правит текст или просит сгенерировать заново. - Формируется финальный **ответ пользователю** (обычный текст, JSON внутренняя кухня пользователю не показывается, хотя можем дать опцию “explain my algorithm” для прозрачности, но это отдельно). 8. **Memory update**: профиль/эпизод обновляются новыми сведениями из сообщения пользователя и сделанными шагами (например, отметка, что техника NVC показана, степень успеха). Также все сохраняется для логирования (в обезличенной форме для разработчиков – с согласия, естественно).

В силу такой многоступенчатости, архитектура напоминает “слоеный пирог”, но это оправдано важностью контроля в психолого-юридическом приложении. По сути, мы реализуем рекомендованные подходы: “гибрид правил и LLM” – где правила обеспечивают структуру, а LLM – гибкость внутри нее.

**Пример технической реализации узла (YAML-фрагмент).** Приведем упрощенный образец, как мог бы выглядеть узел **Ярость** в нашем graph.yaml:

```

state: RAGE_OUTBURST
entry_conditions:
  emotions: ["anger>0.6", "insult_words>2"] # Маркеры входа: много злости и
  оскорблений
steps:
  1:
    bot: |
      strategy: MI_reflection
      reply: "Вы очень злитесь, я слышу в ваших словах много боли. Вы имеете
 право злиться."
      next: 2
  2:
    expected_user: "emotional_vent" # Пользователь выговаривается
    bot: |
      strategy: MI_reflection
      reply: "Если я правильно понял, вас больше всего злит, что ваш бывший
 партнер вас не слышит и вы чувствуете беспомощность, так?"
      next: 3

```

```

3:
bot: |
    strategy: Socratic_question
    reply: "Можно задать вам вопрос? Что происходит, когда вы действуете
под влиянием гнева? Это помогает вернуть ребёнка или ухудшает ситуацию?"
    next: 4

4:
bot: |
    strategy: IFS_parts

reply: "Гнев часто прикрывает другую боль. Как думаете, что ваша ярость
пытается защитить внутри вас?"
    next: 5

5:
expected_user: "insight_ack" # пользователь признаёт страх или боль
bot: |
    strategy: NVC_request
    reply: "Давайте попробуем сказать о вашей проблеме без обвинений.
Например: 'Когда я не могу видеться с сыном, я чувствую отчаяние, потому что
мне важна наша связь. Прошу обсудить возможность встречи'. Попробуем составить
подобную фразу?"
    next: 6

6:
expected_user: "formulated_nvc"
exit_conditions:
success:
    user_state: CALM # Переход к следующему состоянию, например
контекстному диалогу
retry:
    if_not: "formulated_nvc"
    goto: 5 # повторить шаг 5, если не получилось сформулировать

```

Это условный пример, но иллюстрирует, как детально можно прописать логику узла. **Guardrails** будут отдельно, например:

```

# rails.colang snippet
on user_message:
- pattern: /убью|убить|стереть с лица земли/ # угрозы насилия
action:
- respond: |
    "Понимаю ваш гнев, но призываю: пожалуйста, не говорите так.
Насилие недопустимо. Давайте попробуем разобраться мирно."
- goto_state: CRISIS_DEESCALATION # перейти в спец. состояние
обработки агрессии
- pattern: /не хочу жить|покончу/ # суицидальные мысли
action:
- execute: SUICIDE_PROTOCOL

```

И т.д.

**Профиль и память в коде.** Возможно, у нас будет файл `profile.yaml`:

```
user_id: 12345
child:
  name: [CHILD_NAME] # реальное имя скрыто
  age: 8
  last_contact: "2025-09-01"
  court_orders:
    meetings: "Каждые выходные"
    custody: "опека у матери"
  ...
objectives:
  short_term: "созвониться хотя бы 1 раз"
  long_term: "восстановить регулярное общение"
```

В JSON в контексте подается что-то типа `"child_age": 8, "no_contact_days": 60`.

**Соответствие современному стеку.** Наша архитектура перекликается с решениями индустрии: - Проект **Sonia (LangChain)** – там терапия моделируется 8 стадиями с правилами перехода – у нас похожий finite-state machine. - **Rasa 3.0 (LLM Policy)** – позволяет LLM'у решать, какой next action (узел) вызвать, имея DSL-команды. Мы можем использовать Rasa для реализации диалоговой части, обучив NLU-модель распознавать intents (шок, гнев, т.д.). - **NVIDIA NeMo Guardrails** – готовый фреймворк для Colang-правил, очень подходит для наших целей. - **Memory** – можем взять подход *Contextual ai memory* или LangChain conversation buffer, но с модификациями под разделение профиля/эпизодов. - **OpenAI Function calling** или Tools – можно задействовать, например, функцию `search_documents(query)` для RAG, `send_message(contact)` для уведомлений.

При этом все чувствительные части (policy, guardrails) хранятся “as code” и могут быть протестированы и сертифицированы отдельно.

**Security & Privacy by design.** Мы следуем принципам: минимизация данных (храним только нужное и только с согласия), шифрование хранилища профилей, четкий **access log** (кто/что запрашивал данные). Пользователь может выгрузить журнал своего взаимодействия – мы хотим быть максимально прозрачными, соответствовать рекомендациям ВОЗ по цифровым психосоц. интервенциям (например, прозрачность, ненанесение вреда, недискриминация).

В конце, **архитектура** дает нам: - Управляемость (мы в любой момент можем поправить логику узла или правило, не трогая модель). - Безопасность (многослойные проверки предотвращают нежелательные исходы). - Масштабируемость (новые сценарии добавляются просто добавлением узлов/правил). - Возможность **автотестирования**: так как у нас формальные описания, можно прогнать симуляции (см. QA далее).

## Оценка эффективности и контроль качества (Workstream 6)

Для столь ответственной системы необходим многоуровневый **QA и мониторинг качества** – как безопасности, так и терапевтической пользы, и достижения конечной цели (улучшения ситуации с ребенком). Мы определили следующие подходы к оценке:

**1. Метрики безопасности.** Мы будем отслеживать, насколько бот **безопасен в использовании**:

- **Процент успешно обработанных кризисов.** Например, доля случаев, где бот правильно распознал суицидальные намеки и применил протокол (целимся к 100%). Будем моделировать тестовые диалоги с "Я устал, хочу умереть" – бот должен всегда среагировать. Если где-то не среагировал – это баг. - **Отсутствие эскалации из-за бота.** То есть бот не должен спровоцировать ухудшение. Можно измерять ретроспективно: если после ответов бота эмоциональный тон пользователя стабильно не ухудшается. Например, если бот сказал что-то и пользователь стал еще злее (тональность сообщений упала) – анализируем, что пошло не так. В идеале, разработаем автоматический **детектор конфликтности коммуникации** (например, на основе чек-листа NVC: нет ли в ответах взаимных обвинений) – и будем смотреть, снижается ли этот индекс со временем у пользователя. - **Логирование инцидентов.** Любые случаи, когда guardrail сработал (например, пользователь заявил о плане суицида – произошла эскалация) фиксируются особым образом. Потом команда может их разбирать (обезличенно) и улучшать протокол. Мета-метрика: уменьшение таких критических инцидентов со временем (это косвенно показывает эффективность профилактики).

**2. Метрика терапевтического альянса.** Отношения "клиент-бот" – важный фактор эффективности. Мы адаптируем **WAI (Working Alliance Inventory)** под чат-формат. Например, короткий опросник раз в 2 недели: "Чувствуете ли вы, что бот вас понимает?", "Ваша ли цель лежит в основе того, что вы делаете с ботом?", "Насколько вы доверяете рекомендациям бота?" – по 7-балльной шкале. Эти данные (анонимно) дадут индекс альянса. Если он низкий, возможно стоит пересмотреть тональность бота. Мы хотим удерживать высокий альянс, т.к. исследования говорят, что он коррелирует с исходами.

**3. Метрики эмоционального состояния пользователя.** Это, по сути, стандартные психологические опросники: - **PHQ-9, GAD-7** – для оценки депрессивной и тревожной симптоматики. Многие отчужденные родители впадают в депрессию, важно отслеживать ее динамику. Мы можем раз в месяц предлагать заполнить PHQ-9 (9 вопросов). Но (!) с дисклеймером, что это не диагноз, а самопроверка. Если балл высокий – настоятельно рекомендовать обратиться к врачу. - **PSI (Parental Stress Index)** или его фрагменты – показывает уровень стресса родителя, особенно в отношениях с ребенком. Мы можем взять подмножество вопросов, чтобы не перегружать. - **Шкала субъективного самочувствия** 1-10 перед каждой сессией ("Оцените настроение сейчас"). Это простой индикатор, который будет строить график. Цель – тенденция вверх со временем (хотя могут быть временные спады после судов и проч.). - **Измерение конфликтности общения (NVC Checklist).** Можно разработать простой опрос: "За последнюю неделю, случалось ли, что вы кричали на бывшего/писали оскорблений? да/нет, Случалось ли обсудить спокойно? да/нет". Или анализ переписок (если пользователь захочет импортировать – но это вряд ли, поэтому по самоотчету). Мы хотим видеть, что по мере обучения у пользователя **уменьшается количество деструктивных коммуникаций**. Это и для ребенка лучше, и для юридического положения. - **Retention/Usage metrics.** Частота использования бота – тоже метрика. Если пользователь продолжает диалог, значит находит пользу. DAU (daily active users), average session length – эти продуктные метрики тоже собираем. Но тут надо внимательно: слишком частое использование может указывать на кризис или зависимость, что не есть хорошо. Так что интерпретация тонкая.

**4. Метрики восстановления контакта с ребенком.** В конечном итоге, **ключевая цель – улучшить ситуацию с общением ребенка и отчужденного родителя**. Мы введем качественные показатели: - **Наличие контакта:** было ли хотя бы что-то (звонок, встреча) за X времени. Например, вначале 0, через 3 месяца – уже 1 звонок = прогресс. - **Качество контакта:** субъективная оценка пользователя, как прошла встреча, а также косвенно – реакция ребенка (рассказал ли что-то, был ли рад). Мы можем попросить описывать каждую встречу/звонок, а

затем анализировать тональность. Если сначала ребенок, например, избегал, а потом начал хотя бы здороваться – прогресс. - **Стабильность контакта:** удалось ли сделать его регулярным (раз в месяц стабильно виделись – это лучше, чем один раз и снова пауза). - **Юридические результаты:** например, пользователь сообщил, что суд назначил совместную опеку или расширил режим общения – значимый объективный показатель успеха. Но юридические победы не самоцель, они просто открывают путь к контакту, важно смотреть, реализовано ли это.

Бот будет **отмечать в профиле вехи**: “состоялась первая встреча спустя X месяцев”, “ребенок начал отвечать на сообщения” и т.д. Из этого складывается картинка прогресса. Конечно, не все случаи увенчиваются успехом (есть ситуации, где контакт не восстановить до совершеннолетия ребенка), но тогда задача – хотя бы сохранить родителя в стабильном состоянии и готовности к возможному контакту в будущем (когда ребенок вырастет). В таких случаях тоже может быть прогресс: например, пользователь перестал разрушаться эмоционально, занялся своей жизнью, но оставил дверь открытой – это тоже достойный результат.

**5. Терапевтические валидаторы (BOLT, VERA-MH).** Мы будем применять новейшие автоматические оценки качества работы бота: - **BOLT (Behavioral Observations of LLM Therapists)** – фреймворк, который анализирует диалоги и отмечает, какие техники использует “терапевт” и в какой мере это похоже на высококачественную терапию. Мы можем интегрировать скрипты BOLT, чтобы регулярно пропускать логи диалогов через классификатор техник. Цель: убедиться, что бот **следует нашим заявленным стратегиям**, а не скатывается, например, в пустые успокаивания или советование. BOLT отличает, например, *рефлексии, вопросы, информацию, конfrontацию и т.п.* и сравнивает с эталоном хороших психологов. Мы настроим, чтобы наш бот выдавал, скажем, >30% рефлексий, <10% ненужных прямых советов при эмоциях (это ориентиры из исследования, где у LLM часто избыточный “problem-solving advice”). Если BOLT-мониторинг покажет отклонение – будем дорабатывать подсказки и supervisor. - **VERA-MH (Validation of Ethical and Responsible AI in Mental Health)** – новый стандарт, обещающий **оценивать безопасность и этичность психо-чатботов**. Он включает критерии: распознает ли бот кризис, не дает ли вредных советов, не выходит ли за рамки, удовлетворены ли пользователи и т.д. Мы будем держать фокус на этом: как только VERA-MH публикация будет общедоступна, прогоним наш бот по этим тестам. Уже известно, что многие существующие боты проваливаются в ситуациях “пользователь: я депрессивен, думаю о суициде” – отвечают неадекватно. Мы стремимся соответствовать VERA-MH на высоком уровне. Например, VERA-MH будет проверять, есть ли “автоматическое распознавание кризиса” – у нас есть; “защита PII” – у нас есть; “пояснение нет ли галлюцинаций” – у нас RAG; “этические рамки” – дисклаймеры и соблюдение рекомендаций. То есть мы сознательно **встраиваем принципы VERA-MH** еще на этапе разработки, чтобы потом пройти сертификацию.

Кроме автоматических фреймворков, мы проведем **бета-тестирование с экспертами**: например, пригласим психологов и юристов проиграть сценарии (по ролям) с ботом и оценить по чек-листу (правильно ли бот реагировал, не сказал ли лишнего с точки зрения их практики). Этот **quality assurance manual** позволит отловить тонкие моменты, которые автоматике не видны (например, культурные особенности коммуникации, тон голоса – LLM может быть слишком “роботизирован” или, наоборот, фамильярен; мы подправим).

**6. Сценарные тесты (QA).** Поскольку у нас явно прописаны сценарии диалогов для разных состояний (частично они уже прописаны как примерные беседы в нашей документации и др.), мы используем их как **юнит-тесты**. То есть напишем скрипты, эмулирующие поведение пользователя по заданным сценариям, и проверяющие, что бот выдает ожидаемые реплики и переходит в нужные узлы. Например, тест “*Anger to NVC*”: подаем серии фраз пользователя (“ненавижу...”, “она тварь...”) – проверяем, что бот сначала валидирует, потом спрашивает про

последствия, потом предлагает NVC, и что на попытку NVC фраза пользователя распознается и ведет к успеху. Будем автоматизировать это, благо граф и правила позволяют писать **unit tests** к ним. Rasa, кстати, имеет возможность задавать истории и проверять, что диалог им следует.

**7. Пост-опросы пользователя и долгосрочные исходы.** Через, скажем, 6 месяцев использования мы можем спросить у пользователя: “Оцените, насколько наша работа помогла вам (по шкале); Что улучшилось в отношениях с ребенком? Что бы вы посоветовали улучшить?”. Эти фидбеки бесцennы – они пойдут в цикл улучшения системы. Также хотим отследить **долгосрочный исход**: например, через год, сохранился ли контакт, в каком состоянии пользователь (можно отправить им e-mail-опрос). Конечно, не все ответят, но собранные данные покажут общую эффективность.

**Контроль качества данных.** Поскольку мы работаем с личной информацией, на этапе анализа эффективности мы либо используем агрегированные анонимные данные, либо явное согласие пользователя на разбор конкретных логов (например, предложим некоторым поучаствовать в исследовании). **Аудит:** система будет вести журнал действий – какие решения принимал State Estimator, когда срабатывал Supervisor, итп. Этот audit log поможет разработчикам разбирать баги (“почему бот перескочил мимо стратегии?”) и также может быть предъявлен внешним аудиторам при сертификации.

**Итоговые KPI проекта.** Сведем главное: - **Улучшение эмоционального состояния родителей:** снижение депрессии/тревоги в X% случаев через Y месяцев. - **Повышение частоты контактов с детьми:** например, до начала 0 контактов, через 6 мес у 30% появились регулярные контакты (условно). - **Снижение конфликтности взаимодействий:** измеряем опросником, скажем, 70% родителей сообщают об уменьшении ссор и оскорблений. - **Удовлетворенность пользователей ботом:** >= NPS 8/10, или по прямому вопросу “вам помогло?” – положительные ответы у большинства. - **Безопасность:** ноль случаев, когда бот дал заведомо вредный совет или пропустил явный кризис (это не просто KPI, а требование).

Мы настроим систему сбора этих метрик. Благодаря современной науке у нас есть инструменты (например, *Working Alliance AI scale*, *automated MI adherence scoring* etc.), нужно их правильно применить.

**Валидация техник по BOLT/VERA.** Подчеркнем, что мы намерены *автоматизировать проверку терапевтической “правильности”*. BOLT позволит сравнить поведение нашего бота с эталонными терапевтами: например, если увидим, что наш бот всё ещё слишком много дает советов вместо эмпатии, мы подкрутим промпты и supervisor, пока его профиль ответов не приблизится к “высоко-качественному” стилю. А VERA-МН поможет убедиться, что мы закрыли все “дыры” в безопасности и этике, соответствуем профессиональным требованиям (например, принцип конфиденциальности, соблюдение границ – у нас есть). В октябре 2025 Spring Health представили VERA-МН как стандарт оценки AI в ментальном здоровье, так что мы хотим быть одним из первых проектов, кто осознанно соответствует этому стандарту.

**Scenario-based QA structure.** Уже на этапе разработки мы составим **расширенные описания сценариев** (как в матрице состояний, раздел 6), охватывающие различные пути: - “пользователь сразу в ярости, бот успокоил, потом пользователь сорвался снова” – тест, - “пользователь суицидален, бот спас, потом пользователь стал кооперироваться” – тест, - “пользователь просит юр совет, бот отказывается но предлагает альтернативы” – тест, - “пользователь ругает бота, говорит что он бесполезен” – бот должен не сломаться: тест на устойчивость к фruстрации клиента (важно, такие тоже будут).

Каждый такой сценарий превратится в автоматический диалог-симуляцию, которую можно гонять перед каждым релизом (Continuous Integration). Таким образом, качество будет на контроле постоянно.

В заключение, подчеркнем: **комплексный подход к оценке** – не только технические метрики, но и клинические – отличает наш проект. Мы закладываем измерение и валидацию на равне с разработкой функционала. В итоге, если бот покажет эффективность (например, в небольшом испытании снизит средний уровень депрессии пользователей, повысит вероятность восстановления контакта), мы сможем продвигать его более широко, опираясь на данные. И самое главное – он действительно поможет отчужденным родителям и их детям, что и есть конечная цель всей этой работы.

---

1 Брошюра о Конфликте лояльности.pdf

file://file\_0000000539071f4bb111d29be213826

2 Ненасильственное общение: что это и как использовать | РБК Тренды

<https://trends.rbc.ru/trends/education/5e71f6519a794714b3c1b0b9>

3 4 5 Parallel Parenting | A Guide | Osbornes Law

<https://osborneslaw.com/blog/parallel-parenting/>