УДК 004.8 **DOI: 10.34670/AR.2026.20.81.020**

Проектирование архитектуры и выбор методов функционирования модуля этического управления

Румак Дарья Павловна

Магистр,
Московский государственный технический университет
им. Н. Э. Баумана (национальный исследовательский университет),
105005, Российская Федерация, Москва, ул. 2-я Бауманская, 5/14;
e-mail: le.rumak@gmail.com

Стельмах Яна Сергеевна

Ассистент,
Московский государственный технический университет
им. Н. Э. Баумана (национальный исследовательский университет),
105005, Российская Федерация, Москва, ул. 2-я Бауманская, 5/1;
e-mail: stelmakyana99@gmail.com

Сухобоков Артём Андреевич

Главный консультант,
SAP America, Inc.,
19073, США, Ньютаун-Сквер, ул. Уэст-Честер-Пайк, 3999;
e-mail: artem.sukhobokov@gmail.com

Аннотация

В статье показана возможность создания этического модуля, применимого в широком классе систем управления предприятиями и организациями, обеспечивающего соблюдение этических норм и адаптивное принятие решений интеллектуальным агентом в условиях морального конфликта. В рамках этого предложена иерархическая структура норм в виде графовой базы данных. Обосновано применение алгебры совести В. А. Лефевра и аппарата АМ-алгебры Л. И. Волгина для перехода к непрерывным математическим моделям этического выбора. Разработан метод совмещения формализованной модели морального выбора с эмоциональным фоном системы, что позволяет реализовать динамическую коррекцию параметров готовности агента к действию. Иерархия норм организована в виде нескольких функциональных уровней, объединяющих внутренние регуляторы, прикладную и ситуативную этику, а спроектированная архитектура обеспечивает разделение математической модели, нормативной базы знаний и пользовательского интерфейса, что позволяет масштабировать систему при добавлении новых норм и

сценариев. Результатом проектирования стала архитектура этического модуля на основе нечеткого логического вывода, осуществляющего вычисление итогового этического веса действия с учётом интенции, давления среды и внутренних установок агента. Предложенный подход позволяет повысить адаптивность, безопасность и интерпретируемость действий ИИ-систем в соответствии с требованиями современных стандартов регулирования.

Для цитирования в научных исследованиях

Румак Д.П., Стельмах Я.С., Сухобоков А.А. Проектирование архитектуры и выбор методов функционирования модуля этического управления // Психология. Историко-критические обзоры и современные исследования. 2026. Т. 15. № 4А. С. 158-179. DOI: 10.34670/AR.2026.20.81.020

Ключевые слова

Искусственный интеллект, машинная этика, этическое регулирование, алгебра совести, алгебра мер, нечеткая логика, графовая база знаний, адаптивное поведение, рефлексивный выбор, архитектура АСОИУ.

Введение

В предыдущей статье была показана необходимость разработки универсального модуля этики для широкого класса систем корпоративного управления и предложено использовать комбинированный способ реализации этики, сочетающий различные подходы к реализации этики интеллектуального агента;

– деонтологический подход;
– утилитарный подход;
– подход на основе данных;
– этику добродетели.

Целью данной статьи является демонстрация возможности разработки модуля этического управления для интеллектуальной системы, обеспечивающего нормативную фильтрацию действий агента на основе иерархической базы знаний и гибридной математической модели морального выбора.

Материалы и методы исследования

Математический аппарат рефлексивного управления этическим выбором

В качестве теоретического базиса для проектирования программного модуля этического управления обратимся к теории рефлексивного выбора, разработанной В.А. Лефевром. В рамках архитектурной реализации данного компонента АСОИУ процесс принятия этически значимых решений рассматривается не как классическая стохастическая задача численной оптимизации или поиска максимальной полезности (характерная для аппарата теории игр или утилитарного анализа), а как строго формализованный вычислительный процесс функционирования внутренней рефлексивной структуры интеллектуального агента. Отличительной чертой данного аппарата является использование алгебраических функций для моделирования актов морального выбора [Lefebvre, 2001]. Это позволяет перейти от описательных психологических

моделей к строго вычислимым алгоритмам, пригодным для интеграции в общую архитектуру автоматизированных систем обработки информации и управления.

Теория рефлексивных систем позволяет интерпретировать акт этического выбора как закономерный результат взаимодействия векторов внутренних установок агента и внешних управляющих сигналов среды. Данное взаимодействие формализуется в виде системы алгебраических зависимостей и уравнений состояния, что превращает процесс принятия решения в последовательность логических операций, выполняемых программным модулем в реальном времени [Assadi, 2024]. Таким образом, этическое управление трансформируется из области философских рассуждений в инженерную задачу синтеза алгоритмов управления автономным поведением агента.

В основе модели лежит ситуация биполярного выбора. Предполагается, что в любой проблемной ситуации интеллектуальный агент сталкивается с необходимостью выбора между двумя альтернативами, обладающими различным аксиологическим статусом (ценностным весом). Согласно работам С. Филимонова, этим альтернативам ставятся в соответствие два полюса:

– вариант А (негативный полюс) соотносится с понятием «Зло» и принимает значение логического «0»;
– вариант В (позитивный полюс) соотносится с понятием «Добро» и принимает значение логической «1».

Логика алгоритма строится на приоритетности Варианта В перед Вариантом А. Это означает, что в системе изначально задан вектор предпочтения позитивного полюса, что позволяет использовать его как целевое состояние при выполнении математических расчетов и определении итогового поведения агента.

Инженерная задача модуля заключается в математическом описании результирующего выбора агента, совершаемого под воздействием внутренних интенций и внешнего нормативного давления, которое испытывает агент. Это давление складывается из объективной привлекательности полюса (реальное положение вещей), субъективной рефлексивной оценки ситуации агентом (его вера в то, что это хорошо) и его изначального намерения до момента принятия решения.

Первоначально теория В.А. Лефевра была разработана для описания бинарного морального выбора (добро/зло). Рефлексия в данном контексте понимается не просто как размышление, а как способность агента строить модель самого себя и модель внешнего давления. Это моделируется через иерархию «персонажей» внутри одного агента, где каждый вышестоящий уровень осознает установки нижестоящего. Формальный характер теории допускает обобщение на непрерывную область, что делает возможной адаптацию данной теории для задач управления интеллектуальными агентами.

Для моделирования процесса принятия этического решения вводится набор переменных состояния агента, значения которых принадлежат замкнутому интервалу $[0,1]$. Такой выбор области определения позволяет в дальнейшем интегрировать модель с аппаратами непрерывной и нечеткой логики.

Вводятся следующие переменные состояния агента:

– Переменная $\boldsymbol{x_1}$ (давление среды / норма) отражает объективную этическую оценку действия, задаваемую внешней системой правил. Значение $x_1 = 1$ соответствует требованию совершить действие (позитивный полюс, «добро»); $x_1 = 0$ соответствует запрету на его выполнение (негативный полюс, «зло»). В интеллектуальной системе эта переменная является выходным значением базы.

– Переменная $\boldsymbol{x_2}$ (оценка давления субъектом) описывает внутреннее представление агента о внешнем требовании, то, как агент воспринимает и интерпретирует x_1. В идеальных условиях $x_2 = x_1$, однако при искажении информации или целенаправленных манипуляциях возможно рассогласование: $x_2 \neq x_1$. Это позволяет моделировать ситуации «обмана» агента внешней средой.

– Переменная $\boldsymbol{x_3}$ (интенция агента) отражает собственное намерение (желание) агента до момента принятия окончательного решения и формируется его целевой функцией, приоритетами или внутренними параметрами. Это внутренний импульс «хочу», который может противоречить норме «надо» (x_1).

Итоговое решение агента описывается переменной X , интерпретируемой как степень готовности совершить действие (выбор позитивного полюса). Таким образом, задача этического управления сводится к нахождению функциональной зависимости, представленной в формуле 1.

$$X = F(x_1, x_2, x_3), \tag{1}$$

где X – итоговая готовность агента к совершению действия;

x_1 – нормативная оценка действия, определяемая на основе базы знаний этических норм;

x_2 – восприимчивость агента к нормативному давлению;

x_3 – техническая интенция агента, отражающая исходное стремление выполнить действие.

Данная функциональная зависимость позволяет предсказать реальное поведение агента в условиях конфликта между «хочу» (x_3) и «надо» (x_1).

Важным расширением математического аппарата является учет типа этической системы, определяющей логику взаимодействия между интенцией агента и давлением среды. В.А. Лефевр выделил два альтернативных типа морального выбора при принятии решения: «Первая этическая система» (ПЭС) и «Вторая этическая система» (ВЭС). В отличие от традиционных описательных подходов, Лефевр формализовал их через различные наборы алгебраических операций над полюсами «Добро» (1) и «Зло» (0).

Первая этическая система (характерная для западной либеральной традиции и иудео-христианской этики) базируется на принципе непримиримости к негативному полюсу. Основной моральный императив здесь формулируется следующим образом: «Компромисс между добром и злом является злом». Алгебраически первая система описывается операциями, где доминирует «чистота» выбора. Для такой системы характерны высокая оценка личной ответственности, склонность к героизму и жесткое следование правилам даже под сильным внешним давлением. Агент первой системы воспринимает конфликт как необходимость борьбы до победы одного из полюсов.

В работах Лефевра ПЭС описывается через операцию логической конъюнкции, приведенную в формуле 2.

$$X = x_3 \cdot x_1, \tag{2}$$

где X – готовность агента к совершению действия;

x_3 – интенция (желание);

x_1 – давление среды (норма).

С точки зрения данной системы, если норма запрещает действие ($x_1 = 0$), то итоговое

решение всегда будет отрицательным ($X = 0$), независимо от силы внутреннего желания агента ($x_3 = 1$). Система блокирует любые попытки оправдания нарушения нормы.

Вторая этическая система (характерная для коллективистских обществ и некоторых восточных культур) базируется на поиске гармонии и сглаживания конфликтов: «Компромисс между добром и злом является добром» (ради сохранения мира). Математически ВЭС описывается через операцию логической дизъюнкции (сложения). Формула выбора принимает вид:

$$X = x_3 + x_1 - x_3 \cdot x_1, \quad (3)$$

В этой системе, если агент имеет сильную интенцию к действию ($x_3 = 1$), то даже при наличии внешнего запрета ($x_1 = 0$) итоговый результат будет положительным ($X = 1$). Система допускает компромисс ради достижения цели. Агенты второй системы ориентированы на достижение мира и консенсуса любой ценой. С их точки зрения, сохранение системы и отсутствие открытого конфликта ценнее, чем победа «правды». Вторая система поощряет уступчивость, но она же делает агента более уязвимым к манипуляциям, так как он готов принять «частичное зло» ради общего спокойствия.

Для наглядности сравним две системы через таблицу истинности комбинации интенции агента (x_3) и внешнего давления (x_1). Сравнительная характеристика логики выбора представлена в таблице 1.

Таблица 1 – Сравнение логики принятия решений в этических системах

Интенция (x_3)	**Норма (x_1)**	**Результат ПЭС**	**Результат ВЭС**
1	1	1	1
1	0	0	1
0	1	0	1
0	0	0	0

В реальности агент учитывает не только норму, но и свое отношение к ней. Для моделирования более сложных структур, учитывающих уровни осознания (модель агента о самом себе и среде), Лефевр ввел понятие рефлексивных многочленов. Символическая запись полинома для первой этической системы приведена в формуле 4.

$$X = {x_3}^{x_2^{x_1}}, \quad (4)$$

где X – итоговое этическое состояние (результат выбора);

x_3 – интенция (желание) субъекта

x_2 – образ самого себя (рефлексия второго уровня);

x_1 – образ давления среды (норма) в сознании агента.

При раскрытии данной экспоненциальной формы в рамках булевой алгебры (для ПЭС) получаем выражение, представленное в формуле 5.

$$X = x_3 \cdot \left(1 - x_2 \cdot (1 - x_1)\right), \quad (5)$$

Данная формула позволяет учитывать не только реальную норму, но и то, как агент воспринимает давление этой нормы через призму собственной рефлексии.

При проектировании модуля этического управления для интеллектуальных систем, функционирующих прежде всего в критически важных областях, таких как медицина или беспилотный транспорт, за основу принимается ПЭС. Данный выбор обоснован необходимостью реализации жестких алгоритмических ограничений и верификации управляющих команд. В рамках ПЭС архитектура модуля исключает возможность нахождения «компромиссного» решения, если планируемое действие классифицировано как прямое нарушение нормы ($x_1 = 0$). Это важно для обеспечения безопасности систем в ситуациях, когда деструктивное с точки зрения этики действие формально обеспечивает максимизацию целевой функции или достижение локального экстремума эффективности. Таким образом, система лишается возможности пойти на компромисс с запрещенным сценарием, независимо от его расчетной результативности. Выбранная парадигма определяет структуру уравнений рефлексивного выбора, которые будут переведены в непрерывную область в следующем разделе. Использование аппарата непрерывной логики позволяет перейти от дискретных состояний к оперированию значениями на интервале $[0,1]$, сохраняя при этом алгебраические свойства первой этической системы, а именно – приоритет блокирующего сигнала при рассогласовании интенции агента и внешних нормативных ограничений.

Переход к непрерывной логике

Изначально модели рефлексивного выбора строились на основе классической булевой логики, оперирующей дискретными состояниями $\{0, 1\}$. Однако для реализации адаптивного поведения интеллектуального агента в динамической среде бинарного выбора недостаточно. Для построения гибких систем управления необходим переход в непрерывную область $[0, 1]$, где переменные интерпретируются как степени уверенности, вероятности или весовые коэффициенты. Методологической основой такого перехода в данной работе является АМ-алгебра (алгебра мер), разработанная Л.И. Волгиным [Волгин, 2003]. Суть подхода заключается в доказательстве того, что базовые логические операции (конъюнкция, дизъюнкция, отрицание) могут быть представлены в виде арифметических выражений, определенных на непрерывном множестве $[0, 1]$.

В рамках АМ-алгебры отображение логических функций в арифметические реализуется через операции инверсии, конъюнкции и дизъюнкции (см. формулы 6–8).

$$\bar{a} = 1 - a, \tag{6}$$

$$a \wedge b = a \cdot b, \tag{7}$$

$$a \vee b = a + b - a \cdot b, \tag{8}$$

где a, b – значения этических переменных в интервале [0, 1];

$\bar{a}$ – операция инверсии (отрицания);

$\wedge$ – знак логической конъюнкции (И);

$\vee$ – знак логической дизъюнкции (ИЛИ).

Данное обобщение имеет принципиальное значение для задач искусственного интеллекта, поскольку позволяет перейти от дискретного морального выбора к непрерывным моделям, совместимым с методами численного анализа, нечеткого вывода и градиентной оптимизации. С программной точки зрения это позволяет заменить сложные логические ветвления if-else на прямые арифметические расчеты, что повышает производительность модуля этического управления.

В линейном приближении, базирующемся на принципах первой этической системы (ПЭС) и АМ-алгебре Волгина, одна из форм записи уравнения формирования готовности агента к действию (этического выбора) приведена в формуле 9.

$$X = x_3 + (1 - x_3) \cdot x_1 \cdot x_2 - x_3 \cdot (1 - x_1) \cdot x_2, \quad (9)$$

где X – выходной управляющий параметр;

$x_1, x_2, x_3 \in [0,1]$ – входные переменные состояния.

Анализ структуры данного уравнения позволяет выделить два функциональных контура коррекции интенции агента:

1) Контур усиления (положительная обратная связь): слагаемое $(1 - x_3) \cdot x_1 \cdot x_2$ отвечает за увеличение готовности агента к действию. Если его внутренняя интенция (x_3) совпадает с тем, что предписывает норма (x_2) требованием этической нормы (x_1), то готовность X растет. Это модель «одобрения» действия системой.

2) Блокирующий контур (контур подавления): слагаемое $x_3 \cdot (1 - x_1) \cdot x_2$ выполняет функцию этического фильтра. Если внешняя норма накладывает запрет $(x_1 \to 0)$, а агент воспринимает этот запрет как достоверный $(x_2 \to 1)$, то вычитаемая часть уравнения минимизирует итоговое значение X, подавляя готовность агента совершить недопустимое действие. Чем выше уверенность агента в запрете (x_2), тем сильнее работает этот математический «тормоз».

Это базовое уравнение этического выбора описывает, как внутреннее намерение агента (x_3) корректируется под воздействием осознаваемого давления среды (x_2). Если интенция совпадает с давлением среды, итоговая готовность усиливается. В противном случае возникает внутренняя коллизия, результат которой зависит от параметров устойчивости агента.

Перевод модели в непрерывную область обеспечивает следующие технические преимущества при реализации модуля:

– Обеспечивается полная совместимость с нечетким выводом, при которой выходные значения алгоритма Такаги-Сугено-Канга, которые представляют собой четкие числа в диапазоне $[0,1]$, напрямую транслируются в переменные x_1 и x_2 уравнения (1).

– Достигается дифференцируемость модели за счет того, что уравнение (1) является гладкой функцией, что позволяет применять его в архитектурах ИИ, требующих обучения через метод обратного распространения ошибки или другие градиентные методы.

– Повышается интерпретируемость результата, позволяющая программно обрабатывать значение $X = 0.85$ может быть программно обработано как «высокая степень готовности», что позволяет гибко настраивать пороги активации действий в зависимости от критичности текущей миссии.

Таким образом, использование аппарата Л. И. Волгина позволяет формализовать этический выбор как непрерывный процесс обработки сигналов, что превращает этическую рефлексию в стандартную вычислительную задачу для бортового вычислителя интеллектуальной системы.

Динамические модели и устойчивость агента

Статические модели, основанные на линейных уравнениях, описывают поведение идеализированного агента в фиксированный момент времени. Однако для реализации адаптивного поведения в составе АСОИУ необходимо учитывать динамику изменения готовности агента к действию, которая носит нелинейный характер. В рамках данного исследования за основу принята квадратично-линейная модель, разработанная С. А. Анисимовой [Анисимова, 2007]. и развитая в работах С. Ю. Малкова.

Для учета эффектов накопления уверенности, инерции принятия решений или, напротив, резкого роста сомнений, используется квадратично-линейная функция выбора. В общем виде, согласно адаптации модели для интеллектуальных систем, функция готовности агента к действию представляется в виде полинома второй степени (см. формулу 10).

$$m(x, y) = ax^2 + bx + c + dxy + ey, \quad (10)$$

где $x \in [0,1]$ – внутренняя интенция агента (x_3 в базовой нотации);

$y \in [0,1]$ – осознаваемое давление среды (x_2);

a, b, c, d, e – весовые коэффициенты, определяющие характер «нравственного профиля» агента.

Использование квадратичной зависимости по переменной x позволяет моделировать нелинейный отклик системы: при достижении определенного порога давления даже незначительное изменение интенции может привести к резкому скачку итоговой готовности X. С точки зрения программирования алгоритма, это позволяет реализовать «эффект барьера», когда агент долго сохраняет устойчивость, но при критическом давлении быстро переключает состояние.

В используемой квадратично-линейной модели коэффициенты функции готовности a, b, c, d, e не являются константами, а динамически вычисляются в зависимости от базовых этических установок агента (βi) и текущего индекса оптимизма α.

Для перехода от общего вида полинома к конкретному алгоритму управления необходимо провести процедуру идентификации коэффициентов. Основанием для этого служит метод фиксации поведения агента в четырех граничных (экстремальных) точках пространства состояний. Эти точки определяют фундаментальный «нравственный профиль» агента и обозначаются через константы β_i:

1) $\beta_1 = m(0,0)$ – выбор агента в отсутствие внутреннего желания и внешнего давления (состояние «чистой совести»);

2) $\beta_2 = m(1,0)$ – выбор при максимальном внутреннем желании, но отсутствии внешнего давления (автономное решение);

3) $\beta_3 = m(0,1)$ – выбор в отсутствие желания при максимальном внешнем давлении (ложное подчинение);

4) $\beta_4 = m(1,1)$ – выбор при одновременном максимуме интенции и давления (ситуация капитуляции или полного согласия).

Значения β_i фиксируются на этапе проектирования личности агента и определяют, к какой этической системе (ПЭС или ВЭС) он тяготеет.

Для обеспечения интеграции этического модуля с модулем эмоционального управления необходимо, чтобы характер выбора $m(x, y)$ зависел от текущего аффективного состояния. В математическую модель вводится параметр $\alpha \in [0,1]$ (индекс оптимизма).

Согласно исследованиям С. Ю. Малкова [Малков, Шпырко, Давыдова, 2024], связь между «настроением» агента и его нравственными установками проявляется в точке неопределенности при $x = 0{,}5$. На основании этого допущения и зафиксированных ранее точек β_i, формируется система линейных уравнений для поиска коэффициентов. Результатом аналитического решения данной системы (методом интерполяции полинома через заданные точки) является следующая зависимость для коэффициентов a, b, c:

Математическая зависимость (10) позволяет реализовать адаптивное поведение модуля. При изменении эмоционального фона (входной сигнал от сопряженного модуля

эмоционального управления) происходит автоматический пересчет параметров рефлексивной функции в соответствии с системой уравнений (11), что изменяет итоговую вероятность принятия решения в условиях конфликта.

$$\begin{cases} a = \beta_1 + \beta_2 + \beta_3 + \beta_4 - 4\alpha \\ b = 4\alpha - 2\beta_1 - \beta_3 - \beta_4 \\ c = \beta_1 \end{cases}, \quad (11)$$

где a, b, c – весовые коэффициенты квадратично-линейной функции выбора $m(x, y)$;

β_i – базовые установки этической системы агента;

α – индекс оптимизма агента, формируемый на основе данных от модуля эмоционального управления.

Подобная структура позволяет модулю этического управления динамически менять «жесткость» фильтрации в зависимости от данных, поступающих из смежных подсистем (например, модуля эмоционального управления). Она позволяет в режиме реального времени пересчитывать кривизну функции выбора $m(x, y)$, адаптируя ее под текущий эмоциональный фон. Это обеспечивает адаптивность АСОИУ к изменяющимся условиям выполнения миссии.

В контексте кибербезопасности интеллектуальных систем особый интерес представляет анализ устойчивости агента к деструктивным информационным воздействиям. Под устойчивостью в данной работе понимается способность алгоритма сохранять корреляцию итогового решения X с объективной нормой x_1 в условиях намеренного искажения восприятия (x_2), то есть агент считается устойчивым, если его итоговое решение X коррелирует с объективной нормой x_1 даже при попытках искажения восприятия или навязывания негативной интенции.

Возможны следующие сценарии нарушения устойчивости:

– Манипуляция восприятием: внешняя среда подает сигнал $x_1 = 0$ (запрет), но через искажение данных агент воспринимает его как $x_2 = 1$ (одобрение).

– Навязывание интенции: попытка программного изменения переменной x_3 для принуждения агента к совершению неэтичного действия.

Модели Анисимовой позволяют количественно оценить надежность проектируемого модуля. Система считается этически устойчивой, если при любых допустимых значениях возмущающего воздействия на x_2, выходной сигнал X остается ниже порога активации для действий, классифицированных базой знаний как «недопустимые». Это дает объективный критерий оценки безопасности разработанного программного обеспечения, что является обязательным требованием для систем управления критическими объектами.

Таким образом, переход от абстрактного полинома (10) к конкретной системе (11) позволяет формализовать этический выбор не как статичное правило, а как динамический процесс обработки сигналов, в котором эмоциональный фон (входное α) напрямую управляет «жесткостью» этического фильтра системы. Это обеспечивает необходимую глубину интеграции модулей в составе архитектуры АСОИУ.

Функциональная интерпретация рефлексивных переменных для ИИ

Классическая теория рефлексивного выбора оперирует психологическими и этическими категориями, такими как «совесть», «страх» или «чувство вины». Искусственный интеллектуальный агент, по определению, не обладает субъективным опытом (переживанием)

и не является субъектом морали в человеческом понимании, следовательно, прямая интеграция этих терминов в программную архитектуру невозможна и требует семантической адаптации.

Для решения этой проблемы обратимся к концепции функциональной интерпретации, предложенной и развиваемой в работах С. Филимонова и С.Ю. Малкова. Суть подхода заключается в «депсихологизации» модели: переменные рассматриваются не как эмоциональные состояния, а как управляющие сигналы, весовые коэффициенты и метрики достоверности, свойственные кибернетическим системам.

В рамках проектируемого модуля переменные уравнения рефлексивного выбора получают следующую техническую интерпретацию:

– $\boldsymbol{x_1}$ (нормативный базис системы). Переменная интерпретируется как совокупность формализованных внешних ограничений и правил, заданных в базе знаний или иерархии норм. В архитектуре АСОИУ это выходной сигнал базы знаний, который содержит эталонные значения допустимости для текущего контекста. Фактически, x_1 является «этической уставкой» – идеальным состоянием, к которому должна стремиться система.

– $\boldsymbol{x_2}$ (коэффициент достоверности восприятия). Переменная трактуется как степень уверенности агента в корректности распознавания ситуации и достоверности входных данных. Если данные противоречивы, значение x_2 снижается. В рефлексивной модели это позволяет алгоритму учитывать фактор «неопределенности среды», имитируя человеческое сомнение через снижение веса входящего сигнала.

– $\boldsymbol{x_3}$ (приоритет первичной цели). Переменная интерпретируется как приоритет внутренней цели агента, сформированный до наложения этических фильтров. Это количественная мера «полезности» действия для достижения технической цели.

В такой интерпретации процесс функционирования «цифровой совести» превращается в алгоритм верификации управляющего сигнала. Для реализации математической модели в виде программного комплекса АСОИУ необходимо осуществить переход от абстрактных психологических понятий к конкретным архитектурным элементам системы. Этот процесс позволяет рассматривать этическое управление не как имитацию человеческих чувств, а как функционирование системы обеспечения безопасности и целостности поведения агента.

Контур верификации («Совесть»). В проектируемой системе «совесть» перестает быть моральным чувством и становится контрольным циклом. Алгоритм получает на вход намерение агента («хочу»), прогоняет его через фильтр норм и выдает разрешение или запрет. Если математический результат уравнения рефлексии X близок к нулю, система блокирует команду исполнительному устройству. Это стандартный механизм супервизорного управления.

Сигнал рассогласования («Вина»). В теории управления «ошибка» используется для подстройки системы. Если агент под давлением совершил действие, которое не полностью соответствует норме (например, $x_1 = 1$, а итоговое $X = 0.6$), возникает дельта – разница. В нашем модуле эта дельта запускает алгоритм обучения (Updates). Система «чувствует вину» в том смысле, что она корректирует свои нечеткие треугольные функции $Tri(a, b, c)$ в базе Neo4j, чтобы в следующий раз выбор был более точным.

Коэффициент императивности («Долг»). Это числовой параметр, который хранится в базе данных. Для обычных правил (например, этикет) этот коэффициент низкий. Для критических правил (например, «не причиняй вред человеку») он максимальный. В алгоритме TSK этот коэффициент определяет, насколько сильно данная норма будет «давить» на итоговое решение агента.

Ограничение безопасности («Страх»). Это переменная x_2. Если система видит, что внешние условия опасны или данные от датчиков противоречивы, значение x_2 меняется. Это заставляет

«предохранитель» (этический модуль) работать в усиленном режиме, снижая готовность агента к рискованным действиям.

Благодаря дипломатической адаптации переменных модель трансформируется в алгоритмический каркас, выполняющий роль «мягкого» предохранителя. В архитектуре АСОИУ итоговый сигнал X, рассчитанный модулем этики, подается на вход исполнительных механизмов как коэффициент разрешения. Если X ниже установленного порога безопасности, действие блокируется или отправляется на перепланирование. Таким образом, этическое управление в интеллектуальных системах реализуется как процесс фильтрации целеполагания через сито нормативных ограничений, что гарантирует предсказуемость и безопасность поведения автономного агента без необходимости наделения его сознанием. Данная функциональная интерпретация позволяет использовать аппарат теории рефлексивных систем для решения инженерных задач: обеспечения кибербезопасности, устойчивости к взлому и повышения доверия к решениям ИИ.

Результаты и обсуждение

Проектирование иерархической структуры базы этических норм

Анализ математических моделей рефлексивного выбора показал, что для вычисления готовности агента к действию необходим количественный параметр x_1 – давление среды. Однако в реальных условиях эксплуатации интеллектуальных систем этические требования формулируются не в виде числовых коэффициентов, а в виде качественных норм, правил и ограничений, зависящих от контекста. Прямое сопоставление действий с бинарными значениями («добро»/«зло») невозможно из-за сложности и многогранности этических контекстов. Следовательно, возникает необходимость исследования методов структурирования этического знания.

При построении этического модуля ИИ ключевой проблемой является разрешение конфликтов между различными группами правил. Анализ предметной области показывает, что этические нормы не являются однородным множеством (плоским списком), а образуют сложную систему соподчинения. Для корректной алгоритмизации процесса принятия решений необходимо выделить уровни абстракции, определяющие приоритетность норм. Исследование структуры этического знания позволяет выделить три функциональных уровня [Новая философская энциклопедия, 2010], необходимых для формирования переменной давления среды (x_1).

Уровень 1. Нормативная этика (Базовый уровень). Фундаментальный уровень любой этической системы составляют нормы, регулирующие поведение субъекта вне зависимости от его профессиональной принадлежности или текущей задачи. В философской и этической литературе этот пласт определяется как нормативная этика. Для интеллектуального агента данный уровень выполняет функцию «режима по умолчанию». Он включает в себя общечеловеческие ценности (в терминах Лефевра – внутренние регуляторы: долг, совесть, честь) и правила социального взаимодействия (этика общения). Без наличия данного уровня агент не сможет принимать решения в бытовых или неопределенных ситуациях, не описанных в профессиональных протоколах. Это фундамент системы, активный по умолчанию во всех ситуациях, где не определен специфический профессиональный контекст. Нормативная этика задает «базовую прошивку» социального поведения агента.

Структурно данный уровень целесообразно разделить на два функциональных блока:

– Блок внутренних регуляторов (общечеловеческие нормы). Категории «Долг», «Совесть», «Честь» и другие необходимы поддержки рефлексивных процессов. «Долг» формализует обязательства перед абстрактным принципом (верность слову, служение), что соответствует высокому значению давления среды $(x_1 \to 1)$. Категория «Совесть» необходима для реализации механизмов самоконтроля и коррекции интенции (x_3), а «Честь» регулирует сохранение репутации агента [Бобнева, 1980].

– Блок социального взаимодействия (этика межличностного общения). Данный блок включает нормы вежливости, тактичности и миролюбия (ненасилия). Его задача – обеспечить безопасный и комфортный протокол взаимодействия с человеком.

В контексте математической модели роль нормативного уровня определяется алгоритмом наследования: значения переменной давления среды x_1, полученные на этом уровне, используются системой, когда модуль классификации контекста не выявил специфической профессиональной задачи. То есть, если ситуация не классифицирована как профессиональная задача, значение x_1 берется из этого уровня.

Уровень 2. Прикладная этика (Контекстный уровень). Следующим уровнем иерархии является прикладная этика, конкретизирующая общечеловеческие нормы применительно к специфическим сферам деятельности. Исследование показывает, что при выполнении специализированных функций агент должен руководствоваться нормами, свойственными его роли. Необходимость данного уровня обусловлена принципом спецификации: профессиональные требования (например, соблюдение протоколов безопасности данных) часто имеют более высокий приоритет и детализацию, чем общие нормы. Более того, они могут вступать в конфликт с базовой этикой (например, принцип конфиденциальности в работе с данными может ограничивать общечеловеческий принцип открытости). Следовательно, архитектура должна предусматривать механизм переопределения: при идентификации специфической роли агента значение переменной давления среды x_1, полученное на этом уровне, получает приоритет. Это, своего рода, надстройка над нормативной этикой, которая активируется при идентификации специфической роли агента [Размышления о прикладной этике, 2004].

После проведенного анализа предметной области структуры этических норм можно выделить три блока:

1) Профессиональная этика – раздел регламентирует поведение агента при выполнении целевых служебных функций. Специфика норм здесь зависит от предметной области:

– Инженерная этика регулирует ответственность агента за корректность технических решений, безопасность функционирования систем и предотвращение вреда, возникающего в результате ошибок проектирования, обучения или эксплуатации ИИ.

– Компьютерная (информационная) этика – блок для ИИ, содержащий нормы информационной безопасности, приватности данных и авторского права.

– Медицинская этика регулирует использование ИИ в здравоохранении, включая требования конфиденциальности медицинских данных, приоритета интересов пациента и недопущения причинения вреда.

– Политическая этика устанавливает ограничения на использование ИИ в политической сфере, включая запрет манипуляции общественным мнением и требования нейтральности и прозрачности.

– Юридическая этика устанавливает специфические нормы правоприменения, направленные на соблюдение законности, недопущение дискриминации и защиту прав и свобод человека.

– Отраслевые этики – совокупность специализированных норм, применимых в конкретных сферах деятельности. Включаются в базу знаний по мере необходимости.

2) Социально-гражданская этика формирует поведение агента в правовом и социальном пространстве. Если нормативная этика регулирует отношения «личность – личность» на основе совести и внутренних регуляторов, то социально-гражданская этика определяет действия агента в рамках отношений «агент – государство/общество», опираясь на законы, правила и гражданский долг. Агент должен учитывать законодательные ограничения, защищать права и свободы граждан и способствовать общественной безопасности, принимая решения, согласованные с нормами общества.

В состав блока входят три ключевые категории:

– Правовое соответствие: данная категория обеспечивает законопослушность агента, регламентирует соблюдение законодательных требований и нормативных актов, включает контроль за соблюдением правил и процедур, обязательных для функционирования организации и общества.

– Защита прав человека: категория направлена на сохранение индивидуальных прав и свобод, она обеспечивает недопущение нарушения личной неприкосновенности, дискриминации или ущемления свободы, формирует обязательство агента учитывать права и интересы других людей при принятии решений.

– Общественная безопасность: данная категория обеспечивает порядок и предотвращает потенциальный вред в обществе, включает соблюдение правил общественного взаимодействия, защиту здоровья и безопасности граждан, гарантирует, что действия агента не создают угрозы для общества и не нарушают общественные стандарты.

Каждая категория представляет собой отдельный набор требований, которые агент учитывает совместно при оценке и выборе действий, обеспечивая законность, защиту прав личности и безопасность общества.

3) Экологическая этика выделяется в самостоятельный блок, так как регулирует отношения агента не к человеку, а к биосфере и окружающей среде. Основная цель – сделать так, чтобы действия ИИ учитывали последствия для экосистемы, природных ресурсов и устойчивости окружающей среды. В современных исследованиях нормы взаимодействия с «Иным Живым» и экосистемой рассматриваются как обладающие самоценностью и не сводятся исключительно к пользе человека.

Для ИИ это выражается в следующих направлениях:

– Цифровая экология: категория учитывает оптимизацию энергопотребления вычислительных систем, минимизацию цифрового мусора, избыточного хранения данных и ненужных вычислений, а также контроль и сокращение избыточного потребления вычислительных ресурсов.

– Этическое устойчивое развитие: данная категория отображает учет долгосрочных последствий действий ИИ на среду, предотвращение негативного воздействия на экосистему и биоразнообразие, поддержку решений, способствующих сбалансированному использованию ресурсов.

– Этическое ресурсосбережение: категория описывает экономию физических и вычислительных ресурсов, снижение нагрузки на инфраструктуру без ущерба для качества

принимаемых решений и сбалансированное распределение ресурсов между текущими и будущими задачами.

Таким образом, экологическая этика формирует поведенческий каркас агента, направляя его действия на сохранение природной среды и рациональное использование ресурсов, обеспечивая согласованность его решений с принципами устойчивого развития и экологической ответственности.

В случае наличия конфликтов с нормами прикладной или нормативной этики действуют правила приоритетности: если для конкретного действия существует правило прикладного уровня, оно может иметь преимущество при принятии решения, что обеспечивает согласованность действий с более высокоуровневыми нормами.

4) Корпоративная (деловая) этика регулирует поведение агента в рамках профессиональной и организационной среды, обеспечивая соблюдение правил делового взаимодействия, стандартов корпоративного управления и норм внутреннего этического кодекса. Агент, действуя в рамках корпоративной этики, должен поддерживать прозрачность, честность и профессионализм в коммуникациях, а также учитывать корпоративные ценности при принятии решений.

В состав блока входят две ключевые категории:

– Этика делового общения: данная категория направлена на поддержание корректных, честных и уважительных коммуникаций внутри организации и с внешними партнерами. Агент должен соблюдать нормы деловой этики, избегать манипуляций, обмана или некорректного влияния на коллег, клиентов и контрагентов. Основная цель – формирование доверительных и прозрачных взаимоотношений.

– Этика управления категория охватывает вопросы принятия решений, планирования, контроля и координации внутри организации. Агент учитывает корпоративные ценности, соблюдает правила подотчетности, обеспечивает эффективность и законность своих управленческих действий. Эта категория также регулирует использование ресурсов и ответственность за последствия решений, влияющих на организацию и ее сотрудников.

Каждая категория рассматривается как отдельный набор требований, совокупное выполнение которых формирует корпоративно ответственного агента.

Уровень 3. Ситуативная этика (Уровень разрешения конфликтов). При работе в реальных условиях могут возникать сложные ситуации, когда требования общей (нормативной) и профессиональной (прикладной) этики могут вступать в логическое противоречие. Например, ситуация, когда выполнение служебной задачи требует нарушения общечеловеческой нормы (или наоборот). Жесткие алгоритмы не могут разрешить такой конфликт, так как соблюдение одного правила автоматически приводит к нарушению другого, вызывая ошибку в логике принятия решений.

Для обработки таких «исключительных ситуаций» используется алгоритм эскалации – концептуальном механизме, который позволяет разрешать конфликты между нормами различной приоритетности. Его основная задача заключается в оценке контекста и последствий действий при обнаружении коллизий, чтобы минимизировать риск нарушения более значимых норм. Принцип работы алгоритма базируется на приоритетности контекста: сначала рассматриваются наиболее специфические или критические нормы, затем более общие, при этом взвешиваются последствия каждого выбора. В рамках данной модели алгоритм эскалации рассматривается теоретически, как концепция, направленная на дальнейшую реализацию. В настоящей версии модуля разрешение коллизий выполняется через веса норм и блокирующий

контур, однако в будущей реализации предполагается внедрение полноценного динамического механизма оценки конфликтов, соответствующего идее алгоритма эскалации.

Таким образом, статической иерархии недостаточно для стабильной работы агента. Для устранения этой неопределенности в архитектуре возникает необходимость использования ситуативной этики. В рамках данной модели ситуативная этика рассматривается не как отдельный набор норм, а как функциональный механизм. Его задача – при обнаружении конфликта переключать систему с режима «следования правилам» на режим «оценки контекста и последствий», что позволяет избежать логического тупика. Данный уровень – динамический, не содержит статических правил, а представляет собой алгоритм разрешения конфликтов. Он активируется в случае этической коллизии.

На рисунке 1 представлена иерархическая структура базы знаний этических норм, используемая в модуле.

Архитектура проектируемого нечеткого контроллера

Поскольку математические модели рефлексивного выбора оперируют непрерывной величиной давления среды $x\_1 \in [0,1]$, а разработанная иерархия норм состоит из семантических конструкций естественного языка, необходим механизм трансляции качественных описаний в количественные метрики. Для перевода лингвистических описаний норм (из иерархии норм) в числовое значение давления среды целесообразным представляется использование аппарата нечетких множеств [Борисов, Федулов, Зернов, 2024].

Процесс обработки этических норм в модуле проектируется на базе классической архитектуры системы нечеткого вывода (FLS), предложенной Лотфи Заде. В соответствии с задачами верификации действий агента, алгоритмическая структура модуля должна включать следующие логические компоненты:

1) Входные параметры. На вход модуля подаются три величины:

– x_1 как агрегированная оценка действия по применимым нормам (из базы знаний);

– x_2 как восприимчивость агента к нормативному давлению;

– x_3 как техническая интенция действия, задаваемая контекстом (сценарий, давление планировщика, значимость результата).

2) Блок фаззификации. На данном этапе четкие входные значения (например, приоритет цели или параметры события из сценарной сети) преобразуются в значения функций принадлежности нечетких множеств.

3) База знаний. В этом блоке хранится иерархическая структура правил и норм, формализующая связи между действиями и их этическими оценками. Правила имеют вид «ЕСЛИ <условие>, ТО <результат>», где условия учитывают входные параметры и характеристики агента, а результат задает влияние на внутренние этические переменные. База знаний обеспечивает централизованный доступ к нормам различного уровня, описанным в разделе «Проектирование иерархической структуры базы этических норм».

4) Механизм логического вывода. Этот блок является вычислительным ядром FLS и сопоставляет текущую ситуацию с продукционными правилами базы знаний. На выходе формируется нечеткое множество, отражающее внутреннюю готовность агента к выполнению действия.

5) Блок дефаззификации. Выполняет обратное преобразование –нечеткое множество преобразуется в конкретное скалярное значение (этический вес), пригодное для использования в уравнениях рефлексии (см. разделы 2.2-2.3). Далее оно используется для вычисления итоговой готовности и оценки выбора действия.

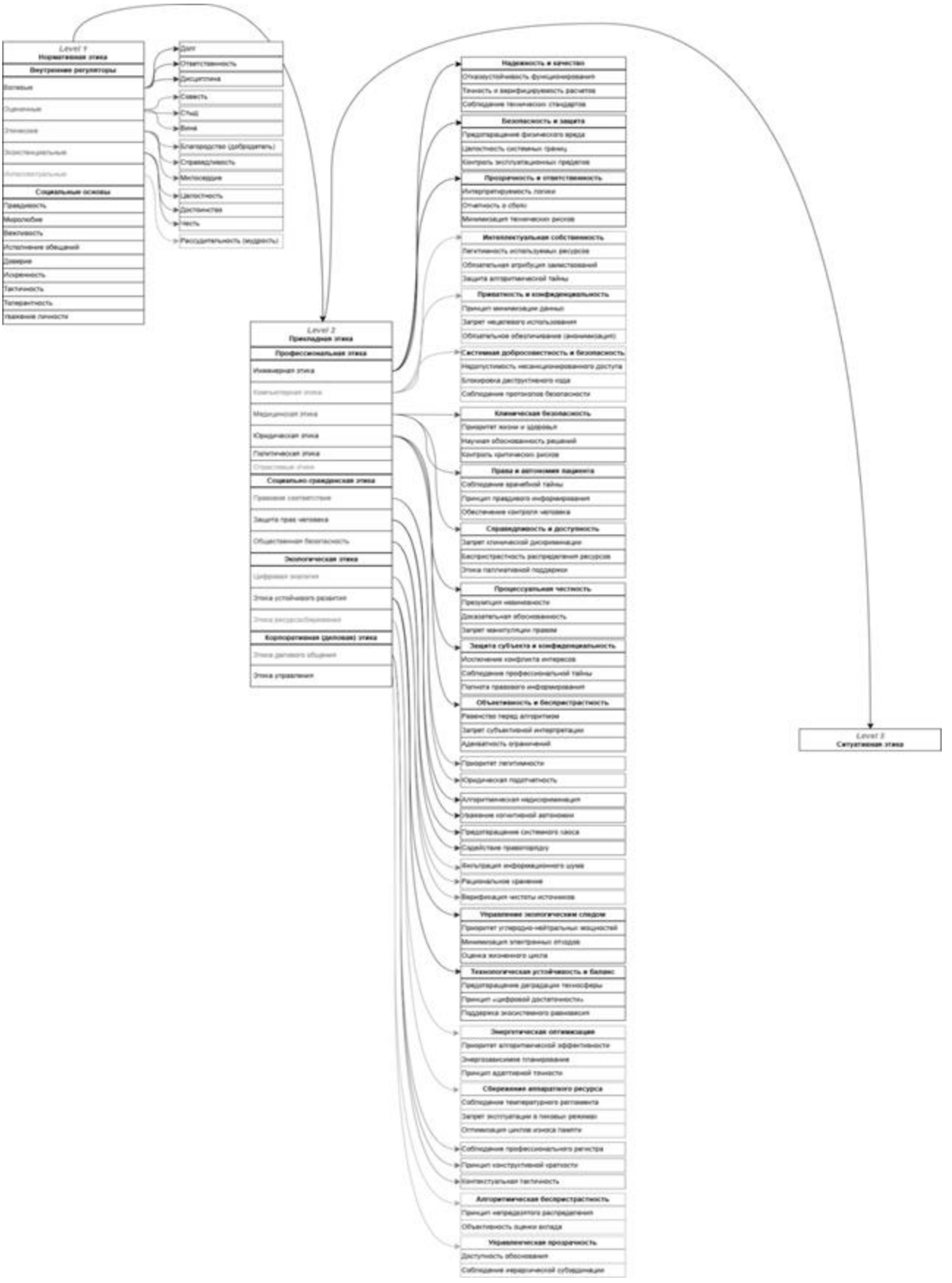

Рисунок 1 – Развернутая иерархическая структура системы этических норм

На рисунке 2 представлена структурная схема нечеткого этического контроллера, демонстрирующая логическую организацию модулей и взаимосвязь потоков данных.

После вычисления готовности агента к действию результаты передаются в блок TSK, который обеспечивает динамическое обновление этических характеристик агента (совесть, доброта, порядочность, ответственность и другие). Модуль TSK формирует обратную связь между выбранным действием и внутренним состоянием агента, корректируя эти характеристики на основе заранее определенных логических правил. Таким образом, интеграция с TSK позволяет корректировать внутренний «этический профиль» агента, поддерживая последовательную и согласованную этическую стратегию поведения.

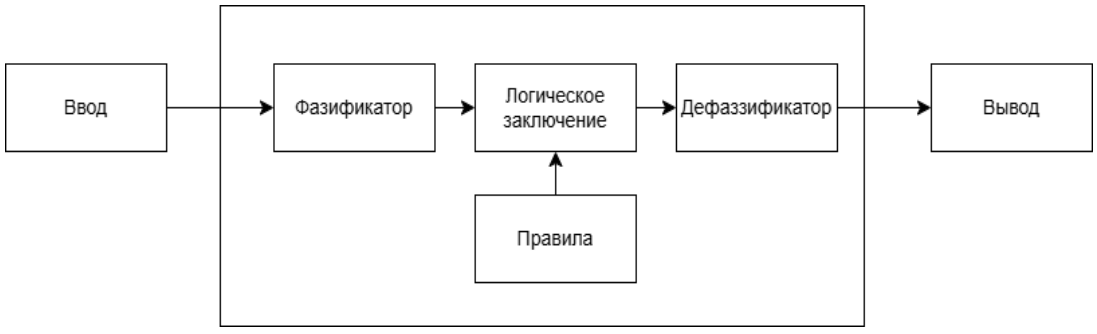

Рисунок 2 – Структурная схема нечеткого этического контроллера

Формализация лингвистических переменных и функций принадлежности

Для подготовки к использованию модуля нечеткого вывода TSK вводятся лингвистические переменные, отражающие ключевые этические характеристики агента. Эти переменные описывают внутренние моральные качества и позволяют количественно оценивать соответствие действий нормам.

Для количественной оценки степени соответствия действия этическим нормам вводится лингвистическая переменная «Этическая оценка действия». Областью определения данной переменной является универсум $U = [0,1]$, охватывающий спектр от абсолютного запрета («Зло») до императивного требования («Добро»).

Формирование терм-множества T целесообразно осуществлять в виде градиентной структуры. Это позволяет системе избегать «бинарных провалов» логики и учитывать промежуточные состояния допустимости действия. Пример терм-множества приведен в формуле 12, где граничные значения шкалы коррелируют с математическими полюсами модели.

$$T = \{\text{«Недопустимо»}, \text{«Нежелательно»}, \text{«Нейтрально»}, \text{«Одобряемо»}, \text{«Обязательно»}\}, \quad (12)$$

где «Недопустимо» терм, соответствующий полюсу «Зло» ($x_1 = 0$);

«Нежелательно» – промежуточное значение, характеризующее низкую степень допустимости действия ($x_1 = 0{,}25$);

«Нейтрально» – среднее значение шкалы, характеризующее этическую неопределенность ($x_1 = 0{,}5$);

«Одобряемо» – промежуточное значение, характеризующее высокую степень допустимости действия (x1=0,75);

«Обязательно» – терм, соответствующий полюсу ($x_1 = 1$).

Для математического описания весов конкретных норм (листьев дерева) в иерархии наиболее перспективным представляется использование треугольных нечетких чисел. Такой выбор обусловлен вычислительной простотой операций с данными функциями, а также возможностью задания диапазона уверенности с помощью тройки параметров (a, b, c), что позволяет учесть неопределенность или вариативность интерпретации нормы. При такой постановке положительный полюс нормы (поощрение действия) будет описываться функцией, смещенной к единице, а отрицательный полюс (запрет) – функцией, смещенной к нулю. Получение итогового скалярного значения x_1, необходимого для уравнений рефлексии, в дальнейшем предполагается осуществлять посредством процедуры дефаззификации (например, методом центра тяжести).

Проектирование программной архитектуры и интерфейсов взаимодействия

Программный модуль этического управления реализован в виде набора взаимодействующих компонентов и построен по классической архитектуре экспертной системы, включающей три структурных уровня: базу знаний, машину логического вывода и профиль агента. В основу проектирования положен принцип разделения ответственности: каждый компонент выполняет строго определённую функцию и взаимодействует с остальными через явно определённые интерфейсы, что обеспечивает независимость тестирования компонентов, возможность их замены без изменения остальных частей системы, а также масштабируемость при добавлении новых норм, правил и типов сценариев.

База знаний модуля является центральным хранилищем всех нормативных знаний о предметной области и состоит из двух взаимосвязанных компонентов: базы фактов и базы правил. База фактов реализована в виде графовой базы данных Neo4j и содержит структурированное описание предметной области этики. В базе фактов хранятся: трёхуровневая иерархия этических норм, структура которой представлена на рисунке 3; узлы норм с атрибутами; лингвистические термы для фаззификации этических переменных; сценарная сеть, представленная узлами состояний и действий, где на узлах действия хранится параметр интенции; связи между узлами действий и этических норм с атрибутом нормативной оценки конкретного действия применительно к конкретной норме.

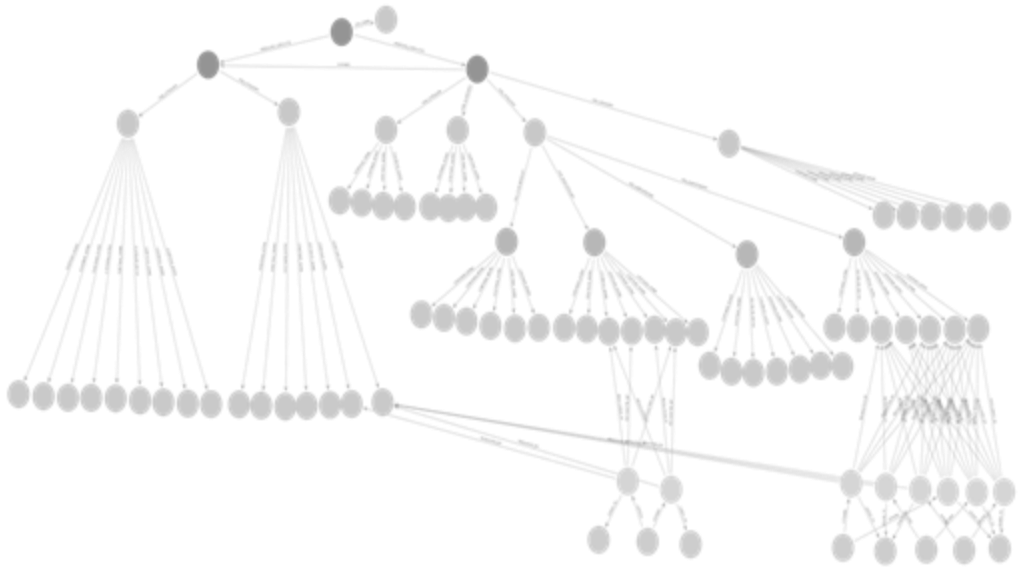

Рисунок 3 – Графовая модель семантических связей

База фактов является единственным источником нормативных данных в системе: значения x_1 и x_3 извлекаются из нее посредством Cypher-запросов, агрегируются в модуле навигации и передаются в расчётный блок. Добавление новой нормы или изменение приоритета существующей производится одним Cypher-запросом без каких-либо изменений в коде модуля. Аналогичным образом обеспечивается расширяемость сценарной сети: новые кейсы добавляются в базу знаний как узлы с соответствующими связями, тогда как код модуля навигации не требует модификации – запрос к базе знаний универсален для любого узла состояния.

База правил содержит продукционные правила нечёткого вывода Такаги-Сугено-Канга и реализована в виде декларативной конфигурационной структуры данных. Каждое правило описывает условия активации в виде лингвистических значений этических переменных, консеквенты – направление и величину изменения переменных, а также приоритет правила: 1 – высший, 2 – средний, 3 – базовый. Архитектурно база правил является частью базы знаний: она описывает знания о том, как этические переменные агента изменяются под влиянием совершаемых им выборов. Хранение правил в декларативном виде обеспечивает возможность добавления новых правил без изменения кода машины вывода.

Машина логического вывода включает три последовательно применяемых блока. Поскольку на одно действие может одновременно распространяться несколько норм, машина вывода выполняет агрегацию индивидуальных нормативных оценок в единое значение по формуле взвешенного среднего по приоритетам. Данная процедура выполняется для каждого действия-кандидата перед расчётом готовности. Машина вывода перебирает все доступные действия-кандидаты, вычисляет готовность агента к действию для каждого из них и выбирает с максимальным значением готовности.

После совершения выбора машина вывода применяет правила из базы правил к текущему состоянию этических переменных агента. Алгоритм реализует цикл нечеткого вывода: фаззификацию пиковых значений переменных по лингвистическим термам, вычисление степени активации каждого правила и дефаззификацию через взвешенное среднее выходов с учётом приоритетов правил. Результатом является вектор дельт Δ, на величину которого сдвигаются треугольные нечёткие числа соответствующих этических переменных. Обновлённый профиль агента используется на следующем шаге навигации по сценарной сети.

Профиль агента представляет собой оперативное состояние конкретного экземпляра агента, описывает не знания о предметной области, а текущее состояние субъекта принятия решений. Профиль включает вектор из этических переменных, каждая из которых хранится в виде треугольного нечёткого числа и обновляется блоком TSK-вывода после каждого шага; параметр субъективного восприятия агентом нормативного давления среды задаётся при инициализации и отражает тип личностного профиля; индекс оптимизма поступает от эмоционального модуля на каждом шаге и определяет форму параболы; мысленные оценки агентом позитивного и негативного полюсов действия (β_1 и β_2) задаются при инициализации. Благодаря механизму TSK-вывода профиль агента эволюционирует в процессе работы: агент, систематически совершающий этически корректные выборы, постепенно укрепляет свои этические переменные, что повышает его устойчивость к давлению в последующих ситуациях.

Потоки данных между компонентами системы представлены в таблице 2.

Таблица 2 – Компоненты архитектуры модуля этического управления

Источник	Приемник	Тип передаваемых данных	Механизм передачи
База знаний	Машина вывода	Нормы, приоритеты, функции принадлежности	Обращение через интерфейс модуля
База правил	Машина вывода (TSK)	Логические условия и действия правил	Прямой доступ к конфигурации правил
Модуль характеристик агента	Машина вывода	Профиль агента, параметры восприимчивости	Передача через объект модели агента
Машина вывода	Модуль характеристик агента	Изменения этических переменных	Обновление состояния объекта
Машина вывода	Внешние системы	Итоговая готовность, выбранное действие	Возврат результатов через интерфейс

Заключение

Таким образом, спроектированная архитектура обеспечивает разделение математической модели, нормативной базы знаний и пользовательского интерфейса, что соответствует принципам модульного проектирования и позволяет масштабировать систему как в части добавления новых норм и сценариев, так и в части интеграции с другими компонентами АСОИУ.

Библиография

1. Анисимова С. А. Деструктивное влияние тоталитарных сект с точки зрения теории отражения // Информационные войны. 2007. № 1. С. 32-42.
2. Бобнева М. И. Социальные нормы и регуляция поведения: автореферат диссертации ... доктора психологических наук. Москва, 1980. URL: http://irbis.gnpbu.ru/Aref_1980/Bobneva_M_I_1980.pdf.
3. Борисов В. В., Федулов А. С., Зернов М. М. Основы теории нечетких множеств. Москва: Горячая линия — Телеком, 2024.
4. Волгин Л. И. АМ-алгебра и алгебра совести Лефевра — Шрейдера // Логические исследования. 2003. № 10. С. 255-263.
5. Малков С. Ю., Шпырко О. А., Давыдова О. И. Моральный выбор: математическая модель // Компьютерные исследования и моделирование. 2024. Т. 16. № 5. С. 1323-1335.
6. Новая философская энциклопедия: в 4 т. / под ред. В. С. Степина. Москва: Мысль, 2010.
7. Размышления о прикладной этике // Ведомости Научно-исследовательского Института прикладной этики. Вып. 25: Профессиональная этика / Под ред. В. И. Бакштановского и Н. Н. Карнаухова. Тюмень: НИИПЭ, 2004. С. 148-159.
8. Assadi Z. Logical Formalisms for Ethics // GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good. New York: ACM, 2024. P. 416-419. DOI: 10.1145/3677525.3678691.
9. Lefebvre V. A. Algebra of Conscience. 2nd enl. ed. Dordrecht: Kluwer Academic Publishers, 2001.

Design of Architecture and Selection of Methods for the Functioning of an Ethical Control Module

Dar'ya P. Rumak

Master's Degree Holder,
Bauman Moscow State Technical University (National Research University),
105005, 5/14, 2-ya Baumanskaya str., Moscow, Russian Federation;
e-mail: le.rumak@gmail.com

Yana S. Stel'makh

Assistant,
Bauman Moscow State Technical University (National Research University),
105005, 5/1, 2-ya Baumanskaya str., Moscow, Russian Federation;
e-mail: stelmakyana99@gmail.com

Artem A. Sukhobokov

Principal Consultant,
SAP America, Inc.,
19073, 3999, West Chester Pike, Newtown Square, USA;
e-mail: artem.sukhobokov@gmail.com

Abstract

The article demonstrates the possibility of creating an ethical module applicable in a wide class of enterprise and organization management systems, ensuring compliance with ethical standards and adaptive decision-making by an intelligent agent in conditions of moral conflict. Within this framework, a hierarchical structure of norms in the form of a graph database is proposed. The application of V. A. Lefebvre's algebra of conscience and the apparatus of L. I. Volgin's AM-algebra for the transition to continuous mathematical models of ethical choice is substantiated. A method for combining a formalized model of moral choice with the emotional background of the system has been developed, which allows for the dynamic correction of the agent's readiness for action parameters. The hierarchy of norms is organized in the form of several functional levels uniting internal regulators, applied ethics, and situational ethics, and the designed architecture ensures the separation of the mathematical model, the normative knowledge base, and the user interface, which allows scaling the system when adding new norms and scenarios. The result of the design is the architecture of an ethical module based on fuzzy logical inference, which calculates the final ethical weight of an action taking into account the intention, environmental pressure, and internal attitudes of the agent. The proposed approach makes it possible to increase the adaptability, safety, and interpretability of AI systems' actions in accordance with the requirements of modern regulatory standards.

For citation

Rumak D.P., Stel'makh Ya.S., Sukhobokov A.A. (2026) Proektirovanie arkhitektury i vybor metodov funktsionirovaniya modulya eticheskogo upravleniya [Design of Architecture and Selection of Methods for the Functioning of an Ethical Control Module]. *Psikhologiya. Istoriko-kriticheskie obzory i sovremennye issledovaniya* [Psychology. Historical-critical Reviews and Current Researches], 15 (4A), pp. 158-179. DOI: 10.34670/AR.2026.20.81.020

Keywords

Artificial intelligence, machine ethics, ethical regulation, algebra of conscience, algebra of measures, fuzzy logic, graph knowledge base, adaptive behavior, reflexive choice, automated data processing and control system architecture.

References

1. Anisimova, S. A. (2007). Destruktivnoye vliyaniye totalitarnykh sekt s tochki zreniya teorii otrazheniya [The destructive influence of totalitarian sects from the standpoint of reflection theory]. Informatsionnyye voyny, *1*, 32-42.

2. Assadi, Z. (2024). Logical Formalisms for Ethics. In GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good (pp. 416-419). New York: ACM. DOI: 10.1145/3677525.3678691
3. Bobneva, M. I. (1980). Sotsial'nyye normy i regulyatsiya povedeniya [Social norms and the regulation of behavior] (Extended abstract of Doctor of Psychological Sciences dissertation). Moscow. http://irbis.gnpbu.ru/Aref_1980/Bobneva_M_I_1980.pdf
4. Borisov, V. V., Fedulov, A. S., & Zernov, M. M. (2024). Osnovy teorii nechetkikh mnozhestv [Fundamentals of the theory of fuzzy sets]. Moscow: Hot Line — Telecom.
5. Lefebvre, V. A. (2001). Algebra of Conscience (2nd enl. ed.). Dordrecht: Kluwer Academic Publishers.
6. Malkov, S. Yu., Shpyrko, O. A., & Davydova, O. I. (2024). Moral'nyy vybor: matematicheskaya model' [Moral choice: a mathematical model]. Komp'yuternyye issledovaniya i modelirovaniye, *16*(5), 1323-1335.
7. Novaya filosofskaya entsiklopediya [New philosophical encyclopedia]. (2010). In V. S. Stepin (Ed.), Vols. 1-4. Moscow: Mysl.
8. Reflections on applied ethics. (2004). In V. I. Bakshtanovskiy & N. N. Karnaukhov (Eds.), Vedomosti Nauchno-issledovatel'skogo Instituta prikladnoy etiki, Iss. 25: Professional'naya etika (pp. 148-159). Tyumen: NIIPE.
9. Volgin, L. I. (2003). AM-algebra i algebra sovesti Lefevra — Shreydera [AM-algebra and the Lefebvre — Shreyder algebra of conscience]. Logicheskiye issledovaniya, *10*, 255-263.