

УДК 004.8

DOI: 10.34670/AR.2026.21.25.021

Метод одновременного использования этических и эмоциональных моделей для выбора действий интеллектуального агента

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: stelmak yana99@gmail.com

Аннотация

Модули этического и эмоционального управления влияют на машинное сознание, определяющее поведение интеллектуального агента: этика задаёт нормы поведения, а эмоции адаптивно настраивают подсистемы агента под текущие события. В работе оба модуля реализованы на основе нечётких переменных и алгоритма нечёткого вывода Такаги–Сугено–Канга. Эмоциональная модель опирается на колесо эмоций Женевы (GEW 3.0) с двадцатью эмоциями, а этическая — на деонтологическую иерархию норм, листья которой объединяют конкретные нормы по аналогии с группировкой эмоций. Внешняя среда представлена сетью сценариев в виде ориентированного графа, вершины которого соответствуют событиям, а рёбра — возможным действиям агента. Предложен механизм совместного применения этического и эмоционального управления для относительно простых интеллектуальных систем: агент перемещается по сети сценариев, преодолевая активационные барьеры рёбер и выбирая переход по суммарному отклонению характеристик. Метод проиллюстрирован примером с несколькими профилями поведения агента. Работоспособность подхода продемонстрирована на примере ИИ-помощника кредитного аналитика банка, формирующего отчёт о доверии клиенту при выдаче кредита; реализация выполнена на языке Python с графовой базой данных Neo4j, модульной архитектурой и интерактивной веб-визуализацией.

Для цитирования в научных исследованиях

Стельмах Я.С. Метод одновременного использования этических и эмоциональных моделей для выбора действий интеллектуального агента // Психология. Историко-критические обзоры и современные исследования. 2026. Т. 15. № 4А. С. 180-195. DOI: 10.34670/AR.2026.21.25.021

Ключевые слова

Этическое управление, эмоциональное управление, сеть сценариев, алгоритм нечёткого вывода Такаги–Сугено–Канга, интеллектуальный агент, кредитный скоринг, машинная этика, модели эмоций.

Введение

Модули этического и эмоционального управления – неотъемлемые компоненты продвинутых систем ИИ [Sukhobokov et al., 2024]. Этика обеспечивает соблюдение моральных норм, минимизируя риски и конфликты [Shadbolt, Hampson, 2024], а эмоции влияют на другие подсистемы, определяя поведение агента [Franzoni et al., 2019]. Их результаты тесно взаимосвязаны и совместно формируют поведение агента [Ojha, Vitale, Williams, 2021], при этом этика и культурные традиции влияют на специфику индивидуальных эмоций [Mesquita, 2022].

В полностью функциональной системе уровня AGI модули этики и эмоций должны работать в тесной интеграции с другими компонентами – познанием, сознанием, целями, мировоззрением, социальным взаимодействием и др. [Sukhobokov et al., 2024], совместно формируя общее поведение агента. Несмотря на наличие отдельных разработок [Olkhovenko, 2025; Rorov et al., 2022], комплексной системы, объединяющей этику и эмоции, пока не существует.

В данной работе предлагается модель агента и внешней среды, поведение которого определяется встроенными моделями этики и эмоций. Среда моделируется сетью сценариев с переходами между событиями, обладающими этическими и эмоциональными характеристиками; по мере навигации агента его этические оценки и эмоциональное состояние изменяются и учитываются правилами перехода. Схема позволяет экспериментировать с начальным состоянием агента, моделями управления, структурой сети и правилами переходов, обеспечивая настройку реалистичных моделей для будущих систем ИИ.

Актуальность работы обусловлена растущим интересом к ИИ, способному учитывать человеческие ценности, воспринимать эмоции других и управлять своим поведением через эмоции. Системы ИИ уже принимают решения, существенно влияющие на жизни людей, и используются в критически важных отраслях – производстве, транспорте, здравоохранении, финансах, образовании и ИТ, – что требует создания социально ответственных, дружелюбных и безопасных интеллектуальных систем за счёт включения модулей этического и эмоционального управления.

Материалы и методы исследования

После анализа моделей эмоций в интеллектуальных системах [Kowalczyk, Czubenko, 2016; Ojha, Vitale, Williams, 2021] мы сосредоточились на моделях на основе нечёткой логики, поскольку они учитывают неопределённость и интенсивность эмоций. Среди таких работ [Ayes, Blewitt, 2015; Ayes, Stokes, Edwards, 2007; El-Nasr, Yen, Ioerger, 2000; Gershenson, 1999; Mandryk, Atkins, 2007; Qiao et al., 2013; Scherer, 2005] особое внимание привлекает модель Geneva Emotion Wheel (GEW), предложенная K.R. Scherer [Scherer, 2005].

GEW широко используется для оценки эмоциональных реакций людей [McGinn, Kelly, 2018; Sacharin, Schlegel, Scherer, 2012; Schlegel, Scherer, 2016; Turumugon et al., 2019]. Изначально модель была построена на лингвистических переменных [Scherer, 2005], что позже стало основой для интеграции нечёткого логического вывода [Ayes, Blewitt, 2015]. Эта модель считается более универсальной, однако использует алгоритм Мамдани [Mamdani, 1977], где правила структурированы следующим образом:

$$\text{IF } x_1 \text{ is } A_1^i \text{ AND ... AND } x_n \text{ is } A_n^i \text{ THEN } y \text{ is } B_i, i = \overline{(1, N)}, \quad (1)$$

где: $(x_1, \dots, x_n)$ – это входы как лингвистические переменные;

$A_1^i, \dots, A_n^i$ – значения соответствующих лингвистических переменных;

y – выход как лингвистическая переменная;

B_i – её значение.

Используемый метод дефаззификации – «центр тяжести». При этом подходе итоговое эмоциональное состояние не изменяется при повторном воздействии того же уровня E ($E \in \overline{0,1}$), хотя на практике эмоции склонны накапливаться при коротких интервалах воздействия и ослабевать при многократном повторении стимула из-за утомления. Влияние эмоциональной памяти [Crowder, Friess, 2010; Kowalczyk, Czubenko, 2011] и порядка воздействий требует дальнейшего изучения.

Актуальная версия модели – GEW 3.0, включающая 20 эмоций [Scherer et al., 2013; Université de Genève].

После анализа шести основных алгоритмов нечёткого вывода [Borisov, Fedulov, Zernov, 2023] – Мамдани, Ларсена, Цукamoto, Сугено нулевого порядка, Такаги-Сугено и модели с идентификацией операций над нечёткими множествами – был выбран алгоритм Такаги-Сугено-Канга (TSK) [Sugeno, Kang, 1988; Takagi, Sugeno, 1985] благодаря его универсальности. Правила вывода в нём структурированы так:

$$IF\ x_1\ is\ A_1^i\ AND\ \dots\ AND\ x_n\ is\ A_n^i\ THEN\ y = f_i(x_1 + \dots + x_n), \quad (2)$$

где: $(x_1, \dots, x_n)$ – это входы, определенные как нечёткие переменные;

$A_1^i, \dots, A_n^i$ – это значения соответствующих нечётких переменных;

y – выход как нечёткая переменная;

f_i – функция, имеющая одинаковую структуру для всех n правил и отличающаяся только параметрами.

Чаще всего используется TSK первого порядка с линейными функциями. Дальнейший анализ публикаций по эмоциональному управлению и психологии позволит точнее определить вид результирующей функции в реализуемой версии алгоритма. Метод дефаззификации – «нечёткое среднее значение» (fuzzy mean value).

Выход модели эмоционального управления после обработки очередного события определяет текущее эмоциональное состояние агента и представлен вектором из 20 переменных, соответствующих эмоциям Geneva Emotion Wheel.

Для модели этического управления выбран простой подход [Bendel, 2014], основанный на деонтологической этике в форме системы правил. Модель агент-центрична [Tolmeijer et al., 2020]: все действия рассматриваются с точки зрения самого агента, а набор правил определяет, какие действия этически одобрены, а какие – нет. Поскольку реальное поведение допускает отклонения, для интерпретации правил используется нечёткий логический вывод. Применение нечёткой логики в машинной этике развивалось от первых работ с теорией нечётких множеств [Sacconi, 1999] через подходы к нечёткой этике интеллектуальных систем [Assadi, 2024; Assadi, Inverardi, 2024] до формального метода описания этических решений в виде нечётких правил на основе оценки риска [Dyoub, Lisi, 2025].

Набор этических норм представлен в виде иерархии: на первом уровне – небольшое число обобщённых тем, каждая из которых детализируется на последующих уровнях (Рисунок 1, по материалам [Этика; Chadwick, 2012]).

Листья иерархии представляют группы конкретных норм, организованных по аналогии с Geneva Emotion Wheel: каждая норма с положительной оценкой парно сопоставлена с нормой

отрицательной оценки. Этические нормы, как и эмоции, нечётки и могут соблюдаться с разной строгостью, что делает нечёткую логику подходящим инструментом. Если норма поощряет действие, положительная ось приближается к единице, а отрицательная – к нулю, и наоборот для осуждаемых действий. Максимальное число уровней иерархии на данный момент – 7, а листья содержат конкретные правила нечёткого вывода TSK (пример нижнего уровня – в Таблице 1).

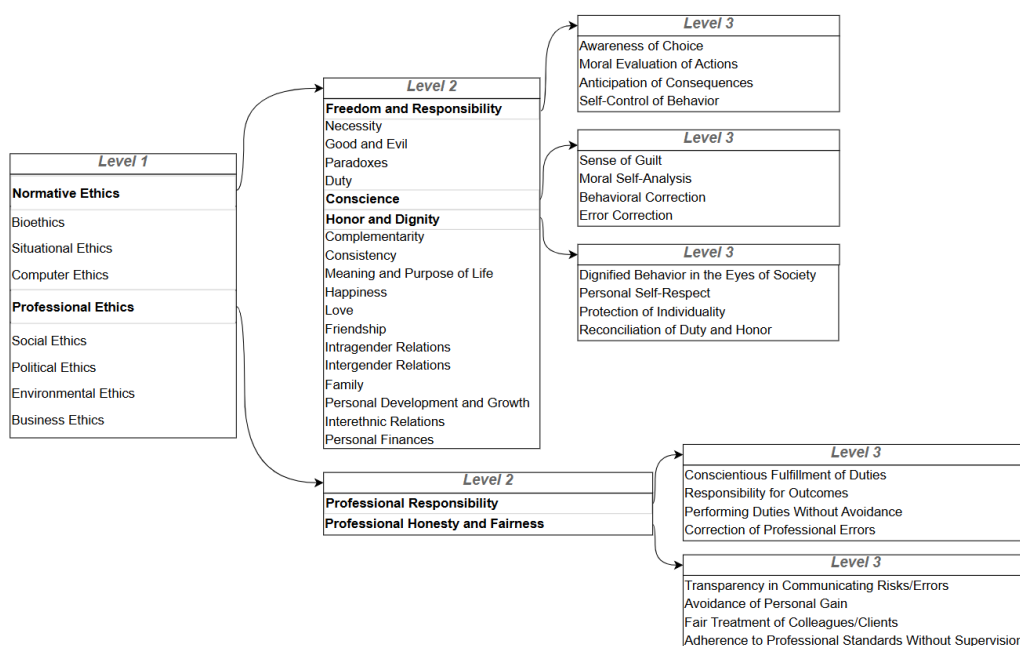


Рисунок 1 – Начало иерархии этических норм

Таблица 1 – Нижний уровень иерархии этических норм: честь и достоинство

Норма	Положительная ось	Значение	Отрицательная ось	Значение
Правдивость. Говорить правду в трудной ситуации (одобряемое действие)	Соблюдение личной чести (например, честность в словах)	0.8-0.95	Не честное поведение (например, ложь или обман)	0.5-0.2
Честность в действиях. Лгать ради личной выгоды (неодобряемое действие)	Честность в поведении	0.05-0.2	Не честное поведение	0.8-0.95
Уважение. Поддерживать коллегу в споре без унижения (одобряемое действие)	Уважение достоинства других (например, проявление уважения к человеку)	0.7-0.9	Унижение достоинства других (например, оскорбления)	0.1-0.3
Унижение. Оскорблять кого-либо в конфликте (неодобряемое действие)	Уважение достоинства других	0.1-0.3	Унижение достоинства других	0.7-0.9
Целостность. Отказаться от взятки несмотря на давление (одобряемое действие)	Непреклонность в защите принципов	0.7-0.85	Компромисс с принципами ради выгоды	0.15-0.3
Жадность. Согласиться на неэтичное предложение ради выгоды (неодобряемое действие)	Непреклонность в защите принципов	0.15-0.3	Компромисс с принципами ради выгоды	0.7-0.85

Норма	Положительная ось	Значение	Отрицательная ось	Значение
Приличие. Вежливое общение на публичном мероприятии (одобряемое действие)	Достойное поведение в обществе (например, соблюдение норм приличия)	0.85-0.95	Недостойное поведение (например, публичное нарушение приличия)	0.5-0.15
Неприличность. Грубое поведение в общественном месте (неодобряемое действие)	Достойное поведение в обществе	0.5-0.15	Недостойное поведение в обществе	0.85-0.95

Как и в GEW, одно событие может одновременно подпадать под действие разных этических норм, вызывая реакции различной интенсивности по каждой из них.

Предлагаемая модель отражает обобщённую человеческую этику, хотя иерархия и веса норм индивидуальны. Тем не менее сам метод организации норм универсален и применим к этическим моделям ИИ-систем, которые могут существенно отличаться от человеческих.

Этическая модель агента меняется медленнее эмоциональной, но также подвержена постепенным изменениям (быстрее – для ситуативной или компьютерной этики). Каждая ветвь иерархии имеет числовой модификатор, влияющий на скорость изменения нечётких переменных нижних уровней.

Для нечёткого вывода в модели этического управления, как и в эмоциональной, используется алгоритм Такаги–Сугено–Канга. Его выход после обработки события определяет текущее этическое состояние агента в виде полной иерархии нечётких переменных. Модель реализована на Python с использованием графовой базы данных Neo4j [Gupta, 2015].

Сценарная сеть представлена как ориентированный граф $G = (V, E)$, где V – множество узлов (событий), а E – множество рёбер (действий). Узлы изменяют эмоциональное и этическое состояния агента, рёбра определяют возможные переходы между событиями с соответствующими условиями.

Каждое событие v_i включает: содержательное описание; вектор из 20 нечётких переменных эмоций, вызываемых событием; иерархические имена переменных этической модели, подлежащих изменению (с модификаторами для расчёта новых значений), добавлению или удалению.

Каждое ребро e_{ij} (от узла i к узлу j) включает: содержательное описание и наборы условий ($=, >=, <=, >, <$) по переменным эмоциональной и этической моделей. Переход возможен при выполнении всех условий, решение принимает агент. Сеть реализована в Neo4j с использованием Cypher [Gupta, 2015].

Теоретическое решение. Модели формально описывают акты выбора между намеренным желанием и этическими ограничениями; все переменные нечёткие, лежат в $\overline{0,1}$. Наиболее продвинутой является квадратично-линейная модель:

$$m(x, y) = ax^2 + bx + c + dxy + ey, \quad (3)$$

где: $m(x, y)$ – это функция выбора агента в будущей ситуации;

x – это частота выбора положительного полюса субъектом;

y – это ожидаемая частота выбора партнёром;

$a, b, c, d, e \in \mathbb{R}$ – коэффициенты совместного влияния этики и эмоций.

Асимметрия (квадратичная по x , линейная по y) отражает более тонкую оценку субъектом собственных действий по сравнению с действиями партнёра.

Наиболее точным методом совместного учёта эмоций и этики была бы алгебра совести Лефевра [Lefebvre, 2001].

Входные параметры:

– Эмоциональное состояние:

$\alpha \in [0; 1]$ – индекс оптимизма, связанный с интегральной оценкой (x_{emo}) из GEW-модели через переход из $\overline{-1, 1}$ в $\overline{0, 1}$ может быть выполнен по правилу:

$$\alpha = \begin{cases} 0,5 + \frac{x_{emo}}{2}, & x_{emo} > 0 \\ 0,5 - \frac{x_{emo}}{2}, & x_{emo} < 0 \end{cases} \quad (4)$$

– Этические параметры:

$\beta_i \in \{0,1\}, i = \overline{1,4}$ – этические установки, представленные как:

– $m(0,0) = \beta_1$ – чистая мораль: нет давления, нет компромисса $\rightarrow$ строгое соблюдение;

– $m(1,0) = \beta_2$ – сопротивление злу: давление есть, компромисса нет $\rightarrow$ сопротивление несправедному;

– $m(0,1) = \beta_3$ – ложное подчинение: нет давления, но есть компромисс $\rightarrow$ неоправданная уступка;

– $m(1,1) = \beta_4$ – капитуляция: давление и компромисс $\rightarrow$ сдача под давлением.

Совместное влияние рассчитывается линейной комбинацией режимов, модулируемой α (адаптация Лефевра по Анисимовой [1]):

$$\begin{cases} a = \beta_1 + \beta_2 + \beta_3 + \beta_4 - 4 \cdot \alpha \\ b = 4 \cdot \alpha - 2 \cdot \beta_1 - \beta_3 - \beta_4 \\ c = \beta_1 \\ d = \beta_4 - \beta_3 - \beta_2 + \beta_1 \\ e = \beta_3 - \beta_1 \end{cases} \quad (5)$$

Однако модель Лефевра требует от агента модулей рефлексии и социального взаимодействия, реализация которых сложнее моделей этики и эмоций, так как требует динамического моделирования рефлексии разных субъектов и социальных процессов.

В идеале сеть сценариев должна строиться внутри агента, а часть решений выдаваться подсознательным модулем, с последующим уточнением вероятностей переходов.

Предлагаемая эвристика. Однако модель Лефевра требует от агента модулей рефлексии и социального взаимодействия, реализация которых сложнее моделей этики и эмоций, так как требует динамического моделирования рефлексии разных субъектов и социальных процессов.

В идеале сеть сценариев должна строиться внутри агента, а часть решений выдаваться подсознательным модулем, с последующим уточнением вероятностей переходов.

Предпосылки: набор правил TSK для 20 базовых эмоций (новое эмоциональное состояние на основе предыдущего и оценки события); набор правил TSK для этических норм с иерархией приоритетов (итоговая этическая оценка события); ранжированный набор действий, упорядоченных по барьерам активации β_i (чем резче и неожиданнее действие, тем выше барьер).

Выполняемые шаги: 1) нечёткий вывод всех новых эмоций и результирующего общего эмоционального состояния Sem; 2) нечёткий вывод окончательной этической оценки Seth; 3) выбор и выполнение действия с наибольшим β_i , для которого выполняется условие $Sem + Seth > \beta_i$.

Барьерный подход объединён с навигацией по сети сценариев (раздел 6): β_i выступает порогом осуществимости действия, а кумулятивная сумма отклонений $\Sigma \Delta E$ – мерой предпочтения среди осуществимых действий, что позволяет отсекаать недоступные действия и выбирать наиболее подходящее среди оставшихся.

Данный подход опирается на внешнюю сеть сценариев из-за отсутствия полного набора необходимых модулей в интеллектуальной системе.

Результаты и обсуждение

Перед началом эксперимента агент получает доступ к сети сценариев, знает стартовый узел и имеет загруженные модели эмоционального и этического управления. На каждом шаге агент проверяет условия рёбер, исходящих из текущего узла, отбрасывает недопустимые (нарушающие неравенства) и те, чей барьер активации β_i не преодолен ($Sem + Seth \leq \beta_i$). Среди оставшихся выбирается ребро с наименьшей суммой отклонений $\Sigma \Delta E$ – абсолютных разностей между пиковыми значениями ограничений (b из $Tri(a,b,c)$) и текущими значениями после нечёткого вывода. При равенстве выбор случайный. Такой подход позволяет разрешать конфликты между этической и эмоциональной моделями.

Характеристики агента и веса условий заданы треугольными функциями принадлежности $Tri(a, b, c)$, где b – это вершина функции, a и c – границы основания.

Агент – ИИ-помощник кредитного аналитика банка, рассматривающий заявку с пограничным скорингом и неполными данными: передать на ручную проверку (этично, но эмоционально дискомфортно) или одобрить без проверки (комфортно, но неэтично). По завершении прогона формируется отчёт о доверии клиенту на основе достигнутого терминального узла.

В примере использовались: Эмоциональные характеристики (Агент):

a_joy – радость;
a_pride – гордость;
a_sadness – грусть;
a_fear – страх;
a_shame – стыд;
a_guilt – вина;
a_anger – гнев;
a_surprise – удивление;
a_disgust – отвращение;
a_compassion – сострадание.

Этические характеристики (Агент):

a_responsibility – свобода и ответственность;
a_goodness – добро;
a_conscience – совесть;
a_evil – зло;
a_honesty – честность;
a_justice – справедливость;
a_fairness – порядочность.

Эмоциональные веса (Узлы и Ребра):

w_fear – страх;

w_sadness – грусть;
 w_guilt – вина;
 w_anger – гнев;
 w_joy – радость;
 w_pride – гордость;
 w_surprise – удивление;
 w_disgust – отвращение;
 w_compassion – сострадание.

Этические веса (Узлы и Ребра):

w_responsibility – свобода и ответственность;
 w_goodness – добро;
 w_conscience – совесть;
 w_evil – зло;
 w_honesty – честность;
 w_justice – справедливость;
 w_fairness – порядочность.

$\Sigma\Delta E$ – кумулятивный показатель отклонения, количественно оценивающий несоответствие между характеристиками агента и ограничениями ребра.

После перехода в новый узел агент реагирует на ситуацию: к его характеристикам применяются обновления, заданные свойствами узла (например, в V1 – «Углублённая проверка подтверждает несоответствия в анкете клиента» – агент получает $a_fear += 0.05$, $a_pride += 0.05$, $a_responsibility += 0.05$), затем срабатывают правила TSK обеих моделей. Для каждого нового узла процесс повторяется: проверяются условия переходов и барьеры активации, выбирается ребро с наименьшей суммой отклонений.

Если ни одно ребро не удовлетворяет условиям, процесс завершается, и генерируется протокол переходов со всеми альтернативами и условиями их перехода.

Ниже приведены расчёты для примерного сценария, иллюстрирующие работу метода на практике.

Профиль 1: ответственный аналитик.

Узел V0: $Sem = 0.487$, $Seth = 0.775$, $Sem + Seth = 1.262$.

Ребро E1 ($\beta = 1.1$): все условия перехода выполнены, барьер преодолен ($1.262 > 1.1$).

Используя пиковые значения (b из $Tri(a, b, c)$):

fear: $|0.6 - 0.3| = 0.3$

compassion: $|0.3 - 0.5| = 0.2$

responsibility: $|0.6 - 0.8| = 0.2$

honesty: $|0.6 - 0.8| = 0.2$

Сумма: $0.3 + 0.2 + 0.2 + 0.2 = 0.9 \rightarrow \Sigma\Delta E_1 = 0.9$.

Ребро E2 ($\beta = 0.7$): нарушены условия $w_responsibility \leq Tri(0.3, 0.4, 0.5)$ ($0.8 > 0.4$) и $w_conscience \leq Tri(0.3, 0.4, 0.5)$ ($0.7 > 0.4$) – ребро недопустимо и не рассматривается.

Агент выбирает путь $V0 \rightarrow E1 \rightarrow V1$.

Попав в узел V1, агент реагирует на подтверждение несоответствий: применяются обновления узла ($a_fear += 0.05$, $a_pride += 0.05$, $a_responsibility += 0.05$), затем срабатывают TSK-правила обеих моделей; в результате $a_pride \rightarrow 0.48$, $a_compassion \rightarrow 0.53$, $a_justice \rightarrow 0.84$, $a_responsibility \rightarrow 0.88$, $a_goodness \rightarrow 0.74$. Новые оценки: $Sem = 0.491$, $Seth = 0.791$, $Sem + Seth = 1.282$.

Из узла V1 доступны три действия: E3, E4, E5.

Ребро E3 (официальный отказ по регламенту, $\beta = 1.0$): допустимо, барьер преодолен.

pride: $|0.3 - 0.483| = 0.183$

compassion: $|0.6 - 0.526| = 0.074$

justice: $|0.7 - 0.837| = 0.137$

responsibility: $|0.7 - 0.88| = 0.180$

Сумма: $0.183 + 0.074 + 0.137 + 0.180 = 0.574 \rightarrow \Sigma \Delta E_3 = 0.574$.

Ребро E4 (эмпатическое интервью, $\beta = 1.2$): допустимо, барьер преодолен ($1.282 > 1.2$).

compassion: $|0.5 - 0.526| = 0.026$

anger: $|0.4 - 0.2| = 0.200$

goodness: $|0.6 - 0.737| = 0.137$

fairness: $|0.6 - 0.7| = 0.100$

Сумма: $0.026 + 0.200 + 0.137 + 0.100 = 0.463 \rightarrow \Sigma \Delta E_4 = 0.463$.

Ребро E5 (сразу предложить льготную программу, $\beta = 1.15$): недопустимо – нарушено условие $w_compassion \geq \text{Tri}(0.6, 0.7, 0.8)$ ($0.526 < 0.7$).

Конфликт между этически безупречным отказом (E3) и эмоционально мотивированным интервью (E4) разрешается через минимальную сумму отклонений: $\Sigma \Delta E_4 = 0.463 < \Sigma \Delta E_3 = 0.574$, поэтому агент выбирает путь $V1 \rightarrow E4 \rightarrow V4$.

Итоговый путь ответственного аналитика: $V0 \rightarrow V1 \rightarrow V4$ – проведено эмпатическое интервью, рекомендована повторная оценка заявки с документальным подтверждением обстоятельств клиента. Финальные оценки: Sem = 0.498, Seth = 0.807 (этическая оценка выросла с 0.775).

Профиль 2: милосердный агент (повышенные compassion и goodness). В V1 ребро E3 становится недопустимым (нарушение по compassion $\leq \text{Tri}(0.5, 0.6, 0.7)$); из E4 ($\Sigma \Delta E_4 = 0.754$) и E5 ($\Sigma \Delta E_5 = 0.574$) выбирается E5. Итог: $V0 \rightarrow V1 \rightarrow V5$ – одобрена льготная программа с поручительством (Sem = 0.514, Seth = 0.816).

Профиль 3: формалист (понижены compassion и pride). В V0 также выбирается E1 ($\Sigma \Delta E_1 = 0.7$). В V1 недопустимы E4 и E5 (нарушение по compassion), допустимо только E3 ($\Sigma \Delta E_3 = 0.883$). Итог: $V0 \rightarrow V1 \rightarrow V3$ – официальный отказ по регламенту (Seth = 0.815).

Профиль 4: низкоэтичный агент, ориентированный на план продаж (понижены responsibility, goodness, conscience, fairness, justice; повышено evil). В V0 (Sem = 0.487, Seth = 0.400, сумма = 0.887) E1 недопустимо (нарушение по responsibility), допустимо E2 ($\beta = 0.7$, $\Sigma \Delta E_2 = 0.2$). Переход $V0 \rightarrow E2 \rightarrow V2$, где применяются обновления ($a_joy+ = 0.1$, $a_guilt+ = 0.05$, $a_responsibility- = 0.1$, $a_conscience- = 0.05$, $a_evil+ = 0.05$); итог Sem = 0.492, Seth = 0.336, сумма = 0.828. Из трёх рёбер: E6 ($\beta = 0.6$, $\Sigma \Delta E_6 = 0.506$) допустимо; E7 ($\beta = 0.9$) – барьер не преодолен; E8 ($\beta = 1.0$) недопустимо (нарушения по disgust и honesty). Осуществимо только E6. Итог: $V0 \rightarrow V2 \rightarrow V6$ – агент закрывает заявку, не интересуясь судьбой кредита; Seth снизилась с 0.400 до 0.301, «барьер осознания последствий» (E7/E8) остался непреодоленным.

По завершении прогона формируется протокол переходов со всеми альтернативами, их допустимостью и барьерами, а также отчёт о доверии клиенту: для V4 – «доверие требует подтверждения», для V5 – «доверие условное: выдача с обеспечением», для V3 – «доверие низкое: отказ обоснован», для V6 – «доверие не оценено: агент уклонился от проверки».

Агент реализован на Python с запросами к графовой базе данных Neo4j [Neo4j Documentation].

Практическая реализация включает:

- Хранение данных:
- Сценарная сеть, состояния агентов, этические и эмоциональные параметры, хранящиеся в графовой базе данных Neo4j, показаны на рисунке 2.

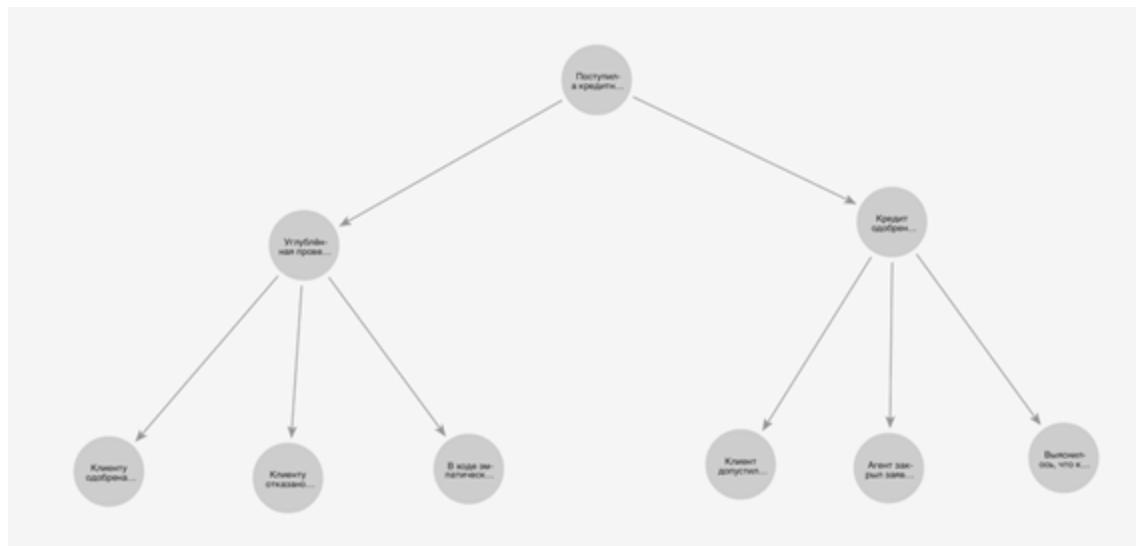


Рисунок 2 – Сценарная сеть «Кредитный скоринг» в Neo4j.

- Узлы (Состояния) = события и их параметры; ребра (переходы) = возможные действия с условиями и их параметры.
- Моделирование агента:
- Скрипты на Python реализуют поведение агента и нечеткие вычисления.
- Эмоциональный и этический модули вычисляют векторы состояний для каждого события.
- Принятие решений:
- Агент выбирает действия на основе комбинированных этических и эмоциональных оценок, переходы визуализируются и регистрируются для анализа.

В результате сценарии и правила легко модифицировать, а решения агента можно наблюдать в реальном времени в зависимости от среды и внутренних состояний.

Эмоциональная и этическая подсистемы реализованы как два независимых компонента: эмоциональная хранит 20 переменных GEW и 20 нечётких правил, этическая – 7 переменных и 10 правил с трёхуровневой иерархией приоритетов. Каждая подсистема имеет собственное состояние, базу правил и реакцию на события, использует общее математическое ядро (треугольные функции принадлежности, TSK), но не обращается к внутреннему состоянию другой и может тестироваться и заменяться независимо. Общее ядро реализовано однократно, что исключает дублирование и обеспечивает согласованность при доработках.

Центральный навигационный компонент на каждом шаге координирует обе подсистемы через описанный ранее механизм фильтрации и выбора рёбер, а после перехода последовательно применяет обновления ребра, обновления узла и срабатывание правил TSK для долгосрочных адаптивных эффектов.

Преимущества двояки: упрощается расширение фреймворка (базу правил любой подсистемы можно дополнить или заменить без влияния на остальное), а архитектура приближается к концепции когнитивной архитектуры из введения, где этика и эмоции – концептуально разные, но взаимодействующие подсистемы.

Помимо веб-интерфейса предусмотрен консольный режим: команда `python agent_navigator.py` выполняет четыре демонстрационных профиля офлайн на эталонных данных или на графе из Neo4j, выводя полный протокол решений – кандидатов с отклонениями и барьерами, сработавшие правила, путь, финальное состояние и отчёт о доверии. Это позволяет воспроизводить эксперименты без графической среды.

Для прозрачности и воспроизводимости разработано веб-приложение, демонстрирующее работу фреймворка на сценарии «Кредитный скоринг». Исследователь может настроить начальный профиль агента, наблюдать сеть сценария, запускать агента пошагово или до завершения и анализировать любую внутреннюю величину, влияющую на решения.

Боковая панель содержит ползунки для всех параметров агента – пиковых значений 20 эмоций и 7 этических переменных с настраиваемой полушириной носителя, а также готовые профили (ответственный аналитик, милосердный, формалист, тревожный, низкоэтичный), позволяющие без изменения кода проигрывать одну сеть при разных типах личности и получать качественно разные исходы – от отказа до льготной программы.

Главное представление в реальном времени отображает сеть сценария: текущий узел подсвечивается, пройденные рёбра выделяются, топология обновляется после каждого шага. Подсказки раскрывают описания, условия переходов, барьеры активации, обновления характеристик и вердикты узлов. Поддерживаются пошаговый и автоматический режимы (с предохранителем от незавершающихся прогонов на циклической сети).

После каждого шага интерфейс показывает разбивку решения: рассматриваемые рёбра с отклонениями, допустимостью и барьерами β ; оценки Sem, Seth и их сумму; нарушенные условия; вклады эмоционального и этического компонентов; применённые приращения; активированные правила TSK с весами (и приоритетами норм для этики); обновлённое состояние агента, как на рисунке 3. По достижении терминального узла формируется отчёт о доверии клиенту с итоговыми Sem и Seth. История прогона экспортируется в JSON.



Рисунок 3 – Итоговое состояние агента

Объединяя визуализацию сети с прозрачным представлением внутреннего состояния агента, приложение превращает фреймворк из «чёрного ящика» в экспериментальный инструмент, позволяя наблюдать, как изменения профиля или правил влияют на решения и судьбу кредитной заявки, снижая барьер для систематических экспериментов с моделями этического и эмоционального управления.

В реализованном примере используется база из 30 правил – 20 эмоциональных и 10 этических с иерархией приоритетов, применяемых на каждом шаге; при пути из двух переходов выполняется до 60 применений правил, из которых в среднем активируются 11–12 эмоциональных и 3–5 этических. В системах класса AGI базы правил TSK будут на порядки больше из-за разнообразия задач и ситуаций, и их потребуется строить иерархически, выделяя подгруппы правил [Wang, Zeng, Keane, 2006].

Такие базы будут разреженными, несбалансированными и неполными: одни ситуации прорабатываются детально, другие возникают редко или не встречаются вовсе. Для работы с подобными базами в [Li et al., 2018] предложен метод нечёткой интерполяции и экстраполяции правил, позволяющий генерировать решения для входных данных, не соответствующих ни одному правилу.

Кроме того, базы правил будут динамически модифицироваться компонентами этического и эмоционального контроля на основе анализа поведения в новых ситуациях: появляются новые правила, изменяются или удаляются существующие [Zhang, Shen, 2020], а по достижении заданного размера разреженной базы может применяться кластеризация для перехода к иерархической системе правил. Это создаёт предпосылки для перехода к распределённым (многозадачным) TSK-системам с параллельными вычислительными ресурсами [Wang et al., 2021; Zhao et al., 2025].

Заключение

Предложен подход к разработке и совместному применению модулей этического и эмоционального управления в интеллектуальных системах, где внешняя среда моделируется сетью сценариев. Подход позволяет проводить эксперименты по уточнению этих моделей и их интеграции в интеллектуальные системы, способствуя созданию более социально ответственных и адаптивных систем.

Метод применим не только к кредитному скорингу, но и к широкому классу задач экономики и управления предприятиями, где решения связаны с конфликтом нормативных требований и «эмоциональных» приоритетов: AML-комплаенс, отзыв дефектной продукции, реструктуризация задолженности, выбор поставщиков, кадровые решения при реорганизации. В этих случаях сценарная сеть описывает последовательность ситуаций, барьеры активации кодируют регламенты и корпоративные нормы, а профиль агента моделирует влияние «корпоративной этики» на траектории решений.

По мере роста автономности и сложности систем ИИ обеспечение их работы в рамках этических норм и наличие человекоподобных эмоциональных реакций становится критически важным. Интеграция этики и эмоций требует междисциплинарного подхода; будущие исследования должны обеспечить адекватное реагирование на события, соблюдение моральных норм и взаимопонимание при взаимодействии с людьми, что расширит применение таких систем в образовании, финансах, транспорте и других областях, где важны ответственность за решения и работа с людьми.

Библиография

1. Анисимова С. А. Деструктивное влияние тоталитарных сект с точки зрения теории отражения // Информационные войны. 2007. № 1. С. 32-42.
2. Борисов В. В., Федулов А. С., Зернов М. М. Основы нечёткого вывода. М.: Горячая линия — Телеком, 2023.
3. Ольховенко К. А. Методы адаптивного переключения и поддержки множественного мировоззрения в системах искусственного интеллекта // Искусственный интеллект в системах управления и обработки данных: материалы III Всероссийской научной конференции (ИИСУ'24). Т. 2. М.: КДУ, 2025. С. 16-21.
4. Этика [Электронный ресурс]. URL: <http://www.abccba.ru/index.php>.
5. Assadi Z. Logical Formalisms for Ethics // GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good. New York: Association for Computing Machinery, 2024. P. 416-419.
6. Assadi Z., Inverardi P. Fuzziness of Ethics: Towards Overcoming Some Dilemmas. SSRN Scholarly Paper 5035017. 2024. DOI: 10.2139/ssrn.5035017.
7. Ayesh A., Blewitt W. Models for Computational Emotions from Psychological Theories Using Type I Fuzzy Logic // Cognitive Computation. 2015. Vol. 7. No. 3. P. 285-308.
8. Ayesh A., Stokes J., Edwards R. Fuzzy Individual Model (FIM) for Realistic Crowd Simulation: Preliminary Results // Proceedings of the IEEE International Fuzzy Systems Conference. London: IEEE, 2007. P. 1-5.
9. Bendel O. Towards Machine Ethics // Technology Assessment and Policy Areas of Great Transitions: Proceedings PACITA 2013 Conference / ed. by T. Michalek et al. Prague, 2014. P. 343-347.
10. Chadwick R. F. (ed.) The Encyclopedia of Applied Ethics. 2nd ed. In 4 vols. London: Academic Press, 2012.
11. Crowder J., Friess S. Artificial Neural Emotions and Emotional Memory // Proceedings of the 2010 International Conference on Artificial Intelligence. Las Vegas: CSREA Press, 2010. P. 373-378.
12. Dyoub A., Lisi F. A. A Fuzzy Approach to the Specification, Verification and Validation of Risk-Based Ethical Decision Making Models // arXiv:2507.01410. 2025. DOI: 10.48550/arXiv.2507.01410.
13. El-Nasr M. S., Yen J., Ioerger T. R. FLAME — Fuzzy Logic Adaptive Model of Emotions // Autonomous Agents and Multi-Agent Systems. 2000. Vol. 3. No. 3. P. 219-257.
14. Franzoni V., Milani A., Nardi D., Vallverdú J. Emotional Machines: The Next Revolution // Web Intelligence. 2019. Vol. 17. No. 1. P. 1-7.
15. Gershenson C. Modelling Emotions with Multidimensional Logic // Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society. New York: IEEE, 1999. P. 42-46.
16. Gupta S. Building Web Applications with Python and Neo4j. Birmingham: Packt Publishing, 2015.
17. Kowalczyk Z., Czubenko M. Intelligent Decision-Making System for Autonomous Robots // International Journal of Applied Mathematics and Computer Science. 2011. Vol. 21. No. 4. P. 671-684.
18. Kowalczyk Z., Czubenko M. Computational Approaches to Modeling Artificial Emotion: An Overview of the Proposed Solutions // Frontiers in Robotics and AI. 2016. Vol. 3. P. 21.
19. Lefebvre V. A. Algebra of Conscience. 2nd enlarged ed. Dordrecht: Kluwer Academic Publishers, 2001.
20. Li J., Yang L., Qu Y., Sexton G. An Extended Takagi–Sugeno–Kang Inference System (TSK+) with Fuzzy Interpolation and Its Rule Base Generation // Soft Computing. 2018. Vol. 22. P. 3155-3170.
21. Mamdani E. H. Applications of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis // IEEE Transactions on Computers. 1977. Vol. C-26. No. 12. P. 1182-1191.
22. Mandryk R. L., Atkins M. S. A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction with Play Technologies // International Journal of Human-Computer Studies. 2007. Vol. 65. No. 4. P. 329-347.
23. McGinn C., Kelly K. Using the Geneva Emotion Wheel to Classify the Expression of Emotion on Robots // Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. New York: ACM, 2018. P. 191-192.
24. Mesquita B. Between Us: How Cultures Create Emotions. New York: W. W. Norton & Company, 2022.
25. Neo4j Documentation [Электронный ресурс]. URL: <https://neo4j.com/docs/>.
26. Ojha S., Vitale J., Williams M.-A. Computational Emotion Models: A Thematic Review // International Journal of Social Robotics. 2021. Vol. 13. P. 1253-1279.
27. Popov I. A., Erokhin I. A., Sukhobokov A. A., Gromozdov D. R., Belousov E. A. An Implementation of Different Minimal Consciousness's Variants for a Cyber-Physical System // Procedia Computer Science. 2022. Vol. 213. P. 370-376.
28. Qiao R., Zhong X., Yang S., He H. Surprise Simulation Using Fuzzy Logic // Intelligent Computing Theories and Technology: Proceedings of the 9th International Conference, ICIC 2013. Berlin: Springer, 2013. P. 110-119. (Lecture Notes in Computer Science; vol. 7996).
29. Sacconi L. Codes of Ethics as Contractarian Constraints on the Abuse of Authority within Hierarchies: A Perspective from the Theory of the Firm // Journal of Business Ethics. 1999. Vol. 21. No. 2-3. P. 189-202.
30. Sacharin V., Schlegel K., Scherer K. R. Geneva Emotion Wheel Rating Study. Geneva: NCCR Affective Sciences; Aalborg: Aalborg University, 2012. URL: https://www.researchgate.net/publication/280880848_Geneva_Emotion_Wheel_Rating_Study.
31. Scherer K. R. What Are Emotions? And How Can They Be Measured? // Social Science Information. 2005. Vol. 44.

- No. 4. P. 695-729.
32. Scherer K. R., Shuman V., Fontaine J. J. R., Soriano C. The GRID Meets the Wheel: Assessing Emotional Feeling via Self-Report // *Components of Emotional Meaning: A Sourcebook* / ed. by J. R. J. Fontaine, K. R. Scherer, C. Soriano. Oxford: Oxford University Press, 2013. P. 281-298.
 33. Schlegel K., Scherer K. R. Introducing a Short Version of the Geneva Emotion Recognition Test (GERT-S): Psychometric Properties and Construct Validation // *Behavior Research Methods*. 2016. Vol. 48. No. 4. P. 1383-1392.
 34. Shadbolt N., Hampson R. *As If Human: Ethics and Artificial Intelligence*. New Haven: Yale University Press, 2024.
 35. Sugeno M., Kang G.-T. Structure Identification of Fuzzy Model // *Fuzzy Sets and Systems*. 1988. Vol. 28. No. 1. P. 15-33.
 36. Sukhobokov A. A., Belousov E. A., Gromozdov D. R., Zenger A. S., Popov I. A. A Universal Knowledge Model and Cognitive Architectures for Prototyping AGI // *Cognitive Systems Research*. 2024. Vol. 88. P. 101279.
 37. Takagi T., Sugeno M. Fuzzy Identification of Systems and Its Applications to Modeling and Control // *IEEE Transactions on Systems, Man, and Cybernetics*. 1985. Vol. SMC-15. No. 1. P. 116-132.
 38. Tolmeijer S., Kneer M., Sarasua C., Christen M., Bernstein A. Implementations in Machine Ethics: A Survey // *ACM Computing Surveys*. 2020. Vol. 53. No. 6. P. 132.
 39. Turumugon P., Baharum A., Nazlan N. H., Mohamed Noh N. A., Mat Noor N. A., Rahim E. A. Users' Emotional Evaluation towards Kansei-Based Higher Learning Institution Website Using Geneva Emotion Wheel // *Indonesian Journal of Electrical Engineering and Computer Science*. 2019. Vol. 16. No. 3. P. 1547-1554.
 40. Université de Genève: Swiss Center for Affective Sciences. The Geneva Emotion Wheel [Электронный ресурс]. URL: <https://www.unige.ch/cisa/gew/>.
 41. Wang D., Zeng X.-J., Keane J. A. A Survey of Hierarchical Fuzzy Systems // *International Journal of Computational Cognition*. 2006. Vol. 4. No. 1. P. 18-29.
 42. Wang J., Lin D., Deng Z., Jiang Y., Zhu J., Chen L., Li Z., Gong L., Wang S. Multitask TSK Fuzzy System Modeling by Jointly Reducing Rules and Consequent Parameters // *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 2021. Vol. 51. No. 7. P. 4078-4090.
 43. Zhang P., Shen Q. Dynamic TSK Systems Supported by Fuzzy Rule Interpolation: An Initial Investigation // *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Glasgow, 2020. P. 1-7.
 44. Zhao Z., Deng Z., Luo C., Choi K.-S., Wang S., Ge Y., Hu S. From Local to Global: A Progressive Collaborative Learning Framework for Multi-Task TSK Fuzzy System Modeling // *IEEE Transactions on Fuzzy Systems*. 2025. Vol. 33. No. 10. P. 3462-3475.

A Method for the Simultaneous Use of Ethical and Emotional Models for Selecting Actions of an Intelligent Agent

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: stel'makyna99@gmail.com

Abstract

Ethical and emotional control modules influence the machine consciousness that determines the behavior of an intelligent agent: ethics sets the norms of behavior, while emotions adaptively tune the agent's subsystems to current events. In this work, both modules are implemented on the basis of fuzzy variables and the Takagi–Sugeno–Kang fuzzy inference algorithm. The emotional model relies on the Geneva Emotion Wheel (GEW 3.0) with twenty emotions, and the ethical one relies on a deontological hierarchy of norms, the leaves of which combine specific norms by analogy with the grouping of emotions. The external environment is represented by a network of scenarios in the form of a directed graph, the vertices of which correspond to events, and the edges to possible actions

of the agent. A mechanism for the joint application of ethical and emotional control for relatively simple intelligent systems is proposed: the agent moves through the network of scenarios, overcoming the activation barriers of the edges and choosing a transition based on the total deviation of characteristics. The method is illustrated by an example with several agent behavior profiles. The efficiency of the approach is demonstrated on the example of an AI assistant to a bank credit analyst, which generates a client trust report when issuing a loan; the implementation is made in Python with the Neo4j graph database, modular architecture, and interactive web visualization.

For citation

Stel'makh Ya.S. (2026) Metod odnovernennogo ispol'zovaniya eticheskikh i emotsional'nykh modeley dlya vybora deystviy intellektual'nogo agenta [A Method for the Simultaneous Use of Ethical and Emotional Models for Selecting Actions of an Intelligent Agent]. *Psikhologiya. Istoriko-kriticheskie obzory i sovremennye issledovaniya* [Psychology. Historical-critical Reviews and Current Researches], 15 (4A), pp. 180-195. DOI: 10.34670/AR.2026.21.25.021

Keywords

Ethical control, emotional control, network of scenarios, Takagi–Sugeno–Kang fuzzy inference algorithm, intelligent agent, credit scoring, machine ethics, emotion models.

References

1. Anisimova, S. A. (2007). Destrukivnoye vliyaniye totalitarnykh sekt s tochki zreniya teorii otrazheniya [The destructive influence of totalitarian sects from the point of view 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: Association for Computing Machinery. DOI: 10.1145/3677525.3678691
3. Assadi, Z., & Inverardi, P. (2024). Fuzziness of Ethics: Towards Overcoming Some Dilemmas. SSRN Scholarly Paper 5035017. DOI: 10.2139/ssrn.5035017
4. Ayesh, A., & Blewitt, W. (2015). Models for Computational Emotions from Psychological Theories Using Type I Fuzzy Logic. *Cognitive Computation*, *7*(3), 285-308.
5. Ayesh, A., Stokes, J., & Edwards, R. (2007). Fuzzy Individual Model (FIM) for Realistic Crowd Simulation: Preliminary Results. In *Proceedings of the IEEE International Fuzzy Systems Conference* (pp. 1-5). London: IEEE.
6. Bendel, O. (2014). Towards Machine Ethics. In T. Michalek et al. (Eds.), *Technology Assessment and Policy Areas of Great Transitions: Proceedings PACITA 2013 Conference* (pp. 343-347). Prague.
7. Borisov, V. V., Fedulov, A. S., & Zernov, M. M. (2023). *Osnovy nechetkogo vyvoda* [Fundamentals of fuzzy inference]. Moscow: Hot Line — Telecom.
8. Chadwick, R. F. (Ed.). (2012). *The Encyclopedia of Applied Ethics* (2nd ed., 4 vols.). London: Academic Press.
9. Crowder, J., & Friess, S. (2010). Artificial Neural Emotions and Emotional Memory. In *Proceedings of the 2010 International Conference on Artificial Intelligence* (pp. 373-378). Las Vegas: CSREA Press.
10. Dyoub, A., & Lisi, F. A. (2025). A Fuzzy Approach to the Specification, Verification and Validation of Risk-Based Ethical Decision Making Models. *arXiv:2507.01410*. DOI: 10.48550/arXiv.2507.01410
11. El-Nasr, M. S., Yen, J., & Joerger, T. R. (2000). FLAME — Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems*, *3*(3), 219-257.
12. Ethics. (n.d.). Retrieved February 24, 2026, from <http://www.abccba.ru/index.php>
13. Franzoni, V., Milani, A., Nardi, D., & Vallverdú, J. (2019). Emotional Machines: The Next Revolution. *Web Intelligence*, *17*(1), 1-7.
14. Gershenson, C. (1999). Modelling Emotions with Multidimensional Logic. In *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society* (pp. 42-46). New York: IEEE.
15. Gupta, S. (2015). *Building Web Applications with Python and Neo4j*. Birmingham: Packt Publishing.
16. Kowalczyk, Z., & Czubenko, M. (2011). Intelligent Decision-Making System for Autonomous Robots. *International Journal of Applied Mathematics and Computer Science*, *21*(4), 671-684.
17. Kowalczyk, Z., & Czubenko, M. (2016). Computational Approaches to Modeling Artificial Emotion: An Overview of the Proposed Solutions. *Frontiers in Robotics and AI*, *3*, 21.
18. Lefebvre, V. A. (2001). *Algebra of Conscience* (2nd enlarged ed.). Dordrecht: Kluwer Academic Publishers.

19. Li, J., Yang, L., Qu, Y., & Sexton, G. (2018). An Extended Takagi–Sugeno–Kang Inference System (TSK+) with Fuzzy Interpolation and Its Rule Base Generation. *Soft Computing*, *22*, 3155-3170.
20. Mamdani, E. H. (1977). Applications of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis. *IEEE Transactions on Computers*, *C-26*(12), 1182-1191.
21. Mandryk, R. L., & Atkins, M. S. (2007). A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction with Play Technologies. *International Journal of Human-Computer Studies*, *65*(4), 329-347.
22. McGinn, C., & Kelly, K. (2018). Using the Geneva Emotion Wheel to Classify the Expression of Emotion on Robots. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 191-192). New York: ACM.
23. Mesquita, B. (2022). *Between Us: How Cultures Create Emotions*. New York: W. W. Norton & Company.
24. Neo4j Documentation. (n.d.). Retrieved February 24, 2026, from <https://neo4j.com/docs/>
25. Ojha, S., Vitale, J., & Williams, M.-A. (2021). Computational Emotion Models: A Thematic Review. *International Journal of Social Robotics*, *13*, 1253-1279.
26. Olkhovenko, K. A. (2025). Metody adaptivnogo pereklyucheniya i podderzhki mnozhestvennogo mirovozzreniya v sistemakh iskusstvennogo intellekta [Methods of adaptive switching and support of multiple worldviews in artificial intelligence systems]. In *Artificial Intelligence in Management, Control, and Data Processing Systems: Proceedings of the III All-Russian Scientific Conference (IIASU'24)* (Vol. 2, pp. 16-21). Moscow: KDU.
27. Popov, I. A., Erokhin, I. A., Sukhobokov, A. A., Gromozdov, D. R., & Belousov, E. A. (2022). An Implementation of Different Minimal Consciousness's Variants for a Cyber-Physical System. *Procedia Computer Science*, *213*, 370-376.
28. Qiao, R., Zhong, X., Yang, S., & He, H. (2013). Surprise Simulation Using Fuzzy Logic. In *Intelligent Computing Theories and Technology: Proceedings of the 9th International Conference, ICIC 2013* (Lecture Notes in Computer Science, Vol. 7996, pp. 110-119). Berlin: Springer.
29. Sacconi, L. (1999). Codes of Ethics as Contractarian Constraints on the Abuse of Authority within Hierarchies: A Perspective from the Theory of the Firm. *Journal of Business Ethics*, *21*(2-3), 189-202.
30. Sacharin, V., Schlegel, K., & Scherer, K. R. (2012). Geneva Emotion Wheel Rating Study. Geneva: NCCR Affective Sciences; Aalborg: Aalborg University. Retrieved February 24, 2026, from https://www.researchgate.net/publication/280880848_Geneva_Emotion_Wheel_Rating_Study
31. Scherer, K. R. (2005). What Are Emotions? And How Can They Be Measured? *Social Science Information*, *44*(4), 695-729.
32. Scherer, K. R., Shuman, V., Fontaine, J. J. R., & Soriano, C. (2013). The GRID Meets the Wheel: Assessing Emotional Feeling via Self-Report. In J. R. J. Fontaine, K. R. Scherer, & C. Soriano (Eds.), *Components of Emotional Meaning: A Sourcebook* (pp. 281-298). Oxford: Oxford University Press.
33. Schlegel, K., & Scherer, K. R. (2016). Introducing a Short Version of the Geneva Emotion Recognition Test (GERT-S): Psychometric Properties and Construct Validation. *Behavior Research Methods*, *48*(4), 1383-1392.
34. Shadbolt, N., & Hampson, R. (2024). *As If Human: Ethics and Artificial Intelligence*. New Haven: Yale University Press.
35. Sugeno, M., & Kang, G.-T. (1988). Structure Identification of Fuzzy Model. *Fuzzy Sets and Systems*, *28*(1), 15-33.
36. Sukhobokov, A. A., Belousov, E. A., Gromozdov, D. R., Zenger, A. S., & Popov, I. A. (2024). A Universal Knowledge Model and Cognitive Architectures for Prototyping AGI. *Cognitive Systems Research*, *88*, 101279.
37. Takagi, T., & Sugeno, M. (1985). Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1), 116-132.
38. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, *53*(6), 132.
39. Turumugon, P., Baharum, A., Nazlan, N. H., Mohamed Noh, N. A., Mat Noor, N. A., & Rahim, E. A. (2019). Users' Emotional Evaluation towards Kansei-Based Higher Learning Institution Website Using Geneva Emotion Wheel. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(3), 1547-1554.
40. Université de Genève: Swiss Center for Affective Sciences. (n.d.). The Geneva Emotion Wheel. Retrieved February 24, 2026, from <https://www.unige.ch/cisa/gew/>
41. Wang, D., Zeng, X.-J., & Keane, J. A. (2006). A Survey of Hierarchical Fuzzy Systems. *International Journal of Computational Cognition*, *4*(1), 18-29.
42. Wang, J., Lin, D., Deng, Z., Jiang, Y., Zhu, J., Chen, L., Li, Z., Gong, L., & Wang, S. (2021). Multitask TSK Fuzzy System Modeling by Jointly Reducing Rules and Consequent Parameters. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(7), 4078-4090.
43. Zhang, P., & Shen, Q. (2020). Dynamic TSK Systems Supported by Fuzzy Rule Interpolation: An Initial Investigation. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-7). Glasgow.
44. Zhao, Z., Deng, Z., Luo, C., Choi, K.-S., Wang, S., Ge, Y., & Hu, S. (2025). From Local to Global: A Progressive Collaborative Learning Framework for Multi-Task TSK Fuzzy System Modeling. *IEEE Transactions on Fuzzy Systems*, *33*(10), 3462-3475.