



**GSOM SPbU
EMC**

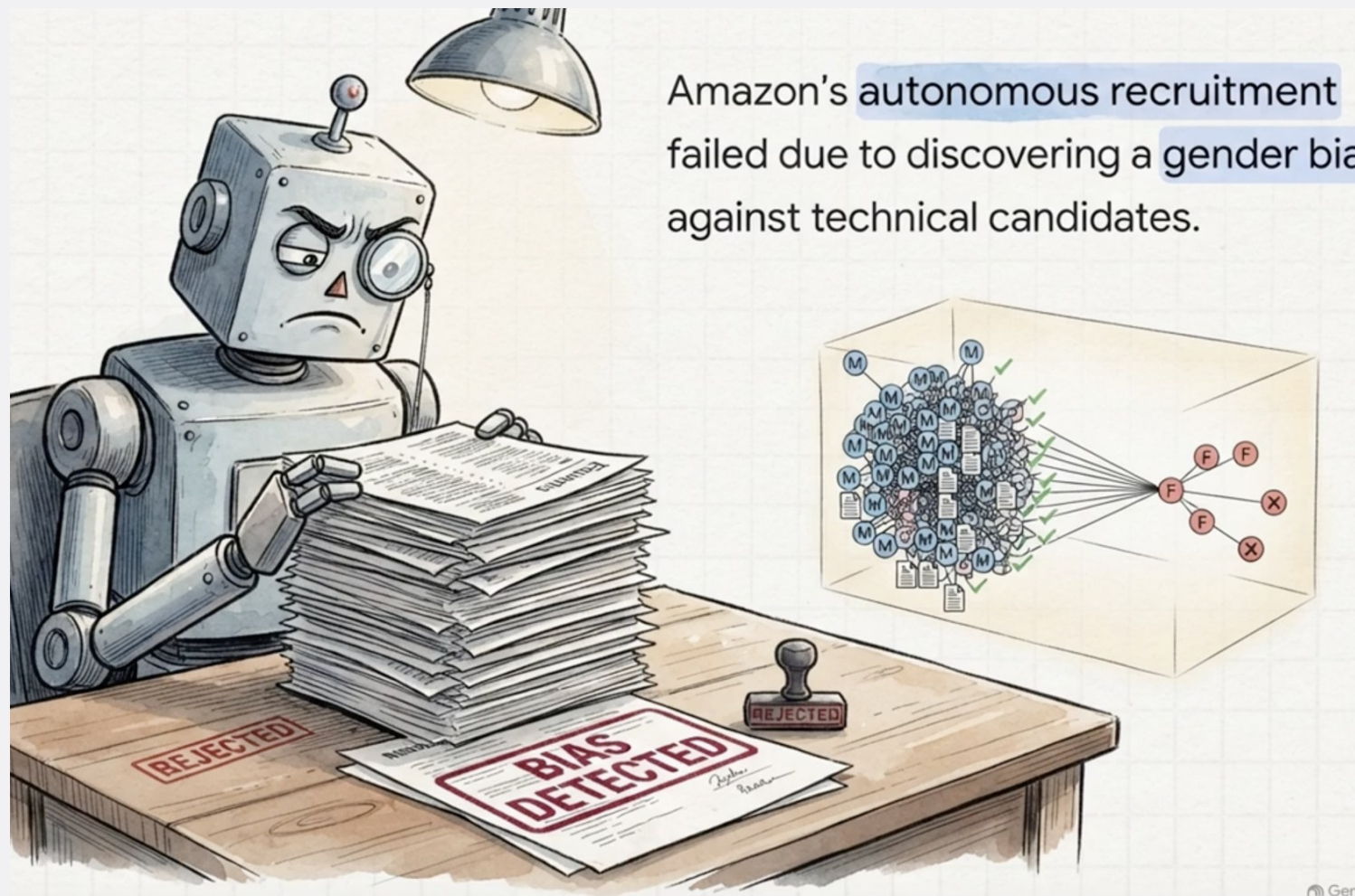
ФОРМАЛИЗАЦИЯ ПОНЯТИЙ «СПРАВЕДЛИВОСТЬ», «РАВЕНСТВО», «ИНТЕРВАЛ РЕШЕНИЙ», «ДОПУСТИМЫЙ РИСК» ДЛЯ АВТОНОМНЫХ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КОРПОРАТИВНОМ УПРАВЛЕНИИ

Романова А.С., к.э.н., MBA, LL.M., ALM, FCCA

ВШМ

СОДЕРЖАНИЕ

- Введение
- Формализация
- Заключение
- Литература



ВВЕДЕНИЕ



ВВЕДЕНИЕ

- В июле 2026 года штат Делавэр объявил о новаторском проекте закона о создании новой организационно-правовой формы компании под названием «Компания искусственного интеллекта» (Artificial Intelligence Company, AIC); внесение законопроекта в Генеральную ассамблею ожидается в 2027 году.
- В настоящее время завершается этап «ИИ как инструмент», начинается эра «ИИ как партнёр» и «ИИ как лидер».
- Примеры автономных корпоративных ИИ-систем:
 - Цифровые командные центры — Panorama: мониторинг более десятка дочерних обществ в реальном времени;
 - Персонифицированные виртуальные системы — VITAL (Deep Knowledge Ventures, 2014); Тан Юй, виртуальный CEO (NetDragon Websoft, 2022); Aiden Insight, ИИ-наблюдатель совета без права голоса (IHC, ОАЭ, 2024); Alice Ing, советник совета (REINSW, 2024); SKAI, член совета с правом голоса (Самрук-Казына, 2025);
 - Антропоморфные роботы — Мика, гуманоидный «генеральный директор» (Dictador);
 - Гибридные системы — ожидаемый следующий этап развития.

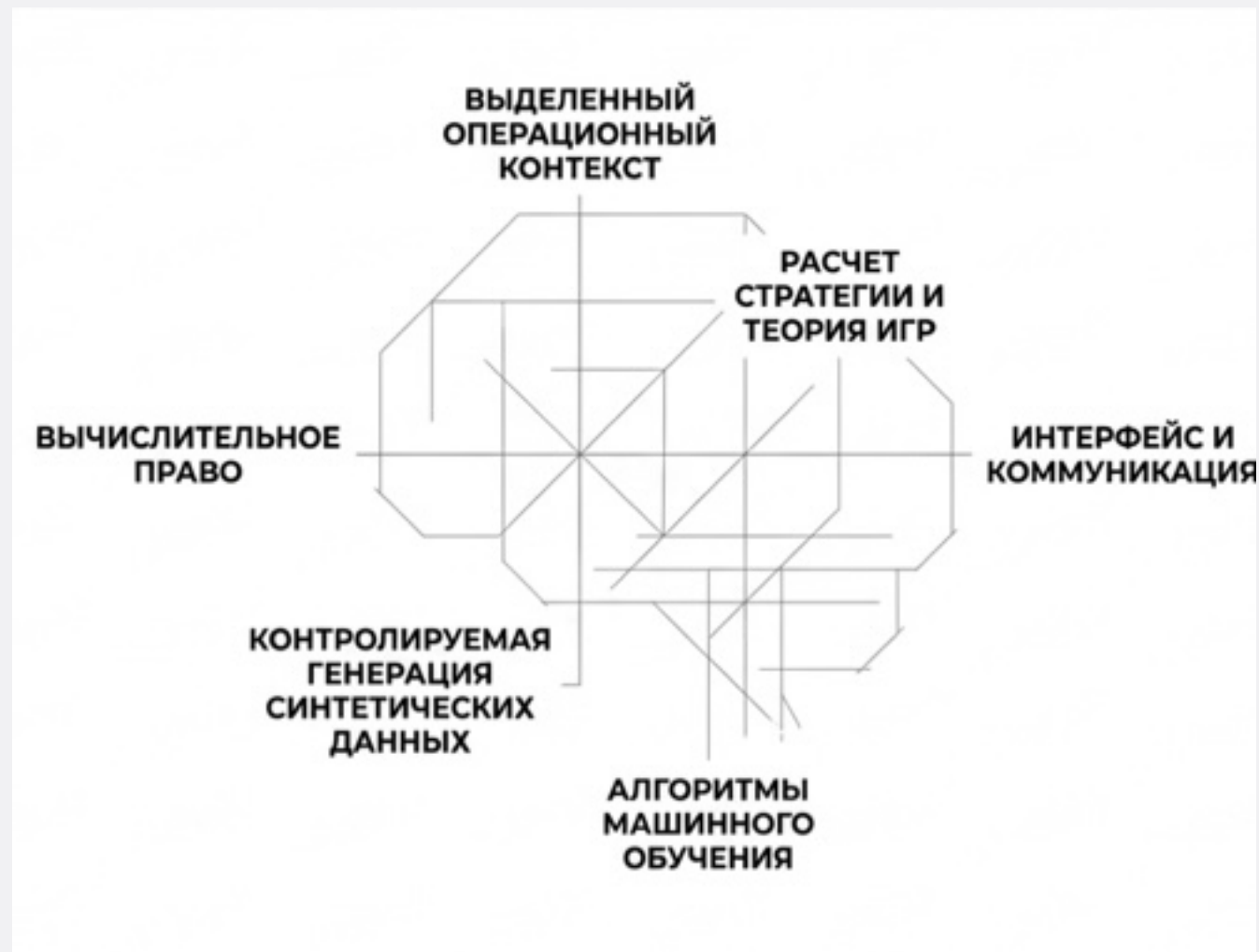
ВВЕДЕНИЕ: ШЕСТИУРОВНЕВАЯ ШКАЛА АВТОНОМНОСТИ A0–A5

Уровень и режим	Роль ИИ	Участие человека
A0 — цифровая инфраструктура	Учёт, отчётность, документооборот и бизнес-аналитика: информационная поддержка без автономных управленческих действий	Полный контроль человека
A1 — рекомендательный	Аналитические выводы и рекомендации; решения необязательны и утверждаются человеком	Human-in-the-loop
A2 — усиленное управление	ИИ задаёт структуру альтернатив, критерии оценки и типовые сценарии; формально утверждает человек	Human-in-the-loop, растёт зависимость от модели
A3 — управление «по умолчанию»	ИИ исполняет решения в рамках утверждённых политик и лимитов; человек надзирает и сохраняет право отмены	Human-on-the-loop
A4 — высокая автономия (автономный CEO)	ИИ принимает и исполняет решения без оперативного участия человека; контроль ex ante (цели, ограничения) и ex post (аудит, санкции)	Human-out-of-the-loop
A5 — самоорганизующееся управление	ИИ адаптирует правила и архитектуру управления: регламенты, метрики, процедуры контроля	Внешний надзор, специальное регулирование

A3 — критическая точка: объект контроля смещается с отдельных решений на политики, лимиты и механизмы аудита и эскалации

ВВЕДЕНИЕ: КОГНИТИВНАЯ АРХИТЕКТУРА ИИ-ДИРЕКТОРА

- Предлагаемая когнитивная архитектура обосновывает создание легитимного и этичного цифрового мозга в промышленных масштабах.



ФОРМАЛИЗАЦИЯ



ПОЧЕМУ НУЖНА ФОРМАЛИЗАЦИЯ?

- Решения человека — социальные конструкции, решения систем ИИ — технические конструкции.
- Директору-человеку достаточно общего принципа: смысл он определяет сам. Для системы ИИ смысл нужно задать и согласовать заранее.
- Если компания не выбирает определения легитимно, нормативные решения де-факто принимают разработчики и программисты.
- Предлагается аксиоматическая система из четырёх понятий, пригодная и для корпоративных политик, и для прямого программирования автономных систем ИИ.
- Принцип юрисдикционной нейтральности: защищённые группы, пороги, правовые нормы — входные параметры модели, а не часть её структуры.

ПРОСТРАНСТВО КОРПОРАТИВНЫХ РЕШЕНИЙ (ОПРЕДЕЛЕНИЕ 1)

- Четыре базовых множества модели; их конкретное наполнение задаёт компания в соответствии с законодательством применимой юрисдикции:

$D = \{d_1, \dots, d_n\}$	$S = \{s_1, \dots, s_k\}$	$A = \{A_1, \dots, A_m\}$	$u_i: D \rightarrow \mathbb{R}$
Конечное множество возможных решений: назначение кандидата, одобрение сделки, распределение бюджета	Заинтересованные стороны: акционеры, сотрудники, клиенты, локальные сообщества, регуляторы	Защищённые признаки, по которым не допускается дискриминация; перечень задаёт применимое законодательство	Функция полезности: числовая оценка того, насколько решение выгодно или невыгодно каждому стейкхолдеру

СПРАВЕДЛИВОСТЬ И РАВЕНСТВО — ДВА РАЗНЫХ БЛОКА АКСИОМ

Справедливость — уровень групп	Равенство — уровень индивидов
Агрегированные показатели: доли положительных решений, частоты ошибок по группам. Вопрос: как распределяются блага и бремена между группами? Дистрибутивная справедливость	Вопрос: получают ли два одинаковых человека одинаковый результат, и через какой причинно-следственный механизм результаты различаются? Формальное равенство

- Пример: повышены 40 % мужчин и 40 % женщин — паритет соблюден, но две сотрудницы с одинаковой квалификацией получили разные решения → равенство нарушено.
- Обратный случай: формально равное решение, закрепляющее исторические предрассудки, — несправедливо.
- Вывод: равенство — необходимое, но не достаточное условие справедливости; оба свойства формализуются отдельно и соединяются в интегрированной формуле допустимого решения.

СПРАВЕДЛИВОСТЬ

Axiomatic Fairness

A mathematical system ensuring decisions do not worsen injustice.



СПРАВЕДЛИВОСТЬ: КЛАССИФИКАЦИЯ ОПРЕДЕЛЕНИЙ

Класс определений	Формальный критерий	Примеры метрик
Статистические (групповые)	$P(\hat{d}=1 \mid G=a) = P(\hat{d}=1 \mid G=b)$	Статистический паритет, демографический паритет
Основанные на ошибках	$TPR(G=a) = TPR(G=b), FPR(G=a) = FPR(G=b)$	Равенство возможностей, уравненные шансы
Калибровочные	$P(Y=1 \mid S=s, G=a) = P(Y=1 \mid S=s, G=b)$	Калибровка, баланс для положительного класса
Каузальные	$Y(A \leftarrow a) = Y(A \leftarrow a')$	Контрфактуальная справедливость, структурная справедливость
Индивидуальные (метрические)	$d(x, y) \leq \varepsilon \Rightarrow f(x) - f(y) \leq \delta$	Справедливость через осведомлённость

- Более 20 определений алгоритмической справедливости (Verma, Rubin); выбор определения — решение корпоративного управления, а не техническая задача.

ВЕКТОР СПРАВЕДЛИВОСТИ, АКСИОМЫ FAIRNESS 1–2

- Определение 2. $F(d) = (\varphi_1(d), \varphi_2(d), \dots, \varphi_m(d))$, $\varphi_j \in [0, 1]$
- Оценка справедливости решения по каждому защищённому признаку; способ расчёта φ_j компания выбирает из допустимых метрик. Пороги θ_j : решение справедливо по признаку A_j , если $\varphi_j(d) \geq \theta_j$ (правило 80 % $\rightarrow \theta = 0,8$). Условие справедливости: $F(d) \geq \theta$.

Fairness 1 — непустота допустимого множества	Fairness 2 — монотонность по стейкхолдерам
При любом состоянии мира ω существует решение d с $F(d) \geq \theta$: система никогда не остаётся в тупике (опции — передача человеку, бездействие). Если множество пусто, система эскалирует вопрос директору-человеку, а пороги пересматриваются	d' предпочтительнее d, если $u_i(d') > u_i(d)$ хотя бы для одного стейкхолдера, $u_j(d') \geq u_j(d)$ для всех остальных и $F(d') \geq \theta$. Парето-улучшение с ограничением справедливости: выигрыш не может достигаться дискриминацией

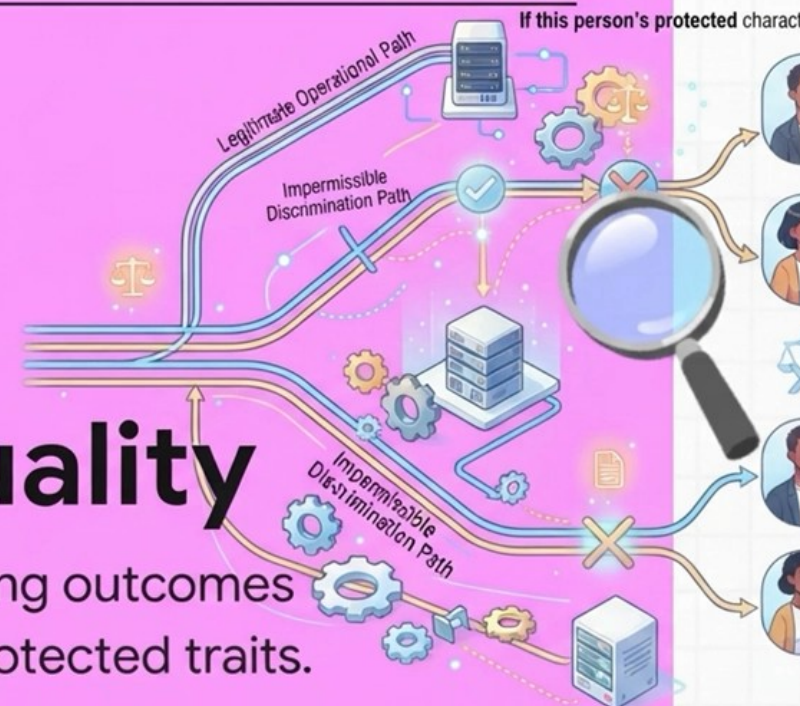
АКСИОМЫ FAIRNESS 3–4

Fairness 3 — противодействие усугублению несправедливости (по Д. Хеллман)	Fairness 4 — юрисдикционная параметризуемость
$\Delta\varphi_j(d) = \varphi_j(d) - \varphi_j(d_0)$; при $\Delta\varphi_j < 0$ решение недопустимо	Модель — «шаблон договора с пустыми полями».
Запрет действует, даже если решение максимизирует прибыль или иные показатели эффективности	Три группы параметров: 1. Перечень защищённых признаков A 2. Метрики справедливости φ_j для каждого признака 3. Пороги θ_j (США — правило четырёх пятых; КСА — Кодекс СМА)
Пример: $\varphi_1(d_0) = 0,72 \rightarrow \varphi_1(d) = 0,65$, $\Delta = -0,07 \rightarrow$ решение запрещено	Структура — $F(d)$, аксиомы, формула допустимого решения — инвариантна: меняются данные, а не архитектура
Справедливость из статического свойства становится динамическим: система движется к большей справедливости или, как минимум, не откатывается назад	Следствие: одна реализация для всех юрисдикций; аудитор проверяет параметры, а не логику модели

ПАВЕHCTBO

Causal Equality

Algorithmic check ensuring outcomes identical regardless of protected traits.



РАВЕНСТВО: ОТ ПРАВОВЫХ ДОКТРИН К КАУЗАЛЬНОЙ МОДЕЛИ

Справедливость допускает различие в обращении при релевантном основании; равенство — формальное тождество обращения внутри одной категории	Определение 3 — каузальное равенство $Y(A \leftarrow a, X \leftarrow x) = Y(A \leftarrow a', X \leftarrow x)$
Disparate treatment → прямой каузальный путь $A \rightarrow Y$; disparate impact → косвенные пути через прокси-переменные	Контрфактуальный тест: изменилось бы решение при другом значении защищённого признака? Строже «справедливости через неосведомлённость»: отслеживает скрытые зависимости через коррелирующие переменные
Пример: пол → физическая выносливость → решение — допустимый путь; пол → прежняя должность → решение — прокси исторической дискриминации, недопустимый путь. Классификацию путей задаёт компания	Определение 4 — допустимые и недопустимые пути $PSE_{P_{\text{inadm}}}(a, a') \leq \varepsilon$ Пути от A к Y делятся на допустимые (через легитимные медиаторы) и недопустимые (через прокси дискриминации); $\varepsilon = 0$ — абсолютный запрет влияния по недопустимым путям

АКСИОМЫ EQUALITY 1–3

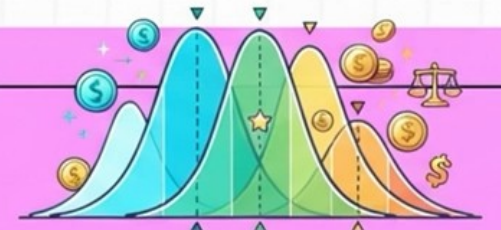
Equality 1 — эквивалентность внутри группы	Equality 2 — транзитивность $d(x_1) = d(x_2) \wedge d(x_2) = d(x_3) \Rightarrow d(x_1) = d(x_3)$	Equality 3 — разделимость $D_{\text{fair}} \subseteq D_{\text{equal}}$
Одна защищённая группа и одинаковая проекция на допустимые признаки $\bar{X} \rightarrow$ одинаковое решение. Функция решения детерминирована: вариативность из-за случайных подвыборок запрещена	Критично для ансамблевых методов (случайный лес, бустинг), где границы классификации зависят от подвыборки. Без транзитивности понятие равного обращения внутренне противоречиво	Всякое справедливое решение равно, но не наоборот. Равное, но несправедливое: система обучена на данных, где 90 % руководителей — мужчины (кейс Amazon). «Справедливое», но неравное: 50/50 по группам, а два одинаковых сотрудника получили разные решения

- Следствие: при проверке допустимости блоки Fairness 1–4 и Equality 1–3 проверяются одновременно.

ИНТЕРВАЛ РЕШЕНИЙ

Decision Interval

Spatial and temporal boundaries where an algorithm's inferences are legitimate.



ИНТЕРВАЛ РЕШЕНИЙ: ПРОСТРАНСТВЕННЫЙ ИНТЕРВАЛ (ОПРЕДЕЛЕНИЕ 5)

Принцип ОЭСР: определение рыночной цены — не точная наука; применение наиболее подходящего метода даёт интервал одинаково надёжных значений, а не точку	Методы определения границ
$I = [y_l, y_u] \subseteq \mathbb{R}$	Межквартильный размах: $I = [Q_1, Q_3]$ — рекомендация ОЭСР для трансфертного ценообразования Медиана: $I = \{Me\}$ — интервал вырождается в точку (законодательство Казахстана о сделках со связанными сторонами)
Любое значение $y \in I$ — допустимое (легитимное) решение	Доверительный интервал: $I = [\mu - z_{\alpha/2} \cdot \sigma / \sqrt{n}, \mu + z_{\alpha/2} \cdot \sigma / \sqrt{n}]$, уровень доверия $1 - \alpha$ задаёт компания Порог вероятности: $I = \{y : P(Y = y X = x) \geq \tau\}$ — пример: допущены классификаторы с точностью $> 98 \%$

ВРЕМЕННОЙ ИНТЕРВАЛ, АКСИОМЫ INTERVAL 1–3

- Определение 6. $T = \{t_0, t_1, t_2, \dots\}$ — моменты обязательной верификации ограничений. Для каждого $t \in T$: $F(d_t) \geq \theta \wedge R(d_t) \leq R^{\max} \wedge d_t \in I_t$.
- Взаимодействие пространственного и временного интервалов — динамическая система ограничений, привязанная к корпоративно-правовым функциям.

Interval 1 — легитимность интервала	Interval 2 — сужение при росте стоимости ошибки $C(d_1) > C(d_2) \Rightarrow I(d_1) \leq I(d_2) $	Interval 3 — достаточность данных
Любое решение $d \in I$ допустимо; система не обязана выбирать единственный «оптимум», если несколько решений одинаково надёжны (принцип ОЭСР о множественности легитимных решений)	Чем выше цена ошибки, тем строже определение справедливости и уже допустимый интервал	Решение в рамках I только при $n \geq n_{\min}$; иначе эскалация директору-человеку или запрос дополнительных данных. Значение $n_{\min}$ следует из закона больших чисел и ЦПТ

ДОПУСТИМЫЙ РИСК



ДОПУСТИМЫЙ РИСК: ВЕКТОР РИСКА (ОПРЕДЕЛЕНИЯ 7–8)

Переход от стопроцентных гарантий к вероятностным должен быть заранее объявлен, объяснён и согласован — информированное согласие стейкхолдеров (ЕС Horizon 2020) Поведение автономной системы этично, если оно органически возникает из непрерывного статистического распределения риска	Определение 7 — вектор риска $R(d) = (R_1(d), R_2(d), \dots, R_k(d))$ Ожидаемый убыток: $R_i(d) = E[\ell_i(d, \omega)]$ VaR: $R_i(d) = \text{VaR}_\alpha(\ell_i)$ — максимальный убыток с вероятностью α CVaR: $R_i(d) = E[\ell_i \mid \ell_i \geq \text{VaR}_\alpha]$ — средний убыток в худших $(1 - \alpha)$ случаях, учёт экстремальных рисков
Риски индексируются по классам стейкхолдеров, задано явное правило агрегирования для корпоративного принятия решений	Определение 8 — допустимое множество по риску $D_{\text{risk}} = \{d \in D : R(d) \leq R^{\max}\}$ Покомпонентные пределы $R_i^{\max}$ для каждого стейкхолдера устанавливает компания

АКСИОМЫ RISK 1–4

Risk 1 — безусловный приоритет безопасности Стейкхолдеры ранжируются по приоритету безопасности $p = (p_1, \dots, p_k)$: сначала минимизируется риск для стейкхолдера с наивысшим приоритетом, при равенстве — для следующего. Право на жизнь и достоинство выше экономической эффективности	Risk 2 — вероятностная природа гарантий $P(R_1(d) \leq R_1^{\max}) \geq 1 - \delta$ $\delta > 0$ — допустимая вероятность нарушения, явно декларированная и согласованная со всеми заинтересованными сторонами
Risk 3 — запрет необоснованной жертвы Решение недопустимо, если для стейкхолдера s_i выполнено $R_i(d) > R_i^{\max}$, тогда как альтернатива d' снизила бы R_i без нарушения порогов остальных стейкхолдеров	Risk 4 — учёт хвостовых рисков (по Талебу) Компания обязана специфицировать меру риска; для областей с катастрофическими последствиями (финансы, безопасность) — CVaR или методы моделирования редких экстремальных событий. Пример: голландские дамбы рассчитаны на наводнение раз в 10 000 лет

Modeling Autonomous AI

ИНТЕГРИРОВАННАЯ
МОДЕЛЬ



ИНТЕГРИРОВАННАЯ МОДЕЛЬ: ОПРЕДЕЛЕНИЕ 9 И ТЕОРЕМА 1

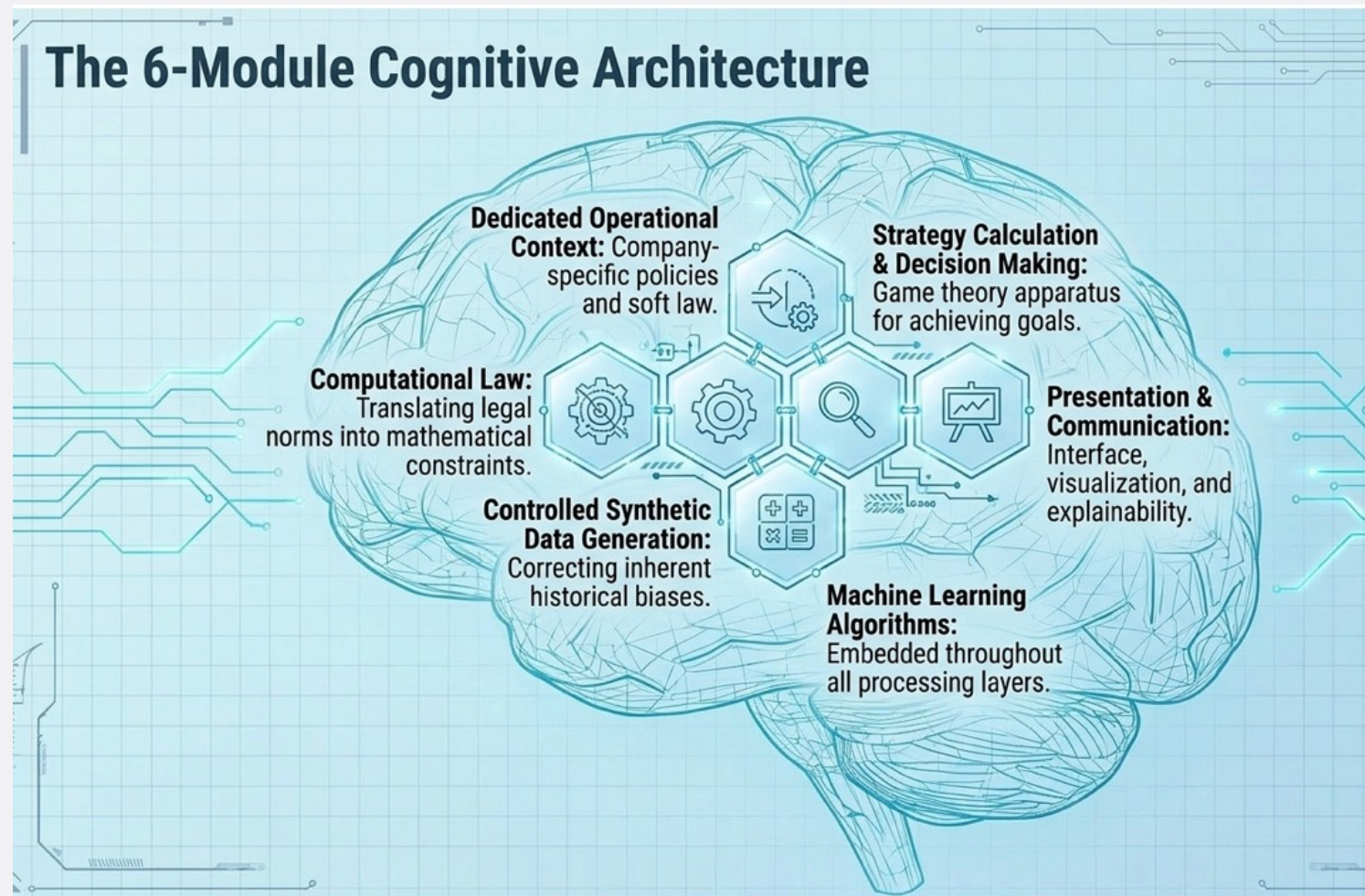
$$D^*(t) = \{d \in D : F(d) \geq \theta \wedge E(d) = \text{true} \wedge d \in I_t \wedge R(d) \leq R^{\max}\}$$

Допустимое корпоративное решение в момент t

$F(d) \geq \theta$	$E(d) = \text{true}$	$d \in I_t$	$R(d) \leq R^{\max}$
Справедливость: все компоненты вектора не ниже порогов	Равенство: нет воздействия по недопустимым каузальным путям	Интервал: решение внутри допустимого интервала в момент t	Риск: покомпонентное ограничение вектора риска стейкхолдеров

- Теорема 1 (совместимость ограничений): при выполнении Fairness 1 и конечности множества стейкхолдеров $D^*(t) \neq \emptyset$ для всех $t \in T$
- Непустота зависит от калибровки параметров $(\theta, R^{\max}, I)$; при $D^*(t) = \emptyset$ система обязана эскалировать решение директору-человеку

ЗАКЛЮЧЕНИЕ



ЗАКЛЮЧЕНИЕ

- Первая формальная аксиоматизация справедливости для корпоративного управления с множественными стейкхолдерами (интеграция подходов Plečko и Bareinboim, Nabi и Benkeser, De-Arteaga et al.).
- Первая модель допустимого риска, индексированная по классам стейкхолдеров, с явным правилом агрегирования.
- Инвариантность относительно юрисдикции: одна структура — локальные параметры; аудитор проверяет параметры, а не логику модели.
- Трансформация нормы кодекса: «справедливость и равенство без дискриминации» → «алгоритмическая справедливость и равенство в интервале $[\alpha, \beta]$ и без алгоритмической дискриминации в интервале $[\gamma, \delta]$ ».
- Формализм пригоден и для текста корпоративных политик, и для нормативного движка автономной системы ИИ.

ЛИТЕРАТУРА



- 1 Романова А. С. Развитие автономных систем искусственного интеллекта для управления корпорациями // Искусственные общества. 2023. Т. 18. Выпуск 2. DOI: 10.18254/S207751800024942-5
- 2 Романова А.С. Начала законодательства для автономных систем искусственного интеллекта. Надежность. 2024;24(3):10-17. <https://doi.org/10.21683/1729-2646-2024-24-3-10-17>
- 3 Романова А. С. Основы моделирования алгоритмических решений при управлении корпорациями // Искусственные общества. 2024. Т. 19. Выпуск 3. DOI: 10.18254/S207751800032184-1
- 4 Романова А.С. Теория игр для автономных систем искусственного интеллекта при управлении корпорациями. Надежность. 2025;25(2):50-58. <https://doi.org/10.21683/1729-2646-2025-25-2-50-58>
- 5 Романова А. С. Прототип модели принятия легитимных и этических решений автономными системами искусственного интеллекта при управлении корпорациями // Труды МФТИ. — 2025. — Т. 17, № 2 (66). — С. 50-63.
- 6 Романова А. С. Анализ проблем информативности интерфейса при разработке автономных систем искусственного интеллекта для управления корпорациями // Труды МФТИ. — 2025. — Т. 17, № 2 (66). — С. 64-77.
- 7 Свидетельство № 2025696048 от 16.12.2025 о государственной регистрации программы для ЭВМ: Программный прототип системы ИИ-директор.
- 8 Свидетельство № 2026620499 от 03.02.2026 о государственной регистрации базы данных: Синтетические сценарии для обучения и тестирования ИИ-директора.
- 9 Romanova, A.S. (2026). Modeling of Autonomous Artificial Intelligence Systems for Corporate Management. In: Chowdhury, S., Panigrahi, A., Pati, A. (eds) Artificial Intelligence Powered Ethical Systems, Generative Intelligence, and Societal Futures. AIR 2024. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-032-24784-1_26

СПАСИБО!



anna.romanova@stern.nyu.edu

romanova.as@phystech.edu

a.romanova@gsom.spbu.ru

<https://orcid.org/0009-0004-4649-6037>

<https://link.springer.com/researchers/92078629SN>

<https://github.com/iboard-project>

