ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Технический Комитет 164

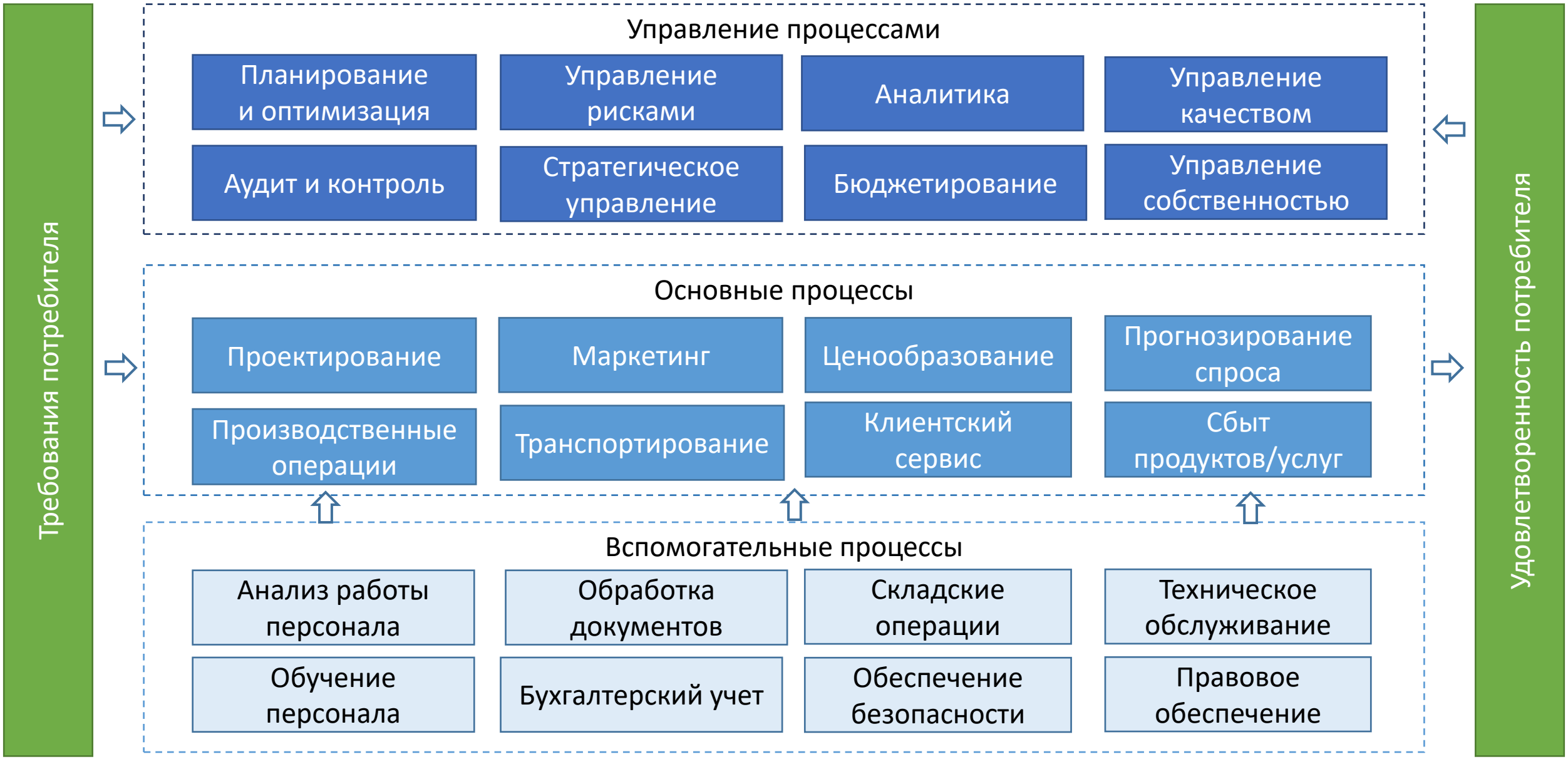
Система оценки соответствия технологий ИИ: отраслевая специфика

Пленарное заседание

XIII Международная научная конференция ИТ-Стандарт 2024

22 мая 2024 года

Применение ИИ в процессах организации (на основе анализа 783 вариантов использования)



Задачи ИИ в организации (на примере ритейла, около 90 вариантов использования)

Бизнес-процессы

Управленческие

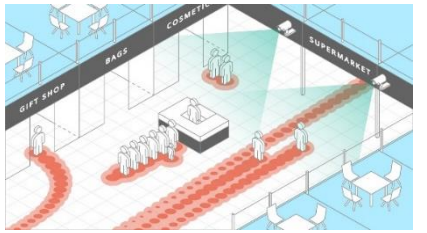
- Обоснование стратегии компании через прогнозирование долгосрочных последствий
- Оценка рисков с использованием прогнозных моделей
- Распознавание документов при внутреннем контроле и аудите

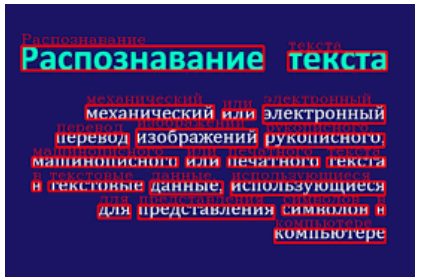
Обеспечивающие

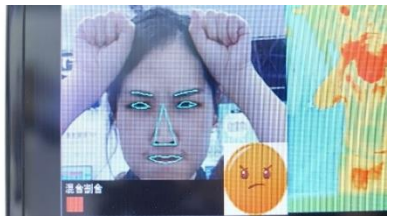
- Анализ переговоров и поведения персонала для контроля качества его работы
- Распознавание документов для формирования электронной базы клиентов и сотрудников
- Мониторинг торгового зала с целью оценки загрузки магазина, обнаружения происшествий
- Анализ дорожных и погодных условий для оптимизации маршрута движения товара, отслеживание этапа логистического цикла и условий перевозки товара
- Выгрузка/погрузка и перемещение товара с помощью роботизированной складской техники

Основные

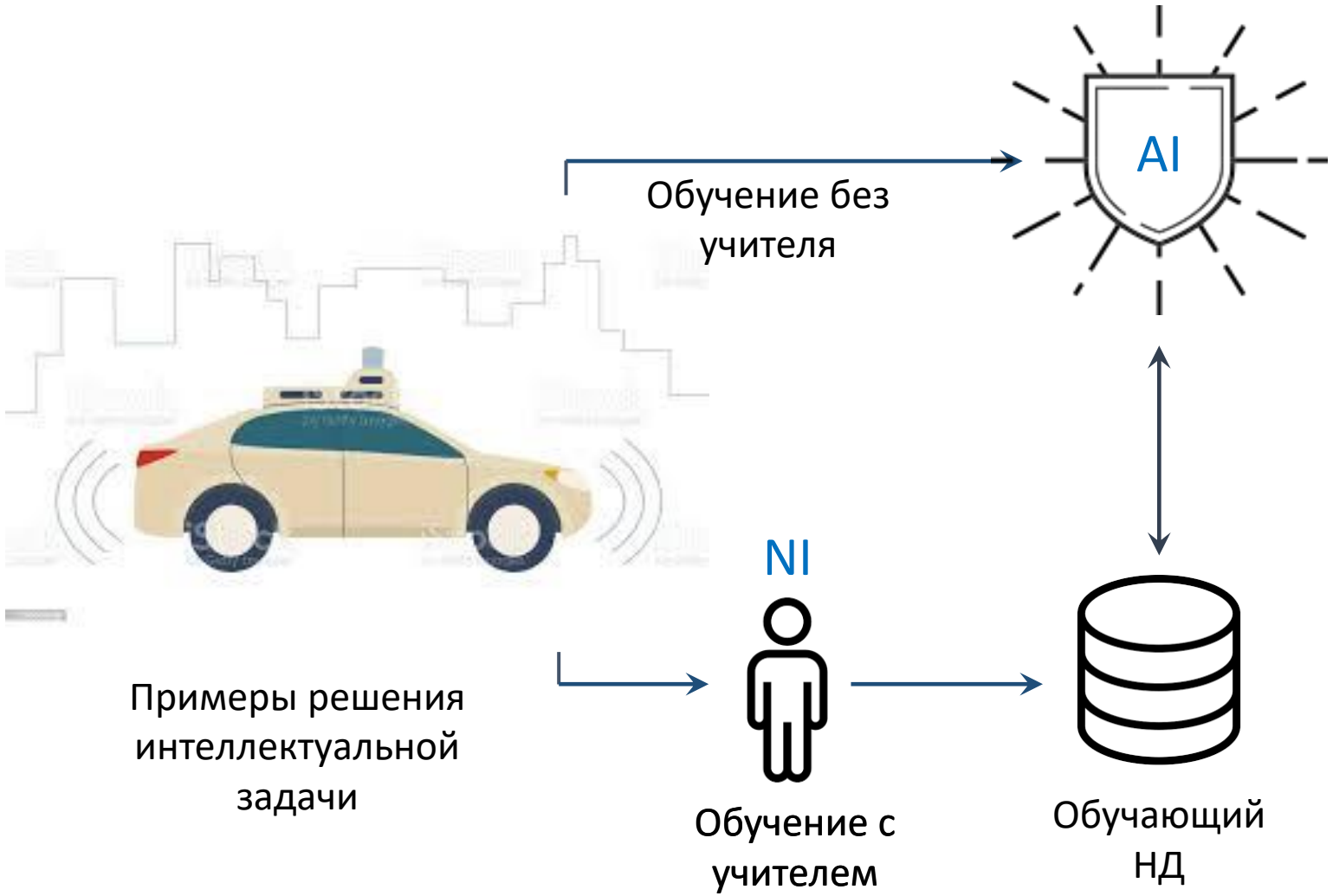
- Анализ цен конкурентов, срока годности продукции для управления ценообразованием
- Распознавание товаров и действий покупателя для оплаты покупок и предотвращения краж
- Анализ информации о покупателе (пол, возраст, эмоции), его действиях для формирования персонализированных предложений и оценки удовлетворенности работой магазина
- Доставка покупок покупателям при помощи дронов, роботов, автономных автомобилей
- Контроль соответствия выкладки продукции на стеллажах заявленному плану-схеме
- Управление запасами с помощью прогнозирования спроса
- Автоматизация клиентского сервиса (поиск товара по фото, роботизированная телефония)







Технологии искусственного интеллекта – технологии обработки данных с использованием методов машинного обучения



Алгоритм системы ИИ принципиально не обладает полной понятностью (объяснимостью) для человека

Плохо предсказуемое поведение системы ИИ в реальных условиях эксплуатации, отсутствие в поведении систем «здравого смысла», подверженность воздействию т.н. «состязательных» атак на исходные данные

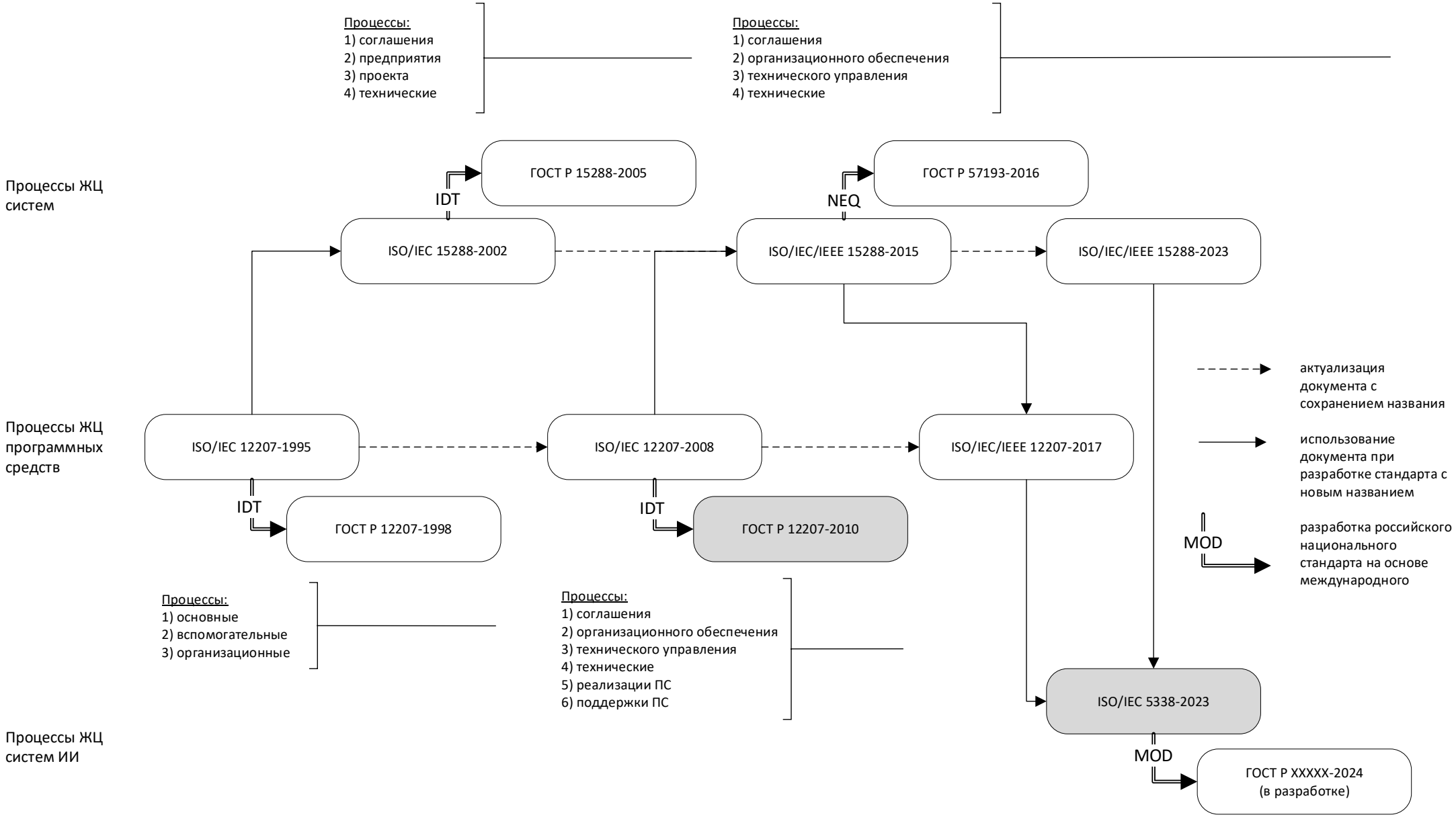
Некорректная работа алгоритмов ИИ наблюдается при определённых (сложно предсказуемых):

- условиях эксплуатации (сочетаниях параметров внешней среды и объекта измерения)

- небольших (не значительных с точки зрения здравого смысла человека) неумышленных или умышленных искажениях исходных данных, подаваемых на вход алгоритма ИИ

- характеристиках наборов данных, используемых для дообучения алгоритмов ИИ на стадии их эксплуатации

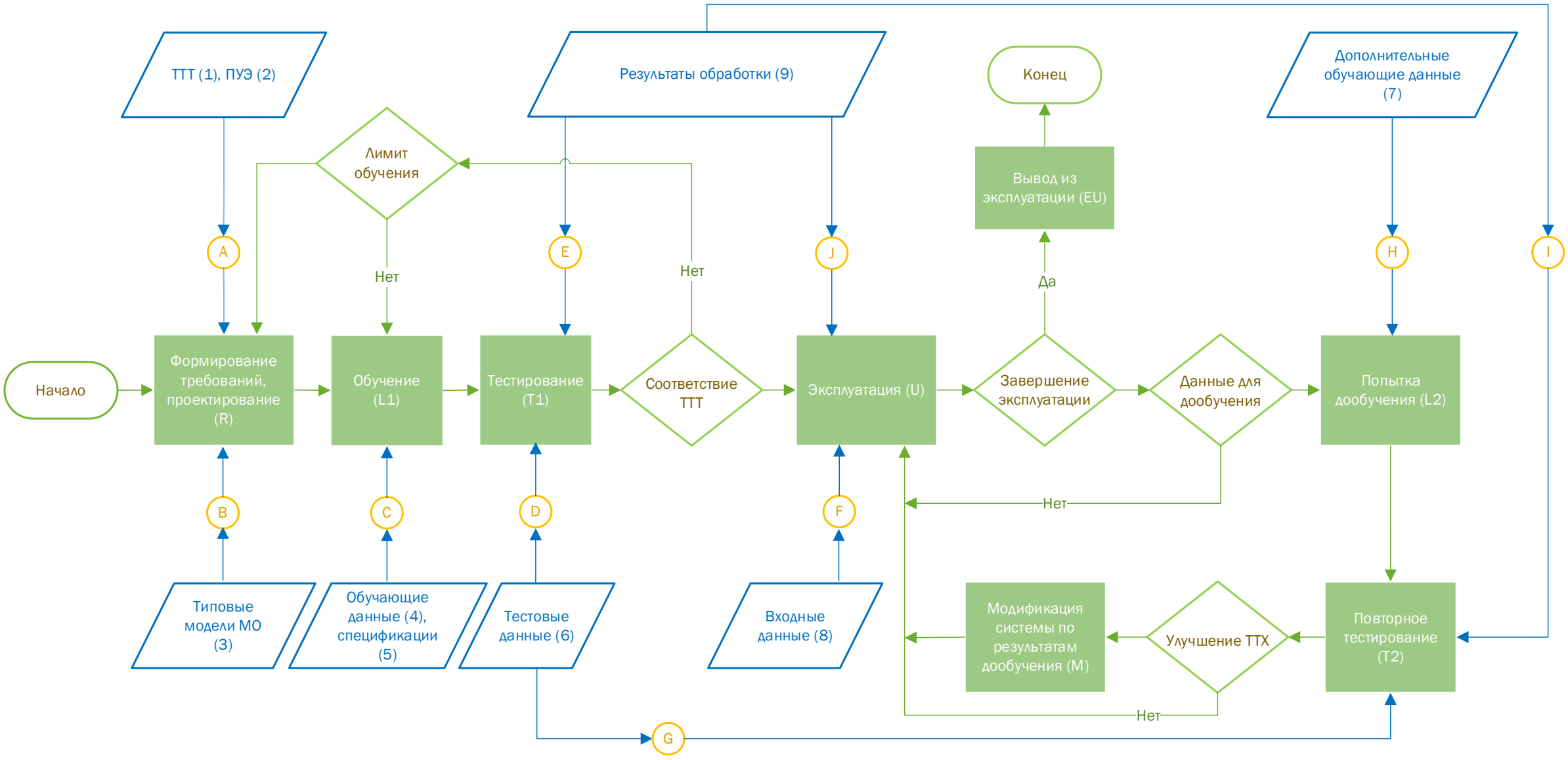
Стандарты ЖЦ систем, программных средств и систем ИИ



Особенности алгоритмов МО

В соответствии с моделью ЖЦ		В соответствии с ISO/IEC/IEEE 5338-2023	
№	Содержание	№	Содержание
1	Обязательность этапа обучения	5	Принципиальная необходимость использования представительных НД для обучения, тестирования, верификации и валидации СИИ. Определение поведения моделей МО не путем программирования (прим.: “not programmed”), а методом изучения на основе данных
2	Неполная интерпретируемость и отсутствие строгих доказательств функциональной корректности алгоритмов МО	4	Принципиальный вероятностный характер поведения СИИ, наличие ограничений на применение формальных методов при верификации корректности моделей МО
		6	Важное значение знаний, определяющих корректность моделей МО
		8	Возможное снижение доверия к СИИ на основе алгоритмов МО, связанное с их меньшей предсказуемостью, понятностью и объяснимостью поведения по сравнению с системами, основанных на интерпретируемых знаниях
3	Возможность дообучения СИИ на стадии эксплуатации	1	Необходимость мониторинга возможных изменений в поведении контролируемого объекта в процессе применения СИИ
		3	Итерационное уточнение требований и поведенческих сценариев СИИ, в том числе – при возникновении непредвиденных ситуаций в процессе их эксплуатации
4	Важность социальной приемлемости применения	7	Необходимость информирования пользователей о возможных рисках применения СИИ для предотвращения избыточного и неоправданного доверия к ним, в том числе – при попытке использования систем для замены человека
5	Необходимость сопоставления с интеллектуальными способностями человека	2	Необходимость контроля качества СИИ, обладающих автономностью, сопоставимой с поведением человека и способных нанести существенный ущерб окружающим
6	Рост конфиденциальности данных при эксплуатации СИИ	-	-

Модель жизненного цикла системы ИИ



Требования в области целостности и конфиденциальности информационных компонент СИИ

Требования целостности (+доступность)

- Недостаточная репрезентативность (объём и вариативность) и точность обучающих и тестовых НД
- Преднамеренное и естественное искажение входных данных
- Преднамеренное внесение изменений (уязвимостей) в обучающие и тестовые НД

Требования конфиденциальности

- Получение доступа к НД злоумышленников и других заинтересованных лиц (например, доступ разработчиков к тестовым НД)
- Компрометация данных в результате недооценки уровня их конфиденциальности на этапе разработки СИИ

Требования к информационным компонентам СИИ

Информационная компонента	Причины, связанные с нарушением целостности n-й компоненты	Причины, связанные с нарушением конфиденциальности n-й компоненты
1. Функциональные характеристики (ФХ)	$i1$ Неполный набор, некорректные пороговые значения и весовые коэффициенты	$c1$ Компрометация ФХ (для систем безопасности и др., предполагающих возможность активного противодействия)
2. Предусмотренные условия эксплуатации (ПУЭ)	$i2$ Неполный перечень факторов эксплуатации, расхождение реальных условий эксплуатации с ПУЭ	$c2$. Компрометация ПУЭ (для систем безопасности)
3. Типовые модели МО, эталонные архитектуры	$i3$ Манипуляции с моделями: размещение злоумышленниками в открытом доступе моделей МО с программными закладками, реализующими НДВ	$c3$ Извлечение моделей: нарушение конфиденциальности сведений о моделях МО, использованных при разработке СИИ
4. Обучающие НД	$i4$ Манипуляции с обучающими НД: «атаки отравления» (poisoning), каузативные (causative) атаки	$c4$ Извлечение сведений об обучающих НД с целью повышения эффективности реализации атак на СИИ и извлечения сведений о объектах, данные которых использовались при обучении СИИ
5. Спецификации требований к моделям МО	$i5$ Неполные и искаженные спецификации, обусловленные заблуждениями относительно законов и закономерностях, присущих предметной области	$c5$ Использование злоумышленниками спецификаций для реализации атак, противоречащих принятым интерпретируемым моделям
6. Тестовые НД	$i6$ Искажение тестовых НД, низкая репрезентативность, смещённость тестовых НД	$c6$ Извлечение сведений о тестовых НД заинтересованными лицами (например, недобросовестными разработчиками)
7. Данные для дообучения СИИ	$i7$ Смещение обучающего НД вследствие эксплуатации в однотипных условиях	$c7$ Компрометация сценариев применения (для ВВСТ, систем безопасности)
8. Исходные данные	$i8$ Прямые манипуляции с входными данными, «состязательные примеры», непрямые манипуляции (воздействие на входы сенсоров СИИ)	$c8$ Извлечение исходных данных: «атаки инверсии модели» (model inversion)
9. Результаты обработки	$i9$ Искажение результатов обработки при их отображении, хранении, использовании в процессе тестирования СИИ и т.п.	$c9$ Извлечение выходных данных: «атаки инверсии модели» (model inversion), подмена результатов тестирования для завышения возможностей или дискредитации СИИ

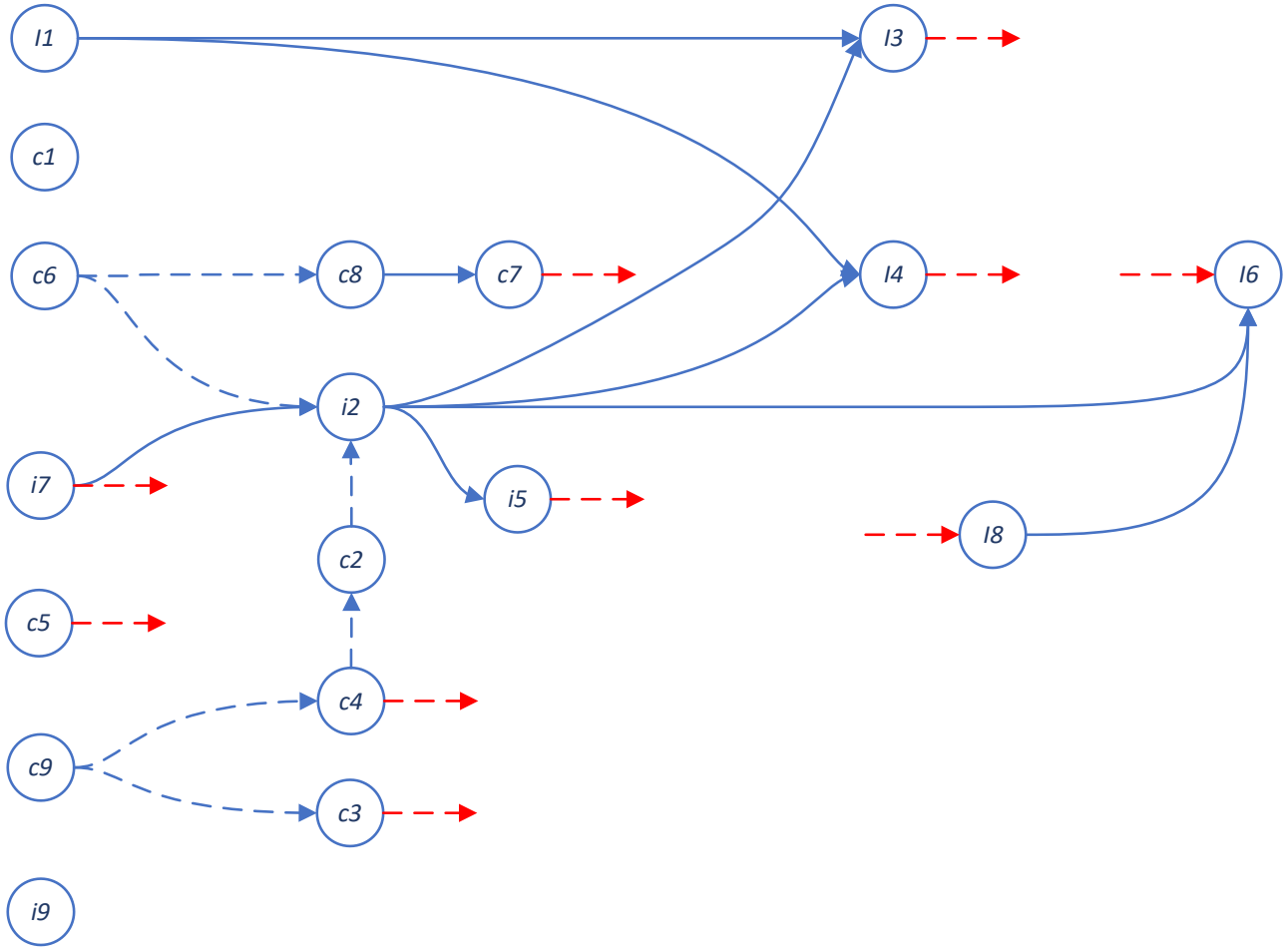
Негативные эффекты, обусловленные несоответствием требованиям к информационным компонентам СИИ

Нарушения функциональности

- 1. Деградация функциональных характеристик СИИ
- 2. Рост погрешности оценки функциональных характеристик

Нарушения конфиденциальности

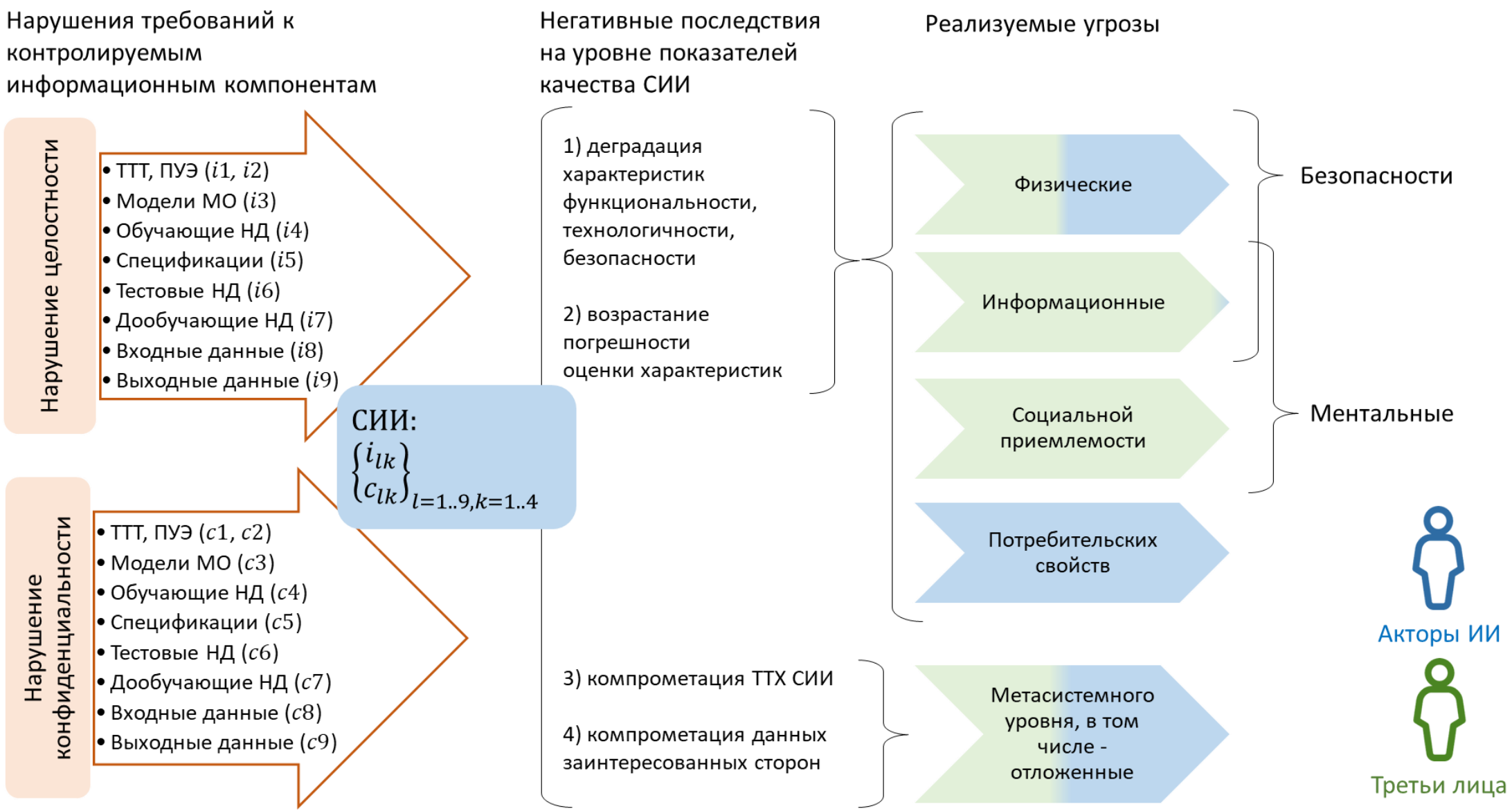
- 3. Компрометация данных о СИИ
- 4. Компрометация данных третьих лиц, использованных при создании СИИ



Риски, обусловленные деградацией функциональных характеристик ИИ

Вид угроз, обусловленных нарушением функциональной корректности СИИ	Категория заинтересованной стороны	
	Лица, непосредственно участвующие в создании и применении СИИ (акторы ИИ)	Третьи лица
1 Угрозы жизни и здоровью людей, экологические угрозы	1.1 Потребители, разработчики и поставщики (собственная безопасность, дополнительные требования гос. регуляторов)	1.2 Общество в целом и регуляторы (безопасность общества и окружающей среды)
2 Угрозы информационной безопасности в отношении заинтересованных сторон	Нет	2.2 Общество в целом и государственные регуляторы (защита персональных данных, предотвращение деструктивных информационно-психологических воздействий)
3 Нарушение этических и других норм «мягкого» права	Нет	3.2 Общество в целом (социальная приемлемость создания и применения СИИ)
4 Неопределенные потребительские свойства, не влияющие непосредственно на безопасность жизни и здоровья людей, экологическую безопасность	4.1 Потребители (функциональные характеристики, определяющие возможность применения СИИ по назначению), разработчики и поставщики (характеристики конкурентоспособности СИИ)	Нет

Модель рисков при создании и применении СИИ



Национальная система оценки соответствия в области искусственного интеллекта

Росстандарт

ТК164 — Технический Комитет 164

ГОСТ Р, ПНСТ

Перспективная программа стандартизации по приоритетному направлению «Искусственный интеллект»

Комплекс национальных стандартов в области ИИ

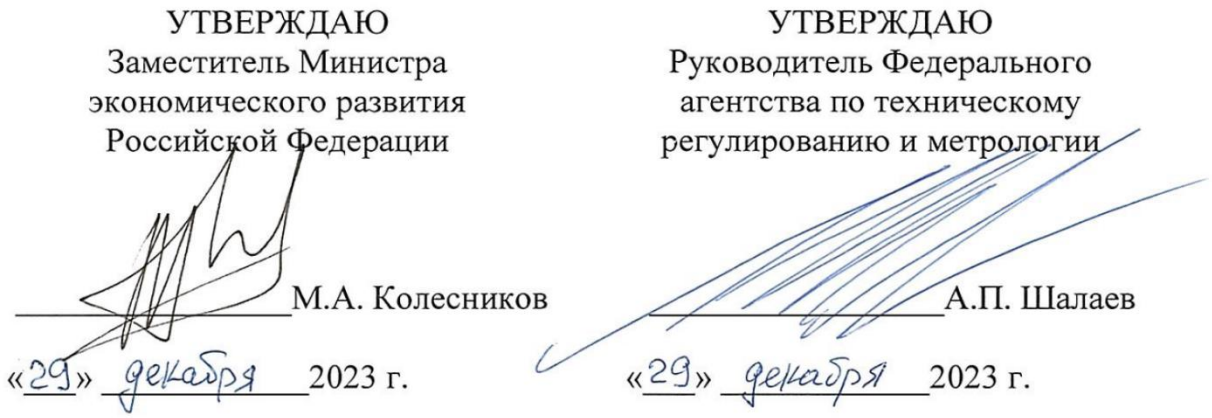
Информационный портал www.tc164.ru

Потребители и регуляторы ИИ

СДС «Интеллометрика»

Разработчики ИИ

Перспективная программа стандартизации по приоритетному направлению «Искусственный интеллект» на 2021-2024 годы



УТВЕРЖДАЮ
Заместитель Министра
экономического развития
Российской Федерации

М.А. Колесников

«29» декабря 2023 г.

УТВЕРЖДАЮ
Руководитель Федерального
агентства по техническому
регулированию и метрологии

А.П. Шалаев

«29» декабря 2023 г.

ПЕРСПЕКТИВНАЯ ПРОГРАММА СТАНДАРТИЗАЦИИ ПО ПРИОРИТЕТНОМУ НАПРАВЛЕНИЮ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ» НА 2021–2024 ГОДЫ

Разработана в рамках федерального проекта «Искусственный интеллект» в декабре 2020 и актуализирована в декабре 2023 года.

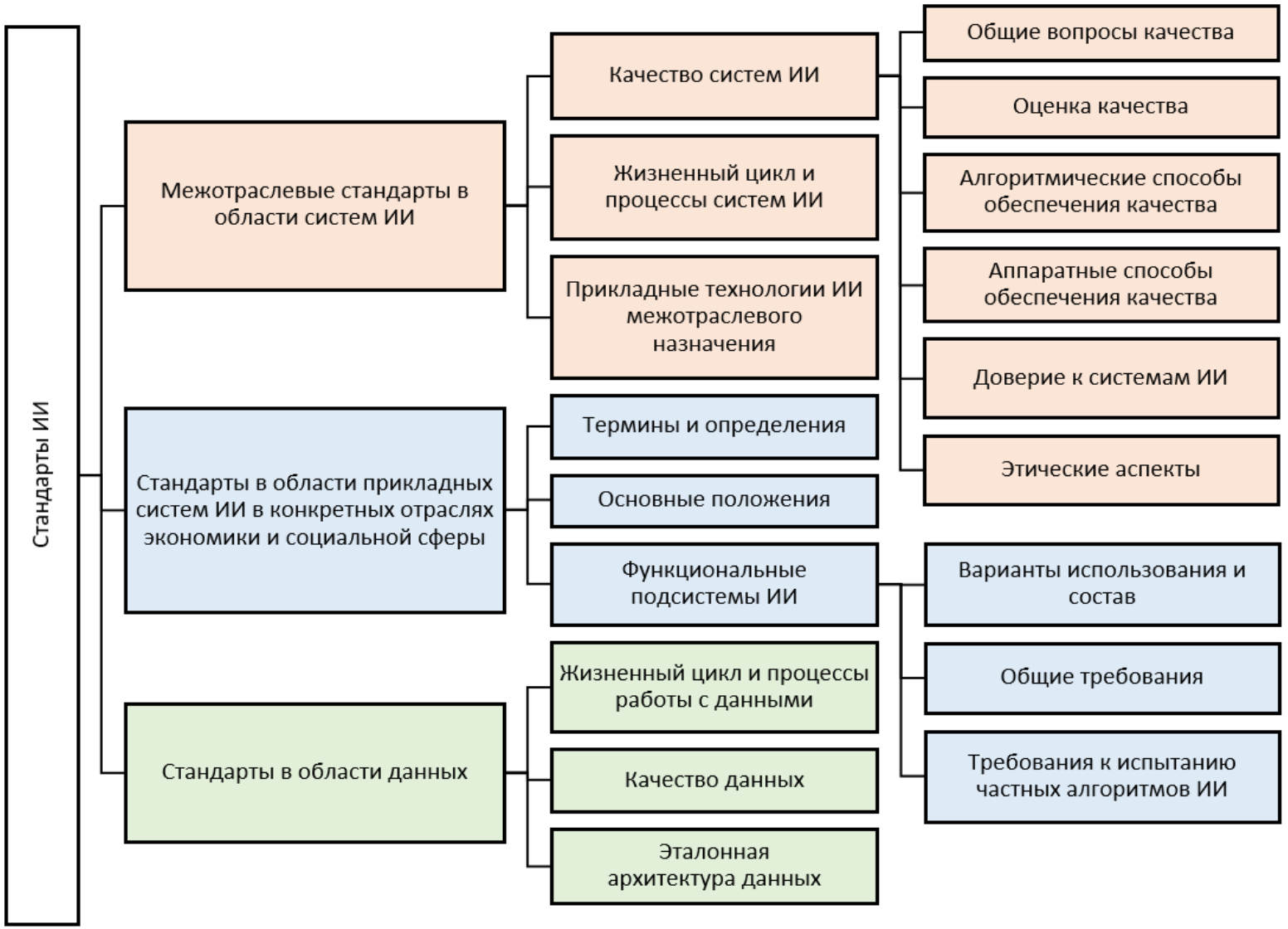
Включает:

1. Описание принципов стандартизации ИИ;
2. Стандарты общего назначения (как разработанные на основе международных, так и разрабатываемые впервые)
3. Метрологические стандарты, направленные на унификацию способов измерения функциональных характеристик прикладных технологий ИИ в основных отраслях экономики и социальной сферы

Основные цели стандартизации ИИ

1) Обеспечение гарантий функциональной корректности СИИ в реальных условиях эксплуатации, в том числе – при дообучении СИИ в процессе эксплуатации и при автоматизации процессов обработки информации, связанных с заменой человека-оператора
2) Разработка методов и средств оценки и подтверждения безопасности СИИ, в том числе – в отношении третьих лиц (не участвующих непосредственно в эксплуатации систем), включая:
   - обеспечение физической безопасности СИИ для окружающих людей, природной среды и материальных активов (например, в случае беспилотного транспорта)
   - обеспечение специальных требований в области информационной безопасности СИИ
   - оценку уровня социальной приемлемости СИИ, в том числе – этических последствий разработки и применения этих систем
3) Обеспечение терминологического единства
4) Унификация форматов представления данных, необходимых для создания и применения СИИ, обеспечение интероперабельности информационных систем
5) Фиксация вариантов использования и лучших практик создания и применения СИИ при решении различных прикладных задач в отраслях экономики и социальной сферы

Комплекс национальных стандартов в области ИИ



Межотраслевые стандарты

- Требования к аппаратным и программно-алгоритмическим средствам, используемым для создания доверенных систем ИИ

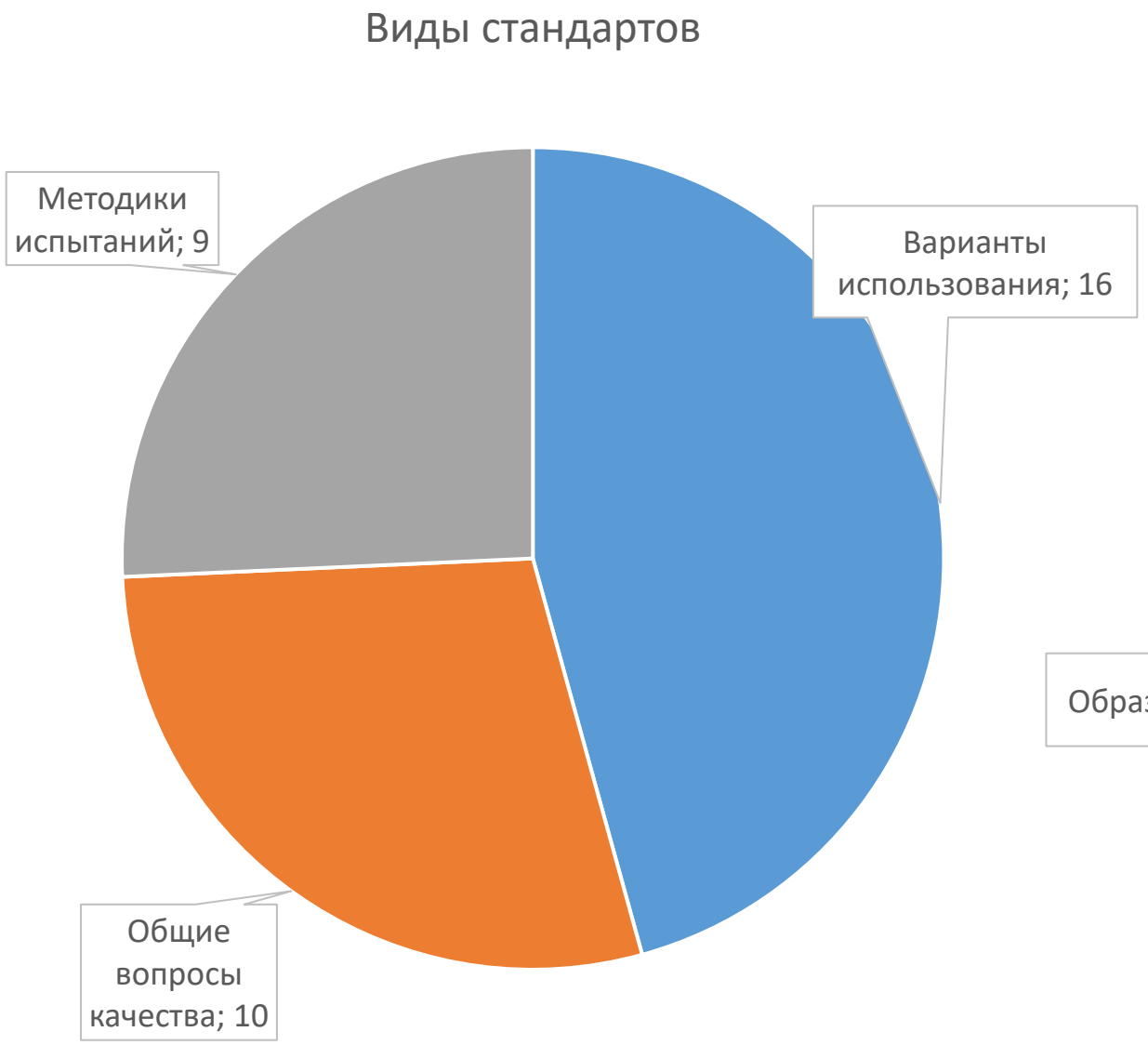
Отраслевые стандарты

- Требования к унифицированным процедурам оценки качества прикладных систем ИИ

Данные

- Требования к данным, используемым для создания доверенных систем ИИ

Стандарты ИИ, утвержденные в 2023 году



Правила функционирования СДС «Интеллометрика»

Зарегистрированы Росстандартом в едином реестре систем добровольной сертификации 26.12.2023, свидетельство № РОСС RU.В2915.04ВШЭ0

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

СИСТЕМА ДОБРОВОЛЬНОЙ СЕРТИФИКАЦИИ В СФЕРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА «ИНТЕЛЛОМЕТРИКА» (СДС «ИНТЕЛЛОМЕТРИКА»)

УТВЕРЖДАЮ

Ректор НИУ ВШЭ

_____Н.Ю. Анисимов

«21» декабря 2023 г.

ПРАВИЛА ФУНКЦИОНИРОВАНИЯ СИСТЕМЫ ДОБРОВОЛЬНОЙ СЕРТИФИКАЦИИ В СФЕРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА «ИНТЕЛЛОМЕТРИКА»

Правила предназначены для применения всеми участниками системы и другими заинтересованными юридическими и физическими лицами.

Устанавливают:

1. Объекты оценки соответствия
2. Организационную структуру и функции участников
3. Принципы функционирования системы
4. Правила и порядок проведения работ по сертификации
5. Требования к экспертам и испытателям системы

Система добровольной сертификации «Интеллометрика»

Зарегистрирована Росстандартом в едином реестре систем добровольной сертификации 26.12.2023 (№ РОСС RU.В2915.04ВШЭ0)

Заявители

Руководящий орган

- Наблюдательный совет
- Комиссия по допуску участников СДС
- Апелляционная комиссия

Транспорт

РОСДОРНИИ

НАМИ

ГЭТ Электротранспорт Санкт-Петербурга

РЖД НИИАС

Здравоохранение

ЦЕНТР ДИАГНОСТИКИ И ТЕЛЕМЕДИЦИНЫ

Образование

Энергетика

МЭИ

Сельское хозяйство

ВИМ

Следственная деятельность

Московская академия Следственного комитета имени А.Я. Сухарева

Специализированная техника

РОССПЕЦМАШ

ДСТ УРАЛ

Промышленность

ФГАУ «ФЦПР ИИ»

Розничная торговля

РУС СОФТ

АЙТИЛЛЕКТ Инструменты для бизнеса

*перечень органов по оценке соответствия не является исчерпывающим

Отраслевые РГ по применению и стандартизации технологий ИИ

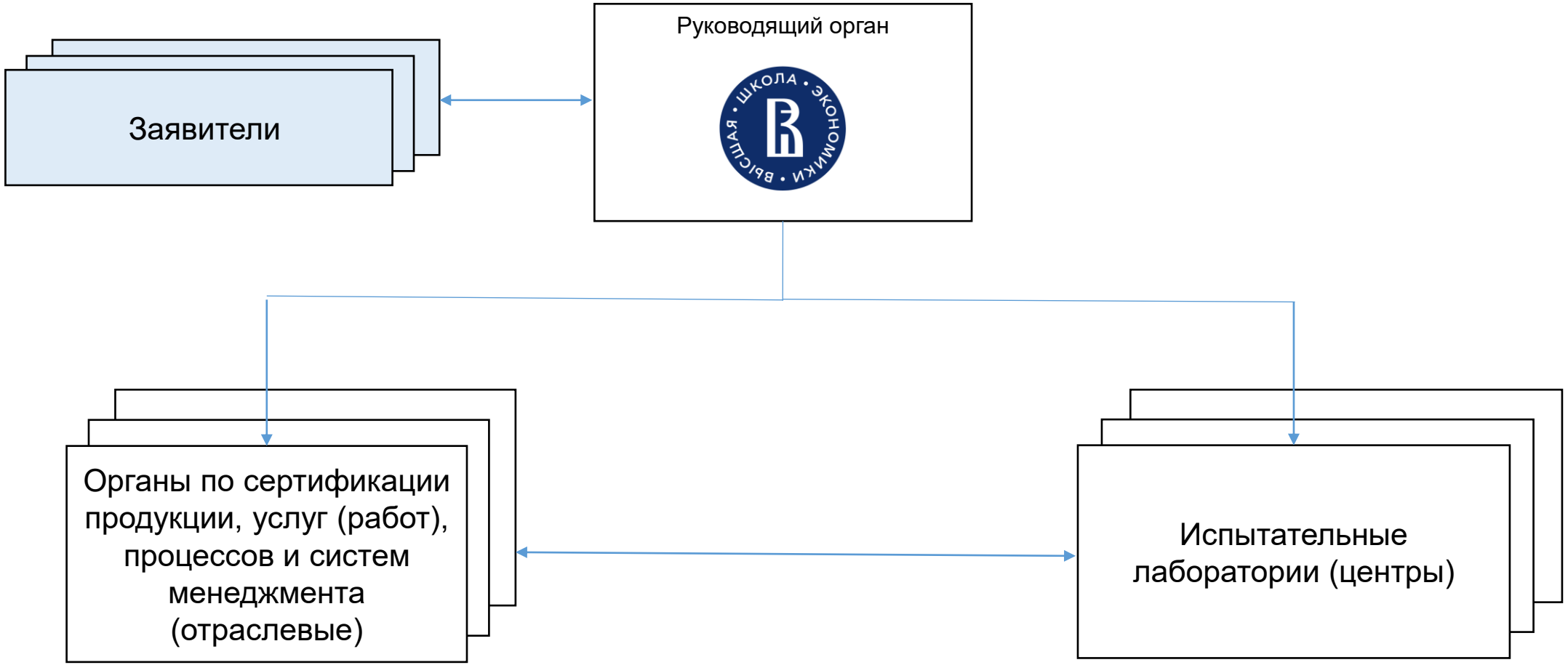
№	Область	Ведущая организация	Смежные ТК
1	Здравоохранение	НПКЦ диагностики и телемедицины Депздрава Москвы	ТК 011 «Медицинские приборы, аппараты и оборудование» ТК 436 «Управление качеством медицинских изделий» ТК 468 «Информатизация здоровья»
2	Образование	НИУ ВШЭ, ВолГУ	ТК 461 «Информационно-коммуникационные технологии в образовании»
3	Гражданская авиация	Союз авиапроизводителей России, ГосНИИАС	ТК 323 «Авиационная техника» ТК 363 «Радионавигация»
4	Специализированная техника	Ассоциация Росспецмаш	ТК 267 «Строительно-дорожные машины и оборудование» ТК 284 «Тракторы и машины сельскохозяйственные»
5	Дорожно-транспортный комплекс	РОСДОРНИИ, НАМИ	ТК 056 «Дорожный транспорт» ТК 057 «Интеллектуальные транспортные системы» ТК 278 «Безопасность дорожного движения» ТК 418 «Дорожное хозяйство»
6	Городской электрический транспорт	Горэлектротранс	ТК 315 «Автомобильный и городской электрический транспорт»
7	Водный транспорт	Отраслевой центр Маринет	ТК 032 «Водный транспорт»
8	Беспилотные летательные аппараты	НИЦ «Институт имени Н.Е.Жуковского»	ТК 323 «Авиационная техника»
9	Промышленность	ФЦПРИИ, СТАНКИН	ТК 070 «Станки»
10	Средства измерений и неразрушающий контроль	ВНИИМ им. Менделеева	ТК 053 «Основные нормы и правила по обеспечению единства измерений» ТК 371 «Неразрушающий контроль»
11	Ж/д транспорт	НИИАС	ТК 045 «Железнодорожный транспорт»
12	Ритейл	НП «РУССОФТ»	
13	Сельское хозяйство	Фед. научный агроинженерный центр ВИМ	ТК 284 «Тракторы и машины сельскохозяйственные»
14	Технические средства охраны	НИЦ «Охрана» Росгвардии	ТК 234 «Системы тревожной сигнализации и противокриминальной защиты»
15	Следственная деятельность	Московская академия СКР	ТК 134 «Судебная экспертиза»

Иерархическая модель объекта стандартизации:

интеграция СДС «Интеллометрика» с отраслевыми системами регулирования

Репрезентативность испытаний СЗИ с ИИ

Архитектура системы сертификации искусственного интеллекта



Профильные организации, выполняющие функции «третьей стороны» при сертификации технологий ИИ

Испытательная лаборатория в области средств измерений на основе технологий ИИ (ФГУП «ВНИИМ им. Д.И. Менделеева»)

ВНИИМ им. Д.И.Менделеева

Объекты испытаний:

Средства измерений, измерительные системы контроля выбросов в атмосферу

Характеристики:

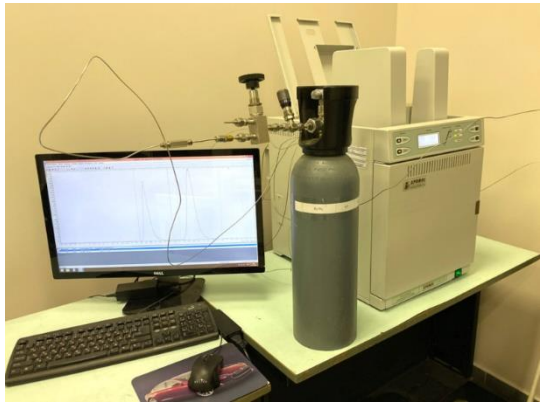
показатели функциональной корректности (ГОСТ Р 59898-2021, ПНСТ 835-2023)

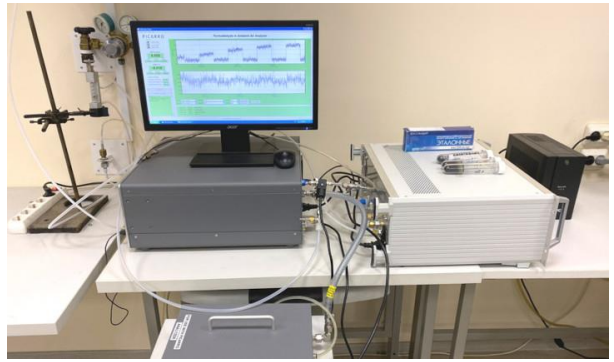
Нормативная база испытаний:

> 15 ГОСТ Р и ПНСТ

Средства испытаний:

1. Программно-аппаратный комплекс «Испытательный стенд программного обеспечения «ИБИС»
2. Государственные эталоны единиц величин





3. Стандартные образцы состава газовых смесей (поверочные газовые смеси)

Испытательная лаборатория в области применения технологий ИИ в дорожно-строительной технике (ООО «ДСТ-УРАЛ»)

Объекты испытаний:

Алгоритмы искусственного интеллекта, предназначенные для автоматизации перемещения ДСТ, идентификации оператора, предиктивной аналитики состояния ДСТ

Характеристики:

показатели функциональной корректности (ГОСТ Р 59898-2021, ПНСТ 835-2023)

Нормативная база испытаний:

> 10 ГОСТ Р и ПНСТ (утвержденных и разрабатываемых)

Средства испытаний:

1. **Наборы данных** для испытания алгоритмов: обнаружения и идентификации препятствий, ландшафтной навигации, управления движением ДСТ и др.
2. **Полигон** для испытания высокоавтоматизированной ДСТ

3. **Дорожно-строительная техника:** бульдозеры, колесные погрузчики и др.

Средства испытания технологий ИИ в энергетике (МЭИ)

Цифровые двойники

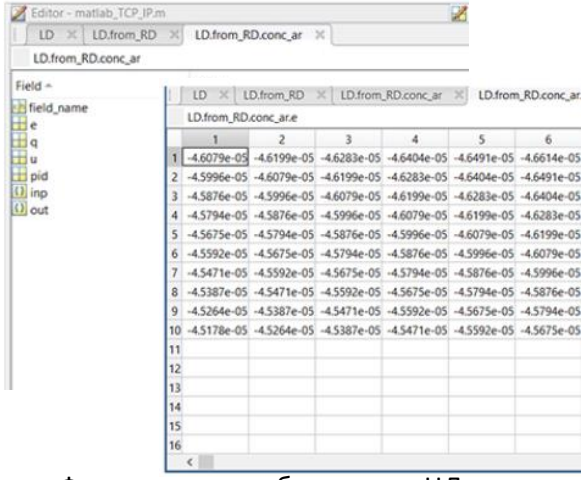
Цифровой двойник Энергосистемы

Применение цифровых двойников для проверки правильности решений синтезируемых ИИ в реальном времени перед их практической реализацией в реальных энергосистемах

Модельные испытательные комплексы и полигоны

Испытательный полигон технологий транспортировки электроэнергии и распределенных интеллектуальных энергосистем

Тестовые наборы данных



Фрагмент разработанного НД

Подготовка НД для прогнозирования изменения технического состояния оборудования, прогнозирования изменения режимов работы электрических сетей; аугментация данных с использованием цифровых двойников

Экспериментальная ТЭЦ

СИИ с гарантированной функциональной корректностью

Для предусмотренных условий эксплуатации могут быть оценены:

- доверительные интервалы и вероятности функциональных характеристик систем ИИ в предусмотренных условиях эксплуатации
- предельные интегральные риски, связанные с некорректной работой систем ИИ
- ресурсы, необходимые злоумышленнику для успешного информационного воздействия на измерительную систему с алгоритмами ИИ (опционально, при наличии активного злоумышленника)

Гарантии экономического эффекта от применения технологий ИИ с подтвержденной функциональной корректностью

Экономия на ФОТ за счет делегирования рутинных задач ИИ

Повышение качества выполнения рутинных задач за счет снижения влияния человеческого фактора

Повышение эффективности омникальной торговли с помощью ИИ (персонализированные продажи)

Снижение затрат за счет предотвращения краж

Оптимизация складских и логистических операций

Налоговые льготы для бизнеса при покупке и внедрении отечественных ИИ-решений

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Технический Комитет 164

Спасибо за внимание

Гарбук Сергей Владимирович

Председатель ТК164

www.tc164.ru