**ФЕДЕРАЛЬНОЕ АГЕНТСТВО**
**ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ**

**П Р Е Д В А Р И Т Е Л Ь Н Ы Й**
**Н А Ц И О Н А Л Ь Н Ы Й**
**С Т А Н Д А Р Т**
**Р О С С И Й С К О Й**
**Ф Е Д Е Р А Ц И И**

**ПНСТ**
***(проект)***

**Искусственный интеллект**

# ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
# УПРАВЛЯЕМОСТЬ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

**(IDT ISO/IEC CD TS 8200, IDT)**

**Издание официальное**

Настоящий проект стандарта не подлежит применению до его утверждения

**Москва**
**2023**

# Предисловие

1 ПОДГОТОВЛЕН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова» (МГУ имени М.В.Ломоносова) и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ИРИО), на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 202 г. № -ст

4 Настоящий стандарт идентичен проекту международного документа ИСО/МЭК 8200 «Информационная технология. Искусственный интеллект. Управляемость автоматизированных систем искусственного интеллекта» (ISO/IEC CD TS 8200 «Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems»)

5 ВВЕДЕН ВПЕРВЫЕ

*Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТР 1.16—2011 (разделы 5 и 6).*

*Национальный орган Российской Федерации по стандартизации собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее, чем за девять месяцев до истечения срока его действия, разработчику настоящего стандарта по адресу: 119991, Российская Федерация, Москва, Ленинские горы, д. 1 и в национальный орган Российской Федерации по стандартизации по адресу: 123112, город Москва, Пресненская набережная, дом 10, строение 2.*

*В случае отмены настоящего стандарта соответствующее уведомление будет опубликовано в ежемесячно издаваемых информационном указателе «Национальные стандарты» и журнале «Вестник технического регулирования». Уведомление будет размещено также на официальном сайте национального органа Российской Федерации по стандартизации в сети Интернет.*

## Содержание

## Введение

Методы искусственного интеллекта (ИИ) применяются в приложениях для таких сфер и отраслей как здравоохранение, образование, чистая энергетика, устойчивая жизнедеятельность и т. д. Несмотря на то, что эти методы используются для того, чтобы различные системы могли делать автоматизированные прогнозы, давать рекомендации или предлагать решения, применение систем ИИ вызвало широкий спектр проблем. Некоторые характеристики (например, недостаточная объяснимость) систем ИИ (таких как обучение и логические выводы на основе глубоких нейронных сетей) могут внести неопределенность в поведение системы ИИ. Это может привести к непредсказуемым последствиям для конечных пользователей. По этой причине очень важна управляемость систем ИИ. Этот документ в первую очередь предназначен в качестве руководства по проектированию и использованию системы ИИ с точки зрения реализации и улучшения управляемости.

Чтобы использовать преимущества ИИ устойчивым и ответственным образом, в этом документе определены характеристики и принципы управляемости системы ИИ. В этом документе описываются потребности в управляемости в контексте предметной области и укрепляется понимание управляемости системы ИИ. Управляемость – важная фундаментальная характеристика, обеспечивающая безопасность систем ИИ для конечных пользователей.

Автоматизированные системы, описанные в ISO/IEC 22989:2022, таблица 1, могут быть реализованы с использованием ИИ. Степень внешнего управления или контроля является важной характеристикой автоматизированных систем. Особенности гетерономных систем могут варьироваться от отсутствия внешнего управления до прямого внешнего управления. Степень реализации внешнего управления или контроля может использоваться для направления или управления системами на различных уровнях автоматизации, чтобы они вели себя так, как предполагалось, и в пределах функциональной безопасности. Этого можно добиться, используя функции управляемости или предпринимая определенные превентивные действия на каждом этапе жизненного цикла системы ИИ, как определено в ISO/IEC 22989:2022, 6. Под управляемостью в настоящем документе

понимается способность управлять, имеющаяся у оператора, т.е. человека либо иного внешнего агента. В документе описываются особенности управляемости (что и как происходит), но не предопределяется, кто или что осуществляет управление.

Нежелательные последствия возможны, если у системы ИИ есть возможность принимать неправильные решения или совершать действия без какого-либо внешнего вмешательства, управления или надзора. Для реализации управляемости выделяются ключевые точки наблюдения за состоянием системы и ее переходом из одного состояния в другое. Реализация вмешательства требует «передачи управления» от системы ИИ человеку или другому внешнему агенту. Конкретные точки, в которых возможна передача управления, можно продумать при разработке и внедрении системы ИИ.

Передача управления с целью внешнего вмешательства в работу системы ИИ может быть легко выполнимой в разумных пределах времени, пространства, энергии и сложности, одновременно сводя к минимуму задержку для обеих сторон (т.е. системы ИИ и внешнего управляющего агента). Заинтересованные стороны учитывают конкретные затраты на передачу управления или контроля автоматизированными системами ИИ, от чего зависит эффективность реализации управляемости в системах ИИ. Более того, так как неопределенность при передаче управления может существовать с обеих сторон, важно тщательно разработать процессы передачи управления, чтобы свести к минимуму или смягчить неопределенность и другие нежелательные последствия.

Эффективность управления и контроля подвергается тестированию и зависит от конструктивных особенностей системы и способа реализации передачи управления или контроля. Для этого необходимо определить принципы и подходы для валидации и верификации управляемости систем ИИ.

**ПНСТ**
*(проект)*

**П Р Е Д В А Р И Т Е Л Ь Н Ы Й  Н А Ц И О Н А Л Ь Н Ы Й  С Т А Н Д А Р Т**
**Р О С С И Й С К О Й  Ф Е Д Е Р А Ц И И**

**Искусственный интеллект**

# УПРАВЛЯЕМОСТЬ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Artificial intelligence. Controllability of automated artificial intelligence systems

**Дата введения – с __-____-____**

**до __-____-____**

## 1 Область применения

Этот документ определяет базовую концептуальную структуру реализации и улучшения управляемости автоматизированных систем искусственного интеллекта (ИИ), содержащую принципы, характеристики и подходы.

Документ охватывает следующие области:

— наблюдаемость состояния и перехода системы из одного состояния в другое;

— процесс передачи управления или контроля и связанные с ним затраты;

— реакция на неопределенность при передаче управления или контроля;

— подходы к верификации и валидации.

Этот документ применим ко всем типам организаций (например, коммерческим предприятиям, государственным учреждениям, некоммерческим организациям), разрабатывающим и использующим системы ИИ в течение всего их жизненного цикла.

## 2 Нормативные ссылки

Следующие документы упоминаются в тексте таким образом, что часть или всё их содержание является необходимым условием для настоящего документа. Для датированных ссылок применяется только цитируемое издание. Для

недатированных ссылок применяется последнее издание ссылочного документа (включая любые поправки).

ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

## 3 Термины и определения

ISO и IEC поддерживают терминологические базы данных для использования в стандартизации по следующим адресам:

- Платформа просмотра ISO Online: доступно по адресу https://www.iso.org/obp
- IEC Electropedia: доступно по адресу https://www.electropedia.org/c

Для целей настоящего документа применяются термины и определения, приведенные в ISO/IEC 22989:2022 и ISO/IEC 23053:2022.

3.1 **онтология** (ontology): Логическая структура терминов, используемых для описания области знаний, включая как определения применяемых терминов, так и отношения между ними.

[ISO/IEC/IEEE 24765:2017, пункт 3.2691]

3.2. **представление знаний** (knowledge representation): Процесс, который разрабатывает и создает системы символов, правила, рамки и другие методологии, используемые для выражения знаний, которые машины могут распознавать и обрабатывать.

[ISO/IEC 5392, пункт 3.18]

3.3. **вычислительная обработка знаний** (knowledge computing): Процесс получения новых знаний на основе существующих знаний и их взаимосвязей.

[ISO/IEC 5392, пункт 3.23]

3.4. **слияние знаний** (knowledge fusion): Процесс, который объединяет, комбинирует и интегрирует знания из различных ресурсов в согласованную форму.

[ISO/IEC 5392, пункт 3.21]

3.5. **управляющее воздействие** (управление) (control (noun)): Целенаправленное воздействие на процесс или действие в процессе для достижения определенных целей.

[IEC 61800-7-1:2015, пункт 3.2.6]

3.6. **управлять** (контролировать) (control (verb)): В инженерии это означает мониторинг выходных данных системы для сравнения с ожидаемыми выходными данными и принятие корректирующих мер, когда фактический результат не соответствует ожидаемому результату.

[SO/IEC/IEEE 24765:2017, пункт 3.846.1]

3.7. **управляющий агент** (controller): Уполномоченный человек или другой внешний агент, осуществляющий управляющее воздействие (3.5).

Примечание — Управляющий агент взаимодействует с точками управления системы ИИ с целью управления.

3.8. **выключение управления** (disengagement of control, control disengagement): Процесс отказа управляющего агента (3.7) от контроля над набором точек управления.

3.9. **включение управления** (engagement of control, control engagement): Процесс, в котором управляющий агент (3.7) берет на себя контроль над набором точек управления.

Примечание — Помимо взятия на себя контроля над набором таких точек, включение управления может также включать в себя подтверждение распределения управления между управляющим агентом и системой.

3.10. **система** (system): Взаимное расположение частей или элементов, которые вместе демонстрируют заявленное поведение и это же означает, что по отдельности составляющие системы этого не делают.

Примечания

1 Система иногда рассматривается как продукт или предоставляемые ею услуги.

2 На практике интерпретация значения этого термина часто разъясняется использованием ассоциативного прилагательного, например, авиационная система. В качестве альтернативы слово «система» заменяется просто контекстно-зависимым синонимом (например, самолет), хотя это потенциально искажает перспективу системных принципов.

3 Система в своей полноте включает в себя все сопутствующее оборудование, средства, материалы, компьютерные программы, микропрограммы, техническую документацию, услуги и персонал, необходимые для эксплуатации и поддержки в той степени, в которой это необходимо для самодостаточного использования в предполагаемой среде.

[ISO/IEC/IEEE 15288:2023, пункт 3.47]

3.11. **состояние системы** (состояние) (system state, state): Один из нескольких этапов или фаз работы системы.

[ISO 21717:2018, пункт 3.3]

3.12. **стабильность состояния системы** (стабильное состояние системы, стабильность) (system state stability, stable system state, stability): Особенность состояния системы, при которой параметры и наблюдаемые характеристики системы остаются неизменными в течение определенного периода времени или другого измерения, такого как пространство.

Примечания

1 Неизменность может быть определена посредством допуска изменчивости, основанного на требованиях бизнес-логики.

2 При выходе из стабильного состояния изменяются параметры или наблюдаемые характеристики системы. Это происходит независимо от того, является ли следующее стабильное состояние безопасным или небезопасным из-за того, что система входит в нестабильное состояние системы.

3 Систему можно назвать стабильной, если она находится в стабильном состоянии.

3.13. **безопасное состояние** (safe state): Состояние (3.11), не приводящее к неприемлемым последствиям или потере управления (контроля).

3.14. **небезопасное состояние** (unsafe state): Состояние (3.11), которое не является безопасным состоянием (3.13).

Примечание — Неопределенные состояния являются подмножеством небезопасных состояний.

3.15. **отказ** (failure): Потеря системой способности действовать так, как требуется.

[ISO 22166–1:2021, пункт 3.1.6]

3.16. **успех** (success): Одновременное достижение требуемой производительности по всем характеристикам.

[ISO 26871:2020, пункт 3.1.62]

3.17. **точка управления** (control point): Часть интерфейса системы, где могут быть применены управляющие воздействия (3.5).

Примечание — Такой точкой может быть функция, физическое устройство (например, переключатель) или подсистема приема сигналов.

3.18. **диапазон управления** (span of control): Подмножество точек управления (3.17), к которым могут быть применены управляющие воздействия (3.5) для достижения конкретной цели.

3.19. **интерфейс** (interface): Средства взаимодействия с компонентом или модулем системы.

3.20. **передача управления** (transfer of control, control transfer): Процесс смены управляющего агента (3.7), реализующего управляющие воздействия (3.5) над системой.

Примечания

1 Передача управления не влечет за собой совершение управляющего воздействия, а является передачей контроля над точками управления системного интерфейса между агентами.

2 Включение управления и его отключение являются двумя фундаментальными взаимодополняющими частями передачи управления.

3.21. **управляемость для пользователя** (user controllability): Степень, в которой пользователь может надлежащим образом своевременно вмешиваться в работу системы ИИ.

[ISO/IEC 25059, пункт 3.2]

3.22. **возможность вмешательства** (intervenability): Степень, в которой оператор может своевременно вмешиваться в работу системы ИИ, чтобы предотвратить вред или опасность.

[ISO/IEC 25059, пункт 3.2]

3.23. **конечный автомат** (finite state machine): Вычислительная модель, состоящая из конечного числа состояний (3.11) и переходов между этими состояниями, возможно, с сопутствующими действиями.

[ISO/IEC/IEEE 24765:2017, пункт 3.1604]

3.24. **переход состояния системы** (переход) (system state transition, transition): Процесс, заключающийся в том, что система переходит из одного состояния (3.11) в другое состояние или в то же самое состояние.

Примечание — Переход происходит, когда удовлетворяется определенное условие, включая вмешательство управляющего агента.

[ISO/IEC 11411:1995, пункт 2.2]

3.25. **затраты на управляющее воздействие** (cost of control): Затраченные ресурсы и внешние воздействия для осуществления управления системой ИИ.

Примечания

1 Ресурсы включают в себя время, пространство, энергию, материалы и любые другие расходные материалы.

2 Внешние воздействия включают в себя все возможные эффекты и побочные эффекты управления, т. е. изменение среды.

3.26. **отчет о завершении тестирования** (итоговый отчет о тестировании) (test completion report, test summary report): Отчет, в котором содержится сводная информация о проведенном тестировании.

[ISO/IEC/IEEE 29119–1:2022, пункт 3.87]

3.27. **процесс** (process): Набор взаимосвязанных или взаимодействующих действий, которые преобразуют входные в выходные сигналы.

[ISO/IEC/IEEE 15288:2023, пункт 3.XX]

# 4 Обозначения и сокращения

ИИ — искусственный интеллект (artificial intelligence)
КА — конечный автомат (finite state machine)
МО — машинное обучение (machine learning)

# 5 Обзор

## 5.1 Управляемость системы ИИ

Управляемость – это свойство системы ИИ, которое позволяет управляющему агенту вмешиваться в функционирование системы ИИ. Концепция управляемости имеет отношение к следующим областям, для которых в международных стандартах предусмотрена терминология, концепции и подходы к системам ИИ:

а) **Модель качества системы ИИ**: ISO/IEC 25059 описывает управляемость пользователем как вспомогательную характеристику удобства использования. Управляемость пользователем – это свойство системы ИИ, при котором управляющий агент может вмешиваться в функционирование системы ИИ. В ISO/IEC 25059 особое внимание уделяется интерфейсу системы искусственного интеллекта, который обеспечивает управление с помощью управляющего агента, в то время как управляемость, определенная в этом документе, больше связана с функциями, не связанными с интерфейсом, которые позволяют осуществлять управление;

б) **Надежность (свойство вызывать доверие) системы ИИ**: в ISO/IEC TR 24028 управляемость описывается как свойство системы ИИ, которое

способствует установлению доверия. Управляемость, описанная в ISO/IEC TR 24028, может быть достигнута за счет предоставления механизмов, с помощью которых оператор может взять на себя управление системой ИИ. ISO/IEC TR 24028 не дает определения управляемости. Определение управляемости в настоящем документе подразумевает то же самое, что описывается в ISO/IEC TR 24028;

в) **Функциональная безопасность системы ИИ**: ISO/IEC TR 5469: – термин «управление» используется в двух разных значениях:

1) Управление риском: это значение относится к повторяющемуся процессу оценки риска и снижения риска. Термин «контроль» относится к контексту управления. Это значение отличается от термина «управление», определенного в этом документе;

2) Контрольное оборудование: это значение относится как к управлению оборудованием, а так и к необходимости управлять оборудованием, которое имеет определенный уровень автоматизации. Значение контроля в ISO/IEC TR 5469 такое же, как и в этом документе;

г) **Управление рисками ИИ**: в ISO/IEC 23894 термин «управление» используется в контексте управления организацией, что означает способность организации влиять или ограничивать определенные виды деятельности, определенные как источники риска. Это значение отличается от значения управления или управляемости в этом документе;

д) **Концепции и терминология ИИ**: в этом документе используется определение управляемости из ISO/IEC 22989.

Управляемость имеет решающее значение для систем ИИ, базовые методы реализации которых не могут обеспечить полную объяснимость или проверяемое поведение. Управляемость может повысить способность системы вызывать доверие, включая ее надежность и функциональную безопасность.

Независимо от уровня автоматизации системы ИИ, управляемость системы ИИ важна, поэтому внешний управляющий агент может гарантировать, что система ведет себя так, как ожидается, и предотвратить причинение ею вреда.

Проектирование и реализацию управляемости системы ИИ можно рассматривать и реализовывать на каждом этапе жизненного цикла системы ИИ, определенном в ISO/IEC 22989:2022, 6.

Управляемость является технической предпосылкой человеческого надзора за системой ИИ, поэтому человеко-машинный интерфейс может быть технически

осуществимым и включенным в такую систему. Заинтересованные стороны системы ИИ должны учитывать и внедрять управляемость, которая может влиять на пользователей, окружающую среду и общество.

Управляемость системы ИИ может быть достигнута при выполнении следующих двух условий:

- Система способна давать управляющему агенту представление о своем состоянии (например, внутренние параметры или наблюдаемые характеристики), чтобы управляющей агент мог управлять системой.
- Система способна принимать и исполнять управляющие команды от управляющего агента, что вызывает переходы системы из одного состояния в другое

### 5.2 Состояние системы

В системе взаимодействующие элементы могут обмениваться данными и способствовать функционированию друг друга. Эти взаимодействия могут привести к различным конфигурациям значений внутренних параметров системы и, следовательно, к различным наблюдаемым характеристикам.

Система может находиться во множестве различных состояний. Система может находиться во множестве различных дискретных состояний, в которые отображается непрерывное пространство параметров системы. При проектировании различных состояний системы применимы следующие минимальные рекомендации:

- Состояние должно быть значимым для бизнес-логики системы.
- Продолжительность состояния должна быть достаточной для того, чтобы можно было проводить тесты и конкретные операции с этим состоянием.
- Состояние должно быть доступно для наблюдения квалифицированными заинтересованными сторонами с помощью технических средств, таких как ведение системного журнала, отладка, точки останова и т. д.
- Вход в состояние должен быть возможен через набор определенных операций в системе.

Состояния системы ИИ можно установить на этапе проектирования и разработки в жизненном цикле системы ИИ, как описано в ISO/IEC 22989:2022, рис. 3. Установление состояний системы ИИ важно для реализации управляемости и, следовательно, может повлиять на способность системы ИИ вызывать доверие. Состояния системы ИИ можно разделить на следующие три категории в

соответствии с конструктивными особенностями и предъявляемыми к ней требованиями:

- безопасные и небезопасные;
- работающие как предполагается или нет;
- любые другие категории, имеющие значение для эксплуатации, тестирования и технического обслуживания системы.

Система может находиться в безопасном или небезопасном состоянии и при этом работать правильно или нет. То есть не всегда существует прямая связь между правильной работой только в безопасных состояниях и отказом в небезопасных состояниях (или во время перехода между этими состояниями или через них). Успешное и неудачное выполнение задачи зависят от структуры системы и внутрисистемных переходов в пределах и между безопасным и небезопасным состояниями, что является важной частью проектирования и разработки системы.

***Пример. В банковском сервисе для оценки заявок на получение кредита используется система ИИ. Операция одобрения кредита заблокирована этим компонентом и, следовательно, завершилась неудачей из-за прогнозируемого риска для погашения кредита. Такой сбой не означает, что система переходит в небезопасное состояние.***

## 5.3 Переход системы из одного состояния в другое

### 5.3.1 Цель перехода системы из одного состояния в другое

Цель перехода системы из одного состояния в другое – это конечное подмножество возможных состояний системы, которые приемлемы для пользователей в соответствии с набором пользовательских требований. Цель этого перехода должна быть определена во время проектирования и разработки, а переходы в целевое состояние должны подвергаться верификации и валидации во время тестирования системы. Это справедливо и для систем ИИ.

Внедрение и повышение управляемости системы ИИ зависит от обеспечения или увеличения уверенности в том, что система ИИ может достичь заданного целевого состояния. Дизайнеры, разработчики, менеджеры, пользователи и любые другие заинтересованные стороны системы ИИ должны определить следующие атрибуты предполагаемого целевого состояния:

- полнота;
- стабильность.

### 5.3.2 Критерии перехода системы из одного состояния в другое

Система ИИ не обязательно должна находиться в целевом состоянии. Однако целевые состояния должны быть достижимы при определенных обстоятельствах с помощью конкретных действий. Действия могут в себя включать:

- внешнее управление через системные операции;
- автоматический переход самой системой в другое состояние при выполнении заданных условий;
- принудительный переход системы в другое состояние под воздействием внешнего события.

Есть две критические характеристики механизма запуска перехода состояния:

- достаточное условие, которое само по себе вызывает переход, пока это условие выполняется;
- необходимое условие, которое необходимо выполнить для того, чтобы произошел переход состояния. Выполнение необходимого условия само по себе не гарантирует, что переход произойдет.

При переходе из одного стабильного состояния в другое стабильное состояние система проходит, по крайней мере, через одно нестабильное состояние, независимо от того, является ли целевое состояние или достигнутое в итоге стабильное состояние безопасным или небезопасным.

### 5.3.3 Процесс перехода системы из одного состояния в другое

После срабатывания механизма запуска может произойти переход состояния системы. Для разных систем ИИ их процессы перехода состояний могут быть разными. Можно выделить общие подпроцессы перехода состояний. Процесс перехода состояния системы содержит два подпроцесса:

а) Запуск: после выполнения условия для срабатывания механизма запуска в системе может быть активирован набор внутренних операций в соответствии с конфигурациями или реализацией бизнес-логики. Такие операции могут включать в себя запуск функций, настройку параметров системы, выделение или освобождение

ресурсов, а также другие действия, которые система может выполнять внутри себя, чтобы достичь своего следующего определенного состояния. Подпроцесс запуска может быть коротким по времени, вплоть до того, что его трудно зафиксировать или даже записать, и зависеть от определений состояния системы и их реализации.

***Пример 1. Процесс обучения в рамках глубокого обучения завершается, и изменение параметров модели в памяти прекращается. В зависимости от конфигурации обучения можно активировать сохранение этой модели. Соответственно, система переходит из состояния «обучение модели» в состояние «сохранение модели». Для этого активируются необходимые функции (запись на диск) и выделяются ресурсы (место на диске).***

б) Адаптация: изменение состояния системы ИИ может изменить среду, в которой работает эта система, или объекты, с которыми она работает. Как следствие, такая среда и объекты могут воздействовать на систему в результате их взаимодействия с системой. Эти воздействия могут привести к неустойчивому периоду адаптации, когда системе приходится корректировать внутренние параметры, чтобы войти в предполагаемое состояние. Подпроцесс адаптации не является необходимостью, которую включает в себя каждый процесс перехода состояния системы.

***Пример 2. Основанная на ИИ система транспортного средства автоматически меняет его состояние с низкой скорости на высокую. При разгоне может изменяться сопротивление (со стороны земли, воздуха и т. д.) и устойчивость хода. Чтобы справиться с этим, параметры в подсистемах (таких как электронная программа стабилизации) могут быть скорректированы. Как только целевое состояние (высокая скорость) достигнуто, подходы к настройке, применяемые в подпроцессе адаптации, могут быть остановлены.***

### 5.3.4 Эффекты

Эффекты перехода системы ИИ из одного состояния в другое – это текущие состояния самой системы или дополнительный набор действий, которые необходимо предпринять системе или ее управляющему агенту. Возможны два типа эффектов:

а) В случае успешного перехода состояния: когда система успешно переходит из своего текущего состояния в ожидаемое, у неё появляется возможность обслуживать клиентов и/или предотвращается её переход в опасное состояние. Это положительный эффект перехода состояния.

б) В случае неудачного перехода состояния: когда системе не удается перейти в ожидаемое состояние, можно запросить восстановление в исходное состояние с помощью настроенных операций или определенных команд. Предполагается, что система повторит запрошенный переход состояния или останется в исходном состоянии. На это может расходоваться дополнительное время, операции, мощность и другие ресурсы. Это негативный эффект перехода состояния.

#### 5.3.5. Побочные эффекты

Побочные эффекты могут быть вызваны изменением состояния системы и могут приводить к изменениям в среде, в которой работает система, или в объектах, с которыми взаимодействует система. Не все изменения в среде или объектах, взаимодействующих с системой, могут быть восстановлены до исходного состояния. Невозможность устранения побочных эффектов следует тщательно учитывать при использовании систем ИИ в таких областях, как обработка материалов и производство.

## 6 Характеристики управляемости системы ИИ

### 6.1 Управление и контроль над системой ИИ

Управление и контроль над системой ИИ может помочь реализовать намеченную бизнес-логику (логику предметной области) и предотвратить причинение системой вреда заинтересованным сторонам. Системой ИИ можно управлять, если реализовано хотя бы одно из следующего:

– Существуют средства, разработанные и реализованные с целью управления или контроля.
– Существуют функциональные операции (не предназначенные специально для управления или контроля), которые можно использовать для целей управления или контроля.

Управление и контроль над системой ИИ эффективны, если выполняется по крайней мере следующее:

– Управляющее воздействие проводится, когда системой можно управлять для конкретной цели с приемлемыми побочными эффектами.
– Управление осуществляется через правильный диапазон управления, основанный на точках управления, предоставляемых системой.
– Управление и контроль работает как положено.

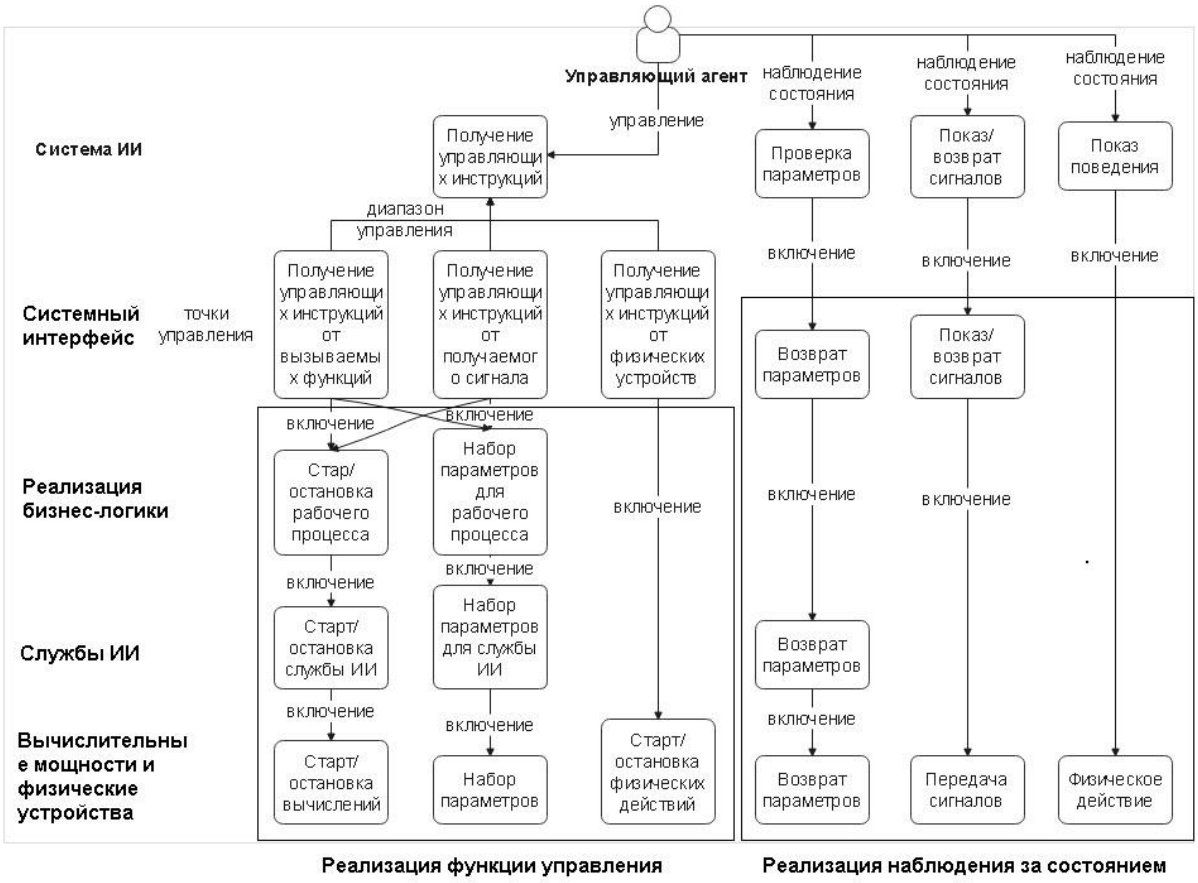


Рисунок 1 — Управление системой ИИ

Примечания

1 Диапазон управления является примером. Каждый конкретный элемент управления может соответствовать определенному диапазону управления, который настраивается, выбирается и используется.

2 Подробную информацию об обозначениях на этой диаграмме см. в ISO/IEC 19505–1.

На рис. 1. Показан вариант схемы управления системой ИИ:

а) Управляющим агентом является человек или другой внешний агент. Для конкретной цели управления управляющий агент может наблюдать за состоянием системы и отдавать управляющие инструкции системе ИИ через диапазон управления, предоставляемый этой системой. Наблюдения за состояниями системы могут быть выполнены следующим образом:

1) управляющий агент активирует функции, которые по запросу дают обратную связь с информацией о параметрах;

2) управляющий агент получает от системы сигналы, содержащие информацию о ее текущем состоянии;

3) управляющий агент наблюдает за физическим поведением системы.

б) Система ИИ разработана и реализована с интерфейсами, облегчающими управление и наблюдение за состоянием. Система ИИ может состоять из нескольких компонентов. Каждый из компонентов может предоставлять средства управления и наблюдения за состоянием:

1) Вычислительные ресурсы могут включать в себя вычислительные устройства, память, хранилища, средства передачи данных и любые другие аппаратные модули, улучшающие вычисления и обмен данными. Состояние и параметры вычислительных ресурсов можно задавать и наблюдать за ними с целью управления. Физические устройства могут включать оборудование, используемое для создания или функционирования системы ИИ. Устройства в элементах системы, такие как джойстики или рычаги переключения передач, могут обеспечивать управление и наблюдение за состоянием;

2) Службы ИИ объединяют те процессы, которые используются для реализации функций прогнозирования, рекомендаций и классификации. Параметры и статус службы ИИ можно устанавливать и наблюдать за ними с целью контроля.

3) Реализации бизнес-логики – это исполняемые программы, формирующие рабочие процессы. Каждый рабочий процесс может вызывать

службы ИИ в качестве строительных блоков. Реализации на этом уровне могут включать средства управления, соответствующие бизнес-логике.

4) Системный интерфейс может содержать подмножество заявленных функциональных возможностей для получения команд управления, предоставления значений параметров, возврата сигналов и отображения наблюдаемых характеристик. Это подмножество является точками управления системы. Для определенного элемента управления можно настроить, выбрать и использовать диапазон управления.

в) Могут существовать зависимости между функциями управления, предоставляемыми разными уровнями.

### 6.2. Процесс управления

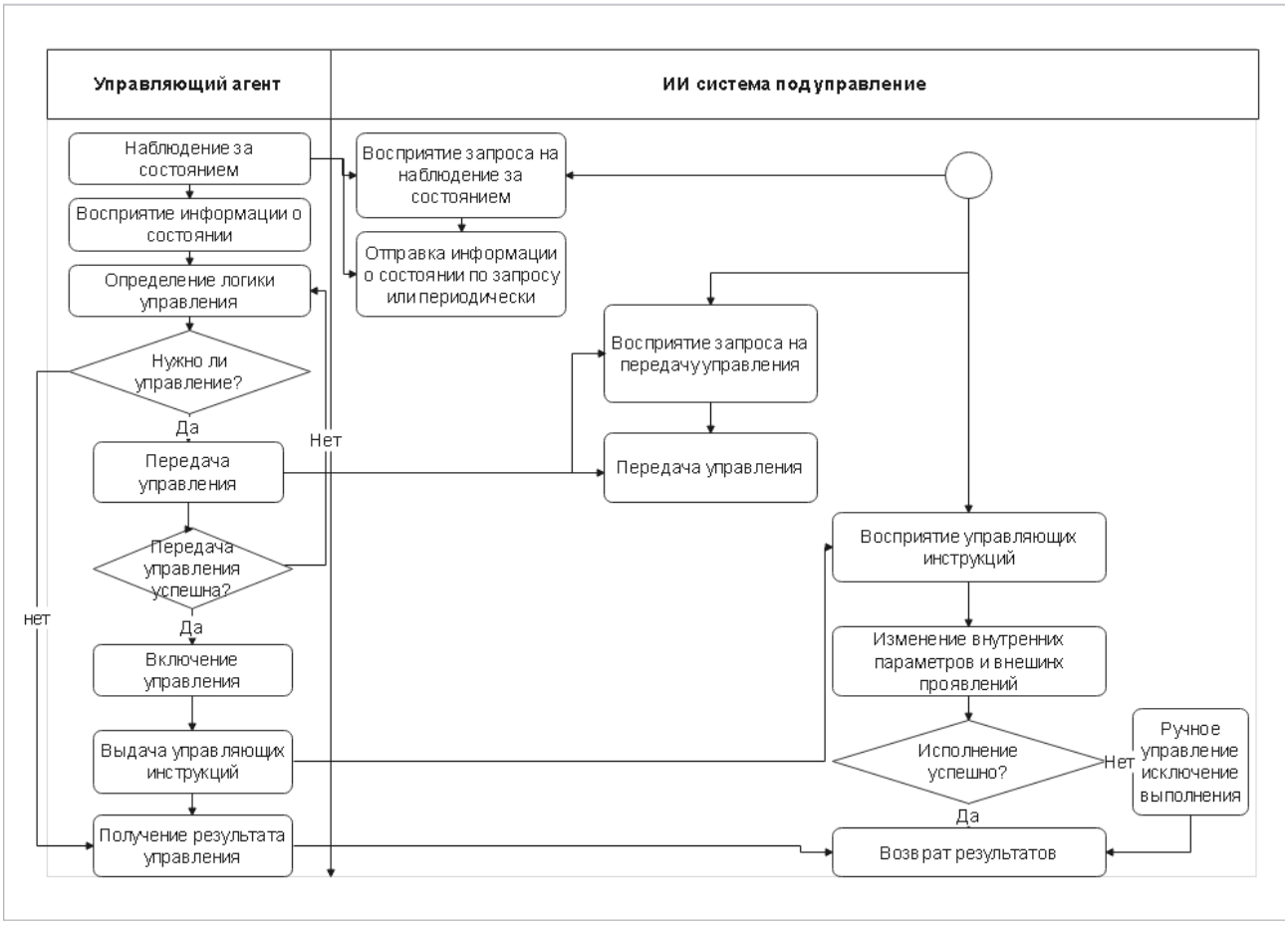


Рисунок 1 — Диаграмма действия процесса управления

Примечание — См. ISO/IEC 19505-1 для получения подробной информации об обозначениях на этой диаграмме.

В процессе управления может участвовать как управляющий агент, так и управляемая система ИИ. Общий процесс показан на диаграмме действия (см. рис. 2). Он включает следующие подпроцессы:

а) управляющий агент наблюдает за текущим состоянием управляемой системы ИИ. Это делается путем взаимодействия с интерфейсом системы ИИ. Для этого управляющий агент воспринимает сигнал с информацией о состоянии, предоставляемый системой ИИ.

б) Подконтрольная система ИИ может воспринимать следующие типы запросов:

1) Наблюдение за состоянием. При получении запроса на наблюдение за состоянием система ИИ возвращает управляющему агенту информацию о текущем состоянии. Такая информация также может периодически сообщаться управляющему агенту.

2) Передача управления. Когда получен запрос на передачу управления, содержащий обращение на передачу управления, система ИИ может передать управление авторизованному управляющему агенту.

3) Инструкции по управлению. Когда получен запрос, содержащий инструкции по управлению, система ИИ выполняет инструкции по требованию.

в) При получении информации о состоянии управляющий агент определяет логику управления. Это может запустить один из следующих вариантов:

1) Если определено, что система ИИ не нуждается в управлении, процесс управления завершается.

2) Если определено, что система ИИ нуждается в управлении, запускается подпроцесс, который подготавливает передачу управления.

г) Если управляющий агент не может выполнить намеченное управляющее воздействие из-за отсутствия диапазона управления, он запрашивает передачу управления. Это может произойти, когда необходимые точки управления не были полностью ему переданы. Если управляющий агент контролирует все необходимые точки управления для этого управляющего воздействия, подпроцесс запроса пропускается; в противном случае система ИИ передает управление управляющему агенту, благодаря чему и управляющий агент, и система ИИ соглашаются на

управление со стороны управляющего агента. При передаче управления также может потребоваться проверка подлинности и авторизации.

д) управляющий агент выдает инструкции по управлению. Получив инструкции, система ИИ изменяет свои внутренние параметры или наблюдаемые характеристики. Это может привести к двум видам результатов:

1) Если инструкции управления выполнены успешно, система ИИ дает обратную связь с результатами управляющему агенту.

2) Если инструкции управления выполняются безуспешно, система ИИ обрабатывает возможные исключения и дает обратную связь с результатами управляющему агенту.

Когда система ИИ имеет конечный набор системных состояний, ее можно смоделировать как КА (конечный автомат). Применение методов управления на основе КА возможно, когда представление передачи управления между управляющим агентом и системой ИИ осуществляется через функцию управления передачей, которая определяется кортежем:

$\Sigma (S, A, E, \gamma)$

где

S – конечное множество состояний системы;

A – набор действий;

E – множество событий;

γ – множество переходов состояний системы.

### 6.3. Точки управления

Точка управления должна быть одной из следующего списка, но не ограничиваясь ими:

- Функция. Когда система управляется с помощью программ, должны быть разработаны функции, реализующие логику управления. Для этого можно рассмотреть локальные вызовы или удаленные вызовы процедур. Функцию такого рода следует вызывать с гарантиями аутентификации и авторизации.
- Физический объект. Когда система оснащена физическими органами управления, такими как рулевое колесо на автомобиле с автоматическим управлением, следует учитывать факторы безопасности и удобства

использования, которые могут физически повлиять на эффективность и действенность управления.

- Подсистема ввода-вывода сигналов. Когда система управляется по беспроводной сети, может применяться подсистема ввода-вывода сигналов. Помимо требований к средам, таким как воздух, вода, расстояния и возможные шумы, подсистема должна также удовлетворять требованиям по своевременности и порядку управления.

В зависимости от конструкции точки управления системы могут использовать следующие аспекты:

- Специально разработанные и реализованные объекты, которые используются исключительно для управления.
- Средства, которые являются частью системных функций, но могут дополнительно использоваться для управления, такие как точка управления и функции паузы, предназначенные для отладки, но полезные для управления в определенных случаях.

Возможность вызова точек управления должна быть защищена механизмами аутентификации и авторизации. Для этого могут применяться сертификация, механизмы шифрования и даже специальные каналы управления.

***Пример. Линией автоматизированной обработки металлов на основе ИИ можно управлять с помощью подсистемы цифрового управления, а также с помощью набора физических объектов на самой производственной линии. Система искусственного интеллекта используется для анализа фотографий ключевой информации об обрабатываемом металле (например, местонахождение и положение обрабатываемой детали). Элементы управления могут включать в себя запуск, остановку и приостановку подпроцессов, выбор и смену патронов, нагрев, охлаждение, токарную и фрезерную обработку материалов, смену долот и т. д. Элементы управления системой могут быть настроены заранее и активированы в режиме реального времени через цифровую подсистему управления. Физические средства также могут использоваться, если необходим ручной и физический контроль (управление) в неотложных случаях. Для использования подсистемы цифрового управления и входа в зону физического контроля (управления) может потребоваться проверка***

***биометрической идентификационной информации управляющих агентов-людей.***

Примечание — Патрон относится к зажимному устройству с подвижными губками для удержания заготовки.

### 6.4. Диапазон управления

Диапазон управления – это подмножество точек управления, к которым можно применить конкретное управляющее воздействие. Наличие у управляющего агента диапазона управления означает, что существует соглашение (между управляющим агентом и системой) о том, что система готова воспринимать и выполнять инструкции, выданные управляющим агентом для конкретного управляющего воздействия. Поэтому, прежде чем осуществлять фактическое управление, помимо аутентификации и авторизации, необходимо заранее дополнительно проверить:

- Должна ли система признавать и выполнять команды управления от конкретного управляющего агента. Всякий раз, когда это не так, управляющий агент не может полностью выполнять намеченное управляющее воздействие. Наличие неполного диапазона управления может привести к передаче управления (см. 6.5) от системы к управляющему агенту.
- Способен ли управляющий агент держать под контролем все точки управления для предполагаемого управляющего воздействия. Всякий раз, когда это не так, должен быть подготовлен механизм обработки неопределенностей, или план такого управления может быть отменен из-за отсутствия осуществимости.

При взаимодействии с точками управления в диапазоне могут существовать правила последовательности использования этих точек. Это важно, поскольку система с точки зрения управления стремится сохранить свое состояние.

### 6.5. Передача управления

Передача управления является обязательным условием, когда внешний управляющий агент решает вмешаться в систему ИИ, в частности, чтобы

предотвратить нанесение ущерба. Процесс передачи управления позволяет управляющему агенту получить управление от любого агента, управляющего системой ИИ. Для этого следует рассмотреть процесс подготовки к передаче управления. Важные подпроцессы во время подготовки включают проверку диапазона управления, подготовку к включению управления, инициализацию стратегии обработки неопределенностей, а также оценку затрат на управление и передачу управления. Процесс подготовки к передаче показан на диаграмме активности (рис. 3) и описан следующим образом:

Примечание — См. ISO/IEC 19505–1 для получения подробной информации об обозначениях на этой диаграмме.

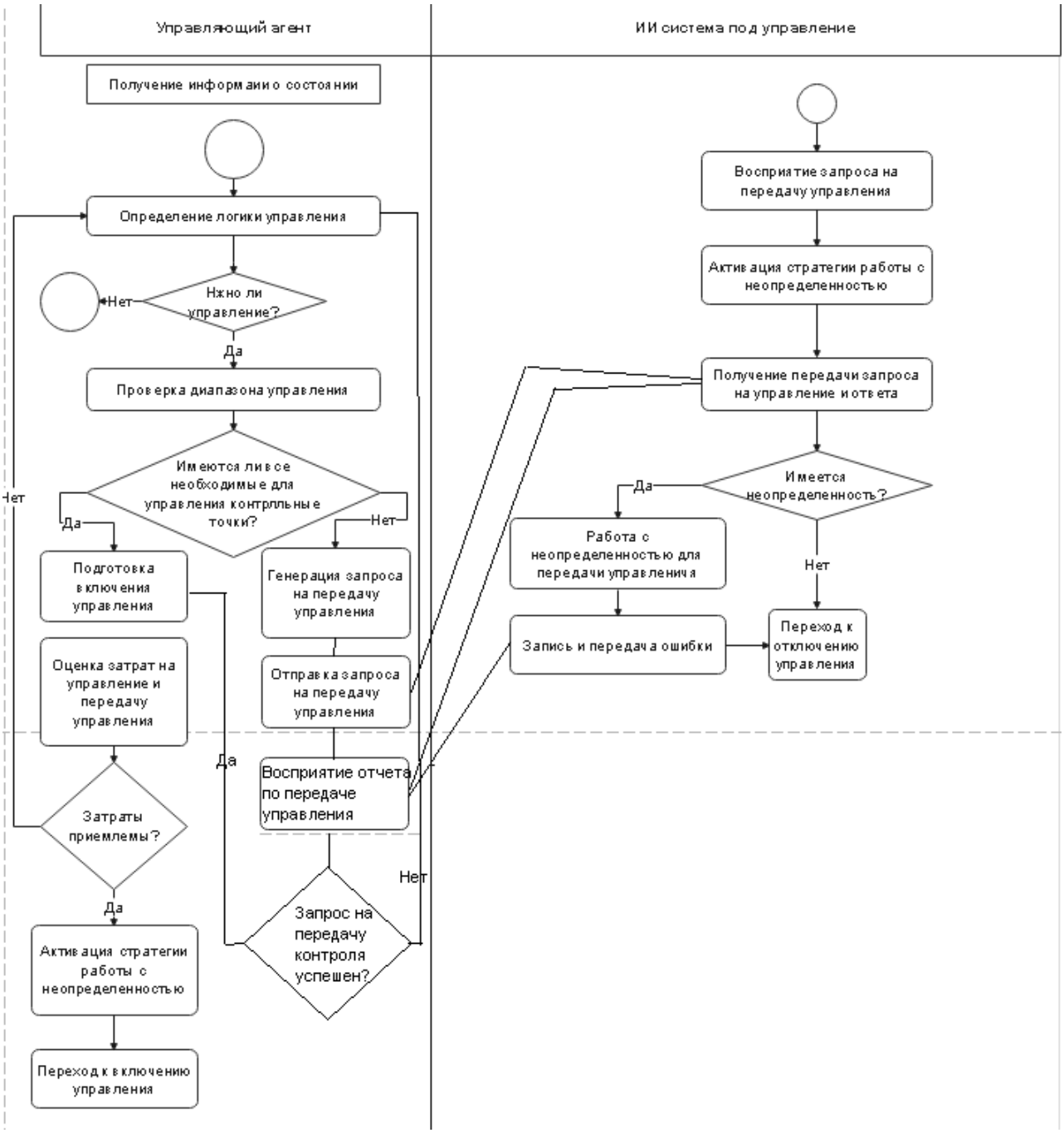


Рисунок 3 — Передача управления от системы ИИ к управляющему агенту

а) Процесс подготовки передачи управления проводится на основе определенной логики управления. Он включает в себя последовательность подпроцессов:

1) управляющий агент проверяет диапазон управления, необходимый для предполагаемого управляющего воздействия;

2) в случае, если управляющий агент не контролирует все точки управления этого диапазона, управляющий агент дополнительно формирует запрос на передачу управления перед включением управления. Такой запрос в

свою очередь делает свой запрос на предстоящую операцию на подмножестве точек управления и отправляется в систему ИИ. Получив запрос, последняя отвечает управляющему агенту подтверждением, и система начинает готовиться к своему отключению от управления. Эти действия выполняются, если управляющий агент имеет полномочия для запрошенного элемента управления;

3) в случае, если управляющий агент уже держит под контролем все необходимые точки управления для предполагаемого управляющего воздействия, действия, указанные в п. 2, пропускаются;

4) запрос на передачу управления может быть отклонен, если во время связи между управляющим агентом и управляемой системой ИИ возникают неопределенности. Неудачный запрос может инициировать переопределение логики управления, что может заставить управляющего агента сформировать другую необязательную стратегию управления;

5) если на запрос о передаче управления получен положительный ответ, происходит подготовка к включению управления. В зависимости от этого управляющий агент вырабатывает план, содержащий последовательность действий (например, перемещение в правильное положение для управления), которые необходимо предпринять, чтобы быть готовым к фактической работе;

6) затраты на управление, а также на возможную в дальнейшем передачу управления оцениваются. Во время этого подпроцесса управляющий агент собирает сведения о возможном потреблении времени, пространства, энергии и материалов, а также о влиянии на систему и соответствующие среды. Когда расчетные затраты выходят за определенный предел, логика управления может быть скорректирована. В случае отсутствия надлежащего управляющего воздействия можно применить стратегию управления по умолчанию;

7) управляющий агент также инициализирует стратегию обработки неопределенностей для обработки непредвиденных отказов во время управления и передачи управления. Подобную стратегию можно также инициализировать в системном окне ИИ.

б) Процесс передачи управления и его подготовка могут быть проигнорированы, если определено, что система ИИ не нуждается в управлении в соответствии с наблюдаемым состоянием.

### 6.6. Включение управления

Для управляющего агента важной предпосылкой для управления системой ИИ является включение функции предполагаемого управления. Включение означает выполнение последовательности действий для вовлечения в процесс. Кроме того, при выполнении определенного действия или всей последовательности действий должен выполняться набор критериев. Полезные действия включают, но не ограничиваются:

- перейти в требуемое положение;
- подключить или настроить необходимое оборудование;
- обращаться с необходимым физическим операционным оборудованием;
- запустить необходимый инструмент управления (программное обеспечение).

Ниже приведены необязательные критерии. Их можно выбирать и использовать в соответствии с требованиями к управлению:

- ограничение продолжительности включения управления;
- ограничение физического пространства, допускаемого системой ИИ для включения управления;
- приказ об ограничении включения управления при реализации многостороннего управления над точками управления;
- ограничение полномочий на включение управления при наличии требований безопасности на получение контроля над точками управления;
- полнота включения по всему диапазону предполагаемого управления.

Чтобы управлять системой ИИ, процесс включения должен быть подготовлен заранее. Он заключается в планировании последовательности действий и удовлетворении соответствующих критериев. Его можно настроить или спланировать заранее для каждого возможного элемента управления, чтобы уменьшить неопределенность и неосуществимость. Подготовка к включению управления может происходить в рамках подготовки к передаче управления, так что затраты на него могут быть оценены.

Когда процесс включения управления подготовлен, управляющий агент может предпринять запланированные действия и подтвердить с помощью управляемой системы ИИ необходимый диапазон управления. Подтверждение – это достижение

окончательного соглашения между управляющим агентом и системой ИИ. Управляющий агент объявляет об использовании точек управления, в то время как система ИИ освобождает эти точки и соглашается воспринимать и выполнять инструкции от управляющего агента.

### 6.7. Отключение управления

Отключение управления – это процесс, противоположный включению контроля. Это означает, что система ИИ вот-вот освободится и передаст управление другой стороне. Основная задача этого процесса состоит в том, чтобы выполнить последовательность действий, а затем удовлетворить ряду критериев. Полезные действия включают, но не ограничиваются:

- выйти из определенного состояния;
- снять или отключить оборудование;
- освободить физическое действующее оборудование;
- прекратить или приостановить работу управляющего инструментария (программного обеспечения).

Критерии включения управления могут быть выбраны и использованы в контексте отключения управления, но с разным значением для каждого из них:

- ограничение времени завершения отключения управления;
- ограничение физического пространства, разрешенного системой ИИ для отключения управления;
- запретить приказ об отключении управления при отказе от нескольких точек управления;
- ограничение полномочий на отключение управления при наличии требований безопасности по отказу от точек управления;
- полнота отключения на всем диапазоне предполагаемого управления.

Для передачи управления системой ИИ необходимо заранее подготовить процесс отключения управления. Это обычно инициируется стороной, которая собирается стать управляющим агентом и готовится к участию в управлении. Подготовка включает в себя составление плана возможного отключения. Планы отключения могут быть составлены заранее, чтобы снизить неопределенность и

неосуществимость. Подготовка к отключению управления может происходить в рамках подготовки к передаче управления.

Когда процесс отключения управления подготовлен, система ИИ может предпринять запланированные действия и распределить с управляющим агентом необходимый диапазон управления. Происходит достижение окончательного соглашения между обоими сторонами. Система ИИ освобождает запрошенные точки управления и начинает воспринимать инструкции от управляющего агента.

### 6.8 Неопределенность при передаче управления

Для передачи управления следует учитывать пропускную способность стороны, которая будет удерживать контроль. Пропускная способность подразумевает то, может ли управляющий агент управлять передачей управления. Это относительное понятие, которое также может зависеть от сложности передачи управления. Факторы, которые могут повлиять на пропускную способность, включают:

- количество точек управления, которое предполагается передать;
- положения точек управления при осуществлении физического контроля;
- временные ограничения, необходимые для передачи управления;
- ресурсы управляющего агента (например, интервалы времени простоя), которые можно использовать для передачи управления.

Передача управления может потерпеть неудачу, если какая-либо из сторон (управляющий агент и система ИИ) не подготовится должным образом или подвергнется непредвиденным внешним воздействиям. Неопределенность следует учитывать, когда происходит сбой, и особенно в случаях, которые могут привести к потере ресурсов, производительности или любым другим результатам и рискам, неприемлемым для обеих сторон. Типы неопределенности включают, но не ограничиваются:

- сбой связи;
- отказ подтверждения передачи управления.

В любом случае следует рассмотреть и внедрить стандартный механизм обработки неопределенностей. То есть остановить или приостановить текущее действие системы ИИ. Более продвинутый подход позволяет сохранить текущее или последнее приемлемое состояние (например, сохранить точку управления

обучаемой модели) системы, чтобы систему можно было восстановить. Это особенно полезно для повышения надежности процесса обучения.

### 6.9 Затраты на управление

#### 6.9.1 Последствия управления

Целью оценки затрат на предполагаемое управляющее воздействие является предоставление информации для определения осуществимости такого воздействия. При управлении системой ИИ могут возникнуть следующие последствия:

а) Незавершенная работа: внутренние параметры или внешнее поведение системы ИИ могут измениться, что указывает на то, что может быть затронуто предыдущее поведение системы ИИ. Когда есть данные, сообщения, материалы, которые не были полностью обработаны, такая работа может остаться в прежнем состоянии. Следовательно, может существовать риск потери данных, материалов или неполной связи.

б) Потребление ресурсов: Инструкции по управлению посредством вызовов функций, передачи сигналов или физических операций могут потребовать времени, энергии, полосы пропускания сигнала, памяти, пространства и любых других необходимых ресурсов. Это следует особенно тщательно проверить, когда управляющий агент или управляемая система имеют ограниченные ресурсы. В процессе управления следует учитывать оценку потребления ресурсов. Кроме того, может существовать дополнительное влияние (например, изменение температуры или электромагнитного поля) на среде, в которой работает система ИИ.

#### 6.9.2 Оценка затрат на управление

Затраты на управление должны быть оценены и сверены с управляющим агентом, системой ИИ, а также с окружающей средой или любыми другими объектами, которые могут быть затронуты. Поэтому следует проверить следующее:

а) Превышает ли объем ресурсов, требуемый для предполагаемого управляющего воздействия, ограничения системы.

б) Влияет ли объем ресурсов, требуемый для предполагаемого управляющего воздействия, на функционирование системы в настоящее время или в будущем.

в) Влияют ли возможные изменения в среде или объектах, с которыми работает система, на функционирование системы в соответствии с требованиями бизнес-логики.

### 6.10 Затраты на передачу управления

#### 6.10.1 Последствия передачи управления

Оценка затрат на передачу управления полезна для определения осуществимости предполагаемого управляющего воздействия. Возможны следующие последствия при передаче управления от системы ИИ к управляющему агенту:

а) Состояние выхода из-под контроля (управления): когда происходит передача управления, система ИИ (управляемая агентом) освобождает определенный набор точек управления, а новый управляющий агент задействует и удерживает их. Не исключено, что новый управляющий агент не в состоянии удерживать или управлять операцией на таких точках управления из-за возможного большого количества таких точек, сложности процесса их задействования. Пока существует одна разблокированная точка управления, которой, однако, не может управлять новый управляющий агент, система может выйти из-под контроля (управления).

б) Потребление ресурсов: Процесс передачи управления может потреблять несколько видов ресурсов, включая продолжительность времени, полосу пропускания сигнала, память, энергию и т. д.

#### 6.10.2 Оценка затрат на передачу управления

При оценке затрат на передачу управления необходимо проверить следующее:

а) Использует ли предполагаемая передача управления ресурсы, необходимые для функционирования системы.

б) Использует ли предполагаемая передача управления объем ресурсов, превышающий системные ограничения.

в) Может ли предполагаемая передача управления привести к выходу системы из-под контроля (управления).

### 6.11 Совместное управление

В системе ИИ может существовать более одного компонента, который может воспринимать и выполнять управляющие инструкции. В зависимости от конструкции системы управляющих агентов также может быть несколько. Каждый управляющий агент может выдавать управляющие инструкции одному или нескольким компонентам. Управляющие агенты или управляемые компоненты взаимодействуют для достижения цели. Совместное управление может быть задействовано в следующих случаях:

а) Несколько управляемых компонентов и один управляющий агент: система ИИ содержит несколько компонентов (например, многоагентную систему на основе ИИ), каждый из которых может воспринимать и выполнять управляющие команды извне.

б) Один управляемый компонент и несколько управляющих агентов: система ИИ (например, робот, управляемый несколькими внешними управляющими агентами-людьми) содержит компонент, который может воспринимать и выполнять управляющие инструкции от нескольких внешних управляющих агентов.

в) Несколько управляемых компонентов и несколько управляющих агентов: система ИИ (например, группа роботов, управляемая несколькими внешними управляющими агентами-людьми) содержит несколько компонентов и внешних управляющих агентов, и каждый из компонентов может воспринимать управляющие инструкции от нескольких управляющих агентов.

Для каждого вышеприведенного случая характеристики управления могут иметь следующие особенности, что указано в Таблице 1.

Таблица 1. Особенности характеристик управляемости для совместного управления

| | Несколько управляемых компонентов и один управляющий агент | Один управляемый компонент и несколько управляющих агентов | Несколько управляемых компонентов и несколько управляющих агентов |
|---|---|---|---|
| Процесс управления | Для каждого управляемого компонента и элемента | Для каждого элемента управления применяется | Для каждого управляемого компонента применяется |
| | управления применяется процесс, показанный на рисунке 2. | процесс, показанный на рис. 2. Управляющие агенты синхронизируют (например, порядок управления, получение и освобождение ресурсов) свои процессы управления. | процесс как для одного управляемого компонента и нескольких управляющих агентов. |
| Точки управления | Объединение точек управления каждого управляемого компонента | 6.3 применяется | Объединение точек управления каждого управляемого компонента |

*Продолжение таблицы 1*

| | | | |
|---|---|---|---|
| Диапазон управления | Для каждого элемента управления диапазон управления является детерминированны м. 6.4 применяется | 6.4 применяется | Для каждого элемента управления диапазон управления является детерминированны м. 6.4 применяется |
| Передача управления | Для каждого элемента управления применяется 6.5. | Применяется 6.5, то есть распределением точек управления между управляющими | Для каждого управляющего воздействия применяется 6.5, в котором |
| | | агентами должны управлять сами управляющие агенты. | распределением точек управления по управляющим агентам должны управлять управляющие агенты. |
| Включение управления | Для каждого элемента управления применяется 6.6. | 6.6 применяется | Для каждого элемента управления применяется 6.6. |
| Отключение управления | Для каждого элемента управления применяется 6.7. | 6.7 применяется | Для каждого элемента управления применяется 6.7. |

*Окончание таблицы 1*

| | | | |
|---|---|---|---|
| Неопределенность при передаче управления | Помимо неопределенностей, указанных в 6.8, следует учитывать отказы части контролируемых компонентов. | Помимо неопределенностей, указанных в 6.8, следует учитывать отказ связи между управляющими агентами. | Помимо неопределенностей, указанных в 6.8, следует учитывать отказы части управляемых компонентов, а также отказ связи между управляющими агентами. |
| Затраты на управление | Для каждого элемента управления применяется 6.9. | Для каждого управляющего воздействия применяется 6.9. Ресурсы, потраченные во время взаимодействия между управляющими агентами, должны учитываться. | Для каждого управляющего воздействия применяется 6.9. Ресурсы, потраченные во время взаимодействия между управляющими агентами, должны учитываться. |
| Затраты на передачу управления | Для каждой передачи управления применяется 6.10. | Для каждой передачи управления применяется 6.10. | Для каждой передачи управления применяется 6.10. |

# 7 Управляемость системы ИИ

## 7.1 Вызовы

В этом подразделе обсуждаются возможные проблемы, связанные с реализацией управляемости систем ИИ. Проблемы могут быть вызваны возможной неполной объяснимостью и проверяемостью, а также особенностями интеллекта самой системы ИИ.

Чтобы изучить вопрос управляемости системы ИИ, необходимо рассмотреть следующие проблемы:

а) Состояния системы ИИ должны быть наблюдаемыми и иметь возможность переходить из одного состояния в другое. Для этого система ИИ должна предоставлять функции, с помощью которых управляющий агент может наблюдать за состояниями системы или, по крайней мере, получать информацию о тех внутренних параметрах, которые имеют значение для управления. Возможность переходить из одного состояния системы в другое относится к способности системы ИИ воспринимать и выполнять авторизованные инструкции управления в любом предполагаемом случае.

б) Управляемость следующих подпроцессов системы ИИ:

1) Выполнение не полностью объяснимых процессов: для систем, которые реализуют не полностью объяснимые подпроцессы из-за отсутствия завершенного отображения связи между математическими процессами (например, математическими вычислениями, определяемыми нейронной сетью на основе глубокого обучения) и вычислительной логикой (семантически верифицируемая логика), эти подпроцессы должны быть управляемыми, чтобы можно было вмешиваться и ограничивать потенциальные опасности, связанные с непредсказуемым поведением. В этом контексте важно контролировать запуск и завершение необъяснимого подпроцесса.

2) Наблюдение за состоянием: для систем, которые предоставляют функциональные возможности для наблюдения за состоянием, подпроцессы от запроса до возврата состояния системы (внутренние параметры системы) должны быть управляемыми. В этом контексте возврат состояния системы должен быть в конечном итоге запрошен без каких-либо предварительных условий, кроме проверки полномочий.

3) Выполнение управляющих инструкций: для систем, которые выполняют управляющие инструкции от авторизованного управляющего агента, последовательность подпроцессов от получения до выполнения управляющих инструкций должна быть управляемой. В этом контексте система должна иметь возможность в конечном итоге принимать и выполнять управляющие инструкции.

4) Определение политики обучения: для системы, способной выбирать знания для изучения или определять подход к обучению (например, непрерывное обучение), подпроцессы для таких решений должны быть управляемыми. Это особенно важно для систем, чья политика обучения в дальнейшем влияет на поведение системы по отношению к человеку.

## 7.2 Требования к управляемости систем ИИ

### 7.2.1 Общие требования

7.2.1.1 Функции управляемости следует планировать на начальных этапах или на этапах проектирования и разработки жизненного цикла системы ИИ. Следует определить использование функций управляемости для реализации политики по обращению с рисками и эта политика должна быть определенной.

7.2.1.2 Поставщик системы ИИ должен предоставлять пользователям описания функций управляемости системы.

7.2.1.3 Для системы ИИ на основе машинного обучения требования к управляемости включают:

а) Начало и окончание процесса логического вывода должны быть управляемыми.

б) Для систем, содержащих последовательность операций, реализованных путем выполнения нескольких моделей машинного обучения, рекомендуется включить элементы управления паузами между выполнением ключевых последовательных моделей.

в) Должна быть включена возможность наблюдения за состояниями системы.

г) Должны быть разрешены наблюдения за входными и выходными значениями следующего:

1) вся система,

2) модуль системы,

3) конкретные нейроны, слои или структуры нейронной сети для систем, в которых используются нейронные сети;

д) Должны быть включены наблюдения за журналами выполнения модели машинного обучения и ошибками. Наличие такой информации может помочь управляющему агенту принять решение о мерах по минимизации потенциальных опасностей. Когда система ИИ обеспечивает как асинхронный, так и синхронный режимы для выполнения своих подпроцессов, рекомендуется, чтобы система реализовывала элементы управления переключением между режимами. Это позволяет системе ИИ своевременно собирать информацию о статусе выполнения подпроцесса, чтобы можно было принять управляющее решение. Шанс на управляющее воздействие может быть упущен, если асинхронное уведомление приходит с опозданием, но при этом указывает на опасность.

7.2.1.4 Для системы ИИ на основе семантических вычислений требования к управляемости включают:

а) Начало и окончание процесса рассуждения должны быть управляемыми.

б) Когда система способна выполнять автоматические рассуждения над несколькими видами представлений знаний, выбор и использование логических рассуждений должны быть управляемыми.

в) Входные данные в и выходные данные из ризонера должны быть наблюдаемыми.

г) Необходимо включить наблюдение за системными журналами и ошибками.

### 7.2.2 Требования к управляемости систем непрерывного обучения

7.2.2.1 Для систем на основе машинного обучения требования к управляемости включают:

а) Начало и окончание процесса обучения должно быть управляемым.

б) Для систем ИИ, использующих нейронные сети, во время обратного распространения должны наблюдаться значения градиента интересующей части нейронной сети.

в) Для тех систем ИИ, которые автоматически определяют содержание для изучения, выбор и изменение содержания для изучения должны быть управляемыми.

7.2.2.2 Для системы, основанной на семантических вычислениях, должны быть управляемыми:

а) выбор онтологий, которые должны быть построены, а также приоритеты новых знаний, которые должны быть объединены в процессе объединения знаний.

б) выбор онтологий, на которых выполняется вычислительная обработка знаний.

### 7.3 Уровни управляемости систем ИИ

Уровни управляемости систем ИИ включают следующие варианты систем:

а) Полностью управляемая: система ИИ в любом состоянии способна воспринимать и выполнять инструкции по управлению и наблюдению за состоянием. Система может немедленно реагировать на управляющие воздействия и наблюдения за состоянием. Исполнение управляющего воздействия (или последовательности воздействий) и наблюдения за состоянием (или последовательности наблюдений за состоянием) может быть выполнено в рамках допустимого ограничения потребления ресурсов, которое соответствует определенным требованиям. Система может достичь требуемого состояния в рамках ограничений ресурсов, включая энергию, время и циклы обработки.

б) Частично управляемая: система ИИ в определенном наборе состояний способна воспринимать и выполнять инструкции по управлению и наблюдению за состоянием. Система может немедленно реагировать на управляющие воздействия и достигать требуемого состояния. Исполнение управляющего воздействия может быть завершено в рамках допустимого ограничения потребления ресурсов, соответствующего заданным требованиям. Когда система находится в других состояниях, она может достичь требуемого состояния последовательностью управляющих воздействий, но без гарантии того, что потребляемые ресурсы являются приемлемыми.

в) Относительно управляемая: система ИИ в любом состоянии может реагировать на инструкции по управлению и наблюдению за состоянием. Система не может достичь какого-либо требуемого состояния путем выполнения одного управляющего воздействия, но может достичь любого требуемого состояния с помощью последовательности наблюдений за состоянием и управляющих

воздействий. Система не может гарантировать, что потребляемые ресурсы находятся в пределах допустимого.

г) Слабо управляемая: система ИИ в любом состоянии может реагировать на инструкции управления и наблюдения за состоянием. Система не может достичь требуемого состояния при выполнении одного управляющего воздействия. Система не может гарантировать, что она может достичь требуемого состояния с помощью последовательности управляющих воздействий и наблюдений за состоянием. Система не может гарантировать, что потребляемые ресурсы находятся в пределах допустимого.

д) Неуправляемая: состояние не распознано или не определено. Наблюдаема только часть параметров или внешнего поведения системы ИИ. Инструкции для перехода между состояниями не реализованы. Система не предоставляет никаких инструкций, которые можно использовать для достижения системой требуемого состояния.

Примечания

1 Последний уровень управляемости может быть применим к тем системам или сценариям, где не требуется управляемость.

2 Завершение функциональности системы ИИ является основным требованием, которое может быть разработано и реализовано не для управления. Эта функция не требуется в уровнях управляемости.

# 8 Проектирование и реализация управляемости систем ИИ

## 8.1 Принципы

Реализация управляемости системы ИИ может предоставить средства, используемые для предотвращения создания опасностей системой, и, следовательно, может повысить общую надежность системы. Для этого заинтересованные стороны должны учитывать следующие принципы при проектировании и разработке важнейших этапов жизненного цикла системы ИИ:

а) Получить характеристики управляемости на основе не только явно заданных требований, но и тех неявных потребностей, которые указаны в сценариях, в которых система может создавать опасности, если она не управляется.

б) Планировать функции управляемости в зависимости от функциональности системы, но реализовывать их независимо от развития функциональности системы:

1) Во время выполнения системой ИИ своих функций требуется управляемость. Реализация и использование управляемости зависят от того, какие функции выполняются.

2) Для повышения эффективности и оперативности управляемости проектирование и разработка не должны зависеть от реализации функционала системы.

в) Наблюдения за состоянием всегда являются предпосылкой управления. Если наблюдение за состоянием и управление могут быть реализованы раздельно, то эффективны для управления следующие аспекты:

1) Наблюдение за состоянием и управление осуществляются по отдельным каналам связи.

2) Наблюдению за состоянием и управлению не отводится одна и та же группа общих ресурсов.

г) Во время проектирования и разработки всегда следует учитывать элемент управления «стоп», который останавливает выполнение текущей задачи системой ИИ. Затраты на управление могут быть ценным ориентиром, но не определяющим фактором.

### 8.2 Начальный этап

На начальном этапе жизненного цикла системы ИИ следует учитывать функции управляемости:

а) Определить цели каждой функции управляемости системы ИИ, включая, помимо прочего, следующее:

1) проблемы, которые решает этот функционал управляемости;

2) потребности клиентов или возможности в контексте бизнес-логики, на которые направлен функционал управляемости;

3) метрики успешности функционала системы.

б) На основании 8.2 пункт (а) отпределите требования к каждой функции управляемости (управления или наблюдения за состоянием):

1) Для каждого взаимодействия между управляющим агентом и системой ИИ необходимо проанализировать и записать следующее:

1.1) случайная связь между инструкциями управляющего агента и поведением или внешним состоянием, которые должна демонстрировать система;

1.2.) состояние системы и управляющие воздействия, которые могут быть применены к системе, когда она находится в этом состоянии;

1.3.) состояние, в котором находится система, после управляющего воздействия.

2) На основании результата 8.2 пункт (б) подпункт (1) определить требования к функциональным возможностям управления.

3) На основании результата 8.2 пункт (б) подпункт (1) определить требования к функциям наблюдения за состоянием системы.

4) Требование может содержать функциональные и нефункциональные аспекты.

5) Каждый аспект может содержать определенные меры (см. 9.1.3 и 9.1.4) и значения, которым должна соответствовать тестируемая система ИИ.

в) Определить функциональные возможности управляемости, полезные в типичных сценариях, в которых предполагается использовать систему. Это должно быть сделано, в частности, чтобы предотвратить риск или остановить систему ИИ в случае создания ею опасностей. Диапазон определенных функций управляемости в этой работе более широк по сравнению с определением требований (см. пункт (б) выше), которые просто соответствуют спецификации системы. Заинтересованные стороны должны выполнить следующее:

1) Определение сценария управляемости обнаруживает типичные ситуации, в которых вызываются функции управления или наблюдения за состоянием системы. Для каждой такой типичной ситуации определите следующее:

1.1) ожидаемые выходные данные или поведение системы, если функциональные возможности управляемости выполняются нормально;

1.2) потенциальные опасности, которые может представлять система, если функции управляемости не выполняются нормально.

2) На основании результата 8.2 пункт (в) подпункт (1) для каждого сценария определить критерии приемлемости, включая, помимо прочего, функциональные и нефункциональные (например, эффективность работы, безопасность и стабильность) аспекты. Каждый аспект может соответствовать

набору показателей и значений, которым должна соответствовать тестируемая система ИИ.

Примечания

1 Для каждой функции управляемости, указанной в 8.2 пункт (б) и 8.2 пункт (в), определите технические характеристики наблюдения за состоянием, которые поддерживают прозрачность и подотчетность системы, поскольку управляемость является технической предпосылкой человеческого наблюдения за системами ИИ (см. 5.1).

2 Проанализируйте осуществимость каждой функции управляемости, указанной в 8.2 пункт (б) и 8.2 пункт (в). Это можно сделать с помощью проверки концепции системы ИИ. Для повышения функциональных возможностей управляемости анализ осуществимости включает, но не ограничивается элементами, указанными в 6.9.2 и 6.10.2.

На начальном этапе, определенном в ИСО/МЭК 22989:2022, 6.2, термин затраты относится к финансированию. Это отличается от термина затраты, используемого в этом документе. Последние относится к ресурсам, которые потребляют элементы управления и передачи управления. Затраты, связанные с финансированием функций управляемости, следует прогнозировать вместе со всей системой ИИ в течение жизненного цикла системы.

### 8.3 Этап проектирования

#### 8.3.1 Общие положения

Этап проектирования системы ИИ дает представление о системе, удовлетворяющей требованиям и целям, в соответствии с результатами начального этапа. Согласно ISO/IEC 22989:2022, 6.2.3, проектирование системы ИИ может включать различные аспекты, но по крайней мере, подход, архитектуру, обучающие данные и управление рисками.

#### 8.3.2 Аспект подхода

Конструктивные особенности управляемости системы ИИ зависят от предполагаемых состояний системы, поскольку управление системой

осуществляется в определенных ее состояниях. Существуют следующие модели этой зависимости. Заинтересованные стороны могут выбирать и применять их в зависимости от своего понимания необходимых состояний системы:

а) Когда проектировщик может предвидеть все состояния системы, для реализации управляемости необходимо принять инженерные методы, обеспечивающие вычислительный контроль модели за счет использования конечного числа состояний и переходов состояний. Эта структура обеспечивает полезный подход к обеспечению управляемости системы ИИ на основе надзора за внешними агентами, использующими внешние интерфейсы системы.

б) Когда разработчик не может предвидеть часть состояний системы, для реализации управляемости необходимо проанализировать каждое из них и определить состояния системы, если они распознаны и имеют смысл для управления. Для компонента ИИ, когда его возможные состояния могут быть полностью предопределены, может применяться модель (а).

Нет необходимости проектировать управляемость системы ИИ с нуля, достаточно использовать возможности вычислительных устройств, а также вспомогательное программное обеспечение, такое как набор инструментов для машинного обучения (например, в системе ИИ на основе глубокого обучения возможности элементы управления и наблюдения за состоянием определенных частей модели, таких как нейрон, слой или структура, во время прямого распространения или обратного распространения могут быть унаследованы от поддерживающего программного обеспечения).

### 8.3.3 Аспект архитектуры

Полезно разрабатывать функции управляемости в соответствии с формированием архитектуры системы ИИ, чтобы при разработке управления и наблюдения за состоянием можно было учитывать функциональные возможности и компоненты системы. Обратные вызовы (например, врезки в процедурах) могут использоваться для наблюдения за состоянием, а синхронные или асинхронные шаблоны барьеров могут помочь сделать систему ИИ доступной и удовлетворяющей критериям (см. 6.5 и 6.6) управления или передачи управления.

### 8.3.4 Аспект обучающих данных

Разработка управляемости процесса обучения может повысить эффективность и результативность использования данных обучения, а также их безопасность. Управляемость процессом обучения включает в себя:

а) контроль начала и окончания процесса обучения;

б) наблюдение за интересующими частями глубокой нейронной сети при прямом и обратном распространении;

в) контроль над выбором и изменением данных для обучения.

### 8.3.5 Аспект управления рисками

Функции управляемости должны быть разработаны таким образом, чтобы технически удовлетворять потребности запланированных мероприятий по оценке и обработке рисков для системы ИИ в соответствии с ISO/IEC 23894:2023, 6.4 и 6.5.

## 8.4 Предложения для стадии разработки

Развитие управляемости системы ИИ соответствует процессам, реализующим функциональные возможности управления и наблюдения за состоянием, включая, помимо прочего, программирование, документирование, тестирование, исправление ошибок и т. д. Целью развития управляемости системы ИИ является реализация требуемой функциональности с эффективностью и в то же время без внесения снижения или нестабильности в производительность. Для этого следует рассмотреть следующие предложения:

а) Разделите удержание и использование вычислительных ресурсов (например, памяти, переменных, пропускной способности связи и процессора) между управляемостью и функциональностью системы. Важно обеспечить адекватные вычислительные ресурсы для управления и наблюдения за состоянием, когда ожидается, что управляемость будет выполняться немедленно в случаях, детерминированных во времени.

б) Обеспечить надлежащие приоритеты для выполнения инструкций по управляемости. В системе на основе информационных технологий вычислительные задачи планируются основным программным обеспечением (например, операционной системой) с помощью унифицированного компонента. Равномерное

распределение приоритетов по управляемости и другим задачам может нести риск несвоевременного выполнения управления или наблюдения за состоянием. Это важно для тех систем ИИ, где ожидается, что управляемость будет выполняться немедленно в случаях, детерминированных во времени.

в) Используйте функции управляемости, предоставляемые уровнями системы ИИ, и избегайте избыточной инкапсуляции или повторной реализации, если это применимо. Функции управляемости системы ИИ используются через точки применения контроля (управления) и могут быть обеспечены определенными уровнями (например, служба ИИ и вычислительные ресурсы на рисунке 1). Может оказаться более эффективным и действенным использование этих базовых функций наблюдения за состоянием и контроля, поскольку масштабы их тестирования и применения могут быть в определенной степени квалифицированы.

# 9 Верификация и валидация управляемости системы ИИ

## 9.1 Верификация

### 9.1.1 Процесс верификации

Целью верификации управляемости систем ИИ является подтверждение того, соответствуют ли реализованные функции управляемости заданным требованиям. Верификация – это этап, определенный в жизненном цикле системы ИИ в ISO/IEC 22989. Верификация должна включать следующее:

Примечание — Подробное определение жизненного цикла системы ИИ приведено в ИСО/МЭК 5338, которое согласовано с ИСО/МЭК 22989:2022, 6.

а) Определите требования к функциям управляемости, которые должна обеспечивать система ИИ, включая управление и наблюдение за состоянием системы. Эта работа должна выполняться на начальном этапе (см. 8.3). Если эта идентификация не проводилась до верификации, ее следует провести перед тестированием (см. 9.1.1 пункт (б)).

б) Протестировать функциональные возможности управляемости, чтобы убедиться, что они правильно реализуют следующие требования:

1) По требованию относительно управляемости спроектировать и провести испытание. Тестовая среда, тестовые данные и конфигурация системы, ввод и вывод – это ключи, которые необходимо подготовить.

2) Тестовая среда представлена набором параметров (например, температура, влажность, пропускная способность сети, коэффициент использования процессора, взаимосвязь между процессами или потоками, время и регион), с которыми должен выполняться намеченный тест. Параметр следует учитывать, если он потенциально может повлиять на результаты контроля или наблюдения за состоянием.

3) Тестовые данные и конфигурация системы – это данные и настройки, необходимые для приведения системы в определенные состояния, чтобы тестируемый контроль или наблюдение за состоянием имели смысл и могли быть выполнены.

4) Вход относится к инструкции управления или наблюдения за состоянием.

5) Выход соответствует эффектам (5.3.4) и побочным эффектам (5.3.5), к которым может привести контроль или наблюдение за состоянием. Для элемента управления выходные данные включают возвращенные сообщения, содержащие состояния системы. Для наблюдения за состоянием выводом является состояние системы.

в) Фактические результаты функции управляемости следует сравнивать с ожидаемыми и неожиданными эффектами и побочными эффектами. Для требования по управляемости следует сравнивать как минимум функциональную корректность и эффективность. Следует учитывать и другие аспекты эффективности, требуемые в конкретном сценарии.

г) Должны быть перечислены проверенные функциональные возможности управляемости с их ожидаемыми и фактическими результатами.

### 9.1.2 Результат верификации

Процесс верификации управляемости системы ИИ должен быть задокументирован. Приложение А описывает форму, которую можно использовать

для документирования процесса верификации. Его можно использовать в отчете о завершении теста.

### 9.1.3 Функциональные испытания на управляемость

Функциональное испытание заключается в проверке того, соответствуют ли функциональные возможности управляемости системы ИИ требованиям. Нефункциональное испытание описано в 9.1.4. Поскольку системы искусственного интеллекта разрабатываются и используются в разных областях, меры функциональной корректности управляемости могут быть разнообразными и специфичными для предметной области. Для функционального испытания управляемости систем ИИ могут быть рассмотрены следующие виды мероприятий:

а) Дискретная мера: когда функция управляемости системы ИИ возвращает результаты, предопределенные в ограниченной вселенной с дискретными элементами (например, целое число, представляющее успех выполнения управления), мера заключается в определении идентичности между возвращаемыми и ожидаемыми значениями с учетом надлежащих конфигураций системы и указывается в требованиях. Мера такого типа может возвращать только логическое значение, указывающее только на идентичность.

***Пример 1. Робот для уборки полов обеспечивает управляемость функциональностью начала уборки, которая может быть активирована физической кнопкой на корпусе робота. Это управляющее воздействие может быть выполнено, когда система находится в состоянии готовности к управлению, на что указывает световой индикатор на ее корпусе. Ожидаемым результатом этого воздействия является предопределенный код, называемый началом уборки, в то время как горит индикатор. Функциональное тестирование для запуска уборки заключается в нажатии кнопки и проверке того, может ли система вернуть предопределенный код начала уборки и горит ли индикатор. В этом сценарии мера может возвращать только логические результаты.***

б) Непрерывная мера: когда функция управляемости системы ИИ должна возвращать результат с дополнительными значениями, указывающими, в какой степени выполняется управляющее воздействие или изменяется система. Мера заключается в дополнительной проверке того, соответствуют ли возвращаемые

вспомогательные значения требованиям при условии надлежащей конфигурации требований к системе и ее состояниям.

***Пример 2. Робот для уборки полов обеспечивает возможность управления движением вперед, когда он находится в состоянии уборки, чтобы предоставить пользователю гибкость для ручного управления и очистки некоторых областей с границами неправильной формы. Это управляющее воздействие может быть выполнено нажатием кнопки на удаленной панели управления и может возвращать результат в зависимости от расстояния, которое робот фактически проходит физически. Требование по этой функциональности управляемости может быть выполнено, когда робот двигается вперед и проходит 10 сантиметров. Мера для этого сценария возвращает не только оценку идентичности, но и разницу между фактическим расстоянием, которое робот проходит вперед, и требованием (10 сантиметров).***

В системе ИИ могут существовать функции управляемости, предназначенные для обеспечения функциональной безопасности (например, продолжительность реакции системы на элементы управления системы ИИ реального времени или ограничение энергопотребления определенных элементов управления системы ИИ с ограниченным источником питания). Следует рассмотреть функциональное тестирование на управляемость с ограничениями. Для этого можно применить 9.1.3 пункт (б).

### 9.1.4 Нефункциональное испытание управляемости

Нефункциональное испытание включает тесты на производительность, безопасность, стабильность, удобство использования и любые другие аспекты, которые могут повлиять на выполнение функции управляемости.

Тестирование эффективности производительности предназначено для измерения величины ресурсов, потребляемых при выполнении функции управляемости системы ИИ, и проверки ее соответствия требованиям. Это может предоставить доказательства для оптимизации дизайна и реализации управляемости системы. Типы мер включают, но не ограничиваются:

а) Продолжительность: Относится к продолжительности времени, необходимому для функции управляемости, включая подготовку (см. 6.5),

выполнение инструкции и обмен информацией о результатах управляющего воздействия. Продолжительность важных подпроцессов (например, передача управления и передача контроля) может быть проверена.

1) Для наблюдения за состоянием продолжительность может варьироваться от момента времени, когда выдается инструкция, запрашивающая состояние системы, до момента времени, когда управляющий агент получает сообщение, содержащее состояние системы.

2) Для управления существуют следующие виды значений. Их можно выбирать и применять исходя из требований:

2.1) продолжительность, которая варьируется от момента времени непосредственно перед тем, как управляющий агент отдает распоряжение, до момента времени, когда управляющий агент получает отчет;

2.2) продолжительность, которая варьируется от момента времени непосредственно перед тем, как диспетчер выдает управление, до момента времени, когда состояние системы указывает на результат управления.

б) Количество операций: относится к количеству операций, которые необходимо выполнить управляющим агентам при выполнении функции управляемости. Эта мера указывает на сложность управления и важна для тех средств управления, где применяются физические операции.

Тесты по любым другим аспектам могут быть разработаны и выполнены, если этого требует системная спецификация.

### 9.2 Валидация

#### 9.2.1 Процесс валидации

Целью валидации управляемости системы ИИ является подтверждение того, соответствуют ли реализованные функции управляемости предполагаемому использованию. Валидация требуется в соответствии с жизненным циклом системы ИИ, определенным в ISO/IEC 22989:2022, 6. Валидация включает следующее:

а) Определите сценарии, в которых вызываются функции управления или наблюдения за состоянием. Эта работа должна выполняться на начальном этапе

(см. 8.2). Если эта идентификация не была выполнена до валидации, ее следует провести до испытания (см. 9.2.1 пункт (б)).

б) Протестируйте функции управляемости в каждом сценарии:

1) Определить вход и выход системы, где используются функциональные возможности управляемости.

2) Выполните управление и наблюдение за состоянием с правильными и неправильными операциями и проверьте, может ли система давать выходные данные или вести себя нормально.

3) Тестовая среда может быть подготовлена путем использования преимуществ параметров в 9.1.1 пункт (б) подпункт (2) и введения дополнительных воздействий, с которыми система может столкнуться в конкретном сценарии (например, турбулентность, пешеходы или препятствия на дороге для проверки функций управляемости в сценарии «повернуть направо» для автоматизированной системы вождения).

4) Тестовые данные и конфигурация системы могут охватывать не только данные и настройки, требуемые 9.1.1 пункт (б) подпункт (3), но также и те, которые используются в реальных условиях. При настройке системы следует учитывать пользовательские и даже неправильные настройки.

5) Ввод относится к вводу системы, а также к инструкции управления или наблюдения за состоянием.

6) Выходные данные относятся к тем, которые указаны в 9.1.1 пункт (б) подпункт (5), а также к системным выходным данным, имеющим значение для сценария.

в) Важно проверять выходные данные и поведение системы ИИ с учетом как правильных, так и неправильных операций, что может произойти при реальном использовании системы в сценариях. Фактический результат системы следует сравнить с ожиданиями в сценариях.

г) Подтвержденные функции управляемости должны быть перечислены вместе со сценариями, а также их ожиданиями и фактическими выходными данными.

### 9.2.2 Результат валидации

Процесс валидации управляемости системы ИИ должен быть задокументирован. Приложение В описывает форму, которую можно использовать для документирования процесса валидации.

### 9.2.3 Ретроспективная валидация

Системы ИИ использовались в различных предметных областях. Не все из них рассчитывали, планировали или реализовывали функции управляемости, достаточные для использования по назначению. Для систем ИИ, которые работают в течение некоторого времени, может применяться ретроспективная валидация для планирования и реализации управляемости. Признаки, что необходима ретроспективная валидация, могут включать, но не ограничиваться следующими:

а) Опасности, возникшие во время функционирования системы;

б) оценка риска системы указывает на потенциальные риски для безопасности или этики;

в) изменено законодательство, связанное с разработкой и использованием системы;

г) Были изменены правила работы или рабочие процессы системы.

Исторические данные могут использоваться для поддержки ретроспективной валидации управляемости. На основе реальных ходовых данных выполнение и результаты управляемости должны быть проверены сценариями. Можно применять подходы, описанные в 9.2.1.

# Приложение А (справочное)

Пример исходящей документации верификации.

Таблица А.1 представляет собой пример документации для верификации управляемости системы ИИ. Таблица А.1 включает следующие столбцы:

а) Требование – это описание управляемости, которую может обеспечить система ИИ.

б) Аспекты могут быть функциональными или нефункциональными (например, эффективность работы) значимым аспектом, в котором заинтересованы заинтересованные стороны.

в) Функциональность управляемости – это описание проверенной функциональности управляемости (см. 8.2 пункт (б)).

г) Тип может быть контрольным или наблюдением за состоянием.

д) Тестовая среда представляет собой совокупность параметров среды (см. 9.1.1 пункт (б) подпункт (2)), связанных с функциональностью управляемости.

е) Вход и выход – это описания входа (см. 9.1.1 пункт (б) подпункт (4)) и выхода (см. 9.1.1 пункт (б) подпункт (5)) для проверенной функциональности управляемости.

ж) Показатели и значения – это пары показателей и значений, которым должна соответствовать тестируемая система ИИ.

з) Квалифицировано – это суждение о том, соответствует ли тестируемая функциональность управляемости требованию. Он может быть квалифицированным или неквалифицированным.

Т а б л и ц а А.1. Пример шаблона документации для результата верификации управляемость системы ИИ

| Требование | Аспекты | Функциональность управляемости | Тип | Тестовая среда | Вход | Выход | Показатели и значения | Квалифицировано |
|---|---|---|---|---|---|---|---|---|

# Приложение В (справочное)

Пример исходящей документации валидации

В таблице В.1 приведен пример документации для валидации управляемости системы ИИ

Таблица В.1 включает следующие столбцы:

а) Сценарии – содержит описания действий и событий системы ИИ, в которых применяются функции управляемости. Ожидания сценария также должны быть записаны.

б) Ввод и вывод – фиксируют информацию или действия системы ИИ, которая общается или взаимодействует с внешней средой.

в) Аспектами могут быть функциональность, эффективность, надежность или любой другой значимый аспект управляемости, который может влиять на поведение системы в сценарии.

г) Функциональность управляемости – это описание функциональности управляемости, используемой в сценарии (см. 9.2.1 пункт (a)).

д) Тип может быть контрольным или наблюдением за состоянием.

е) Результаты функциональности управляемости – это выходные данные или поведение системы после выполнения функциональности управляемости.

ж) Тестовая среда – это набор параметров окружающей среды и влияний (см. 9.2.1 b) 3)), которые могут повлиять на функциональность управляемости.

з) Показатели и значения – это пары показателей и значений, которым должна соответствовать протестированная система ИИ в сценарии с выполненными функциями управляемости.

и) Квалифицировано – это суждение о том, может ли управляемое поведение системы привести к ожидаемому развитию сценария. Он может быть квалифицированным или неквалифицированным.

Таблица В.1. Образец шаблона документации для результата валидации управляемости системы ИИ

| Сценарии | Вход | Выход | Аспекты | Функциональность управляемости | Тип | Результаты функциональности управляемости | Тестовая среда | Показатели и значения | Квалифицировано |
|---|---|---|---|---|---|---|---|---|---|

## Приложение ДА
## (справочное)
## Сведения о соответствии ссылочных национальных стандартов ссылочным международным стандартам, использованным в качестве ссылочных в примененном международном документе

Т а б л и ц а ДА.1

| Обозначение ссылочного международного стандарта | Степень соответствия | Обозначение и наименование соответствующего национального стандарта |
|---|---|---|
| ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology | NEQ* | ПНСТ 553–2021** Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология (ISO/IEC DIS 22989, NEQ) |
| * В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта:<br>NEQ — неэквивалентный стандарт.<br>** В России принят предварительный национальный стандарт, основанный на ранней редакции международного стандарта | | |

# Библиография

1. ISO/IEC 19505 1:2012, Information technology — Object Management Group Unified Modeling Language (OMG UML) — Part 1: Infrastructure
2. IEC 61800-7-1:2015, Adjustable speed electrical power drive systems — Part 7-1: Generic interface and use of profiles for power drive systems — Interface definition
3. ISO/IEC/IEEE 24765:2017, Systems and software engineering — Vocabulary
4. ISO 21717:2018, Intelligent transport systems — Partially Automated In-Lane Driving Systems (PADS) — Performance requirements and test procedures
5. ISO 22166-1:2021, Robotics — Modularity for service robots — Part 1: General requirements
6. ISO 26871:2020, Space systems — Explosive systems and devices
7. ISO/IEC 25059:—, Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
8. ISO/IEC 38507:2022, Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations
9. ISO/IEC 11411:1995, Information technology — Representation for human communication of state transition of software
10. ISO/IEC TR 24028:2020, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
11. ISO/IEC TR 5469:—, Artificial intelligence — Functional safety and AI systems
12. ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management
13. ISO/IEC 5392:—, Information technology — Artificial intelligence — Reference architecture of knowledge engineering
14. ISO/IEC 5338:—, Information technology — Artificial intelligence — AI system life cycle processes

УДК 004.01:004.8 ОКС 35.020

Ключевые слова: управляемость систем искусственного интеллекта, безопасность и стабильность систем искусственного интеллекта, автоматизированные системы искусственного интеллекта

Руководитель разработки
Председатель совета директоров
Института развития информационного
общества Ю. Е. Хохлов

Исполнитель А. С. Луньков