НАУЧНЫЙ СОВЕТ ПРИ ПРЕЗИДИУМЕ РАН
ПО МЕТОДОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И
КОГНИТИВНЫХ ИССЛЕДОВАНИЙ

ПРЕЗИДИУМ
РОССИЙСКОЙ АКАДЕМИИ НАУК

НАУЧНЫЙ СОВЕТ ПО МЕТОДОЛОГИИ
ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
И КОГНИТИВНЫХ ИССЛЕДОВАНИЙ

НСМИИ РАН

109240, г. Москва, ул. Гончарная, д. 12, стр. 1, www.scmai.ru, mail@scmai.ru

МЕТОДИКА ОЦЕНКИ ДОВЕРИЯ К «ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ»

Москва
2022

Алексеев А.Ю., Винник Д.В., Гарбук С.В., Лекторский В.А. (ред.), Черногор Н.Н. Методика оценки доверия к «искусственному интеллекту». М.: Президиум РАН, 2022 г. – 17 с.

Методика оценки доверия к «искусственному интеллекту» разработана в соответствии с пп. 2 Постановления Президиума РАН от 23 ноября 2021 года «Искусственный интеллект в контексте информационной безопасности».

Авторы:

Алексеев Андрей Юрьевич, ученый секретарь и координатор научных программ НСМИИ РАН, доктор философских наук, профессор философского факультета Государственного академического университета гуманитарных наук, г. Москва

Винник Дмитрий Владимирович, соруководитель секции «Нейрофилософия» НСМИИ РАН, доктор философских наук, профессор департамента гуманитарных наук Финансового университета при Правительстве РФ, г. Москва

Гарбук Сергей Владимирович, кандидат технических наук, директор по научным проектам НИУ ВШЭ, председатель технического комитета по стандартизации (ТК164) «Искусственный интеллект», г. Москва

Лекторский Владислав Александрович, председатель НСМИИ РАН, академик РАН, академик РАО, доктор философских наук, профессор, главный научный сотрудник Института философии РАН, г. Москва

Черногор Николай Николаевич, профессор РАН, доктор юридических наук, профессор, заместитель директора Института законодательства и сравнительного правоведения при Правительстве Российской Федерации, г. Москва

1. Доверие к «искусственному интеллекту» (далее – доверие к ИИ) – состояние убеждённости пользователя системы ИИ и других заинтересованных лиц (например, государственных регуляторов в соответствующих прикладных отраслях; разработчиков систем ИИ; страховых компаний) в том, что система $\boldsymbol{S}^a$ может быть использована для решения конкретной прикладной задачи искусственного интеллекта $\boldsymbol{G}$ в заданных (предусмотренных) условиях эксплуатации.

2. Доверие к ИИ устанавливается путём оценки уровня доверия к системе ИИ T^a, для вычисления которого используется выражение:

$$T^a = Q_f^a Q_s^a,$$

где Q_f^a – показатель качества системы $\boldsymbol{S}^a$, под которым понимается уровень соответствия требованиям, установленным потребителем [1]. Угрозы безопасности системе ИИ, влияющие на способность системы выполнять задачи по назначению, также учитываются при определении показателя качества системы. Последовательность определения показателя Q_f^a приведена в пп.6-16;

Q_s^a – уровень безопасности системы $\boldsymbol{S}^a$, обратно пропорциональный значимости угроз безопасности (физических, информационных, социально-психологических, экологических и др.) со стороны системы ИИ для третьих лиц при применении этой системы по назначению. Последовательность определения показателя Q_s^a приведена в пп.17-18.

3. Доверие к ИИ наступает в случае, если значение уровня доверия T^a превышает или равняется установленному порогу Q^* (критериальному уровню доверия). Особенности задания критериального уровня Q^* приведены в пп.19-22.

4. Возможны следующие режимы оценки уровня доверия T^a:

ручной режим, в котором экспертная оценка уровня доверия T^a формируется человеком и определяется его знаниями, опытом, интуицией, воспитанием;

автоматизированный режим, в котором определение T^a поддерживается средствами специальной экспертной системой оценки доверия (*ЭСОД*). Принципы работы *ЭСОД* рассмотрены в пп.23-26, причём в пп.24-26 рассмотрены особенности применения *ЭСОД* для оценки доверия к интеллектуальной системе, рассчитанной на решение неопределённого круга прикладных задач;

автоматический режим, в котором T^a оценивается программно-аппаратным способом и решение о невозможности применения системы ИИ $\boldsymbol{S}^a$ осуществляется без участия человека.

5. В настоящей методике предполагается, что функционально значимые характеристики прикладной системы ИИ $\boldsymbol{S}$ при решении задачи $\boldsymbol{G}$ могут быть заданы в виде конечного набора характеристик:

$$F(\boldsymbol{S}) = \{f_1, f_2, \dots f_N\},$$

где $f_1, f_2, \dots f_N$ – значения частных характеристик системы ИИ, определенные на соответствующих шкалах (числовая шкала, шкала категорий, отношений и др.).

В состав набора функционально значимых характеристик $F(\boldsymbol{S})$ включают различные показатели, связанные с возможностью применения системы ИИ по назначению: показатели, непосредственно характеризующие вероятность успешного решения заданной прикладной задачи обработки данных $\boldsymbol{G}$ (вероятности ошибок первого и второго рода, F-мера, площадь под ROC-кривой и другие); эргономические показатели; показатели ресурсоёмкости; показатели надёжности; показатели процессов жизненного цикла системы и др. Конкретный состав набора характеристик $F(\boldsymbol{S})$ определяют потребители систем ИИ, государственные регуляторы в области безопасности применения прикладных систем ИИ, разработчики систем и другие заинтересованные стороны с учётом отраслевой специфики решаемый прикладных задач обработки данных.

Принимая во внимание, что в с соответствии с национальным стандартом [2] под качеством (quality) системы ИИ подразумевается совокупность характеристик и свойств системы, обусловливающих ее способность удовлетворять установленным или предполагаемым требованиям в соответствии с ее назначением, набор вида $F(\boldsymbol{S})$ может быть определён как качество системы ИИ.

6. Показатель качества системы ИИ Q_f^a определяется с использованием функционала качества $Q_f(M^{ar})$, значение которого зависит от степени соответствия характеристик F^a системы $\boldsymbol{S}^a$ характеристикам F^r эталонной (референтной) системы $\boldsymbol{S}^r$ (M^{ar}), а также от объективной полезности системы $\boldsymbol{S}^a$ для решения заданной интеллектуальной задачи $\boldsymbol{G}$.

7. Степень соответствия M^{ar} определяется с использованием метрики $M^{ab} \in R^1$, определённой на N-мерном пространстве характеристик как правила сравнения функционально значимых возможностей систем $\boldsymbol{S}^a$ и $\boldsymbol{S}^b$, заданных векторами F^a и F^b, соответственно:

$$M^{ab} = M(F^a, F^b) = \sum_i^N v_i \mu_i\left(f_i^a, f_i^b\right),$$

где v_i – коэффициент, характеризующий важность i-й характеристики f_i в контексте решаемой прикладной интеллектуальной задачи;

$\mu_i(f_i^a, f_i^b) \in R^1$ – частная метрика для i-й характеристики, определённая на соответствующей шкале значений.

Метрика M^{ab} обладает следующими свойствами:

$M^{ab} = -M^{ba}$;

$M^{ab} = 0$, если функциональные возможности систем $\boldsymbol{S}^a$ и $\boldsymbol{S}^b$ совпадают;

$M^{ab} > 0$, если функциональные возможности системы $\boldsymbol{S}^a$ превышают возможности системы $\boldsymbol{S}^b$.

8. Функционал качества $Q(M^{ar})$ обладает следующими свойствами:

$Q_f(M^{ar}) \in [0; Q_{fmax}]$;

$Q_f(M^{ar}) = 0$ при фактическом отсутствии у системы $\boldsymbol{S}^a$ функциональных возможностей, значимых для решения интеллектуальной задачи $\boldsymbol{G}$;

$Q_f(M^{ar}) = 1$ при $M^{ar} = 0$;

$Q_f(M^{ar}) = Q_{fmax}$, если дальнейшее улучшение функциональных характеристик F^a не приводит к повышению качества решения задачи $\boldsymbol{G}$, но вызывает дополнительные ресурсные затраты и, соответственно, приводит к снижению эффективности решения задачи $\boldsymbol{G}$. Предельное значение Q_{fmax} определяется ограниченными информационными возможностями сенсоров, поставляющих информацию для системы $\boldsymbol{S}^a$, а также ограниченными возможностями метасистемы (например, объекта управления), получающей информацию от $\boldsymbol{S}^a$. Примерный вид функционала $Q_f(M^{ar})$ показан на рис.1 [3].

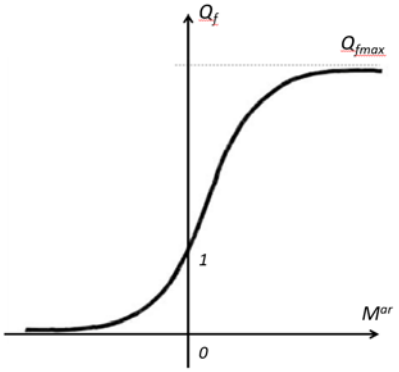

Рис.1. Примерный вид функционала качества системы ИИ.

9. Характеристики $\{f_1, f_2, \dots f_N\}$ могут относиться к разным этапам жизненного цикла (ЖЦ) систем ИИ и устанавливать требования как к процессам ЖЦ систем ИИ, так и непосредственно к самим системам [4]. В общем случае частные метрики $\mu_i(f_i^a, f_i^b)$ оцениваются:

а) на всём ЖЦ системы ИИ (концептуальное проектирование, разработка, изготовление, эксплуатация, включая дообучение, и утилизация) – путём проверки соответствия требованиям, установленным к обучающим наборам данных, используемым при создании систем аппаратным средствам и программному

обеспечению, применяемым типовым архитектурам, способам утилизации систем ИИ и др. (требования к процессам ЖЦ систем ИИ);

б) на стадии эксплуатации (ввод системы в эксплуатацию, повторное тестирование системы ИИ по окончании её дообучения) – путём испытания систем ИИ на представительных наборах тестовых данных (далее – тестовых НД), в том числе, с учетом возможного дообучения систем ИИ в процессе эксплуатации (требования к самим системам ИИ).

10. Различные способы оценки качества систем ИИ по п.9 могут использоваться как по отдельности, так и в совокупности. Для систем ИИ, предназначенных для решения особо ответственных прикладных задач (задачи в области обороны и безопасности, управления критически важной инфраструктурой, связанные с обработкой сведений, составляющих государственную тайну, и др.), способы по п.9 должны применяться только совместно.

11. Оценка соответствия характеристик системы ИИ установленным требованиям и определение частных метрик $\mu_i(f_i^a, f_i^b)$ осуществляется:

путём непосредственной (объективной) проверки выполнения требований. В некоторых случаях (например, для систем ИИ зарубежного производства) данный способ оценки качества может быть нереализуем в полном объёме вследствие недоступности необходимых сведений об особенностях разработки и изготовления системы ИИ;

экспертным путём на основе учёта мнения экспертов, основанного на персональном опыте, репутации разработчика и/или поставщика системы ИИ и других субъективных факторах.

Оценка значений метрик $\mu_i(f_i^a, f_i^b)$ с использованием объективных и субъективных оценок характеристик системы ИИ может быть выполнена с использованием методики Шортлифа [5]. В этом случае в качестве оценки метрики $\mu_i(f_i^a, f_i^r)$ используется значение коэффициента уверенности в том, что на заданном интервале времени и в предусмотренных условиях эксплуатации (при прогнозируемой изменчивости существенных факторов эксплуатации $\{e_1, e_2 \dots e_K\}$, п.15) отклонение характеристики системы ИИ f_i^a от эталонного значения f_i^r не превысит заданное пороговое значение.

12. Требования к процессам ЖЦ систем ИИ (п.9.а) устанавливаются с учётом лучших практик разработки и практической реализации соответствующих систем. В состав требований к процессам ЖЦ должны включаться требования, уверенность в необходимости выполнения которых является достаточно высокой, основанной на ранее полученном опыте создания и применения систем ИИ соответствующего типа и функционального назначения. Избыточные требования к

процессам ЖЦ могут привести к необоснованному сокращению возможных вариантов аппаратно-программной реализации систем ИИ.

13. Оценка соответствия требованиям к самим системам ИИ на стадии эксплуатации (п.9.б) осуществляется путём проведения испытаний систем в соответствии с ГОСТ Р 59898-2021 [2].

14. При проведении испытании систем ИИ на стадии эксплуатации (п.9.б) обеспечение статистической представительности получаемых оценок частных показателей качества $\mu_i(f_i^a, f_i^b)$ обеспечивается за счёт подготовки репрезентативных тестовых НД, на которых могут быть получены необходимые оценки характеристик систем ИИ.

15. Обоснование репрезентативности тестового НД E предполагает выполнение следующих процедур:

а) выявление исчерпывающего перечня внешних по отношению к системе ИИ факторов $\{e_1, e_2 \dots e_K\}$, существенно влияющих на качество работы алгоритмов ИИ для заданной прикладной задачи (существенных факторов эксплуатации системы ИИ). При этом каждый фактор e_j может быть определён на своей специфической шкале (номинальной, отношений и др.), в том числе, с использованием социально-психологических, экономических и др. показателей;

б) обоснование предположений о законах распределения существенных факторов в реальных (предусмотренных, допустимых) условиях эксплуатации на соответствующих шкалах;

в) обеспечение необходимого соответствия статистических характеристик существенных факторов для тестового НД E со статистическими характеристиками этих факторов в предусмотренных условиях эксплуатации.

16. Представительность тестового НД E означает, что её объём и вариативность позволяют с заданной вероятностью γ_E утверждать, что модуль метрики M^{ab}, вычисленной для характеристик системы $\boldsymbol{S}_a$, полученных на тестовом НД ($\widehat{F_E^a}$), и характеристик, апостериорно полученных в реальных условиях эксплуатации ($\widehat{F_\Gamma^a}$), не будет превышать некоторую заданную величину ε_E:

$$P\left(\left|M\left(\widehat{F_E^a}, \widehat{F_\Gamma^a}\right)\right| \le \varepsilon_E\right) = \gamma_E.$$

17. Уровень безопасности системы ИИ Q_s^a определяет допустимость последствий применения системы $\boldsymbol{S}^a$ для решения задачи $\boldsymbol{G}$ с учётом интересов личности, общества и государства. При оценке доверия к ИИ целесообразно учитывать следующие угрозы личности, обществу и государству:

опасное физическое воздействие, приводящее к гибели и ущербу здоровью людей (такие угрозы характерны, например, для беспилотного транспорта, промышленных роботов, высокоавтоматизированной сельскохозяйственной и строительной техники);

деструктивное воздействие на общественное сознание, связанное с распространением недостоверной информации (поисковые системы, социальные сети);

угрозы, связанные с совершением противоправных деяний (технические средства обеспечения безопасности);

опасное воздействие на окружающую природную среду (высокоавтоматизированное нефтехимическое и энергетическое оборудование, объекты инфраструктуры жизнеобеспечения);

угрозы информационной безопасности (информационные системы обработки персональных данных, корпоративные информационные системы);

угрозы, связанные с неэтичным поведением (банковские и страховые автоматизированные системы, медицинские информационные системы, системы подбора персонала и другие.

Особо выделяют возможные социальные риски, связанные с применением систем ИИ:

деградация фундаментальных эмпирических знаний и навыков, непосредственно связанных с элементарными рутинными навыками и процедурами, лежащими в основе тех или иных видов научно-технической деятельности;

неспособность персонала решать задачи управления в случае выхода из строя систем ИИ на длительное время;

размывание ответственности за решения, принятые с участием экспертных систем на основе ИИ;

распространение ошибки объективного вменения в судопроизводстве и деградация ответственности за решения;

десубъективация преимущественно человеческих видов деятельности, таких как психологическое консультирование, врачебная деятельность, образование, наука и др.;

угроза утери гражданами предпринимательского оптимизма и инициативы вследствие значительного роста эффективности фискальной и надзорной деятельности;

неприемлемый рост социальной напряженности вследствие фонового психологического давления от избытка надзорных инструментов, включая распознание и протоколирование сложных форм поведения;

рост безответственного бюрократического творчества вследствие выполнения рутинных операций по документообороту (дебюрократизация), рост неэффективности использования вычислительных мощностей и энергопотребления и другие.

18. Уровень безопасности Q_S^a принимает следующие значения:

$Q_s^a = 0$, если угрозы безопасности, вызванные применением системы ИИ, неприемлемы или не изучены, но потенциально могут нанести неприемлемый ущерб;

$Q_s^a = 1$, если угрозы безопасности принципиально отсутствуют.

Уровень безопасности Q_s^a оценивается экспертным путём по инициативе потребителей системы ИИ и других заинтересованных лиц (п.1) с привлечением представительной группы профильных экспертов, обладающих необходимыми компетенциями в области соответствующей прикладной задачи ИИ.

19. Значение критериального уровня доверия Q^* выбирается в диапазоне $Q^* \in]0; Q_{fmax}]$. Максимальное значение Q_{fmax} может устанавливаться исключительно для систем ИИ, функционирование которых потенциально не может привести к возникновению угроз безопасности личности, общества и государства по п.17.

20. Требуемое значение Q^*, состав функционально значимых характеристик $\{f_1, f_2, \dots f_N\}$ и значения этих характеристик для эталонной системы ИИ $F^r = F(\boldsymbol{S}^r)$ задаются потребителями систем ИИ, государственными регуляторами в области безопасности применения прикладных систем ИИ, разработчиками систем и другими заинтересованными сторонами с учётом:

требований системы, внешней по отношению к системе ИИ (метасистемы);

возможностей квалифицированного человека-оператора по решению заданной прикладной задачи, автоматизация деятельности которого предполагается;

мирового уровня техники в области соответствующих прикладных технологий ИИ;

других соображений.

21. Для систем ИИ, предназначенных для автоматизации деятельности человека, в качестве характеристик F^r референтной системы S^r могут быть приняты возможности квалифицированного человека-оператора, с приемлемым качеством решающего прикладную интеллектуальную задачу $\boldsymbol{G}$. Выбор критериального значения в диапазоне $Q^* \in [1; Q_{fmax}]$ означает, что к создаваемым системам будет предъявляться требование функционального соответствия или превосходства над квалифицированным человеком-оператором.

22. Оценка F^r для систем ИИ, предназначенных для замены человека-оператора, должна осуществляться с привлечением группы экспертов, причём размеры этой группы при заданной тестовой выборке могут быть определены с использованием коэффициента конкордации, характеризующего степень согласованности мнений экспертов [6].

23. Экспертная система *ЭСОД* является информационной системой, обеспечивающей реализацию функции оценки уровня доверия T^a посредством сбора, хранения, обработки и выдачи информации о параметрах Q_f^a , Q_s^a , Q^*. *ЭСОД* поддерживает когнитивные механизмы оценки доверия и способна лишь частично заменить эксперта (экспертов), так как доверие, вера, убеждение, знание – это приватные феномены сознания. Применение *ЭСОД* в автоматическом режиме возможно применительно к оценке доверия к системе $\boldsymbol{S}^a$ в контексте её использования для решения конкретной прикладной задачи ИИ $\boldsymbol{G}$ в заданных (предусмотренных) условиях эксплуатации (п.1), то есть при условии снижения критерия доверия до уровня отождествления когнитивной деятельности с механизмами ее имитации.

24. *ЭСОД* может использоваться для оценки уровня доверия к интеллектуальной системе $\boldsymbol{S}^a$ безотносительно конкретной прикладной задачи (для неопределённого круга задач). При этом разнообразие метода оценки уровня доверия должно превышать разнообразие оцениваемого предмета, поэтому предпочтительна реализация *ЭСОД* методами комплексного теста Тьюринга [7]. Этот тест основан на тьюринговой игре в имитацию интеллектуального (диалогового) поведения и включает более сотни её версий и версии версий, которые имитируют когнитивные компетенции мышления, понимания, творчества, сознания, самосознания, морального вменения, личности, общества и другие *x*-компетенции.

Каждый тест представим пятью функциями:

$\boldsymbol{\varphi_{tw}}^{\mathbf{1}}$ – интеррогативная функция, состоящая из типовых вопросов, которые задает эксперт (тьюринговый судья, interrogator, наблюдатель) при исследовании системы $\boldsymbol{S}^a$ на предмет доверительного приписывания ей *x*-компетенции (мышления, понимания, креативности, осознавания и пр.):

$\boldsymbol{\varphi_{tw}}^{\mathbf{2}}$ – дефинитная функция, обеспечивающая определение когнитивной функции с учетом вычислительных возможностей ее реализации;

$\boldsymbol{\varphi_{tw}}^{\mathbf{3}}$ – конструкторская функция, раскрывающая принципы работы машины (механизма), способного реализовать когнитивную функцию исследуемого частного теста;

$\boldsymbol{\varphi_{tw}}^{\mathbf{4}}$ – критическая функция, отражающая дискуссию по поводу возможности или невозможности компьютерной реализации изучаемой когнитивной функции, а также угроз и последствий этого;

$\boldsymbol{\varphi_{tw}}^{\mathbf{5}}$ – конститутивная функция, позиционирующая этические, эстетические, ценностные, юридические, экономические, политические, правовые и другие отношения эксперта к системе $\boldsymbol{S}^a$.

Эти функции сведены в таблице (Приложение № 1, [7, С.113 – 115]).

Комбинация этих функций обеспечивает сложностную аппроксимацию когнитивных функций системы $\boldsymbol{S}^{a}$.

25. В *ЭСОД* основной компонентой является *база доверия*. Основные атрибуты выборки из этой базы данных следующие:

Тезис: уникальный код тезиса, который аутоинкрементарно добавляется в базу данных;

Класс: код класса иерархического классификатора;

Наименование тезиса: краткое обозначение содержания тезиса;

F: операции с тезисом, в сигнатуру которых входит “+” (подтверждение тезиса), “-” (опровержение тезиса), “н” (концептуальная нормализация тезиса) и др. обозначения операций (см. пп.26);

Ref: ссылка на родительский тезис, если он имеется;

Содержание тезиса: текст произвольной длины, подробно раскрывающий суть тезиса;

Библиография: ссылка на первоисточник в виде текста статьи (книги), записи доклада, выступления.

Тезисы *базы доверия* представимы произвольными мультимедийными источниками: текстовыми, графическими, фото-, аудио- и видео-записями воображаемых и реальных дискуссий, включая мероприятия НСМИИ РАН (https://scmai.ru, [9]).

Единицей оценки доверия является тезис и его подтверждение (“+”) либо опровержение (“-”).

Критическая функция $\boldsymbol{\varphi_{tw}}^{\mathbf{4}}$ лежит в основании оценки доверия. Описания частных тестов Тьюринга включают достаточно обширный библиографический *объем доверия* как отношение количества знаков дискуссии к общему количеству знаков статьи. Так, в описании оригинального теста Тьюринга этот «объем доверия» составляет 50% ÷ 73% (9 тезисов), в тесте Серля объем доверия составляет 72 % (6 тезисов), в тесте Блока он достигает 90,2 % (9 тезисов).

Конститутивная функция $\boldsymbol{\varphi_{tw}}^{\mathbf{5}}$ логически выражается пропозициональным отношением. Ее предпочтительно изучать аппаратом модальных логик.

26. Оценка доверия к системам ИИ, рассчитанным на решение неопределённого круга прикладных задач, реализуется следующими методами:

Идентификация тезиса: распознавание уникальной точки зрения, мнения, положения, теории, концепции, подхода, направления, которые значимы на предмет оценки доверия к ИИ.

Концептуальная нормализация предназначена для формирования тематически нейтральных терминов, которые не вызывают когнитивного

диссонанса при вовлечении социогуманитарных и естественно-научных понятий в технический дискурс.

Редукция: приведение предмета дискуссии (тезиса, аргумента) к уже изученному положению или к более простому предмету.

Кодификация – выделение, группирование, упорядочение тезисов и их компонент.

Формализация – использование специальной символики, примером которой является язык комплексного теста Тьюринга.

Унификация – приведение различных, разнородных, неоднородных и гетерогенных положений дискуссии к единообразной системе.

Апробация – объективная оценка результатов исследования по доверию, проводимое как правило в условиях массового обсуждения в рамках масштабного мероприятия.

Верификация – подтверждение тезиса на основе представления объективных свидетельств, субъективных мнений и интерсубъективных оценок.

Валидация – системная деятельность по определению количественных значений показателей доверия методами Дельфи, теории игр (см. также пп.21).

Стандартизация – системная деятельность по установлению правил и характеристик оценки доверия к системам ИИ для многократного использования.

Первый и последний методы являются точками входа и выхода в *ЭСОД*. Экспертиза доверия ИИ начинается от выявления значимых положений по поводу компьютерного моделирования когнитивных функций и завершается системной работой интеллектуальной системы *ЭСОД*, которая на выходе определяет одно из двух значений: можно ли системе ИИ доверять или ей доверять нельзя?

Если следует вывод о недоверии системе ИИ, то *ЭСОД* выдает рекомендацию: что надо сделать для формирования доверительной системы ИИ.

Список литературы

1. ГОСТ Р ИСО 9001-2015. Системы менеджмента качества. Требования/ Москва, Стандартинформ, 2015 – 32 с.

2. ГОСТ Р 59898-2021 Оценка качества систем искусственного интеллекта. Общие положения. М.: Стандартинформ. 2021 – 37 с.

3. Garbuk S.V. Intellimetry as a way to ensure AI trustworthiness//The Proceedings of the 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI). Limassol, Cyprus, 06-10.10.2018. С.27-30.

4. ГОСТ Р 59276-2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. М.: Стандартинформ. 2020 – 20 с.

5. Каляев И.А., Мельник Э.В. Доверенные системы управления/Мехатроника, автоматизация, управление. 2021, т.22, №5. С.227-236.

6. Гарбук С.В., Бакеев Р.Н. Конкурентная оценка качества технологий интеллектуальной обработки данных. Проблемы управления. 2017, №6. С.50-62.

7. Алексеев А.Ю. Комплексный тест Тьюринга: философско-методологические и социокультурные аспекты. – М.: ИИнтеЛЛ, 2013. – 304 с. – ISBN 978-5-98956-007-3; https://aintell.info/16/book.pdf

8. Лекторский В.А., Васильев С.Н., Макаров В.Л., Хабриева Т.Я., Кокошин А.А., Ушаков Д.В., Валуева Е.А., Дубровский Д.И., Черниговская Т.В., Семёнов А.Л., Зискин К.Е., Любимов А.П., Целищев В.В., Алексеев А.Ю. Человек и системы искусственного интеллекта / Под ред.акад. РАН В.А. Лекторского. — СПб.: Издательство «Юридический центр», 2022. — 328 с.

9. Сайт Всероссийской междисциплинарной конференции «Искусственный интеллект: проблема доверия», 27 – 28 апреля 2021 г., Президиум РАН, г. Москва; URL: https://scmai.ru/2022/04/27-28

Приложение 1

Таблица 1. Комплексный тест Тьюринга

№ пп	Наименование теста	$\varphi_{tw}{}^{1}$ Основной вопрос теста	Автор теста	$\varphi_{tw}{}^{2}$ Дефинитная функция	$\varphi_{tw}{}^{4}$ Критическая (полемическая) Функция	$\varphi_{tw}{}^{3}$ Конструирующая функция
1	Оригинальный тест	Может ли компьютер мыслить?	А.Тьюринг [Turing,, 1950]	Диалоговый интеллект	Компьютер может мыслить	УЦВМ и нейронная сеть
2	Тест на здравый смысл	Может ли компьютер разумно рассуждать?	Дж.Маккарти [McCarthy, 1984]	Здравый смысл (глубинный)	Экспертные системы не способны отвечать на вопросы с глубоким пониманием смысла	Компьютер с эволюционно-эпистемологическими механизмами роста знаний
		Может ли компьютер банально рассуждать?	Д. Деннетт [Dennett, 1984]	Здравый смысл (поверхностный)	Экспертные системы не способны отвечать на банальные вопросы	Компьютер с эволюционно-эпистемологическими механизмами роста знаний
3	Китайская нация	Возможен ли компьютинг общественного сознания?	Н. Блок [Block, 1978]	Глобальное сознание	Ноосфера невозможна	Интернет
4	Параноидальный тест	Может ли компьютер функционировать, как псхически здоровый человек?	К. Колби [Colby, Hilf, Weber ,1971]	Психическая аномалия (паранойя)	Искусственные системы способны имитировать только психические отклонения	Экспертные системы
5	Китайская комната	Может ли компьютер понимать?	Дж. Серль [Searle,1980]	Понимание	Компьютер не может понимать	*Я* и компьютер с базой «данных»
6	Субкогнитивный тест	Может ли компьютер обладать «подсознанием»?	Р. Френч [French, 1990]	Подсознание	Компьютер не может жить как человек	Программы тестирования подсознания
7	Эмоциональный тест	Может ли компьютер испытывать эмоции?	А.Сломан [Sloman, 1998]	Любовь	Компьютер может любить	Человекоподобные агенты
8	Инвертированный тест	Может ли компьютер приписывать ментальность другому?	С. Ватт [Watt, 1996]	Атрибуция когниций	Компьютер не может приписывать ментальное	Машина Тьюринга, тестирующая машину Тьюринга
9	Гендерный тест	Может ли женщина (мужчина) мыслить?	Ю. Генова [Genova, 1994]	Гендерная детерминация интеллекта	Интеллектуальный сексизм	Машина Тьюринга
		Может ли компьютер испытывать сексуальное влечение?	Р. Хопкинс [Hopkins, 1998]	Либидо	Компьютер испытывает сексуальное влечение	Секс-машина
10	Креативный тест	Может ли компьютер творить?	С.Брингсйорд [Bringsjord, 2001]	Творчество	Компьютер не может творить	Программы генерации литературных, изобразительных и музыкальных произведений

11	Тест Гёделя	Способен ли компьютер к самоописанию?	Р. Пенроуз [Penrose, 1989]	Самоописание	Компьютер не способен к самоописанию	Аутоформальная система
		Способен ли компьютер к самоорганизации?	Дж. Лукас [Lucas, 1961]	Самоорганизация	Компьютер не способен к самоорганизации (самопрограммированию)	Ауторепрезентативная система
12	Индуктивный тест	Можно ли индуктивно аргументировать возможность ИИ?	Дж. Мур [Moor, 1976]	Необходимые условия интеллекта	Интеллектуальный компьютер – индуктивно-собирательное понятие	Машина Тьюринга, собирающая факты прохождения ТТ
13	Новый тест	Можно ли дедуктивно аргументировать возможность ИИ?	Н. Блок [Block, 1980]	Достаточные условия интеллекта	Глобальная «машина знаний» – это не искусственный интеллект	Глобальная машина Блока
14	Кибериадный тест	Может ли компьютер творить миры?	Дж. Баресси [Barresi, 1987]	Творение	Компьютер способен репродуцировать искусственных людей, общества, миры	Футуротехнологии
15	Тотальный тест	Может ли компьютер быть неотличимым от человека?	С. Харнад [Harnad, 2001]	Человек и реальность	Компьютер способен дублицировать человека	НБИКС
16	Самый тотальный тест	Может ли развитие компьютера не отличаться от естественной эволюции?	П. Швайзер [Schweizer, 1998]	Человек и общество	Компьютер способен к развитию, подобного социобиологической эволюции	НБИКС и природа
17	Тест личности	Может ли компьютер быть личностью?	Дж. Поллок [Pollok, 1995]	Личность	Компьютер может быть личностью	Артилекты проекта «Оскар»
18	Тест зомби	Может ли бессознательный компьютер казаться сознательным?	Р. Кирк [Kirk, 1971]	Бессознательное (внутри)/ сознательное (снаружи)	Философские зомби возможны	Компьютерные модификации художественных образов зомби
		Могу ли я быть компьютерным зомби?	Д. Чалмерс [Chalmers, 1995]	Компьютерый зомби	Компьютерный Я-зомби возможен	Компьютер на базе «новой теории информации»
19	Интроспективный тест	Может ли компьютер обладать «иным» сознанием?	А. Клифтон [Clifton, 2003]	«Иное» сознание	Компьютер способен обладать нечеловеческой психикой	Проект сильного ИИ
20	Интерактивный тест	Может ли компьютер обрабатывать идеи?	В.К. Финн [Финн, 2009]	Идея как динамичное понятие	Рационализация общественной жизни посредством интеллектуальных технологий	Квазиаксиоматическая ДСМ-система
21	Комплексный тест	Может ли компьютер ВСЁ?	А. Алексеев [Алексеев, 2006]	Когниции частных тестов	Интеграция частных тестов Тьюринга	Комплексная машина Корсакова-Тьюринга

Список дополнительной литературы

[Алексеев, 2006] Алексеев, А.Ю. Возможности искусственного интеллекта: можно ли пройти тесты Тьюринга? [Текст] / А.Ю. Алексеев // Искусственный интеллект: Междисциплинарный подход / под ред. Д.И. Дубровского и В.А. Лекторского. – М.: ИИнтеЛЛ, 2006. – 448 с. – С. 223-243

[Финн, 2009] Финн В.К. К структурной когнитологии: феноменология сознания с точки зрения искусственного интеллекта //Вопросы философии №1, 2009

[Barresi, 1987] Barresi, J. 1987. Prospects for the Cyberiad: Certain Limits on Human Self-Knowledge in the Cybernetic Age, Journal for the Theory of Social Behavior 17, pp. 19–46.

[Block, 1978] Block, N. 1978. Troubles with functionalism. Minnesota Studies in the Philosophy of Science 9:261-325. Reprinted in Readings in the Philosophy of Psychology (MIT Press, 1980).. URL : http://perso.univ-rennes1.fr/pascal.ludwig/troubles.pdf

[Block, 1981-2] Block, N. 1981, Psychologism and Behaviorism, Philosophical Review 90, pp. 5–43.

[Bringsjord, 1996] Bringsjord, S. 1996, The Inverted Turing Test is Provably Redundant. Psycoloquy 7(29).. URL : http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?7.29.

[Chalmers, 1993] Chalmers, D. J. 1993. The Conscious Mind: In Search of a Fundamental Theory. URL : http://jamaica.u.arizona.edu/~chalmers/book/tcm.html

[Clifton, 2003] Clifton A., 2003. The Introspection game. Or, Does The Tin Man Have A Heart. URL : http://cogprints.ecs.soton.ac.uk/archive/00003483/01/Introspection_game(@

[Colby, 1981] Colby, K.M. 1981, Modeling a Paranoid Mind, Behavioral and Brain Sciences 4(4), pp. 515–560.

[Dennett, 1984] Dennett, D. C. 1984. The Age of Intelligent Machines: Can Machines Think? In (M. Shafto, ed) How We Know. Harper & Row.. URL : http://www.kurzweilai.net/articles/art0099.html?printable=1

[Genova, 1994-2] Genova, J. Turing's Sexual Guessing Game, Social Epistemology, 1994, 8(4), pp. 313–326.

[Harnad, 2001] Harnad, Stevan (2001) Minds, Machines and Turing: The Indistinguishability of Indistinguishables. Journal of Logic, Language, and Information 9(4):pp. 425-445.. URL : http://cogprints.ecs.soton.ac.uk/archive/00002615

[Hopkins, 1998] Hopkins, Patrick D. Sex/machine: readings in culture, gender, and technology. Indiana University Press, 1998

[Kirk, 1974-1] Kirk, R. 1974. Sentience and behaviour. Mind 81:43-60.

[Lucas, 1961] Lucas, J.R. (1961) Minds, Machines and Goedel. Philosophy 36:pp. 112-127.. URL : http://cogprints.ecs.soton.ac.uk/archive/00000356

[McCarthy, 1984] McCarthy, J (1984) Some Expert Systems Need Common Sense.. URL : http://cogprints.ecs.soton.ac.uk/archive/00000423

[Penrose, 1989] Penrose, R. 1989. The Emperor's New Mind. Oxford University Press.

[Polger, 2011] Polger Tom, Zombies. URL : http://host.uniroma3.it/progetti/kant/field/zombies.htm#Top

[Schweizer, 1998] Schweizer, P. 1998. The truly total Turing Test. Minds and Machines 8, pp. 263–272.

[Searle, 1980] Searle, J. R. 1980. Minds, brains and programs. Behavioral and Brain Sciences 3, pp. 417–424.. URL : http://www.cs.ucsb.edu/~cs165a/articles/Searle.html

[Sloman, 1998] Sloman, Aaron (1998) What Sort of Architecture is Required for a Human-Like Agent?, in Foundations of Rational Agency. Kluwer Academic Publishers.. URL : http://cogprints.ecs.soton.ac.uk/archive/00000411/>

[Turing, 1950] Turing, A. (1950), 'Computing Machinery and Intelligence', Mind 59(236), pp. 433–460.

[Watt, 1996] Watt, S. 1996. Naive Psychology and the Inverted Turing Test, Psycoloquy 7(14). URL : http://psycprints.ecs.soton.ac.uk/archive/00000506