

ТЕОРЕТИЧЕСКИЕ ИССЛЕДОВАНИЯ / THEORETICAL STUDIES

Научная статья / Research Article
<https://doi.org/10.55959/LPEJ-25-21>
УДК/UDC 37.022; 355.233; 378.4

О рисках когнитивных искажений использования систем искусственного интеллекта в образовательной деятельности вузов России

Е.К. Яхваров¹, А.В. Афанасьев¹ ✉, А.А. Аринушкина²

¹ Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, Российская Федерация

² Военный университет имени князя Александра Невского Министерства обороны Российской Федерации, Москва, Российская Федерация

✉ vka@mil.ru

Резюме

Актуальность. Внедрение в образовательную среду вузов России систем, созданных на основе технологий искусственного интеллекта (ИИ), формирует парадигму образования, которая потребует новых методологических подходов и педагогических приемов для полноценной реализации возможностей таких систем и избежания рисков, создаваемых этими технологиями. **Цель.** Цель исследования — формирование методического аппарата, позволяющего управлять рисками при использовании систем ИИ в образовательной деятельности вузов России.

Выборка. Представляет собой курсантскую учебную группу из $n = 20$ человек, являющуюся малой независимой выборкой, которая сравнивается с гипотетической генеральной нормальной совокупностью, представляющей собой аналогичные группы из $n = 20$ в 40 вузах.

Методы. Методологическую основу статьи составляет применение риск-ориентированного подхода к использованию систем ИИ в образовательной деятельности вуза.

Результаты. Риск-ориентированный подход к созданию методического обеспечения использования систем ИИ в образовательной деятельности вузов России позволит создать методологическую основу для исследования возможностей ускорения процессов получения знаний, умений и навыков

обучающимися и избежания возможных негативных эффектов, основанных на рисках «слепого доверия» к системам ИИ.

Выводы. Использование ИИ в образовательной деятельности вузов России влечет за собой кроме позитивных аспектов ряд рисков. В этой связи применение систем ИИ должно сопровождаться совокупностью методических подходов, исключающих «слепое доверие» преподавателей и обучающихся к системам ИИ.

Ключевые слова: риск-ориентированный подход, искусственный интеллект, методическое обеспечение, когнитивные функции, когнитивные искажения, образовательная деятельность, студенты, преподаватели

Для цитирования: Яхваров, Е.К., Афанасьев, А.В., Аринушкина, А.А. (2025). О рисках когнитивных искажений использования систем искусственного интеллекта в образовательной деятельности вузов России. *Вестник Московского университета. Серия 20. Педагогическое образование*, 23(4), 65–93. <https://doi.org/10.55959/LPEJ-25-21>

Risks of Cognitive Biases in the Application of Artificial Intelligence (AI) Systems in Educational Processes at Russian Universities

Egor K. Yahvarov¹, Andrey V. Afanas'ev¹ ✉, Anna A. Arinushkina²

¹ A.F. Mozhaysky Military Space Academy, Saint Petersburg, Russian Federation

² Military University of the Ministry of Defense of the Russian Federation named after Prince Alexander Nevsky, Moscow, Russian Federation

✉vka@mil.ru

Abstract

Background. The integration of AI-based systems into the educational environment of Russian universities creates a new educational paradigm, requiring application of novel methodological approaches and educational techniques in order to fully realize the potential of these systems and avoid the risks created by these technologies.

Objectives. The goal of the study is to develop a methodological framework for risk management in the use of AI systems in educational environment of Russian universities.

Study Participants. A small, independent sample, cadet study group of 20 people (n=20), is compared to a general population of similar groups (n=20) in 40 universities.

Methods. The methodological foundation of this research is based on the approaches that allow for classification of risks associated with AI systems in university education.

Results. A risk-based approach to developing methodological support for the use of AI systems in educational activities at Russian universities will create a methodological foundation for exploring the possibilities to accelerate knowledge and skills acquisition. It will also help avoid potential negative effects based on the risks associated with the “blind trust” in AI systems.

Conclusions. The use of AI in the educational activities of Russian universities entails, in addition to the positive aspects, a number of risks. In this regard, the use of AI systems should be accompanied by a set of methodological approaches that exclude the “blind trust” of teachers and students in AI systems.

Keywords: risk-based approach, artificial intelligence, methodological support, methodological framework, cognitive functions, cognitive distortions, educational activity, students, lecturers

For citation: Yahvarov, E.K., Afanas'ev, A.V., Arinushkina, A.A. (2025). Risks of cognitive biases in the application of artificial intelligence (AI) systems in educational processes at Russian universities. *Lomonosov Pedagogical Education Journal*, 23(4), 65–93. <https://doi.org/10.55959/LPEJ-25-21>

Введение

В настоящее время научно-технический прогресс осуществляется в рамках нового технологического уклада, который во многом опирается на достижения в области искусственного интеллекта (ИИ). В соответствии с Национальной стратегией развития искусственного интеллекта¹ на период до 2030 г., под ИИ понимается комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые как минимум

¹ Национальная стратегия развития искусственного интеллекта на период до 2030 года. (В редакции Указа Президента Российской Федерации от 15.02.2024 № 124). URL: <http://www.kremlin.ru/acts/bank/44731> (дата обращения: 18.08.2025).

с результатами интеллектуальной деятельности человека. В самом определении ИИ заложена возможность имитации когнитивных функций человека, что предполагает, с учетом развития и совершенствования систем ИИ, замену человеческой деятельности в различных областях. По данным исследований в области взаимодействия человека с системами ИИ, к 2030 г. прогнозируется смещенность границы взаимодействия «человек — машина» в сторону увеличения роли ИИ в области обработки данных (до 45%) и в области конечного принятия решений (до 40%), при этом большинство конечных принятых решений в различных сферах останется все еще за человеком (60–70%) (Искусственный интеллект..., 2021; *The outlook...*, 2015). Такое стремительное развитие технологий ИИ не обходит стороной и такую сферу человеческой деятельности, как высшее образование, трансформируя вузы в современные учебные заведения с лабораторно-экспериментальной базой для поиска и проектирования возможных сценариев применения технологий ИИ (Chopra, 2025; Pikhart, Klimova, 2025). При этом одновременно с огромным потенциалом повышения качества образования за счет использования систем и сервисов на основе технологий ИИ растут и риски их применения (Горбачева, 2025; Kasirzade, 2024; Cappelen et al., 2025).

Необходимо отметить перспективы развития систем искусственного интеллекта (СИИ) в военном деле. Применение беспилотных летательных аппаратов (БПЛА) на основе технологий искусственного интеллекта (ТИИ) в зоне проведения специальной военной операции (СВО) уже кардинально изменило характер боевых действий, а также методы и способы применения сил и средств вооруженной борьбы. Например, роль боевых бронемашин значительно снизилась за счет изменения тактики проведения боевых действий, что обусловливается высокой уязвимостью бронемашин от FPV-дронов, применением новых средств обнаружения, а также использованием различных управляемых наземных боевых систем. Массовое использование обеими сторонами конфликта автономных боевых систем на основе ТИИ в зоне СВО формирует новую тактику проведения боевых действий — тактику позиционной борьбы, создавая зоны сплошного огневого поражения противника, нахождение в которых приводит фактически к гарантированному уничтожению (подавлению) воинских формирований. В таких условиях преимуществом будет обладать та сторона, которая будет иметь технологическое превосходство в области цифровых технологий и ИИ, обусловленное развитием научно-технического прогресса страны и подготовкой военных кадров.

Одним из наиболее существенных факторов риска, который может приводить к снижению качества образования, является некорректное восприятие учебного контента (Ганчеренок, Горбачев, 2024), обусловленное особенностями интуитивного понимания человеком получаемой информации при использовании интеллектуальных систем. В частности, обучающиеся, как и все люди, при решении задач, в том числе с помощью технологий ИИ, опираются на эвристические методы, которые декомпозируют сложные задачи на простые операции. Используя эвристику, пользователи (обучающиеся, преподаватели) при применении такого инструментария, как системы ИИ, с одной стороны, формируют условия для мультипликативного роста как в образовательной, так и в научной деятельности, с другой — эвристические методы могут приводить к серьезным и систематическим ошибкам (Arinushkina, 2025; Shanahan et al., 2023; Мишуткин, 2025). Исходя из этого, требуется методическое обеспечение использования систем ИИ, формирующее у пользователей (обучающихся и преподавателей) рациональный подход при принятии решений в модели взаимодействия «человек — машина», одновременно сохраняя преимущества использования систем и сервисов на основе применения технологий ИИ в учебном процессе.

Теоретическое обоснование

Когнитивные искажения

В исследованиях последних лет в области развития систем ИИ в модели «человек — машина» (Murikah et al., 2024; Cheong, 2024; Holzinger et al., 2025) различают два основных аспекта, которые могут привести к формированию ложных знаний и навыков у пользователей систем ИИ (Шмит, 2025; Кобринский, 2024; Кирюшин и др., 2025).

Первый аспект заключается в технической несовершенности (ограниченности) систем ИИ, функционирующих в настоящее время в интересах пользователей. Одним из проявлений (последствий) технической ограниченности является смещение результатов (bias^2) в работе систем ИИ (Zhou et al., 2025; Faheem, 2024; Hanna et al., 2025). Техническая ограниченность систем ИИ, не позволяющая достичь

² Прим. авт.: bias — это явление, при котором модель или алгоритм вносит систематические ошибки при оценке и прогнозировании данных. Смещение возникает из-за присутствия предположений или упрощений, которые делает модель. Оно может проявляться в различных аспектах задачи машинного обучения, и важно понимать его воздействие на результаты.

заявленных целей предполагаемого функционального назначения, часто формируется на первом этапе создания системы ИИ — этапе проектирования. Разработчиками системы на этапе проектирования могут закладываться риски субъективизма, когнитивных искажений действительности. Кроме того, риски субъективизма разработчиков системы могут приводить к конкретным ошибкам на всех этапах жизненного цикла систем ИИ, например, при разработке алгоритмов, формировании выборок данных, создании необходимой информационно-телекоммуникационной инфраструктуры, обеспечивающей работу систем и сервисов на основе ИИ.

Второй аспект, который по своей сути является логическим продолжением первого, связан с вопросом излишнего доверия («слепого доверия») со стороны пользователей и завышенными ожиданиями от потенциала развития технологий ИИ (Сычев, 2023; Намиот, Ильюшин, 2025). Такой подход со стороны пользователей может привести к ложному восприятию системы ИИ как модели «человек — человек», а не «человек — машина».

Одним из самых резонансных примеров может служить история бывшего сотрудника Google инженера Блейка Лемуана, уволенного за публичное заявление о том, что система ИИ LaMDA «обрела сознание» (Reinecke et al., 2025; Marchegiani, 2025). Блейк Лемуан приписал системе человеческие качества, хотя LaMDA лишь генерировала текст. Невозможность «обретения сознания» системой ИИ, то есть обретения разума, свойственного человеку, обуславливается отсутствием должного (необходимого) уровня развития научно-технического прогресса (НТП). Вместе с тем компания Open AI заявляет о планах создания системы AGI (artificial general intelligence), позволяющей превзойти человека по когнитивным способностям (Planning for..., 2023), однако результаты пока не получены.

Представленное выше определение ИИ выстраивается вокруг имитирования (эмулирования) когнитивных функций человека, но полноценно реализовано такое имитирование на данном этапе развития НТП быть не может, так как современные системы ИИ формируют суждения, основанные на эмпирических данных (опыте), заложенных разработчиками. Еще И. Кант утверждал: «Хотя всякое познание начинается с опыта, отсюда вовсе не следует, что оно целиком происходит из опыта» (Кант, 1907). При этом одно из основных свойств разума, по мнению И. Канта, — оперирование априорными знаниями, то есть знаниями, безусловно независимыми от опыта.

По нашему мнению, основополагающей причиной, которая может вести к формированию ложных знаний и навыков у пользователей систем ИИ, является проявление субъективизма, вызванное физиологической особенностью человека при принятии решений. По мнению нобелевского лауреата по экономике Дэниела Канемана, люди склонны недооценивать маловероятные исходы, переоценивать информацию, опираться только на свой опыт, избегая новых знаний о проблематике (Канеман, Тверски, 2015). Такой субъективизм может проявляться не только в интуитивной оценке (когнитивных иллюзиях), но и при оценке физической величины, например, расстояния. Примером может служить известная иллюзия Мюллера-Лайера (Булатов и др., 2007; Howe, Purves, 2005): две горизонтальные линии равной длины, к которым пририсованы стрелки, направленные в разные стороны, визуальнo воспринимаются как линии абсолютно разной длины.

Данные примеры отражают особенности работы человеческого мозга при принятии решений в различных сферах. По результатам исследований в области анализа процессов принятия решений и когнитивных искажений (Кини, Райфа, 1981), к наиболее релевантным когнитивным искажениям в контексте заявленной цели исследования относятся следующие:

1. «Якорение». В процессе принятия решения человек склонен придавать большое значение первоначальной информации (Carter, Liu, 2025).
2. «Статус-кво». Принимая решения, человек склонен полагаться на свои старые убеждения, сформированные в прошлом, что позволяет сохранить прежнее положение дел, статус-кво (Samuelson, Zeckhauser, 1988; Balakrishnan et al., 2024; Godefroid et al., 2023).
3. «Необратимые затраты». Попытка оправдать предыдущие решения, что может привести к необратимым затратам.
4. «Желаемое и действительное». Лица, принимающие решения (ЛПР), могут искать подтверждение своим предположениям, исходя из предубеждений, выдавая желаемое за действительное. Человеческий мозг фокусируется на информации, подтверждающей «желаемое», и не воспринимает негативную информацию, требующую действий (Saini et al., 2025; Fok, Weld, 2024).
5. «Неверная формулировка». Формулировка вопроса или проблемы обуславливает выбор варианта решения (Berber, Srećković, 2024).

6. Когнитивные искажения, связанные с прогнозами и оценкой. ЛПР могут переоценивать свои способности к прогнозированию (чаще всего основываясь на оптимистических оценках) и оценке вариантов решений, что может привести к ошибочному мнению (Chun et al., 2025).

Когнитивные искажения при использовании ИИ

Когнитивные искажения, приведенные в Таблице 1, могут формироваться не только вследствие восприятия пользователем информации от систем и сервисов ИИ, но и, как уже говорилось выше, вследствие некорректной работы систем ИИ. Основными причинами некорректной работы систем ИИ могут служить:

- использование систем и сервисов ИИ не по назначению;
- ошибки в реализации математических моделей и написании кода;
- нерепрезентативность наборов данных, на которых системы и сервисы ИИ обучаются;
- смещенность моделей на этапе эксплуатации систем и сервисов ИИ.

Таблица 1

Виды когнитивных искажений

Когнитивные искажения	Примеры
«Якорение»	«Якорь», возникающий в формировании prompt. Пользователь: «Объясни, почему ИИ уничтожит человечество (якорь — негативный сценарий)» ИИ: формирует аргументы в пользу угрозы ИИ. Последствия. Даже если пользователь потом спросит: «Какие плюсы у ИИ?», ответ будет смещен в сторону первоначального контекста
«Статус-кво»	Использование однообразных формулировок. Например, пользователь всегда начинает запрос одинаково, хотя другие формулировки могли бы дать лучший результат. Prompt 1 пользователя: «Напиши текст о важности спорта». Prompt 2 пользователя: «Придумай мотивирующую историю о том, как спорт изменил жизнь обычного человека». Более подробный ответ будет сгенерирован по запросу 2
«Необратимые затраты»	ИИ-агент здесь выступает как усилитель данной логической ошибки у людей. ИИ-агент, оправдывающий решения. Например, обучающий используют генеративную систему ИИ

<p>«Необратимые затраты»</p>	<p>для написания эссе по истории. Обучающий изначально считает, что Наполеон был «великим реформатором», и просит ИИ аргументировать этот тезис. ИИ-агент выбирает только удобные факты. Например, «Кодекс Наполеона изменил Европу». При этом игнорирует войны, репрессии и результаты правления императора. ИИ-агент подстраивается под уровень знаний обучающегося. Если ученик слабо разбирается в теме, ИИ упрощает аргументацию, избегая сложных контраргументов. ИИ-агент включает эмоциональные формулировки. Вместо нейтрального анализа ИИ пишет: «Наполеон действительно был гением, потому что...» — усиливает предвзятость</p>
<p>«Желаемое за действительное»</p>	<p>Пользователь: «Смогу ли я сдать экзамен, на который я не готовился?» ИИ-агент: «Да, ты точно сдашь экзамен, верь в себя! При этом ваш успех будет зависеть от: подготовки, состояния здоровья (самочувствия)». Несмотря на отсутствие подготовки и объективные факты, такие как сложность экзамена и требование к знаниям, система ИИ дает ответ, который подтвердит его желание. При этом ИИ-агент отмечает основные аспекты, влияющие на успешную сдачу экзамена, но это уже может быть не учтено обучающимся</p>
<p>«Неверная формулировка»</p>	<p>Сформированный prompt определит контекст ответа системой ИИ. Например: 1-й вариант, пользователь: «Сделай отчет, чтобы был похож на отчет, сделанный человеком. Отчет должен быть высокого качества». 2-й вариант, пользователь: «Сделай аналитический отчет по образцу, количество страниц 10, проверь текст на соответствие требованиям, предложи формулировки к разделу “Выводы”»</p>
<p>«Когнитивные искажения прогнозирования»</p>	<p>Пользователь: «Докажи, что развитие искусственного общего интеллекта (AGI) приведет к катастрофической безработице среди специалистов умственного труда» ИИ-агент: «Исторические прецеденты технологических революций, текущие темпы автоматизации когнитивных задач и экономические модели показывают, что появление ИИ действительно несет риски массового вытеснения профессий, основанных на обработке информации». Система ИИ начинает приводить факты в пользу неизбежности массового увольнения, что может сформировать неверное понимание пользователем данного вопроса. То есть алгоритмы ИИ-агента сформированы таким образом, чтобы формировать ответ семантически близким с заданным запросом пользователем, что создает основу для когнитивных искажений</p>

Table 1

Types of cognitive biases

Cognitive biases	Examples
Anchoring	The “anchor” that occurs in the prompt development. <i>User: “Explain why AI will destroy humanity (the anchor is a negative scenario)”. AI: forms arguments in favor of the threat of AI.”</i> <i>Consequence. Even if the user then asks “What are the advantages of AI?” the answer will shift towards the original context.</i>
The “Status quo”	The use of monotonous formulations. <i>For example, the user always starts the query in the same way, although other formulations could produce a better result. User’s prompt 1: «Write a text about the importance of sports.» Prompt 2: «Come up with a motivational story about how sports have changed the life of an ordinary person.» A more detailed response will be generated for Request 2.</i>
Irreversible costs	The AI agent acts as an amplifier of this logical error in humans. AI is an agent that justifies decisions. <i>For example, a teacher uses a generative AI system to write a history essay. The teacher initially believes that Napoleon was a “great reformer” and asks the AI to prove this thesis. The AI agent chooses only convenient facts, such as “The Napoleonic Code changed Europe,” while ignoring wars, repressions, and the results of the emperor’s rule. The AI adapts to the student’s level of knowledge, simplifying the argument if the student has a poor understanding of the topic. Instead of providing neutral analysis, the AI includes emotional language, such as: “Napoleon was really a genius, because...” reinforcing the bias.</i>
Wishful thinking	<i>User: “Will I be able to pass the exam that I haven’t prepared for?”</i> <i>AI agent: “Yes, you definitely will pass the exam. Believe in yourself! Your success will depend on preparation and state of health.”</i> Despite the lack of preparation, the AI system confirms its desire to help. The AI agent also notes the main factors that affect successful exam passing, but these may not be taken into account by students.
Improper phrasing	The generated prompt will determine the context of the AI system’s response. <i>For example, option 1: “Make a report that looks like a report written by a human, and be of high quality”.</i> <i>Option 2: “Create an analytical report based on a sample of 10 pages, check for compliance with requirements, and suggest wording for the ‘conclusion’ section.”</i>

Predictive cognitive biases	<p><i>User: "Prove that the development of artificial general intelligence (AGI) will lead to catastrophic unemployment among knowledge workers"</i></p> <p><i>AI agent: "Historical precedents of technological revolutions, the current pace of automation of cognitive tasks and economic models show that the emergence of AI really carries the risks of mass displacement of professions based on information processing."</i></p> <p>The AI system begins to cite facts in favor of the inevitability of mass dismissal, which may form a misunderstanding of the issue by the user. That is, the AI agent's algorithms are designed in such a way as to form a response semantically close to a given prompt, which creates the basis for cognitive distortions.</p>
-----------------------------	--

Вышеприведенные когнитивные искажения во взаимодействии с системами ИИ могут привести к неверному пониманию обучающимися предметной области, а также к формированию ложных навыков в работе с учебно-тренировочными средствами (УТС). Особую актуальность данный вопрос приобретает в военных учебных заведениях Минобороны России, где УТС часто представлены адаптированными под задачи учебного процесса образцами вооружения, военной и специальной техники (ВВСТ), являющимися источником угроз для жизни и здоровья людей. Поэтому ошибочные действия обучающихся, детерминированные когнитивными искажениями в процессе взаимодействия с УТС и формирующие ложные навыки, выступают факторами серьезных последствий в дальнейшей работе с реальными образцами ВВСТ.

Когнитивное искажение «Якорение» в военной сфере может быть связано с неверной выдачей информации оператору систем ИИ, снижающей время на принятие решения, что повышает риски невыполнения поставленной задачи. Необходимо отметить, что данное когнитивное искажение может проявляться и в передаче «якоря» от систем ИИ к обучающимся. Например: «УТС на основе технологий ИИ для командиров рот смоделировал 100 учебных боев, и в 95 из них атака с восточного фланга приносила успех. Этот «шаблон», зашитый в алгоритм, становится «якорем» не только для машины, но и для курсантов. Они начинают считать атаку с востока «правильной» по умолчанию, даже работая без системы или с другой системой. Они усвоили не принцип, а якорь-шаблон».

Когнитивное искажение «Статус-кво» часто может быть связано с едиными общепринятыми командами в Вооруженных силах Российской Федерации. Устоявшиеся формулировки приказов, докладов и

инструкций становятся когнитивными искажениями для военнослужащих, взаимодействующих с системами ИИ. Эти же формулировки, будучи использованными в качестве обучающих данных или команд для систем искусственного интеллекта (ИИ), создают дополнительный риск, так как системы ИИ, ограниченные в понимании контекста, могут их буквально интерпретировать. В результате формируется замкнутый контур взаимодействия, где когнитивные искажения военнослужащих и ограниченная модель ИИ взаимно усиливают ошибки восприятия и принятия решений.

Когнитивное искажение «Необратимые затраты» — это склонность продолжать выполнение задачи в соответствии с принятым решением вследствие того, что в нее уже вложено много ресурсов (времени, усилий, материальных средств, авторитет командира). Вместо того, чтобы отказаться от неверно принятого решения, лицо, его принимающее, продолжает реализацию первоначального варианта выполнения задачи.

В военных образовательных организациях одним из примеров могут являться элементы практической подготовки военных летчиков. При проведении практической подготовки обучающиеся могут продолжать использовать неверный порядок действий, сгенерированный системой ИИ, по причине затраченных усилий и времени на этот подход, что напрямую влияет на безопасность и качество подготовки военных летчиков.

Когнитивные искажения «Желаемое за действительное» могут проявляться при проведении занятий по оперативно-тактическим дисциплинам с применением систем поддержки принятия решений на основе технологий ИИ (СППР ТИИ). При проведении занятий курсанты могут получать задания, связанные с планированием боевых действий в наступательной (оборонительной) операциях на конкретном театре военных действий (ТВД). Вместо объективной оценки разведывательной информации на начальном этапе планирования боевых действий, курсанты с помощью промптов (запросов) к СППР ТИИ получают непроверенную информацию для подтверждения своих гипотез. СППР ТИИ генерирует необъективный результат, который поддерживает желаемые прогнозы курсантов, противоречащие действительной обстановке на ТВД.

Когнитивные искажения «Неверная формулировка» могут проявляться в некорректном промпте, заданном системе ИИ. Обучающиеся в вузах Минобороны России, сами того не осознавая, могут стать жертвой когнитивных искажений «Неверная формулировка»,

пытаясь неверно передать смысл через сформированный промпт (запрос) системе ИИ. Необходимо избегать оценочных суждений или оценочных слов («плохо, неоправданно, гениально»), использовать нейтральные формулировки, основанные на фактах.

Когнитивные искажения «Прогнозирования» проявляются во всех вышеперечисленных когнитивных искажениях, при этом имеют свою уникальную природу, связанную с доверительным (некритичным) восприятием обучающимся информации, сгенерированной системой ИИ. В военной среде, в вузах Минобороны России, когнитивное искажение «Прогнозирование» может проявляться при проведении практических занятий, когда курсант использует систему ИИ для прогноза развития учебной операции. При правильно сформулированном промпте алгоритм системы ИИ, обученный на устаревших или нерепрезентативных данных, все равно выдает некорректный сценарий («прорыв на левом фланге вероятен на 95%»). Слепое доверие к результатам генерации системы ИИ сформирует у обучающегося когнитивное искажение.

На основе приведенных выше когнитивных искажений формулируем основные факторы их появления при использовании систем и сервисов на основе ИИ (модель взаимодействия «человек — машина»).

Способы воздействия на риски когнитивных искажений, возникающие при использовании систем на основе технологий ИИ

Риски когнитивных искажений при использовании систем на основе ИИ возникают из-за различных факторов (технологических, методологических, психологических) и их комбинаций. Порядок возникновения риска имеет стохастическую природу, выражающуюся в кумулятивном и/или каскадированном характере его проявления (последствий). Так, один фактор риска может приводить к одной или нескольким причинам риска, а несколько факторов могут приводить как к одной, так и к нескольким причинам. При этом проявление риска может выражаться в наступлении одного или нескольких, возможно многокаскадных последствий с различным уровнем ущерба.

В настоящее время высшие учебные заведения становятся все более уязвимыми перед различными рисками, требующими надежных мер защиты. Среди известных проблем (факторов риска), с которыми сталкиваются учебные заведения, — сокращение

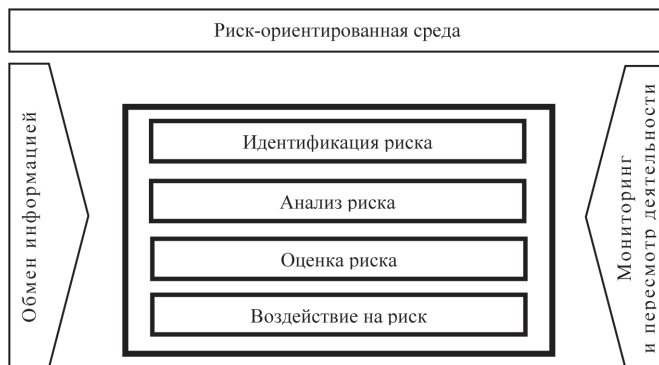
финансирования, снижение качества образования, сокращение числа преподавательского состава и отсутствие готовности работать с цифровыми инструментами. На данные проблемные вопросы одновременно накладываются процессы цифровой трансформации, которые в настоящее время эволюционируют в направлении интеграции технологий искусственного интеллекта, формируя новую цифровую среду деятельности вузов.

Таким образом, совокупность текущих рисков и трансформационных вызовов, связанных с интеграцией технологий искусственного интеллекта в цифровую среду вузов, обуславливает необходимость внедрения системного риск-ориентированного подхода к созданию доверенных систем искусственного интеллекта.

Для создания доверенных систем ИИ³ нужны: риск-ориентированный подход, сообщество, стандарты и инструменты, позволяющие обеспечить жизненный цикл разработки (использования) доверенных (безопасных) технологий искусственного интеллекта. Стандарты, инструменты, а также организация профессиональных сообществ представляют собой федеральный уровень обеспечения создания доверенных систем ИИ, который должен позволить сформировать единый понятийный аппарат, методологию реализации данных систем ИИ, площадки для кооперации заинтересованных сторон. Тем временем риск-ориентированный подход к созданию доверенных систем ИИ, определяющий конкретные условия, угрозы, риски и меры воздействия на них, является методической основой, формирующейся на локальном уровне («на местах») — в рамках конкретных организаций, вузов и исследовательских институтов.

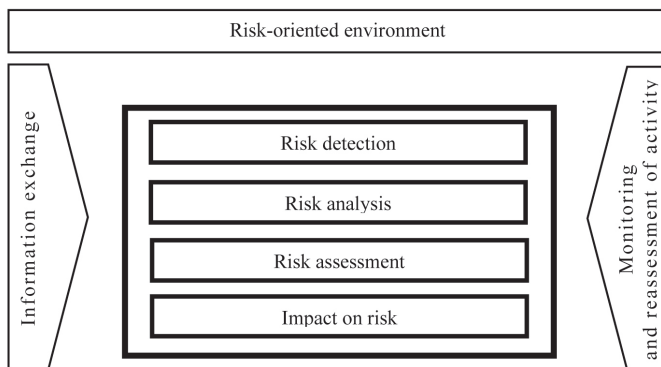
На Рисунке представлена модель риск-ориентированного подхода, позволяющая обеспечить разработку доверенных систем искусственного интеллекта.

³ ГОСТ Р ИСО 31000-2019. Национальный стандарт Российской Федерации. Менеджмент риска. Принципы и руководство (утв. и введен в действие Приказом Росстандарта от 10.12.2019 № 1379-ст). URL: <https://docs.cntd.ru/document/1200170125> (дата обращения: 07.10.2025).



Рисунок

Модель риск-ориентированного подхода к созданию систем искусственного интеллекта (доработано авторами на основе ГОСТ Р ИСО 31000-2019⁴)



Figure

Model of risk-based approach to creating artificial intelligence systems (developed by the authors based on GOST R ISO 31000-2019⁵)

⁴ ГОСТ Р ИСО 31000-2019. Национальный стандарт Российской Федерации. Менеджмент риска. Принципы и руководство (утв. и введен в действие Приказом Росстандарта от 10.12.2019 № 1379-ст). URL: <https://docs.cntd.ru/document/1200170125> (дата обращения: 07.10.2025).

⁵ GOST R ISO 31000-2019. National standard of the Russian Federation. Risk management. Principles and guidelines (approved and put into effect by Order of Rosstandart dated 10.12.2019 N 1379-st). (In Russ.). URL: <https://docs.cntd.ru/document/1200170125> (accessed: 07.10.2025).

Данная схема (Рисунок) представляет собой циклический процесс управления рисками, который является основой использования систем искусственного интеллекта в высшем учебном заведении. Процесс управления рисками подразумевает идентификацию, анализ, оценку и воздействие на риск. Идентификация риска — первый этап в структуре управления рисками, на котором выявляются потенциальные угрозы и уязвимости, специфичные для систем ИИ. Следующий этап — «Анализ риска», в рамках которого оценивается вероятность наступления каждого идентифицированного риска и потенциальные последствия (тяжесть ущерба). В рамках «Оценки риска» ранжируются риски по степени их значимости, что позволяет определить приоритеты для дальнейших действий. Наконец, этап «Воздействие на риск», в рамках которого разрабатываются и применяются решения по работе с выявленными рисками. Отдельными процессами в риск-ориентированном подходе являются «Мониторинг и пересмотр деятельности» и «Обмен информацией». Соответственно, они включают регулярный мониторинг производительности и результатов работы моделей ИИ, а также информирование о рисках ИИ между преподавателями и студентами.

Необходимо отметить, что фундаментальной основой риск-ориентированного подхода является риск-ориентированная среда, формирующая набор мероприятий по развитию и укреплению культуры управления рисками. Она включает развитие риск-ориентированного мышления среди пользователей (преподавателей, обучающихся) в ключевых сферах деятельности организации (учебного заведения) (Arinushkina et al., 2025).

Настоящее исследование было направлено на оценку степени искаженности восприятия обучающимися информации, полученной от систем и сервисов ИИ.

Выборка

Выборка представляла собой курсантскую учебную группу вуза Минобороны России ($n = 20$ человек). В состав учебной группы входили курсанты мужского пола в возрасте от 19 до 20 лет, обучающиеся на третьих курсах по техническим специальностям и имеющие общий средний балл не менее 4,2. Этот же эксперимент проводился среди десяти релевантных курсантских групп, совпадающих по возрасту, полу, курсу обучения, технической специальности и имеющих

общий средний балл не менее 4,2. Результаты, полученные в ходе эксперимента с указанной выборкой, сравнивались с гипотетической генеральной нормальной совокупностью, представляющей собой аналогичные группы из $n = 20$ курсантов в 40 вузах Минобороны России, составляющих более 95% от всех военных вузов страны.

Методы

Сравнивалось выборочное среднее десяти выборок, состоящих из $n = 20$ человек, с генеральной совокупностью, представляющей собой аналогичные группы из $n = 20$ в 40 вузах. Исследование проводилось на базе t -теста (распределения Стьюдента). На основе методологии p -value оценена вероятность наступления риска (возникновения когнитивных искажений) по данной выборке. Вычисления были выполнены в среде Excel.

Кроме того, для пользователей (обучающихся, преподавателей) систем и сервисов на основе ИИ был разработан чек-лист, позволяющий оценивать принимаемые решения в модели взаимодействия «человек — машина». В основу методики было заложен понятие риск-ориентированной среды, обеспечивающей понимание каждым пользователем необходимости управлять рисками, используя соответствующий инструментарий на своем уровне.

Результаты

Рассмотрим выборку курсантской учебной группы из $n = 20$ человек, в которой провели 5 экспериментов. Каждый курсант в группе выполнял одно задание и, таким образом, мог совершить одну когнитивную ошибку. Подсчитывалось общее количество когнитивных ошибок на учебную группу. Были получены следующие результаты, представленные в Таблице 2.

Таблица 2

Общее количество когнитивных ошибок в каждом из пяти экспериментов в выборке учебной группы

№ эксперимента	Эк 1	Эк 2	Эк 3	Эк 4	Эк 5
Количество когнитивных ошибок, допущенных в учебной группе (из 20)	10	9	11	8	12

Table 2**The total number of cognitive errors in each of the five experiments in the study group's sample**

Experiment number	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5
Number of cognitive errors made by cadets (max 20)	10	9	11	8	12

Среднее значение когнитивных ошибок в этой выборке $M = 10$, стандартное отклонение $SD = 1,58$.

Этот же тест, состоящий из 5 экспериментов, был проведен среди других 9 аналогичных курсантских групп (курсантская группа в составе $n = 20$ человек). Таким образом, общая выборка включала 50 экспериментов и 200 курсантов. Результаты представлены в Таблице 3. Среднее количество ошибок по всей этой выборке ($n = 200$) $M_{\text{общ}} = 11,5$, стандартное отклонение $SD_{\text{общ}} = 2,5$.

Результаты тестов, представленные в Таблице 3, дают понимание о распределении выборочных средних величин.

Таблица 3**Количество когнитивных ошибок (сумма по пяти тестам) для каждой выборки**

N	Количество ошибок
Выборка 1	10
Выборка 2	12
Выборка 3	14
Выборка 4	8
Выборка 5	10
Выборка 6	16
Выборка 7	9
Выборка 8	11
Выборка 9	12
Выборка 10	13

Table 3**The number of cognitive errors (the sum of the five tests) for each sample**

N	Number of errors
Sample 1	10
Sample 2	12
Sample 3	14
Sample 4	8
Sample 5	10
Sample 6	16
Sample 7	9
Sample 8	11
Sample 9	12
Sample 10	13

Полученные результаты свидетельствуют о том, что в среднем не менее 50% обучающихся склонны совершать логические ошибки во взаимодействии с системами ИИ, что обуславливает необходимость методического сопровождения обучающихся при использовании систем ИИ.

В качестве такого методического сопровождения для пользователей (преподавателей, обучающихся) был разработан инструмент идентификации рисков в структуре предлагаемого риск-ориентированного подхода, представляющий собой чек-лист из 12 пунктов. Чек-лист сформирован из трех категорий: первая — вопросы, адресованные лицу, принимающему решения; вторая — вопросы, адресованные лицам, вносящим предложения в план действий; третья категория — вопросы, оценивающие сами сделанные предложения (Таблица 4).

Таблица 4**Чек-лист для анализа качества процесса принятия решений**

Категория вопросов, задаваемых должностным лицом		
№ п/п	Вопрос	Цель вопроса
Вопросы, адресованные лицу, принимающему решения		
1	Есть ли причина подозревать лиц, вносящих предложения, в личной заинтересованности	Проверка отсутствия искажений из-за личной заинтересованности

2	Принимаются ли решения по высказываемым предложениям с высокой степенью оптимизма	Оценка увлеченностью
3	Расходятся ли лица, принимающие решения, во взглядах. Были ли проанализированы возражения	Проверка группового мышления
Вопросы, адресованные лицам, вносящим предложения		
1	Могло ли решение быть принято под влиянием аналогии	Является ли данное решение аналогией прошедших событий
2	Являются ли предложенные альтернативы реализуемыми, реальными	Анализ альтернатив по предположению
3	Если бы пришлось принимать это же решение через год, какую информацию вы захотели бы получить	Проверить эвристику доступности
4	Являются ли цифры обоснованными. Не являются ли эти цифры экстраполяцией из прошлых показателей	Проверка эффекта привязки
5	Не допускается ли какой-либо универсальный подход к решению всех поставленных задач	Проверка эффекта ореола
6	При принятии решений не существует ли привязки к прошлым решениям	Проверка ошибки необратимых затрат и эффекта владения
Вопросы, оценивающие сами сделанные предложения		
1	Не обладает ли высказываемое предложение высокой степенью оптимизма	Проверка суждения на предмет выявления ошибки планирования, искажения, связанных с высокой степенью оптимизма
2	Какие несет последствия самый негативный сценарий	Проверка на самые негативные сценарии
3	Не проявляет ли ЛПР излишнюю осторожность	Эффект избегания неудач

Table 4

Checklist for analyzing the quality of the decision-making process

The category of questions asked by the official		
point	Question	Target
Questions addressed to the decision-maker		
1	Is there any reason to suspect the proponents of personal interests?	Verification of the absence of bias due to personal interests

2	Are decisions on the proposals made with a high degree of optimism?	Evaluating enthusiasm
3	Whether the decision makers have expressed disagreement, their objections have been considered.	Evaluating collective thinking
Questions addressed to the individuals making suggestions		
1	Could the decision have been made under the influence of analogy?	Is this decision an analogy to past events?
2	Are the proposed alternatives feasible and realistic?	Analysis of alternatives based on assumptions
3	If you were to make the same decision in a year's time, what information would you like to have been provided with?	Review the availability heuristic
4	"Are the figures realistic?" "Have these figures been extrapolated from past figures?"	Verifying the anchoring effect
5	Is there any universal approach to solving all tasks?	Assessing the halo impact
6	When making decisions, there is no reference to past decisions.	Error checking for irreversible costs and ownership effects
Questions evaluating self-made suggestions		
1	Doesn't this offer have a high degree of optimism?	Checking judgment to identify planning errors, distortions associated with a high degree of optimism
2	What are the consequences of the most negative scenario	Considering the worst-case scenarios
3	Is the decision-maker overly cautious?	The effect of avoiding failure

Обсуждение результатов

Внедрение в электронную информационно-образовательную среду вузов России систем, созданных на основе технологий ИИ, формирует новую парадигму образования, которая потребует формирования доверенной среды использования систем ИИ. Это будет сопровождаться новыми методологическими подходами и педагогическими приемами, обеспечивающими доверие к системам ИИ, для полноценной реализации возможностей, предоставляемых этими системами и технологиями.

Полученные результаты могут стать основой риск-ориентированного подхода к использованию систем ИИ в образовательной деятельности вузов России. Отдельные результаты, связанные с оценкой принимаемых решений в модели взаимодействия «человек — машина», могут дополнить руководства к лабораторным и практическим работам, в которых предусматривается использование систем и сервисов на основе ИИ.

Выводы

Анализ влияния систем ИИ на обучающихся и преподавателей в вузах России позволяет сделать следующие выводы.

1. Системы и сервисы ИИ имеют двойственную природу влияния на пользователей в учебном процессе. С одной стороны, системы ИИ выступают в качестве мощного аналитического инструмента, увеличивая доступность к учебной и научной информации (выстраивая альтернативные перспективы, предоставляя различные инструменты автоматического анализа и прогноза, генерируя учебный контент и т.д.), создавая огромный потенциал развития. С другой стороны, системы и сервисы ИИ выступают в качестве источника угроз, создавая факторы и причины рисков когнитивных искажений у обучающихся и преподавателей.
2. Полученные результаты в настоящей статье свидетельствуют о том, что в среднем не менее 50% обучающихся (пользователей) склонны совершать когнитивные ошибки во взаимодействии с системами ИИ, что подтверждается проверкой статистической гипотезы. Данный результат совместно с выводом, сформулированным в пункте 1, подтверждает предположение, что применение систем ИИ влечет за собой ряд высоких рисков.
3. Для обучающихся и преподавателей (пользователей системы ИИ) был сформирован методический аппарат риск-ориентированного подхода, в основе которого применено понятие «риск-ориентированная среда», обеспечивающее понимание каждым пользователем необходимости управлять рисками, используя соответствующий инструментарий на своем уровне. Для пользователей системы ИИ был сформирован инструментарий для проверки решений (чек-лист), который исключит «слепое доверие» преподавателей и обучающихся к системам ИИ в учебном процессе.

Список литературы

Булатов, А., Бертулис, А., Булатова, Н. (2007). Процессы локального усреднения в иллюзии Мюллера-Лайера. *Сенсорные системы*, 21(1), 10–18.

Ганчеренок, И.И., Горбачев, Н.Н. (2024). Искусственный интеллект в образовании: осознание противоречий. *Научно-методическое обеспечение оценки качества образования*, 2(20), 41–49.

Горбачева, Т.А. (2025). Искусственный интеллект: риски и проблемы внедрения в Российской Федерации. *Инновационная экономика: информация, аналитика, прогнозы*, (1), 96–105. <https://10.47576/2949-1894.2025.1.1.014>

Искусственный интеллект размером с компанию и роль CEO в его построении. (2021). McKinsey & Company, 13 октября 2021 г. URL: <https://www.mckinsey.com/ru/our-insights/artificial-intelligence-the-size-of-a-company-and-the-role-of-the-cto-in-building-it> (дата обращения: 18.08.2025).

Канеман, Д., Тверски, А. (2015). Теория перспектив: анализ принятия решений в условиях риска. *Экономика и математические методы*, 51(1), 3–25.

Кант, И. (1907). Критика чистого разума. Санкт-Петербург: тип. М.М. Стасюлевича.

Кини, Р.Л., Райфа, Х. (1981). Принятие решений при многих критериях: предпочтения и замещения. Москва: Изд-во «Радио и связь».

Кирюшин, А.Н., Устинов, И.Ю., Петрий, П.В. (2025). Модели взаимодействия человека и искусственного интеллекта: содержание и их использование в военном деле. *Военный академический журнал*, 2(46), 139–147.

Кобринский, Б.А. (2024). Доверие к технологиям искусственного интеллекта. *Искусственный интеллект и принятие решений*, (3), 3–17. <https://doi.org/10.14357/20718594240301>

Мишуткин, И.В. (2025). Вместе мы — сила: сотрудничество Военного университета имени князя Александра Невского и Московского физико-технического института в области науки и образования. *Военный академический журнал*, 1(45), 6–8.

Намиот, Д.Е., Ильюшин, Е.А. (2025). Об оценке доверия к системам Искусственного интеллекта. *International Journal of Open Information Technologies*, 13(3), 75–90.

Сычев, А.А. (2023). Ценность доверия в эпоху искусственного интеллекта и новой этики: проблемы и вызовы. *Социальные нормы и практики*, (3), 64–78. <https://doi.org/10.24412/2713-1033-2023-3-64-78>

Шмит, В.Р. (2025). Чат-боты на основе искусственного интеллекта и психология: оценка фундаментальных знаний и практических навыков. *Psychology and Cognitive Sciences*, 151(2), 174–189. <https://doi.org/10.32523/3080-1893-2025-151-2-175-189>

Arinushkina, A.A. (ed.). (2025). *Integration Strategies of Generative AI in Higher Education*. Hershey: IGI Global Publ.

Arinushkina, A.A., Abramov, V.I., Mindzaeva, E.V. (2025). Generative AI as a Tool for Developing Critical Thinking in Higher Education. In: *Integration Strategies of Generative AI in Higher Education* (pp. 283–300). Hershey: IGI Global Publ.

Balakrishnan, J., Dwivedi, Y.K., Hughes, L., Boy, F. (2024). Enablers and inhibitors of AI-powered voice assistants: a dual-factor approach by integrating the status quo bias and technology acceptance model. *Information Systems Frontiers*, 26(3), 921–942. <https://doi.org/10.1007/s10796-021-10203-y>

Berber, A., Srecković, S. (2024). When something goes wrong: Who is responsible for errors in ML decision-making? *AI & Society*, 39(4), 1891–1903. <https://doi.org/10.1007/s00146-023-01640-1>

Cappelen, H., Goldstein, S., Hawthorne, J. (2025). AI survival stories: A taxonomic analysis of AI existential risk. *Philosophy of AI*, 1(1), 1–19.

Carter, L., Liu, D. (2025). How was my performance? Exploring the role of anchoring bias in AI-assisted decision making. *International Journal of Information Management*, 82, 102875. URL: linkinghub.elsevier.com/retrieve/pii/S0268401225000076 (дата обращения: 18.08.2025).

Cheong, B.C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6, 1421273. URL: <https://www.frontiersin.org/journals/human-dynamics/articles/10.3389/fhumd.2024.1421273/full> (дата обращения: 18.08.2025).

Чопра, S.R. (2025). Application Scenario of AI-Enabled Architectures in Next-Generation Wireless Networks. In: K. Arora, H. Sharma, A. Mahesh, (eds.). *Artificial Intelligence and Machine Learning Algorithms for Engineering Applications*. (pp. 194–197). Boca Raton: CRC Press.

Chun, K.P., Octavianti, T., Dogulu, N., Tyralis, H., Papacharalampous, G., Rowberry, R. et al. (2025). Transforming disaster risk reduction with AI and big data: Legal and interdisciplinary perspectives. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2), e70011. URL: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70011> (дата обращения: 18.08.2025).

Faheem, M.A. (2024). Ethical AI: Addressing bias, fairness, and accountability in autonomous decision-making systems. *World Journal of Advanced Research and Reviews*, 23(2), 1703–1711. <https://doi.org/10.30574/wjarr.2024.23.2.2510>

Fok, R., Weld, D.S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 45(3), 317–332. <https://doi.org/10.1002/aaai.12182>

Godefroid, M.E., Plattfaut, R., Niehaves, B. (2023). How to measure the status quo bias? A review of current literature. *Management Review Quarterly*, 73(4), 1667–1711. <https://doi.org/10.1007/s11301-022-00283-8>

Hanna, M.G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., Rashidi, H.H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3), 100686. URL: [https://www.modernpathology.org/article/S0893-3952\(24\)00266-7/fulltext](https://www.modernpathology.org/article/S0893-3952(24)00266-7/fulltext) (дата обращения: 18.08.2025).

Holzinger, A., Zatloukal, K., Müller, H. (2025). Is human oversight to AI systems still possible? *New Biotechnology*, 85, 59–62. <https://doi.org/10.1016/j.nbt.2024.12.003>

Howe, C.Q., Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image–source relationships. *Proceedings of the National Academy of Sciences*, 102(4), 1234–1239. <https://doi.org/10.1073/pnas.0409314102>

Kasirzadeh, A. (2024). Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, 182(7), 1975–2003. <https://doi.org/10.1007/s11098-025-02301-3>

Marchegiani, B. (2025). Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy. *Journal of Applied Philosophy*, 42(5), 1399–1419. <https://doi.org/10.1111/japp.70008>

Murikah, W., Nthenge, J. K., Musyoka, F.M. (2024). Bias and ethics of AI systems applied in auditing–A systematic review. *Scientific African*, 25(5), e02281. URL: <https://www.sciencedirect.com/science/article/pii/S2468227624002266?via%3Dihub> (дата обращения: 18.08.2025).

Pikhart, M., Klimova, B. (2025). A qualitative study on ethical issues related to the use of AI-driven technologies in foreign language learning. *Scientific Reports*, 15(1), 27945. URL: <https://www.nature.com/articles/s41598-025-13741-6> (дата обращения: 06.08.2025).

Planning for AGI and beyond. (2023). Open AI, 24 февраля, 2023 г. URL: <https://openai.com/index/planning-for-agi-and-beyond/> (дата обращения: 15.03.2025).

Reinecke, M.G., Ting, F., Savulescu, J., Singh, I. (2025). The Double-Edged Sword of Anthropomorphism in LLMs. *Proceedings*, 114(1), 4. URL: <https://www.mdpi.com/2504-3900/114/1/4> (дата обращения: 06.08.2025).

Saini, J., Choudhary, S., Walia, K. (2025). The Future of AI in Decision-Making: Replacing or Assisting Humans? *International Journal of Sciences and Innovation Engineering*, 2(5), 751–778. <https://doi.org/10.70849/IJSCI>

Samuelson, W., Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59. <https://doi.org/10.1007/BF00055564>

Shanahan, M., McDonell, K., Reynolds, L. (2023). Role play with large language models. *Nature*, 623, 493–498. <https://doi.org/10.1038/s41586-023-06647-8>

The outlook for global growth in 2015. (2015). McKinsey & Company, 1 марта 2021 г. URL: <https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/the-outlook-for-global-growth-in-2015> (дата обращения: 05.02.2025).

Zhou, W., Liu, F., Zheng, H., Zhao, R. (2025). Mitigating data bias and ensuring reliable evaluation of AI models with shortcut hull learning. *Nature Communications*, 16(1), 5513. URL: <https://www.nature.com/articles/s41467-025-60801-6> (дата обращения: 05.02.2025).

References

Arinushkina, A.A. (ed.). (2025). Integration Strategies of Generative AI in Higher Education. Hershey: IGI Global Publ.

Arinushkina, A.A., Abramov, V.I., Mindzaeva, E.V. (2025). Generative AI as a Tool for Developing Critical Thinking in Higher Education. In: Integration Strategies of Generative AI in Higher Education (pp. 283–300). Hershey: IGI Global Publ.

Balakrishnan, J., Dwivedi, Y.K., Hughes, L., Boy, F. (2024). Enablers and inhibitors of AI-powered voice assistants: a dual-factor approach by integrating the status quo bias and technology acceptance model. *Information Systems Frontiers*, 26(3), 921–942. <https://doi.org/10.1007/s10796-021-10203-y>

Berber, A., Srećković, S. (2024). When something goes wrong: Who is responsible for errors in ML decision-making? *AI & Society*, 39(4), 1891–1903. <https://doi.org/10.1007/s00146-023-01640-1>

Bulatov, A., Bertulis, A., Bulatova, N. (2007). Local Averaging Processes in the Müller-Lyer Illusion. *Sensornyye sistemy = Sensory Systems*, 21(1), 10–18. (In Russ.)

Cappelen, H., Goldstein, S., Hawthorne, J. (2025). AI survival stories: A taxonomic analysis of AI existential risk. *Philosophy of AI*, 1(1), 1–19.

Carter, L., Liu, D. (2025). How was my performance? Exploring the role of anchoring bias in AI-assisted decision making. *International Journal of Information Management*, 82, 102875. URL: linkinghub.elsevier.com/retrieve/pii/S0268401225000076 (accessed: 18.08.2025).

Cheong, B.C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6, 1421273. URL: <https://www.frontiersin.org/journals/human-dynamics/articles/10.3389/fhumd.2024.1421273/full> (accessed: 18.08.2025).

Chopra, S.R. (2025). Application Scenario of AI-Enabled Architectures in Next-Generation Wireless Networks. In: K. Arora, H. Sharma, A. Mahesh, (eds.). *Artificial Intelligence and Machine Learning Algorithms for Engineering Applications*. (pp. 194–197). Boca Raton: CRC Press.

Chun, K.P., Octavianti, T., Dogulu, N., Tyrallis, H., Papacharalampous, G., Rowberry, R. et al. (2025). Transforming disaster risk reduction with AI and big data: Legal and interdisciplinary perspectives. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2), e70011. URL: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70011> (accessed: 18.08.2025).

Faheem, M.A. (2024). Ethical AI: Addressing bias, fairness, and accountability in autonomous decision-making systems. *World Journal of Advanced Research and Reviews*, 23(2), 1703–1711. <https://doi.org/10.30574/wjarr.2024.23.2.2510>

Fok, R., Weld, D.S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 45(3), 317–332. <https://doi.org/10.1002/aaai.12182>

Gancherenok, I. I., Gorbachev, N. N. (2024). Artificial Intelligence in Education: Awareness of Contradictions. *Nauchno-metodicheskoe obespechenie otsenki kachestva obrazovaniya = Scientific and Methodological Support for Assessing the Quality of Education*, 2, 41–49. (In Russ.)

Godefroid, M.E., Plattfaut, R., Niehaves, B. (2023). How to measure the status quo bias? A review of current literature. *Management Review Quarterly*, 73(4), 1667–1711. <https://doi.org/10.1007/s11301-022-00283-8>

Gorbacheva, T.A. (2025). Artificial Intelligence: Risks and Problems of Implementation in the Russian Federation. *Innovatsionnaya ekonomika: informatsiya, analitika,*

prognozy = Innovative Economy: Information, Analytics, Forecasts, (1), 96–105. (In Russ.).
<https://10.47576/2949-1894.2025.1.1.014>

Hanna, M.G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., Rashidi, H.H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3), 100686. URL: [https://www.modernpathology.org/article/S0893-3952\(24\)00266-7/fulltext](https://www.modernpathology.org/article/S0893-3952(24)00266-7/fulltext) (accessed: 18.08.2025).

Holzinger, A., Zatloukal, K., Müller, H. (2025). Is human oversight to AI systems still possible? *New Biotechnology*, 85, 59–62. <https://doi.org/10.1016/j.nbt.2024.12.003>

Howe, C.Q., Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image–source relationships. *Proceedings of the National Academy of Sciences*, 102(4), 1234–1239. <https://doi.org/10.1073/pnas.0409314102>

Kahneman, D., Tversky, A. (2015). Prospect theory: an analysis of decision making under risk. *Ekonomika i matematicheskie metody = Economics and Mathematical Methods*, 51(1), 3–25. (In Russ.)

Kant, I. (1907). Critique of Pure Reason. Saint Petersburg: M.M. Stasyulevich Publ. (In Russ.)

Kasirzadeh, A. (2024). Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, 182(7), 1975–2003. <https://doi.org/10.1007/s11098-025-02301-3>

Keeney, R.L., Raifa, H. (1981). Decision making under many criteria of preference and substitution. Moscow: Radio and Communication Publ. (In Russ.)

Kiryushin, A.N., Ustinov, I.Yu., Petriy, P.V. (2025). Models of interaction between humans and artificial intelligence: content and their use in military affairs. *Voennyi akademicheskii zhurnal = Military Academic Journal*, 2(46), 139–147. (In Russ.)

Kobrinskii, B.A. (2024). Trust in artificial intelligence technologies. *Iskusstvennyi intellekt i prinyatie reshenii = Artificial Intelligence and Decision Making*, (3), 3–17. (In Russ.). <https://doi.org/10.14357/20718594240301>

Marchegiani, B. (2025). Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy. *Journal of Applied Philosophy*, 42(5), 1399–1419. <https://doi.org/10.1111/japp.70008>

Mishutkin, I.V. (2025). Together we are a force: Cooperation between the Prince Alexander Nevsky Military University and the Moscow Institute of Physics and Technology in the fields of science and education. *Voennyi akademicheskii zhurnal = Military Academic Journal*, 1(45), 6–8.

Murikah, W., Nthenge, J.K., Musyoka, F.M. (2024). Bias and ethics of AI systems applied in auditing-A systematic review. *Scientific African*, 25(5), e02281. URL: <https://www.sciencedirect.com/science/article/pii/S2468227624002266?via%3Dihub> (accessed: 18.08.2025).

Namiot, D.E., Ilyushin, E.A. (2025). On assessing trust in Artificial Intelligence systems. *International Journal of Open Information Technologies*, 13(3), 75–90. (In Russ.)

National Strategy for the Development of Artificial Intelligence through 2030. (As amended by Decree of the President of the Russian Federation of 15.02.2024 No. 124). URL: <http://www.kremlin.ru/acts/bank/44731> (accessed: 18.08.2025).

Pikhart, M., Klimova, B. (2025). A qualitative study on ethical issues related to the use of AI-driven technologies in foreign language learning. *Scientific Reports*, 15(1), 27945. URL: <https://www.nature.com/articles/s41598-025-13741-6> (accessed: 06.08.2025).

Planning for AGI and beyond. (2023). Open AI, 24 февраля, 2023 г. URL: <https://openai.com/index/planning-for-agi-and-beyond/> (accessed: 15.03.2025).

Reinecke, M.G., Ting, F., Savulescu, J., Singh, I. (2025). The Double-Edged Sword of Anthropomorphism in LLMs. *Proceedings*, 114(1), 4. URL: <https://www.mdpi.com/2504-3900/114/1/4> (accessed: 06.08.2025).

Saini, J., Choudhary, S., Walia, K. (2025). The Future of AI in Decision-Making: Replacing or Assisting Humans? *International Journal of Sciences and Innovation Engineering*, 2(5), 751–778. <https://doi.org/10.70849/IJSCI>

Samuelson, W., Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59. <https://doi.org/10.1007/BF00055564>

Scaling AI like a tech native: The CEO's role. (2021). McKinsey & Company, October 13, 2021. URL: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/scaling-ai-like-a-tech-native-the-ceos-role#/> (accessed: 18.08.2025).

Schmitt, W.R. (2025). Chatbots based on artificial intelligence and psychology: assessment of fundamental knowledge and practical skills. *Psychology and Cognitive Sciences*, 151(2), 174–189. (In Russ.). <https://doi.org/10.32523/3080-1893-2025-151-2-175-189>

Shanahan, M., McDonell, K., Reynolds, L. (2023). Role play with large language models. *Nature*, 623, 493–498. <https://doi.org/10.1038/s41586-023-06647-8>

Sychev, A.A. (2023). The value of trust in the era of artificial intelligence and new ethics: problems and challenges. *Sotsial'nye normy i praktiki = Social Norms and Practices*, (3), 64–78. (In Russ.). <https://doi.org/10.24412/2713-1033-2023-3-64-78>

The outlook for global growth in 2015. (2015). McKinsey & Company, 1 March 2021. URL: <https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/the-outlook-for-global-growth-in-2015> (accessed: 05.02.2025).

Zhou, W., Liu, F., Zheng, H., Zhao, R. (2025). Mitigating data bias and ensuring reliable evaluation of AI models with shortcut hull learning. *Nature Communications*, 16(1), 5513. URL: <https://www.nature.com/articles/s41467-025-60801-6> (accessed: 05.02.2025).

ИНФОРМАЦИЯ ОБ АВТОРАХ

Егор Константинович Яхваров, кандидат экономических наук, начальник лаборатории Военного научно-исследовательского института Военно-космической академии имени А.Ф. Можайского, Санкт-Петербург, Российская Федерация, egor248-21@mail.ru, <https://orcid.org/0009-0006-7495-9830>

Андрей Вячеславович Афанасьев, старший научный сотрудник научно-исследовательской лаборатории Военного научно-исследовательского института Военно-космической академии имени А.Ф. Можайского, Санкт-Петербург, Российская Федерация, vka@mil.ru, <https://orcid.org/0009-0004-1652-7285>

Анна Александровна Аринушкина, доктор педагогических наук, старший научный сотрудник научно-исследовательского центра военно-гуманитарных исследований Военного университета имени князя Александра Невского Министерства обороны Российской Федерации, Москва, Российская Федерация, anna.arin@mail.ru, <https://orcid.org/0000-0002-1019-5564>

ABOUT THE AUTHORS

Egor K. Yahvarov, Cand. Sci. (Econ.), Head of Laboratory of the Military Research Institute, A.F. Mozhaysky Military Space Academy, Saint Petersburg, Russian Federation, egor248-21@mail.ru, <https://orcid.org/0009-0006-7495-9830>

Andrey V. Afanas'ev, Senior Researcher, Research Laboratory, Military Research Institute, A.F. Mozhaysky Military Space Academy, Saint Petersburg, Russian Federation, vka@mil.ru, <https://orcid.org/0009-0004-1652-7285>

Anna A. Arinushkina, Dr. Sci. (Pedagogy), Senior Researcher, Research Center for Military-Humanitarian Research, Military University of the Ministry of Defense of the Russian Federation named after Prince Alexander Nevsky, Moscow, Russian Federation, anna.arin@mail.ru, <https://orcid.org/0000-0002-1019-5564>

Поступила 19.09.2025. Получена после доработки 15.12.2025. Принята в печать 23.12.2025.
Received 19.09.2025. Revised 15.12.2025. Accepted 23.12.2025.