УДК: 004.89

Худайберидева Г. Б., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Кожухов Д. А., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Пименкова А. А., студент-бакалавр кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

ДЕЦЕНТРАЛИЗОВАННЫЕ АВТОНОМНЫЕ ИИ-АГЕНТЫ ДЛЯ ЭТИЧЕСКОГО АУДИТА В ПРОЦЕССЕ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

***Аннотация**: Предложена концепция системы непрерывного этического аудита в разработке программного обеспечения на основе сети взаимодействующих децентрализованных автономных ИИ-агентов. Система осуществляет анализ исходного кода, технической документации и коммуникаций разработчиков в реальном времени на предмет соответствия установленным этическим принципам и регуляторным требованиям. Ключевая инновация заключается в способности системы выявлять неочевидные этические дилеммы посредством глубокого контекстного анализа, превосходя возможности существующих решений, ориентированных преимущественно на формальные проверки. Децентрализованная архитектура обеспечивает устойчивость, масштабируемость и снижает риски, связанные с единой точкой отказа или контроля. Реализация данной концепции способна существенно повысить уровень интеграции этических соображений в жизненный цикл разработки программного обеспечения.*

Ключевые слова: *этический аудит, разработка программного обеспечения, искусственный интеллект, автономные агенты, децентрализованные системы, машинное обучение, обработка естественного языка, контекстный анализ, регуляторные требования, этика ИИ.*

Khudaiberideva G. B.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Kozhukhov D. A.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Pimenkova A. A.
bachelor's student at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

DECENTRALIZED AUTONOMOUS AI AGENTS FOR ETHICAL AUDITING IN THE SOFTWARE DEVELOPMENT PROCESS

Annotation: *The concept of a continuous ethical audit system in software development based on a network of interacting decentralized autonomous AI agents is proposed. The system analyzes the source code, technical documentation, and developer communications in real time for compliance with established ethical principles and regulatory requirements. The key innovation lies in the system's ability to identify non-obvious ethical dilemmas through deep contextual analysis, surpassing the capabilities of existing solutions focused primarily on formal verification. A decentralized architecture provides resilience, scalability, and*

reduces the risks associated with a single point of failure or control. The implementation of this concept can significantly increase the level of integration of ethical considerations into the software development lifecycle.

***Keywords**: ethical audit, software development, artificial intelligence, autonomous agents, decentralized systems, machine learning, natural language processing, contextual analysis, regulatory requirements, AI ethics.*

Введение

Активное внедрение систем искусственного интеллекта в различные сферы человеческой деятельности обусловливает возрастающую потребность в обеспечении их этической корректности и соответствия нормативно-правовым требованиям [1]. Традиционные подходы к этическому аудиту программного обеспечения, особенно систем с компонентами ИИ, зачастую носят эпизодический характер, осуществляются постфактум и требуют значительных ресурсов [2]. Существующие инструменты статического и динамического анализа кода, хотя и полезны для выявления уязвимостей безопасности или ошибок кодирования, не обладают достаточной глубиной для обнаружения сложных этических импликаций, заложенных в алгоритмах или структурах данных [3]. Этические последствия часто проистекают не из явных нарушений, а из контекста применения, неучтенных предубеждений в данных или неочевидных взаимодействий компонентов системы [4].

Современные тенденции указывают на необходимость смещения фокуса с реактивного подхода к этике в сторону проактивного и непрерывного интегрирования этических принципов на всех этапах жизненного цикла разработки ПО [5]. Однако практическая реализация непрерывного этического мониторинга сталкивается с проблемой сложности и ресурсоемкости привлечения человеческих экспертов в режиме реального времени. Это создает предпосылки для исследования возможностей автоматизации этического аудита с использованием технологий искусственного интеллекта.

Ограничения существующих подходов к этическому аудиту ПО

Текущие методологии этического аудита программного обеспечения можно условно разделить на два крупных класса: экспертно-ориентированные и инструментально-ориентированные. Экспертно-ориентированные подходы, включающие комитеты по этике, внешние аудиты и ревью, обладают высокой гибкостью и способностью к рассмотрению сложных дилемм, но страдают от субъективности, высокой стоимости, низкой частоты проведения и неспособности масштабироваться на большие объемы кода или данные [6]. Их дискретный характер противоречит принципам непрерывной интеграции и поставки, ставшим стандартом современной разработки ПО [7].

Инструментально-ориентированные подходы, такие как статические анализаторы кода с заложенными правилами проверки на предвзятость или сканеры соответствия регуляторным требованиям (например, GDPR, HIPAA), предлагают автоматизацию и скорость. Однако их возможности ограничены проверкой формальных, заранее известных паттернов и правил [8]. Они не способны к семантическому пониманию контекста, интерпретации неструктурированных данных (например, коммуникаций в чатах, комментариев кода) или выявлению этических проблем, возникающих из комбинации корректных по отдельности решений [9]. Существующие ИИ-решения в этой области часто представляют собой централизованные модели, анализирующие код или данные в изоляции, без учета динамики разработки и командного взаимодействия [10].

Концепция децентрализованной сети автономных ИИ-агентов

Для преодоления указанных ограничений предлагается концепция системы этического аудита, основанной на взаимодействии множества автономных ИИ-агентов, организованных в децентрализованную сеть. Каждый агент в этой сети обладает специализацией и функционирует относительно независимо, обмениваясь информацией и результатами

анализа с другими агентами для достижения общей цели – непрерывного мониторинга этической корректности процесса разработки.

Архитектура системы предполагает наличие различных типов агентов. Агенты анализа кода ответственны за сканирование исходного текста программ, выявляя паттерны, ассоциированные с потенциальными рисками дискриминации, нарушения приватности или отсутствия прозрачности алгоритмов [11]. Агенты анализа документации обрабатывают технические задания, спецификации требований, пользовательские соглашения и иную текстовую информацию, оценивая ясность, полноту и соответствие этическим гайдлайнам [12]. Агенты мониторинга коммуникаций анализируют переписку в системах управления проектами, чатах и почте разработчиков, выявляя обсуждения, потенциально указывающие на этические компромиссы, давление сроков, игнорирование рисков или непонимание требований [13].

Ключевым аспектом является специализированный тип агента – Агент контекстной интеграции. Его функция заключается в агрегации и синтезе информации, полученной от других агентов. Этот агент строит контекстную модель проекта, связывая фрагменты кода, части документации и элементы коммуникаций в единую смысловую картину [14]. Именно на этом уровне становится возможным выявление неочевидных этических дилемм. Например, комментарий в коде, описывающий упрощение алгоритма проверки данных в целях оптимизации, сам по себе может не нарушать правила. Однако в контексте обсуждения в чате о жестких сроках сдачи и спецификации требований, подчеркивающей критическую важность точности входных данных, это решение приобретает этическую значимость, указывая на потенциальный риск внедрения недостаточно надежной системы [15].

Принципы функционирования и инновационные аспекты

Функционирование сети агентов базируется на нескольких основополагающих принципах. Автономность агентов подразумевает их

способность к целеполаганию, восприятию среды (артефактов разработки), выполнению задач анализа и коммуникации без постоянного прямого управления из центра [16]. Децентрализация означает отсутствие единого управляющего узла; агенты взаимодействуют по принципам peer-to-peer или через распределенные реестры, что повышает отказоустойчивость и устраняет узкие места [17]. Координация достигается через обмен сообщениями по стандартизированным протоколам, содержащими результаты анализа, уровни уверенности и запросы на дополнительную информацию [18].

Главной инновацией системы является ее ориентация на выявление латентных этических проблем через глубокий контекстный анализ. В отличие от инструментов, проверяющих соответствие пунктам чек-листа или известным антипаттернам, предложенные ИИ-агенты способны к интерпретации смысла. Анализ естественного языка, применяемый агентами мониторинга коммуникаций и документации, выходит за рамки простого поиска ключевых слов; он включает анализ тональности, выявление имплицитных утверждений, определение логических связей между высказываниями [19]. Машинное обучение, в особенности методы глубокого обучения (трансформеры), позволяет агентам строить семантические представления текста и кода, выявляя сходства и противоречия между различными артефактами разработки [20]. Способность агента контекстной интеграции устанавливать связи между, казалось бы, разрозненными событиями или решениями является основой для обнаружения комплексных этических рисков, которые остаются невидимыми для традиционных методов [21].

Преимущества децентрализованной архитектуры

Использование децентрализованной архитектуры для системы этического аудита предоставляет ряд существенных преимуществ перед централизованными альтернативами. Устойчивость к сбоям возрастает,

поскольку выход из строя одного или нескольких агентов не парализует работу всей системы; оставшиеся агенты могут продолжать выполнение своих функций и частично компенсировать потерю [22]. Масштабируемость обеспечивается возможностью добавления новых агентов для обработки возрастающих объемов данных или для специализации на новых типах анализа без необходимости перепроектирования центрального ядра [23]. Отсутствие единой точки контроля снижает риски манипуляции системой или целенаправленного сокрытия нарушений.

Децентрализация также способствует прозрачности и доверию. Механизмы консенсуса или верификации, реализуемые в сети агентов (например, через распределенные реестры), могут использоваться для подтверждения фактов обнаружения потенциальных этических проблем и неизменности журналов аудита [24]. Это создает надежную основу для подотчетности процесса разработки. Кроме того, распределенная природа системы лучше соответствует распределенным моделям разработки ПО, характерным для современных команд, работающих в разных временных зонах и географических локациях [25].

Заключение

Концепция децентрализованной сети автономных ИИ-агентов для непрерывного этического аудита в процессе разработки программного обеспечения предлагает новый подход к решению критически важной проблемы обеспечения этичности ИИ-систем. Преодолевая ограничения существующих методов, основанных на эпизодических экспертных оценках или формальных инструментальных проверках, предложенная система обеспечивает постоянный мониторинг артефактов разработки на всех уровнях. Способность сети агентов к глубокому контекстному анализу, синтезу информации из разнородных источников и выявлению

неочевидных этических дилемм представляет собой значительный шаг вперед.

Инновационный потенциал концепции заключается в комбинации децентрализованной архитектуры, обеспечивающей устойчивость и масштабируемость, с возможностями современных методов ИИ для семантического понимания кода, текста и коммуникаций. Это позволяет системе функционировать как неотъемлемая часть процесса разработки, а не как внешний контрольный механизм. Реализация подобной системы способна существенно повысить уровень ответственности разработчиков, минимизировать риски внедрения этически проблемных решений и способствовать формированию культуры ответственной разработки ПО. Дальнейшие исследования должны быть направлены на разработку конкретных архитектурных решений, протоколов взаимодействия агентов, методологий обучения специализированных моделей машинного обучения и формализацию этических онтологий для различных предметных областей.

СПИСОК ЛИТЕРАТУРЫ:

1. Floridi L. et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations // Minds and Machines. – 2018. – Vol. 28. – P. 689–707.
2. Mittelstadt B. Principles alone cannot guarantee ethical AI // Nature Machine Intelligence. – 2019. – Vol. 1(11). – P. 501–507.
3. Gebru T. et al. Datasheets for datasets // Communications of the ACM. – 2021. – Vol. 64(12). – P. 86–92.
4. Selbst A.D. et al. Fairness and Abstraction in Sociotechnical Systems // Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). – 2019. – P. 59–68.
5. Jobin A., Ienca M., Vayena E. The global landscape of AI ethics guidelines // Nature Machine Intelligence. – 2019. – Vol. 1(9). – P. 389–399.

6. Morley J. et al. Ethical Assurance: A practical approach to the responsible design and implementation of AI systems in the public sector // The Alan Turing Institute. – 2021.
7. Hummer W. et al. Continuous auditing and continuous monitoring in a DevOps environment // IEEE Cloud Computing. – 2018. – Vol. 5(3). – P. 86–93.
8. Bellamy R.K.E. et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias // arXiv:1810.01943 [cs]. – 2018.
9. Rakova B. et al. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices // Proceedings of the ACM on Human-Computer Interaction. – 2021. – Vol. 5(CSCW1). – P. 1–23.
10. Schiff D. et al. Principles to Practices for Responsible AI: Closing the Gap // arXiv:2006.04707 [cs]. – 2020.
11. Madaio M.A. et al. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI // Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). – 2020. – P. 1–14.
12. Vakkuri V., Kemell K.-K., Abrahamsson P. ECCOLA — a Method for Implementing Ethically Aligned AI Systems // 2020 IEEE/ACM 42nd International Conference on Software Engineering Workshops (ICSEW). – 2020. – P. 195–202.
13. Storey M.-A. et al. The Emergence of Software Engineering in the GitHub Archive // 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). – 2019. – P. 490–494.
14. Mikalef P., Gupta M. Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance // Information & Management. – 2021. – Vol. 58(3). – P. 103434.

15. Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines // Minds and Machines. – 2020. – Vol. 30(1). – P. 99–120.
16. Wooldridge M. An Introduction to MultiAgent Systems. – 2nd ed. – John Wiley & Sons, 2009. – 484 p.
17. Wang S. et al. A Survey on Consensus Mechanisms and Mining Strategy Management in Blockchain Networks // IEEE Access. – 2019. – Vol. 7. – P. 22328–22370.
18. Ferber J., Weiss G. Multi-agent systems: An introduction to distributed artificial intelligence. – Addison-Wesley, 1999. – 620 p.
19. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv:1810.04805 [cs]. – 2018.
20. Vaswani A. et al. Attention is All you Need // Advances in Neural Information Processing Systems 30 (NIPS 2017). – 2017. – P. 5998–6008.
21. Dignum V. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. – Springer Nature, 2019. – 126 p.
22. Alami H. et al. Blockchain Technology in Healthcare: A Systematic Review // Blockchain in Healthcare Today. – 2019. – Vol. 2. – P. 1–14.
23. Khan L.U. et al. Federated Learning for Edge Networks: Resource Optimization and Incentive Mechanism // IEEE Communications Surveys & Tutorials. – 2020. – Vol. 22(2). – P. 1291–1331.
24. Pilkington M. Blockchain technology: principles and applications // Research Handbook on Digital Transformations. – Edward Elgar Publishing, 2016. – P. 225–253.
25. Sutherland J., Schwaber K. The Scrum Guide. – Scrum.org, 2020. – 19 p.
26. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) // Official Journal of the European Union. – 2016. – L 119. – P. 1–88.