

Организация мониторинга применения кодекса этики в приложениях ИИ

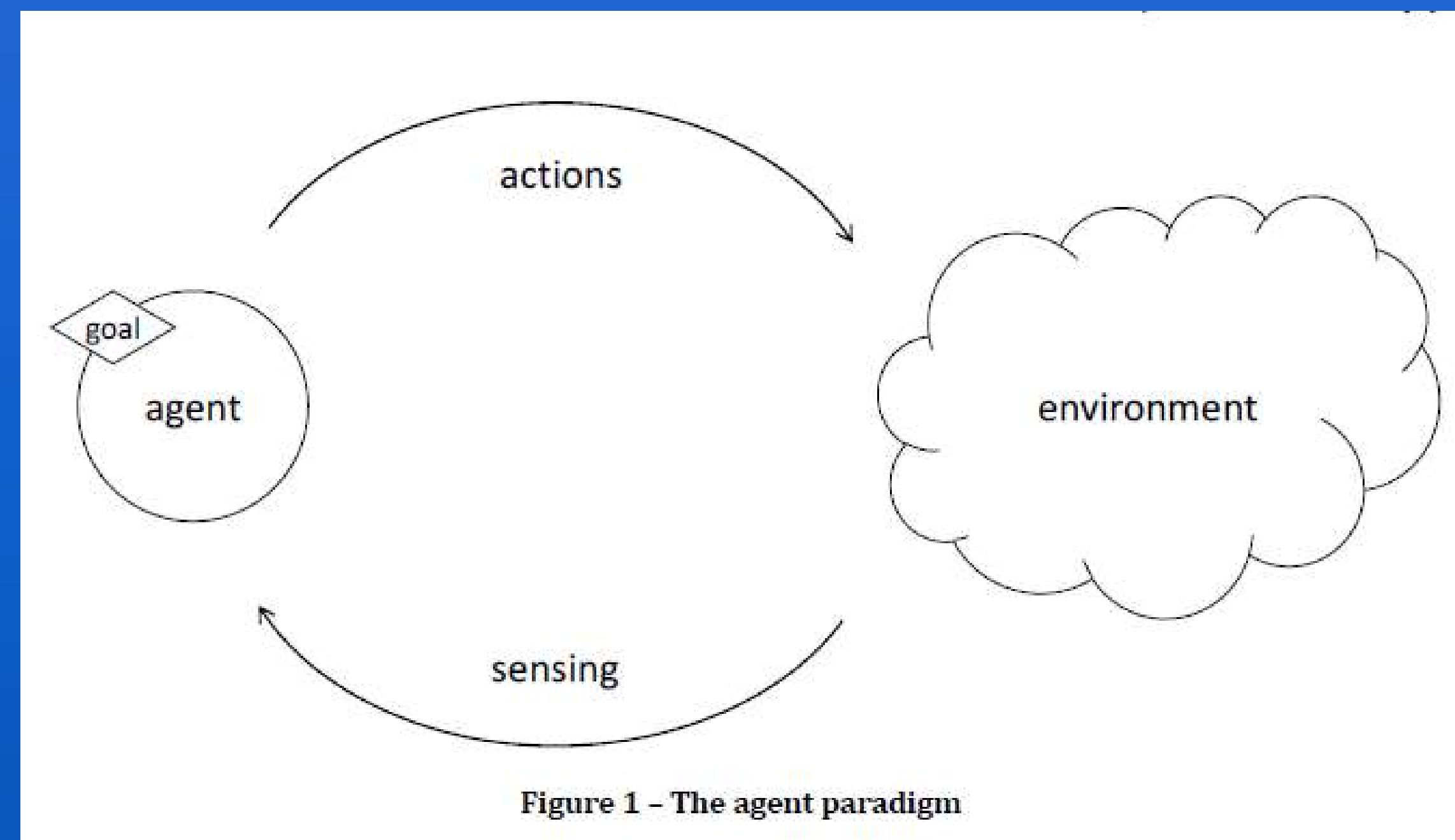
Андрей Кулешов

Уполномоченный по этике ИИ

Институт ИИ на базе МФТИ

Искусственный интеллект

- Способность приобретать, создавать, обрабатывать и применять знания, определенные в форме модели, для выполнения одной или нескольких поставленных задач
- Способность приобретать, создавать, обрабатывать и применять знания, определенные в форме модели, для повышения характеристик в выполнении одной или нескольких поставленных задач

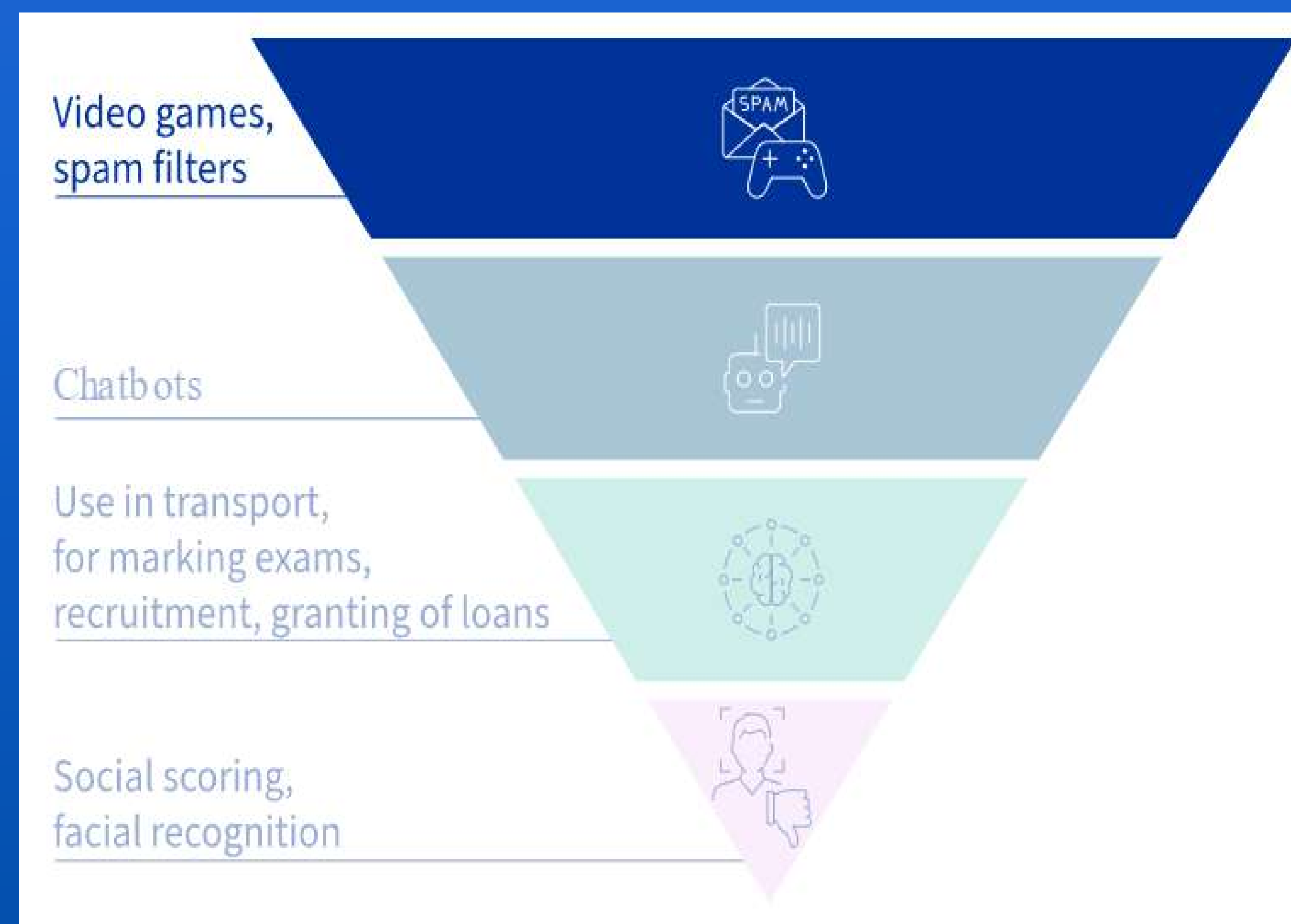


Чужой опыт: ценности

ЮНЕСКО

- (i) уважение прав и основных свобод человека;
- (ii) защита окружающей среды и экосистемы;
- (iii) продвижению разнообразия и инклюзивности;
- (iv) созданию условий для мирного сосуществования во взаимосвязанных обществах

Евросоюз



Чужой опыт: принципы

ЮНЕСКО

- пропорциональность выигрыша от ИИ с и не причинение ущерба;
- приоритет безопасности;
- справедливость и отсутствие дискриминации в соответствии с международными стандартами;
- соответствие принципам устойчивого развития;
- охрана частной жизни и защита данных;
- человеческий надзор за ИИ;
- прозрачность и объяснимость ИИ;
- ответственное использование ИИ;
- поддержка просвещения и грамотности по тематике ИИ;
- участие всех заинтересованных сторон в управлении развитием ИИ.

(принципы ЮНЕСКО лучше)

Евросоюз

- Управление и надзор со стороны человека
- Техническая надежность и безопасность
- Конфиденциальность и управление данными
- Прозрачность: данные, системы и бизнес-модели ИИ должны быть прозрачными.
- Разнообразие, недискриминация и справедливость: избегать несправедливой предвзятости, поскольку это может иметь многочисленные негативные последствия - от маргинализации уязвимых групп до усиления предрассудков и дискриминации
- Социальное и экологическое благополучие: приносить пользу всем людям, включая будущие поколения
- Подотчетность: Механизмы, обеспечивающие ответственность и подотчетность систем искусственного интеллекта и их результатов

ISO

- Прозрачность и объяснимость
- Справедливость и недискриминация
- Подотчетность и профессиональная ответственность
- Безопасность: охватывает как соображения физической безопасности, так и более широкие проблемы общественной безопасности, связанные с внедрением искусственного интеллекта
- Качество данных и защита конфиденциальности
- Управление рисками: выявлять, оценивать и устранять риски и возможности, связанные с ИИ, на протяжении всего жизненного цикла системы
- Ориентированность на человека и социальную ответственность
- Экологическая устойчивость и будущие поколения

CODE OF ETHICS AND STANDARDS OF PROFESSIONAL CONDUCT

PREAMBLE

The CFA Institute Code of Ethics and Standards of Professional Conduct are fundamental to the values of CFA Institute and essential to achieving its mission to continue to lead the investment profession globally by promoting the highest standards of ethics, education, and professional excellence for the ultimate benefit of society. High ethical standards are critical to maintaining the public's trust in financial markets and in the investment profession. Since their creation in the 1960s, the Code and Standards have promoted the integrity of CFA Institute members and served as a model for measuring the ethics of investment professionals globally, regardless of job function, cultural differences, or local laws and regulations. All CFA Institute members (including holders of the Chartered Financial Analyst® [CFA] designation) and CFA® candidates have the personal responsibility to embrace and uphold the provisions of the Code and Standards and are encouraged to notify their employer of this responsibility. Violations may result in disciplinary sanctions by CFA Institute. Sanctions can include revocation of membership, revocation of candidacy in the CFA Program, and revocation of the right to use the CFA designation.

THE CODE OF ETHICS

Members of CFA Institute (including CFA charterholders) and candidates for the CFA designation ("Members and Candidates") must:

- Act with integrity, competence, diligence, and respect and in an ethical manner with the public, clients, prospective clients, employers, employees, colleagues in the investment profession, and other participants in the global capital markets.
- Place the integrity of the investment profession and the interests of clients above their own personal interests.
- Use reasonable care and exercise independent professional judgment when conducting investment analysis, making investment recommendations, taking investment actions, and engaging in other professional activities.
- Practice and encourage others to practice in a professional and ethical manner that will reflect credit on themselves and the profession.
- Promote the integrity and viability of the global capital markets for the ultimate benefit of society.
- Maintain and improve their professional competence and strive to maintain and improve the competence of other investment professionals.

STANDARDS OF PROFESSIONAL CONDUCT

I. PROFESSIONALISM

A. Knowledge of the Law. Members and Candidates must understand and comply with all applicable laws, rules, and regulations (including the CFA Institute Code of Ethics and Standards of Professional Conduct) of any government, regulatory organization, licensing agency, or professional association governing their professional activities. In the event of conflict, Members and Candidates must comply with the more strict law, rule, or regulation. Members and Candidates must not knowingly participate or assist in and must dissociate from any violation of such laws, rules, or regulations.

B. Independence and Objectivity. Members and Candidates must use reasonable care and judgment to achieve and maintain independence and objectivity in their professional activities. Members and Candidates must not offer, solicit, or accept any gift, benefit, compensation, or consideration that reasonably could be expected to compromise their own or another's independence and objectivity.

C. Misrepresentation. Members and Candidates must not knowingly make any misrepresentations relating to investment analysis, recommendations, actions, or other professional activities.

D. Misconduct. Members and Candidates must not engage in any professional conduct involving dishonesty, fraud, or deceit or commit any act that reflects adversely on their professional reputation, integrity, or competence.

E. Competence. Members and Candidates must act with and maintain the competence necessary to fulfill their professional responsibilities.

II. INTEGRITY OF CAPITAL MARKETS

A. Material Nonpublic Information. Members and Candidates who possess material nonpublic information that could affect the value of an investment must not act or cause others to act on the information.

Источник вдохновения

Предмет обсуждения

- Утилитарный подход Кодекса
 - Профессиональное решение...
 - ... которое может негативно затронуть интересы человека/общества...
 - ... без их ведома/понимания
- Особенности собственно систем ИИ
 - «парадигма агента»
 - контекстность
- Особенности «акторов» в жизненном цикле ИИ

Кодекс этики ИИ в РФ

- Приоритет развития технологий ИИ в защите интересов и прав людей и отдельного человека
- Ответственность при создании и использовании ИИ – всегда несет человек
 1. Риск-ориентированный подход
 2. Ответственное отношение
 3. Предосторожность
 4. Непричинение вреда
- Применять по назначению и внедрять там, где это принесёт пользу людям
- Интересы развития технологий ИИ выше интересов конкуренции
- Прозрачность и правдивость в информировании об уровне развития технологий ИИ, их возможностях и рисках

Практические примеры

Робототехника

- Причинение физического вреда
- Злоупотребление и использование в криминальных целях
- Алгоритмы «черного ящика»
- Проблема установления ответственности

Генеративный ИИ

- Распространение и усиление дискриминации
- Глубокие фейки, медиа продукция вводящая в заблуждение
- Галлюцинации генеративного ИИ
- Риски безопасности, фишинг
- Изменения на рынке труда

Мониторинг: задачи для Комиссии



Задача и этические вопросы

- Методические материалы для информирования акторов ИИ о распознавании ситуаций, попадающих под действие Кодекса
- Сбор и классификация задач влекущих этические вопросы

Контекст

- Идентификация контекста принятия решений
- Определение уровня ответственности акторов ИИ в зависимости от контекста задачи
- Рекомендации по действиям в зависимости от контекста

Принятие решения

- Методические материалы по документированию принимаемых этических решений
- Публикация модельных решений о соответствии Кодексу в выборке релевантных контекстов
- Предоставление информации в РГ о принятых решениях

Анализ и обратная связь

- Сбор информации о последствиях принятых решений, возможно в рамках определенной модели в зависимости от контекста
- Анализ судебной практики
- Совершенствование методических материалов
- Публикация «белой книги» по этичному использованию ИИ

Элементы системы мониторинга

- Письменные процедуры и рекомендации
- Организационные структуры - комиссия по этике, институт уполномоченных в организациях участниках – *положение о комиссии и об уполномоченных*
- Обучение и просвещение по вопросам соответствия кодексу этики для акторов жизненного цикла систем ИИ. Привлечение руководителей организаций-подписантов.
- Механизмы мониторинга и обеспечения соответствия
 - Документирование принятия этических решений – *паспорт этического соответствия*
 - Системы добровольной отчетности – *форма отчётности*
 - Аудиты и оценки для выявления областей, требующих улучшения - обратная связь от Комиссии к уполномоченным
 - Системы мониторинга для выявления потенциальных нарушений. Создание надежных и конфиденциальных механизмов отчетности, таких как горячие линии по вопросам этики или системы онлайн-отчетности, позволяет сотрудникам сообщать о нарушениях или проблемах анонимно
 - Меры стимулирования/Дисциплинарные меры направленные на соблюдение требований, последовательно применяемые ко всем сотрудникам
- Документирование, оценка рисков на уровне комиссии и рекомендации по управлению ими
- Обратная связь, оценка и обновление кодекса для обеспечения эффективности и актуальности в свете меняющихся нормативных актов и организационных потребностей

Меры управления рисками

- (i) предотвращение,
- (ii) минимизация
- (iii) компенсация ущерба

Этический паспорт



Принятие этически
состоятельных решений в
команде которая работает с ИИ

1. Контекст использования
2. Вопросы этического соответствия
3. Кто отвечает за решение
 - Разработчик
 - Поставщик данных
 - Эксперт который оценивал характеристики
 - Работник надзора/регулирования
 - ...

Как действовать этично

Задokumentировать принятое решение

Обратиться к уполномоченному

Проинформировать публику

Отказаться от решения

«Дисассоциироваться» если организация действует неэтично

- Эскалировать до ЛПР
- Уволиться
- Проинформировать Комиссию







Спасибо за внимание!

a.kuleshov@phystech.ai