Азиломарские принципы искусственного интеллекта

Искусственный интеллект (ИИ) уже приносит каждый день пользу людям по всему миру. Его дальнейшее развитие на основе приведенных ниже принципов создаст потрясающие условия для помощи и расширения возможностей людей на протяжении будущих веков и десятилетий.

Исследования

1) **Цель исследований**. Целью исследований в сфере ИИ должно быть создание не неуправляемого, а полезного интеллекта.

2) **Финансирование исследований.** Инвестиции в ИИ должны сопровождаться финансированием исследований по обеспечению его полезного использования, которые должны дать ответы на самые острые вопросы в области компьютерных технологий, экономики, права, этики и социальных исследований, таких как:

- Как мы можем сделать будущие системы ИИ очень надежными, чтобы они делали то, что мы хотим, без сбоев или взломов?

- Как мы можем увеличить благополучие человечества посредством автоматизации при одновременном сохранении имеющихся у нас ресурсов и ценностей?

- Как мы можем развивать наши правовые системы, чтобы сделать их более справедливыми и эффективными, идущими в ногу с ИИ и учитывающими риски, связанные с ИИ?

- Какими ценностями должен быть связан ИИ, и какой статус он должен иметь с правовой и этической точек зрения?

3) **Связь науки и политики**. Должно быть установлено конструктивное и полезное взаимодействие между исследователями в сфере ИИ и теми, кто принимает решения о регулировании ИИ.

4) **Культура исследований**. В среде исследователей и разработчиков в сфере ИИ следует поощрять развитие культуры сотрудничества, доверия и прозрачности.

5) **Отказ от гонки**. Команды, разрабатывающие системы ИИ, должны активно сотрудничать между собой и не пытаться победить за счет игнорирования стандартов безопасности.

Этика и ценности

6) **Безопасность**. Системы ИИ должны быть безопасными и надежными на протяжении всего срока их эксплуатации, а также быть контролируемыми насколько это возможно и применимо.

7) **Прозрачность неудачи**. Если система ИИ причиняет вред, всегда должна быть возможность понять причину этого.

8) **Прозрачность правосудия**. Любое участие автономной системы в процессе принятия судебных решений должно сопровождаться предоставлением убедительных объяснений, которые могут быть перепроверены людьми из компетентных органов власти.

9) **Ответственность**. Разработчики продвинутых систем ИИ играют ключевую роль в формировании нравственных последствий использования ИИ, неправильного использования ИИ и действий ИИ; они имеют возможностью и несут обязанность влиять на такие последствия.

10) **Схожесть ценностей**. Высоко автономные системы ИИ должны быть разработаны таким образом, чтобы их цели и поведение были схожи с человеческими ценностями на протяжении всей их работы.

11) **Человеческие ценности**. Системы ИИ должны разрабатываться и работать таким образом, чтобы быть совместимыми с идеалами человеческого достоинства, его прав и свобод, многообразия культур.

12) **Конфиденциальность личных данных**. Учитывая способность систем ИИ анализировать и использовать личные данные, люди должны иметь права на доступ к своим личным данным, управление ими и осуществление контроля за их использованием.

13) **Свобода и неприкосновенность частной жизни**. Применение ИИ к персональным данным не должно необоснованно ограничивать реальную или предполагаемую свободу людей.

14) **Совместная выгода**. Технологии ИИ должны приносить пользу и расширять возможности как можно большего числа людей.

15) **Совместное процветание**. Экономическое процветание, достигнутое благодаря ИИ, должно широко использоваться в интересах всего человечества.

16) **Человеческий контроль**. Люди должны сами выбирать, как использовать системы ИИ для достижения своих целей, и использовать ли их для этого вообще.

17) **Устойчивость**. Власть, получаемая благодаря контролю над высокоразвитыми система ИИ, должна уважать и улучшать, а не подрывать социальные и гражданские процессы, от которых зависит здоровье общества.

18) **Гонка вооружений на основе ИИ**. Следует избегать гонки вооружений в разработке смертельного автономного оружия.

Долгосрочная перспектива

19) **Предостережение об ограничениях.** При отсутствии консенсуса об ином, нам следует избегать уверенных предположений относительно верхних пределов будущих возможностей ИИ.

20) **Важность.** Продвинутый ИИ может повлечь за собой коренное изменение в истории жизни нашей планеты, поэтому он должен разрабатываться и управляться с соответствующим вниманием и способностями.

21) **Риски.** Риски, создаваемые системами ИИ, особенно катастрофические или экзистенциальные риски, должны предвидеться, а их наступление минимизироваться за счет усилий, сопоставимых с ожидаемым последствием реализации этих рисков.

22) **Рекурсивное самосовершенствование.** Системы ИИ, разрабатываемые с возможностью рекурсивного самосовершенствования или самовоспроизведения с последующим быстрым увеличением их количества или качества, должны отвечать строгим критериям безопасности и контроля.

23) **Общее благо.** Суперинтеллект должен разрабатываться только для служения широко разделяемым этическим идеалам и на благо всего человечества, а не одного государства или организации.