КТО НЕСЕТ ОТВЕТСТВЕННОСТЬ ЗА ПРИМЕНЕНИЕ ОРУЖИЯ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ?

И.Ю. Ларионов
Санкт-Петербургский государственный университет
Санкт-Петербург

АННОТАЦИЯ

В докладе рассматривается проблема морального агентства (moral agency) в связи с применением технологий искусственного интеллекта для автономно действующих военных систем (smart weapon). Дается анализ и оценка проработанных концептуальных стратегий решения поставленной проблемы: «Рекомендации по этичному использованию искусственного интеллекта» (Минобороны США) и «Кодекс этики в сфере искусственного интеллекта» (Альянс в сфере искусственного интеллекта РФ). Концептуальный фокус направлен на исследование проблем определения «искусственного морального агента» в современных исследовательских программах, развиваемых на основании морально-философских теорий субъекта ответственности. Проблематика акцентирована на вопросы прозрачности и отслеживаемости (traceability) работы систем искусственного интеллекта со стороны остальных участников взаимодействий. Также показано значение проблемы двойного эффекта («double effect») в применении искусственного интеллекта в случаях, если машина начала наносить незапланированный ущерб.

Ключевые слова: моральная агентность, искусственный интеллект, ответственность, высокоточное оружие.

WHO IS RESPONSIBLE FOR THE USE OF WEAPONS WITH ARTIFICIAL INTELLIGENCE?

I.Yu.Larionov
Saint-Petersburg State University
Saint-Petersburg

This report examines the problem of moral agency in relation to the using of artificial intelligence technologies in the autonomous military systems (smart weapons). I analyze the conceptual strategies in "Recommendations on the Ethical Use of Artificial Intelligence" (Department of Defense, USA) and "Code of Ethics for Artificial Intelligence" (Artificial Intelligence Alliance of Russia). My conceptual focus is to investigate the issue of defining the "artificial moral agent" in contemporary research programs based on the ethical and philosophical theories of the subjective responsibility. The problematic focuses on the problems of transparency and traceability of artificial intelligence systems. I also show the importance of the "double effect" problem in the application of artificial intelligence in cases where a machine could inflict any unplanned damage.

Keywords: moral agency, artificial intelligence, responsibility, precision-guided munition.

В докладе рассматривается проблема морального агентства (moral agency) в связи с применением технологий искусственного интеллекта для автономно действующих военных систем (smart weapon).

При этом обсуждение далее ограничивается областью этики высоких технологий или машинной этики, в рамках которой алгоритмы искусственного интеллекта сталкиваются с системой общественных отношений, но не касается вопроса нравственных ограничений войны.

Проработанные концептуальные стратегии решения проблемы моральной агентности в применении искусственного интеллекта в современном вооружении мы находим в США – «Рекомендации по этичному использованию искусственного интеллекта» (Минобороны США, Пентагон). Похожий документ разработан и в России – «Кодекс этики в сфере искусственного интеллекта» (Альянс в сфере искусственного интеллекта), однако относится ко всем сферам жизни; разработка военного кодекса еще ведется.

Документ Пентагона исходит из того, что «оппоненты (adversaries) и конкуренты» Америки – Китай и Россия – уже осознали потенциал искусственного интеллекта и приняли соответствующие национальные стратегии. Таким образом, налицо научно-техническая конкуренция стран. В обоих документах указывается, что научно-технический прогресс является ценностью, его дальнейшее движение следует

стимулировать, поэтому этическое регулирование требуется как область постоянного обсуждения проблем ответственного использования новой техники. Общества развитых стран поддерживает продолжение разработок искусственного интеллекта, несмотря на временную нормативную неопределенность, которая их сопровождает, и связанные с этим риски.

Концептуальный фокус направлен на исследование проблем определения «искусственного морального агента» в современных исследовательских программах, развиваемых на основании морально-философских теорий субъекта ответственности.

В документе Минобороны США отдельно отмечен важный принцип: систему с искусственным интеллектом нельзя считать автономной по умолчанию (AI is not the same thing as autonomy). Искусственный интеллект, далее, сам по себе не является ни положительным, ни отрицательным. Таким образом, этические принципы могут относиться только к людям (американским военным, а также гражданским служащим Минобороны США) и должны быть направлены на недопущение при использовании искусственного интеллекта:

- злого умысла,
- неосторожности, в том числе в таких ее проявлениях, как халатность, недосмотр, легкомыслие и небрежность.

Российский «Кодекс этики в сфере искусственного интеллекта» вводит понятие «участники отношений в сфере искусственного интеллекта», далее по тексту именуемых «Акторами ИИ». Состав этих акторов раскрывается в пункте 1.3 раздела 2: разработчики, заказчики, операторы данных, эксперты, изготовители, эксплуатанты и операторы самих систем ИИ, разработчики всего объема нормативной базы, а также «иные лица, действия которых потенциально могут повлиять на результаты действий СИИ и т.п.». Таким образом, оба документа в качестве агентов искусственного интеллекта рассматривают именно людей. Речь об искусственном моральном агентстве не идет.

Можно сделать существенный вывод: при разработке этики искусственного интеллекта боевых систем, по всей вероятности, не учитывается (или находится вне сферы открытого обсуждения) ни сфера целенаправленного создания искусственного морального агента, ни тенденция наделения искусственного интеллекта моральным статусом на том основании, что он уже участвует во взаимодействии, описываемом нами в нравственных терминах, а использование военной и связанной с ней техники, очевидно, в них описывается.

В связи с этим вполне последовательно концептуальная проблематика документов оказывается акцентирована на вопросы прозрачности работы систем искусственного интеллекта.

Министерство обороны США декларирует, что использование им систем искусственного интеллекта является ответственным (responsible), беспристрастным (equitable), отслеживаемым (traceable), надежным (reliable) и управляемым (governable).

Ценности, описываемые в российском документе, можно обобщенно сформулировать так: гуманизм, свобода (человека), законность, «недискриминация», ответственное руководство рисками, предосторожность, информационная безопасность.

В целом, разработчики обеих стран ориентируются на два фокуса, две группы ценностей: права человека («беспристрастность» в американском документе относится к проблемам дискриминации) и ответственность агентов использования ИИ. Характерным является, например, следующее развернутое рассуждение из документа Минобороны США: люди (human beings) должны быть ответственными в отношении разработки, размещения, применения и перспектив (outcomes) систем искусственного интеллекта, «выказывая при этом надлежащий уровень рассудительности, здравого смысла (appropriate levels of judgment)».

Относительно принципа управляемости Минобороны США заявляет, что их системы искусственного интеллекта устроены таким образом, что всегда остается возможность их отключения – как автоматического, так и с вмешательством человека, – в случаях, когда они наносят ненамеренный («unintended») ущерб, приводят к незапланированной (опять «unintended») эскалации конфликта и т.п. Таким образом мы видим, что Пентагон здесь оставил место для использования т.н. принципа двойного эффекта («double effect») в применении искусственного интеллекта. Мы не ставим целью критику этой доктрины, а обращаем внимание на то, что ее применение, опять же, сводит вопрос ответственности к человеческим агентам, т.к. речь не идет о перенесении таковой на саму систему ИИ или объявление нанесения «незапланированного» вреда несчастным случаем.

Отсутствие подобного принципа в российском документе (напротив, в нем подробно и много говорится о минимизации рисков и категорической недопустимости причинения вреда человеку) не является показательным, т.к. этот набор принципов не является специализированно военным.

Доклад подготовлен в рамках проекта РНФ № 22-28-00379 «Трансформации морального агентства: этико-философский анализ».