DOI: 10.17803/2542-2472.2021.19.3.061-065

АВТОНОМНЫЕ БОЕВЫЕ СИСТЕМЫ: ЭТИКО-ПРАВОВОЙ ПОДХОД

Бахтеев Дмитрий Валерьевич, доцент кафедры криминалистики Уральского государственного юридического университета, кандидат юридических наук, доцент
ул. Колмогорова, д. 54, г. Екатеринбург, Россия, 620034
krim@usla.ru

Аннотация. *В статье рассматривается современное состояние технологии искусственного интеллекта применительно к автономным боевым системам. Основным юридическим вопросом в данном дискурсе является установление причины и ответственности за причинение вреда такой системой, интеграция таких ситуаций в систему военного права.*

Ключевые слова: *государство; право; международная безопасность; национальная безопасность; угроза; вызов; ответственность; искусственный интеллект; автономные боевые системы.*

AUTONOMOUS COMBAT SYSTEMS: AN ETHICAL AND LEGAL APPROACH

Dmitriy V. Bakhteev, Cand. Sci. (Law), Associate Professor, Department of Criminalistics, Ural State Law University
ul. Kolmogorova, d. 54, Yekaterinburg, Russia, 620034
krim@usla.ru

Abstract. *The paper examines the current state of artificial intelligence technology as applied to autonomous combat systems. The main legal issue in this discourse is the establishment of the cause and responsibility for harm caused by such a system, the integration of such situations into the system of military law.*

Keywords: *state; right; international security; National security; a threat; call; a responsibility; artificial intelligence; autonomous combat systems.*

Одним из главных научно-технических достижений первых двух десятилетий XXI в. вполне уместно назвать реставрацию технологий искусственного интеллекта. Такие технологии можно сравнить с ядерными: они обе в качестве своей исторической предпосылки и источника практической реализации имели развитие компьютерной техники. Ядерные исследования сначала привели к появлению, испытаниям и военному использованию ядерного оружия, приведшему к гибели десятков тысяч людей (атомные бомбардировки Хиросимы и Нагасаки в 1945 г.), однако в результате привели также к прорыву в области энергетики в виде атомных электростанций и последовавшему распространению электричества, что, в свою очередь, послужило дальнейшему развитию техники, в том числе компьютерной.

Технология искусственного интеллекта, разумеется, не способна привести к таким тяжким последствиям (по крайней мере, за небольшой отрезок времени), однако если продолжать названное сравнение, то сейчас общество находится виртуально в начале 1940-х гг.: до старта Манхэттенского проекта остается несколько лет, однако колоссальные перспективы технологии уже очевидны, то есть искусственный интеллект еще не вызвал существенных проблем, однако их предметная область уже вполне намечена, что ставит перед обществом (и юридическим сообществом) нелегкую задачу оценки рисков и выработки моделей их предотвращения/преодоления. Согласимся с абсолютно верным высказыванием А. В. Незнамова, согласно которому «если существует хотя бы 0,1 % вероятности того, что экзистенциальные риски есть, — юристы этими рисками обязаны заняться. А значит, область права в XXI в. ждут серьезные изменения»[1].

Несмотря на то что пока существуют системы искусственного интеллекта, ориентированные на решение частных конкретных задач, отдельные задачи к настоящему моменту времени являются для них невозможными. Так, системы компьютерного зрения на основе искусственных нейронных сетей не могут ни распознать предложенные оптические иллюзии, ни создать новые: «Единственные из известных нам оптических иллюзий были созданы эволюцией (к примеру, рисунки глаз на крыльях бабочки) или художниками-людьми. И в обоих случаях люди играли решающую роль в обеспечении обратной связи — люди могут видеть иллюзию»[2]. Исследование, проведенное в США, указывает, что все современные интеллектуальные системы при распознавании внешности представителей азиатской и негроидной рас имеют повышенный риск совершения ложно-положительной ошибки[3]. Разумеется, это объясняется тем, что датасеты, использованные для обучения таких систем, содержали больше изображений европеоидов, однако результаты работы таких систем уже успели назвать расистскими.

При массовом распространении систем распознавания (к примеру, внешности человека или транспортных средств), возможны случаи противодействия (внешние факторы ошибки) или внутренние сбои самой системы (внутренние факторы ошибки).

В силу наличия таких характеристик, как скрытые слои, глубокое обучение и т. п., системы искусственного интеллекта могут оказаться для человечества «черным ящиком»: механизм принятия ими решений неочевиден, известен лишь результат, для проверки которого может и не найтись корректного алгоритма. Вследствие этого может возникнуть юридический вопрос установления причины нанесенного вреда, без решения которого затруднено или в принципе невозможно установление пределов ответственности систем искусственного интеллекта. Так, к примеру, должен быть решен вопрос отграничения осознанного решения системы искусственного интеллекта от результатов неправильного обучения и от объективного сбоя или внешнего вмешательства. Данная проблема характерна для любых сфер человеческой деятельности, в которых будут функционировать именно киберфизические, овеществленные системы искусственного интеллекта (автономные транспортные средства, роботы, дроны и т. д.).

Существует прецедент: в марте 2020 г. американский дрон Kargu-2, работая в автономном режиме, без получения команд от оператора-человека, убил путем самоподрыва участника Ливийской национальной армии. Отметим, однако, что

[1] Цит. по: *Агапов И.* Искусственный интеллект в законе // Стандарт: Цифровая трансформация, ИТ, коммуникации, контент. 2018. № 3 (182). С. 10–14.

[2] *Williams R. M.* Optical Illusions Images Dataset / R. M. Williams, R. V. Yampolskiy // arxiv.org. 2018. 30 sep. Updated: 16.10.2018. URL: https://arxiv.org/pdf/1810.00415.pdf (accessed: 20.09.2021).

[3] NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software // National Institute of Standards and Technology: official site. Updated: 09.01.2020. URL: https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software (accessed: 20.09.2021).

информационная автономность была неполной: список потенциальных целей был загружен в систему дрона изначально. Однако идентификация цели на наведение на нее осуществлялись автоматически.

Представляется, что проблему использования искусственного интеллекта на поле боя можно рассматривать в двух аспектах.

Во-первых, сознательное использование искусственного интеллекта для исполнения приказов, нарушающих установленные международным правом правила ведения войны.

Во-вторых, несознательное нарушение таких правил, когда перед такой системой ставится задача, но не ограничиваются способы ее достижения, что приводит к тому, что боевой искусственный интеллект выбирает наиболее эффективный результат, не заботясь об иных, кроме как выполнение поставленной задачи, последствиях.

Автономные интеллектуальные системы, по сути, представляют собой третью, после животного и человека модель отношения к боевым действиям. В основе такой модели можно установить критерий эмпатии: животные вполне способны на конфликты и месть, однако в большинстве случаев не готовы передавать негативные эмоции членам сообщества, не участвовавшим в первоначальном непосредственном конфликте. Современные интеллектуальные системы, не обладая эмпатией, способны к стратегическим и последовательным действиям. Человек, как показывает история, не имеет ни социальных, ни биологических действенных ограничений к ведению войны, что не исключает стремления к минимизации таких потребностей. Одним из направлений такой минимизации является ограничение к использованию боевых автономных интеллектуальных киберфизических систем.

Ряд сотрудников Google уволились в знак протеста, еще 4000 подписали петицию с просьбой отказаться от контракта с военными. Более 1000 ученых в области искусственного интеллекта, этики и информационных технологий обратились к Google с открытым письмом с требованием прекратить работы над проектом и поддержать полный запрет автономного оружия[4], что привело к согласию компании ответственно подходить к таким сложным вопросам (хотя бы формальному)[5].

С точки зрения современных этических воззрений, разработка и тем более использование интеллектуального (автономного) оружия должно быть ограничено на как можно более высоком уровне, поскольку в перспективе такие технологии могут создать угрозу, сравнимую с ядерной, и потому требующие не меньших по масштабу средств сдерживания и взаимного контроля на межгосударственном уровне.

Группы негативных с точки зрения этики ситуаций, возникающих с использованием транспортных систем на основе искусственного интеллекта (беспилотных транспортных средств), сводятся к следующему. Каждая ложно-отрицательная ошибка может приводить к гибели людей и иным тяжким последствиям, что также неизбежно влечет утрату доверия к технологии, к примеру, в виде атаки боевым дроном некомбатантов или мирного населения. Ошибки такого рода могут быть вызваны как программно-аппаратными сбоями, так и вмешательством третьих лиц: системы распознавания лиц или объектов могут быть обмануты незначительными изменениями изображения[6]. Незаметные наклейки на дорожном знаке «Стоп» заставляют систему машинного зрения в автономном транспортном средстве автомобиле

[4] Open Letter in Support of Google Employees and Tech Workers // The International Committee for Robot Arms Control: official site. URL: https://www.icrac.net/open-letter-in-support-of-google-employees-and-tech-workers/ (accessed: 20.09.2021).

[5] *Бочкарева Н.* Google написала этический кодекс об искусственном интеллекте // TechFusion.ru. 2018. 10 июня. URL: https://techfusion.ru/google-napisala-eticheskij-kodeks-ob-iskusstvennom-intellekte/ (дата обращения: 20.09.2021).

[6] См.: Intriguing properties of neural networks / C. Szegedy, W. Zaremba, I. Sutskever [et al.] // arxiv.org. 2013. 21 dec. Updated: 19.02.2014. URL: https://arxiv.org/pdf/1312.6199.pdf (accessed: 20.09.2021).

принять его за «Уступи дорогу»[7]. Такого рода действия называются состязательными (adversarial) атаками и могут в перспективе использоваться для провокаций или маскировки целей в военных действиях.

Одним из ключевых цивилизационных рисков ближайшего будущего является распространение использования боевых автономных киберфизических систем. Автор не считает, что удастся остановить процесс совершенствования таких систем, однако задача юриспруденции — разработать жесткие правила использования такого оружия.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-29-16001 «Комплексное исследование правовых, криминалистических и этических аспектов, связанных с разработкой и функционированием систем искусственного интеллекта».

БИБЛИОГРАФИЯ

1. *Агапов И.* Искусственный интеллект в законе // Стандарт: Цифровая трансформация, ИТ, коммуникации, контент. — 2018. — № 3 (182). — С. 10–14.
2. *Бочкарева Н.* Google написала этический кодекс об искусственном интеллекте // TechFusion.ru: сайт про технологии для людей. — 2018. — 10 июн. — URL: https://techfusion.ru/google-napisala-eticheskij-kodeks-ob-iskusstvennom-intellekte/ (дата обращения: 20.09.2021).
3. Intriguing properties of neural networks / C. Szegedy, W. Zaremba, I. Sutskever [et al.] // arXiv.org. — 2013. — 21 dec. — Updated: 19.02.2014. — URL: https://arxiv.org/pdf/1312.6199.pdf (accessed: 20.09.2021).
4. NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software // National Institute of Standards and Technology: official site. — Updated: 09.01.2020. — URL: https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software (accessed: 20.09.2021).
5. Open Letter in Support of Google Employees and Tech Workers // The International Committee for Robot Arms Control: official site. — URL: https://www.icrac.net/open-letter-in-support-of-google-employees-and-tech-workers/ (accessed: 20.09.2021).
6. Robust Physical-World Attacks on Deep Learning Visual Classification / K. Eykholt, I. Evtimov, E. Fernandes [et al.] // arXiv.org. — 2017. — 27 jul. — Updated: 10.04.2018. — URL: https://arxiv.org/pdf/1707.08945.pdf (accessed: 20.09.2021).
7. *Williams R. M.* Optical Illusions Images Dataset / R. M. Williams, R. V. Yampolskiy // arXiv.org. — 2018. — 30 sep. — Updated: 16.10.2018. — URL: https://arxiv.org/pdf/1810.00415.pdf (accessed: 20.09.2021).

REFERENCES

1. Agapov I. Iskusstvennyy intellekt v zakone [Artificial intelligence in law]. *Standard: Digital transformation, IT, Communications, Content.* 2018;3(182):10-14 (In Russ.).
2. Bochkareva N. Google napisala eticheskiy kodeks ob iskusstvennom intellekte [Google wrote a code of ethics about artificial intelligence]. TechFusion.ru: a site about technologies for people. 2018, Jun 10. Available at: https://techfusion.ru/google-napisala-eticheskij-kodeks-ob-iskusstvennom-intellekte/ [Accessed: 20 Sept 2021] (In Russ.).
3. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv.org. 2013. 21 dec. Updated: 19.02.2014. Available at: https://arxiv.org/pdf/1312.6199.pdf [Accessed: 20 Sept 2021].

[7] См.: Robust Physical-World Attacks on Deep Learning Visual Classification / K. Eykholt, I. Evtimov, E. Fernandes [et al.] // arxiv.org. 2017. 27 jul. Updated: 10.04.2018. URL: https://arxiv.org/pdf/1707.08945.pdf (accessed: 20.09.2021).

4. NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software. National Institute of Standards and Technology: official site. Available at: https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software [Accessed: 20 Sept 2021].
5. Open Letter in Support of Google Employees and Tech Workers. The International Committee for Robot Arms Control: official site. Available at: https://www.icrac.net/open-letter-in-support-of-google-employees-and-tech-workers/ [Accessed: 20 Sept 2021].
6. Eykholt K, Evtimov I, E Fernandes, et al. Robust Physical-World Attacks on Deep Learning Visual Classification. arXiv.org. 2017. 27 jul. Available at: https://arxiv.org/pdf/1707.08945.pdf [Accessed: 20 Sept 2021].
7. Williams RM. Optical Illusions Images Dataset. arXiv.org. 2018. 30 sep. Available at: https://arxiv.org/pdf/1810.00415.pdf [Accessed: 20 Sept 2021].