Профессор Е. Н. Пашенцев о злонамеренном использовании искусственного интеллекта и новых вызовах для информационно-психологической безопасности России

В Военной академии Генерального штаба Вооруженных Сил Российской Федерации 16 мая 2019 г. прошел научный семинар на тему «Многополярность как фактор мировой стабильности и безопасности Российской Федерации». В мероприятии приняли участие представители Государственной Думы Федерального Собрания Российской Федерации, Объединенного штаба Организации договора о коллективной безопасности, высших учебных заведений и научно-исследовательских организаций Министерства обороны Российской Федерации. На семинаре выступили представители Военной академии Генерального штаба ВС РФ, Московского государственного университета им. М. В. Ломоносова, Финансового университета при Правительстве Российской Федерации, Института актуальных международных проблем Дипломатической академии Министерства иностранных дел России, Института проблем безопасности СНГ, Академии управления Министерства внутренних дел Российской Федерации и других вузов и организаций. В ходе научного семинара обсуждались современные концепции развития многополярного мира, процессы трансформации основных центров силы, факторы, влияющие на безопасность Российской Федерации, и другие вопросы.

В своем докладе «Злонамеренное использование искусственного интеллекта и новые вызовы для информационно-психологической безопасности России» доктор исторических наук, профессор, ведущий научный сотрудник Института актуальных международных проблем Дипломатической академии МИД РФ, профессор МГУ им. М. В. Ломоносова Евгений Николаевич Пашенцев представил свое видение актуальных и перспективных проблем информационно-психологической безопасности России в контексте растущего злонамеренного использования искусственного интеллекта (ЗИИИ).

Общая крайне напряженная ситуация в мире представляет несомненную угрозу национальной безопасности России. Важным элементом этой угрозы являются быстро

растущие угрозы использования ИИ для манипулирования общественным сознанием на международном уровне. Проф. Пашенцев определил международную информационно-психологическую безопасность (МИПБ) как защищенность системы международных отношений от негативных информационно-психологических воздействий, связанных с разнообразными факторами международного развития. Среди последних он выделил целенаправленную деятельность различных государственных, негосударственных и наднациональных акторов по частичной/полной, локальной/глобальной, кратковременной/долгосрочной, латентной/открытой дестабилизации международного положения с целью получения конкурентных преимуществ вплоть до физического уничтожения противника.

В рамках гибридной войны с помощью материальных средств воздействия в различных сферах (экономической, политической, военной и др.) субъекты международных отношений осуществляют негативное *опосредованное и непосредственное* воздействие на общественное сознание противника, а также нередко и на свое собственное состояние, своих союзников, нейтральных акторов. Злонамеренное использование искусственного интеллекта способно стать очень серьезной угрозой информационно-психологической безопасности России, поскольку наша страна взаимодействует с внешним миром через множество разносторонних связей на государственном, групповом и индивидуальном уровнях. Их невозможно и нецелесообразно полностью контролировать, а тем более разорвать, поскольку от этого пострадает прежде всего сама Россия. Целенаправленное использование ИИ может резко увеличить эффективность информационно-психологических операций против России, что требует системного анализа этой проблемы.

По мнению Е. Н. Пашенцева, ЗИИИ может позволить преступным акторам более успешным образом, чем до сих пор:

- спровоцировать общественную реакцию на **несуществующий фактор** общественного развития в интересах заказчика информационно-психологического воздействия. Целевая аудитория видит то, **что не существует.**

- представить **ложную интерпретацию существующего фактора** общественного развития и таким образом вызвать искомую целевую реакцию. Аудитория видит то, что существует, но в **ложном свете.**

- существенным и опасным образом усилить (уменьшить) общественную реакцию на реальный фактор общественного развития. Аудитория видит то, что существует, но **реагирует неадекватным образом.**

Е. Н. Пашенцев предложил *следующую классификацию* ЗИИИ по степени реализации его возможностей:

• существующая практика ЗИИИ;

• существующие возможности ЗИИИ, которые еще не были использованы на практике (такая вероятность связана с широким спектром быстро развивающихся новых возможностей ИИ – не все они сразу входят в спектр реализованных возможностей ЗИИИ;

• будущие возможности ЗИИИ на основе текущих разработок и будущих исследований (оценка должна быть дана на ближайшую, среднесрочную и долгосрочную перспективы);

• неопознанные риски, также известные как «неизвестное в неизвестном». Не все разработки в сфере ИИ можно точно оценить. Готовность встретить неожиданные скрытые риски имеет решающее значение.

Важно и необходимо использовать независимые команды разных специалистов и сам ИИ для оценки возможностей ЗИИИ.

Е. Н. Пашенцев также полагает возможным следующие варианты классификации ЗИИИ:

• по территориальному охвату: местный, региональный, глобальный;

• по степени нанесения ущерба: незначительный, значительный, крупный, катастрофический;

• по скорости распространения: медленный, быстрый, стремительный;

• по форме распространения: открытый, скрытый.

Среди возможных угроз ЗИИИ (см. подробнее: Bazarkina and Pashentsev, 2019), которые могут вызвать серьезное дестабилизирующее воздействие на социально-политическое развитие той или иной страны и системы международных отношений, включая сферу МИПБ, Е. Н. Пашенцев назвал следующие:

• *Рост комплексных всеохватывающих систем с активным или ведущим участием ИИ* повышает риск злонамеренного перехвата контроля над такими системами. Многочисленные объекты инфраструктуры, например, роботизированные самообучающиеся транспортные системы с централизованным управлением посредством ИИ, могут стать удобной мишенью для высокотехнологичных терактов. Перехват контроля над системой управления транспортом в крупном городе может привести к многочисленным жертвам. Это, несомненно, вызовет панику и создаст информационно-психологический климат, облегчающий дальнейшие враждебные действия.

• *Перепрофилирование коммерческих систем искусственного интеллекта.* Коммерческие системы могут быть использованы во вред (даже не всегда намеренно). Возможно использование беспилотных летательных аппаратов или автономных транспортных средств для доставки взрывчатых веществ и организации аварий. Серия серьезных катастроф, особенно с участием известных лиц, может иметь международный резонанс и нанести ущерб МИПБ.

• *Удаленные во времени и пространстве атаки*. Объекты физических атак будут находиться все дальше от атакующего в результате автономной работы с использованием ИИ (Brundage et al., 2018, p. 28). Эффект неожиданности от таких атак может оказать дестабилизирующее воздействие на систему международных отношений. Например, возможна дистанционная синхронизация срабатывания ядерных устройств в разных странах мира без непосредственного участия человека. О необходимости сохранения контроля над боевым использованием AI заявляют официальные лица во всех странах, владеющих современными технологиями. Им можно верить. Никакое правительство, реакционное или прогрессивное, не желает лишиться контроля над своим оружием. Подобное нельзя утверждать в отношении всех негосударственных акторов: например, группа техноконфессиональных маньяков, озабоченных ликвидацией человечества, будет иметь определенные и растущие шансы преуспеть в силу совершенствования ИИ, создания его сложных трансграничных систем, большей доступности новейших технологий и ряда других факторов.

• *Создание "deepfakes"*. "Deepfake" (от deep learning – «глубокое обучение» и fake – «подделка») – метод синтеза человеческого изображения и/или голоса на основе использования ИИ. Жертвами создания порно- "deepfakes" уже стали актрисы Скарлетт Йоханссон, Мэйси Уильямс, Тейлор Свифт, Мила Кунис и многие другие знаменитости. Любители "deepfakes" начали использовать технологию для создания цифровых видео мировых лидеров, в том числе президентов Владимира Путина и Дональда Трампа, бывшего президента США Барака Обамы и кандидата в президенты Хиллари Клинтон. При соответствующей подготовке "deepfakes" в рамках ИПП могут спровоцировать финансовую панику, торговую или «горячую» войну. Видео, где премьер-министр Беньямин Нетаньяху или другие правительственные чиновники Израиля говорят, например, о предстоящих планах захвата иерусалимской Храмовой горы и мечети Аль-Акса, могут распространиться, как лесной пожар, на Ближнем Востоке (The Times

of Israel, 2018). Потенциально опасно распространение технологии "deepfake" и тем, что люди не захотят доверять никаким видео- или аудио документам (Waddel, 2018).

- *Технология "Fake People"*. После продажи первого произведения искусства, созданного ИИ, в начале 2018 г. алгоритмы глубокого обучения теперь работают с портретами несуществующих людей. Компания NVIDIA недавно поделилась результатами работы генеративной конкурентной сети (generative adversarial network – GAN), обученной самостоятельно генерировать изображения людей (Karras, Laine and Aila, 2018). Сегодня нейросеть каждую секунду генерирует лица несуществующих людей в большом разрешении. И нет проблемы приказать ей создать, например, несуществующего внебрачного ребенка известной личности, чтобы устроить провокацию. Фамильное сходство на картинке будет стопроцентно убедительным.

- *Установка и закрепление повестки дня.* Исследования осуществленные в США показали, что боты составили более 50% всего интернет-трафика в 2016 г. Организации, которые искусственно продвигают контент, могут манипулировать повесткой дня: чем чаще люди видят определенный контент, тем более важным они его считают (Horowitz et al., 2018, p. 5 – 6). Ущерб репутации с помощью ботов во время политических кампаний, например, может быть использован террористическими группами для привлечения новых сторонников или организации убийств политиков.

- *Анализ тональности* – класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и, тем самым, мнений авторов об объектах, о которых идёт речь в тексте. Анализ тональности обеспечивается широким спектром источников, таких как блоги, статьи, форумы, опросы и т. д. Это может быть очень эффективным инструментом в ИПП.

- Искусственный интеллект, машинное обучение и анализ тональности текста позволяют *предсказывать будущее путем анализа прошлого* — чем не Святой Грааль для финансового сектора или органов государственного планирования? Но потенциально такая возможность выгодна и для ЗИИИ различными государственными и негосударственными акторами. Особенно велико значение *прогностического оружия*: методов предсказательной аналитики на основе больших данных и с использованием ИИ, которые позволяют, получая данные о будущих событиях, корректировать будущее из настоящего в интересах субъекта воздействия и вопреки объективным интересам объекта такого воздействия. К примеру, программа EMBERS (Early Model Based Event Recognition Using Surrogates – «Распознавание событий на основе ранних моделей с применением суррогатов») была запущена IARPA в 2012 г. Программа прогнозирует значимые события, такие, как социальные беспорядки, вспышки заболеваний, результаты выборов. EMBERS представляет детальные прогнозы, включая дату, место, тип события, характеристику протестного населения, определяя при этом возможную погрешность. Программа оперирует как открытыми источниками информации, такими, как *Twitter*, так и более сложными и качественными информационными продуктами, как экономические индикаторы, обрабатывая около 5 млн. сообщений в день. Только по возможностям гражданского протеста EMBERS дает свыше 50 прогнозов на 30 дней вперед (see: Doyle et al., 2014).

- *Рост угроз от фишинга*, поскольку AI позволяет резко увеличить скорость обработки данных и быстрее реагировать на ожидания людей. Прогресс в автоматизированном целевом фишинге продемонстрировал, что автоматически сгенерированный текст может быть эффективным средством обмана людей, и очень простые подходы могут быть убедительными для людей, особенно когда текст относится к определенным темам, таким как развлечения (Brundage, et al., 2018, p. 46). Основными способами использования искусственного интеллекта хакерами являются фишинг, целевой фишинг и «фишинг по-крупному», т.е. ориентированный

на вышестоящих руководителей, ответственных за принятие решений финансового характера (whaling).

- *Компьютерные игры с использованием AI* также могут повысить эффективность информационно-психологического воздействия, особенно на детские и подростковые аудитории. Что компьютерные игры могут иметь определенную манипулятивную начинку, известно давно, однако анализ использования ИИ в этих целях – одна из перспективных задач исследователей. С точки зрения МИПБ особое внимание должно быть уделено компьютерным играм, получающим массовое распространение во многих странах мира.

- Можно представить, что на основе комбинации техник психологического воздействия, сложных систем ИИ и больших данных в ближайшие годы появятся *синтетические информационные продукты*, которые по своему характеру будут похожи на модульный вредоносный софт. Однако действовать они будут не на неодушевлённые предметы, социальные сети и т. п., а на человека и массы как на психобиофизические существа. В подобном синтетическом информационном продукте будут содержаться программные модули, которые введут массы людей в депрессию. После введения в депрессию наступит период скрытого действия суггестивных программ. Они, апеллируя к привычкам, стереотипам и даже психофизиологии, побудят людей выполнять строго определенные действия (Larina and Ovchinskiy, 2018, p. 126 – 127).

Вместе с тем любая из представленных выше угроз, может быть также эффективнее нейтрализована при помощи AI. Так, например, Swisscom Innovations разработала и обучила систему обнаружения фишинга на основе искусственного интеллекта. Он надежно предсказывает, содержит ли ранее неизвестный веб-сайт фишинг или нет (Bürgi, 2016). Другая антифищинговая программа *Lookout Phishing AI* постоянно сканирует интернет в поисках вредоносных веб-сайтов. *Lookout Phishing AI* обнаруживает ранние сигналы фишинга, защищает конечных пользователей от посещения таких сайтов по мере их выявления и предупреждает целевые организации о возможных угрозах (Richards, J., 2019).

Е. Н. Пашенцев предложил некоторые конкретные меры ответа на ЗИИИ в сфере информационно-психологической безопасности России.

В заключение докладчик подчеркнул, что задача сегодня заключается в отражения угроз со стороны реально существующего и постоянно развивающегося «слабого» искусственного интеллекта, который является угрозой не сам по себе, а в силу действий асоциальных внешних и внутренних акторов, превращающих его в угрозу национальной безопасности России. В не столь отдаленном будущем могут встать проблемы, связанные с «сильным интеллектом», о возможности создания которого в ближайшие десятилетия говорит все больше исследователей.

Источники:

Bazarkina, D., Pashentsev, E., 2019. Artificial Intelligence and New Threats to International Psychological Security. Russia in Global Affairs, Issue 1, pp.147-170.

Brundage, et al., 2018. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Oxford, AZ: Future of Humanity Institute, University of Oxford.

Bürgi, U., 2016. Using Artificial Intelligence to Fight Phishing. Swisscom [online]. Available at: <https://ict.swisscom.ch/2016/11/using-artificial-intelligence-to-fight-phishing/> [Accessed 22 June 2019].

Doyle, A., et al., 2014. Forecasting significant societal events using the EMBERS streaming predicative analytics system. Big Data, Vol. 4, pp. 185–195.

Horowitz, M. C., et al., 2018. Artificial intelligence and international security. Washington: Center for a New American Security (CNAS).

Karras, T., Laine, S., and Aila, T., 2018. A style-based generator architecture for generative adversarial networks. arXiv of Cornell University [online]. Available at: <https://arxiv.org/pdf/1812.04948.pdf> [Accessed 31 January 2019].

Larina, E., and Ovchinskiy, V., 2018. Iskusstvenny? intellekt. Bol'shie dannye. Prestupnost' [Artificial intelligence. Big Data. Crime]. Moscow: Knizhnyj mir.

Richards, J., 2019. What is Lookout Phishing AI? Lookout Blog. <https://blog.lookout.com/lookout-phishing-ai> [Accessed 22 June 2019].

The Times of Israel, 2018. 'I Never Said That!' The High-Tech Deception of 'Deepfake' Videos. The Times of Israel [online]. Available at: <https://www.timesofisrael.com/i-never-said-that-the-high-tech-deception-of-deepfake-videos/> [Accessed 31 January 2019].

Waddel, K., 2018. The impending war over deepfakes. Axios [online]. Available at: <https://www.axios.com/the-impending-war-over-deepfakes-b3427757-2ed7-4fbc-9edb-45e461eb87ba.html> [Accessed 31 January 2019].