МИР 09.11.2023 | Мелисса Парке

Как предотвратить ядерный Армагеддон искусственного интеллекта

Гонка за внедрение ИИ в системы ядерных вооружений началась. Но ни одна из мировых инициатив не реагирует адекватно на возникающие риски.

Это уже не научная фантастика: гонка за внедрение искусственного интеллекта (ИИ) в системы ядерных вооружений началась. Такое развитие событий может повысить вероятность ядерной войны. Правительства стран мира сейчас предпринимают действия, призванные гарантировать безопасность разработок и применения ИИ, поэтому у них есть возможность смягчить эту угрозу. Но если мировые лидеры действительно хотят воспользоваться этой возможностью, они сначала обязаны понять, насколько серьезна угроза.

Несколько недель назад «Большая семерка» в рамках Хиросимского процесса согласовала «Международный кодекс поведения для организаций, разрабатывающих передовые ИИ-системы» для «продвижения безопасного, надежного и вызывающего доверие ИИ во всем мире». Тем временем президент США Джо Байден подписал указ, устанавливающий новые стандарты «безопасности и надежности» ИИ. А Великобритания провела первый глобальный Саммит по безопасности ИИ, призванный гарантировать, чтобы эта технология разрабатывалась «безопасным и ответственным» образом.

Однако ни одна из этих инициатив не реагирует адекватно на риски, возникающие из-за применения ИИ в сфере ядерных вооружений. В утвержденном «Большой семеркой» кодексе поведения и в указе Байдена лишь мимоходом говорится о необходимости защищать население от генерируемых искусственным интеллектом химических, биологических и ядерных угроз. А премьер-министр Великобритании Риши Сунак вообще не упомянул острую угрозу,

исходящую от применения ИИ в сфере ядерных вооружений, хотя он и заявил, что на Саммите по безопасности ИИ было достигнуто общее понимание по поводу рисков, возникающих из-за ИИ.

Никто не ставит под сомнение экзистенциальные риски применения ядерного оружия, которые могут нанести неописуемый урон человечеству и планете. Даже региональная ядерная война убьет сотни тысяч человек напрямую и приведет к множеству страданий и смертей косвенным образом. Одно только изменение климата в результате такой войны грозит миллиардам людей жестоким голодом.

История ядерного оружия полна случаев, когда его едва не применили. И зачастую Армагеддон удавалось предотвратить людям, которые предпочли довериться собственному суждению, а не слепо следовать информации, поставляемой машинами. В 1983 году советский офицер Станислав Петров получил сигнал тревоги от спутниковой системы раннего предупреждения: американские ядерные ракеты направлялись в сторону СССР. Но вместо того, чтобы немедленно предупредить вышестоящее начальство (а это гарантированно спровоцировало бы ядерное «возмездие»), Петров совершенно правильно решил, что это была ложная тревога.

Мог бы Петров принять такое же решение (и была бы у него вообще для этого возможность), если бы в системе был задействован искусственный интеллект? Более того, применение технологий машинного обучения в сфере ядерных вооружений уменьшит контроль человека над решениями об их применении.

Да, безусловно, многие командные, контрольные и коммуникационные задачи были автоматизированы уже после изобретения ядерного оружия. Но по мере прогресса в технологиях машинного обучения процедура принятия решений продвинутыми машинами становится все более непрозрачной. Это называется «проблемой черного ящика» в ИИ. Она затрудняет людям мониторинг работы машин, не говоря уже о об ответе на вопрос, все ли в порядке с этой машиной, и не является ли она запрограммированной таким образом, что это может привести к незаконным или непреднамеренным последствиям.

Простой гарантии, что финальное решение о запуске принимает человек, будет

Простой гарантии, что финальное решение о запуске принимает человек, будет недостаточно, чтобы смягчить эти риски. В исследовании 2017 года психолог Джон Хоули делает вывод, что «люди в целом

недостаточно, чтобы смягчить риски

очень плохо выполняют требования, которые предъявляют надзорные органы к мониторингу автономных систем и вмешательству в их работу».

Кроме того, как показала в 2020 году программа Принстонского университета «Наука и глобальная безопасность», уже сегодня процессы принятия решений лидерами во время ядерного кризиса являются очень поспешными. Даже если ИИ будет использоваться только в сенсорах и для наведения целей, а не для принятия решений о запуске, уже и так сжатый график принятий решений о необходимости удара сократится еще сильнее. Возросшее давление на лидеров повысит риск иррациональных решений или ошибок в расчетах.

Еще один риск возникает при использовании ИИ в спутниках и других системах разведки и обнаружения. Станет еще труднее скрывать ядерное оружие, например, подводные лодки с баллистическими ракетами, которые исторически умели хорошо прятаться. Это может вынудить страны с ядерным оружием применить весь свой арсенал ядерного оружия в начале конфликта, то есть до того, как противник получат шанс нейтрализовать известные ему ядерные системы.

Пока что ни одна инициатива (ни указ Байдена, ни кодекс поведения «Большой семерки») не заходит дальше добровольных обязательств сохранить за людьми контроль над принятием решений о ядерном оружии. Как отметил генеральный секретарь ООН Антониу Гутерреш, критически необходим юридически обязывающий договор о запрещении «летальных автономных оружейных систем».

Крайне важно, чтобы власти и общество начали бороться не только за отказ от использования ИИ в ядерном оружии, но и за ликвидацию такого оружия вообще

Но хотя такой договор является необходимым первым шагом, нужно сделать намного больше. Когда речь заходит о ядерном оружии, попыток спрогнозировать, смягчить или отрегулировать новые риски, исходящие от новых технологий, всегда будет недостаточно. Мы должны полностью исключить эти вооружения из нашего уравнения. Это означает, что все правительства должны обязаться стигматизировать, запретить и ликвидировать ядерное оружие, присоединившись к «Договору о

запрещении ядерного оружия», который открывает понятный путь к миру без такого оружия. Это также означает, что государства с ядерным оружием обязаны немедленно прекратить инвестировать средства в модернизацию и расширение своих ядерных арсеналов, в том под предлогом повышения их «безопасности» или защиты от кибератак. Учитывая непреодолимые риски, возникающие просто из-за самого факта существования ядерных вооружений, подобные усилия фундаментально являются тщетными.

Мы уже знаем, что автономные системы способны снизить порог начала вооруженного конфликта. В случае внедрения ИИ в сфере ядерных вооружений появится новый пласт рисков, хотя уровень опасности уже является неприемлемым. Крайне важно, чтобы власти и общество поняли это, и начали бороться не только за отказ от использования ИИ в ядерном оружии, но и за ликвидацию такого оружия вообще.

(c) Project Syndicate 2023



Мелисса Парке

Мелисса Парке (Melissa Parke) – бывший министр международного развития Австралии, является исполнительным директором «Международной кампании за упразднение ядерного оружия» (International Campaign to Abolish Nuclear Weapons, ICAN), удостоенной Нобелевской премии мира в 2017 году.