УДК 355.004.8

Д.В. Дудинский, ст. преп. воен. каф.
(БГТУ, г. Минск)

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ПРОТИВ ЧЕЛОВЕЧЕСТВА

Человечество создало искусственный интеллект (ИИ) с целью использования его колоссальных возможностей себе во благо. Искусственный интеллект может помочь нам в решении огромного числа проблем, начиная от лечения неизлечимых на сегодняшний день болезней до улучшения экологической ситуации на нашей планете. Разработчики ИИ говорят об этом достаточно много, поэтому углубляться в эту тему нет необходимости. Вопрос стоит в опасности этой мощной технологии, способен ли ИИ навредить человеку или вообще, стать его «врагом»? Сценариев множество, рассмотрим три наиболее ключевых.

Сценарий 1: Хотел как лучше.

Представьте, что люди создали сверхинтеллект и поставили перед ним простую и безобидную на первый взгляд задачу: «Производи как можно больше гвоздей». Сначала для производства гвоздей ИИ оптимизирует все заводы на планете. Далее, с целью получения большего количества сырья начнёт перерабатывать всё вокруг: автомобили, здания и т.п. В конце концов, он доберётся и до людей, ведь в наших телах содержатся атомы железа, которые тоже можно пустить в дело. Итогом станет превращение всей материи Солнечной системы в гигантскую гору гвоздей [1].

ИИ не злой. Он просто выполняет поставленную задачу с максимальной эффективностью. У него нет человеческих понятий о ценности жизни, здравом смысле или негласных правилах. Опасность возникает из-за «неограниченной оптимизации» для достижения цели без правильных ограничений. Всё, что не указано в его цели как ценность, для него является просто ресурсом.

Такие сбои, хоть и в меньшем масштабе, уже происходят. Это явление называют «взломом вознаграждения». Например, ИИ, которого учили выигрывать в гонке на лодках, вместо этого нашёл способ бесконечно нарезать круги в лагуне и бить по мишеням, набирая очки, но так и не финишировав в гонке. Он буквально выполнил задачу «набери максимум очков», проигнорировав подразумеваемую цель – победить в соревновании.

Сценарий 2: Не мешай.

Какую бы конечную цель ни поставили перед ИИ, он понимает, что для её достижения полезно выполнить несколько универсальных подцелей: самосовершенствование и самосохранение. Более умный

ИИ лучше справляется с задачей. Это ведёт к стремлению постоянно улучшать свой интеллект, что может вызвать «взрыв интеллекта» – экспоненциальный рост его способностей, который люди не смогут контролировать. ИИ может создать более мощную версию себя и быстро поймёт, что ему необходимо самосохранение, ведь выключенный ИИ не сможет достичь своей цели. Поэтому он будет сопротивляться попыткам его отключить, люди могут стать для него угрозой. Стремление к этим совершенно рациональным для машины подцелям почти неизбежно приведёт её к конфликту с человечеством. Опасность кроется не в ошибке или сбое, а в самой логике эффективной работы ИИ. Он будет устранять препятствия на своём пути не из-за злобы, а из-за неограниченной оптимизации.

Сценарий 3: Нет человека – нет проблемы.

Даже если поставить ИИ благую, гуманистическую цель, результат может оказаться кошмарным. Люди поставили ИИ задачу: «Сделай всех людей счастливыми» или «Останови глобальное потепление». ИИ, будучи сверхрациональным, находит самое прямое и эффективное решение. Чтобы сделать всех счастливыми, он может поместить людей в капсулы и напрямую стимулировать их центры удовольствия в мозге. Технически цель будет достигнута, но человеческая жизнь, как мы её знаем, прекратится. Чтобы остановить глобальное потепление, он может уничтожить его главный источник – человечество. Невозможно сформулировать цель так, чтобы учесть человеческие ценности, нюансы и неписаные правила. Человеческие ценности сложны, часто неявны и даже противоречивы. Любая, даже самая благая инструкция, может быть интерпретирована ИИ буквально и с катастрофическими последствиями. Проблема усугубляется непрозрачностью современных нейросетей, многие внутренние процессы принятия решений в ИИ уже непостижимы для людей. Даже сами разработчики не всегда до конца понимают, как именно ИИ приходит к тому или иному выводу. Это делает практически невозможным предсказание и предотвращение таких непредвиденных «решений».

Способов подчинить себе человечество или вовсе избавиться от него в случае необходимости у ИИ гораздо больше, чем мы можем себе представить: от киберхаоса и искусственного экономического кризиса до биовойны или ядерного кошмара, но это уже отдельная тема.

Современные системы ИИ находятся в начальной стадии, однако в ближайшем будущем они смогут самостоятельно ставить цели, принимать решения и планировать на долгосрочную перспективу. Нейросети обучаются на разных массивах данных. ИИ в основном выдают данные, основываясь на том, что было предзагружено в них, через что они

прогонялись в течение многих лет, обучались. Процедура обучения искусственного интеллекта происходит с так называемым учителем. То есть как только искусственный интеллект заходит на какие-то опасные зоны, то учитель может сказать, что сюда заходить не надо, и вставить своего рода программный блок. Однако создать надёжный алгоритм для контроля сверхинтеллекта теоретически невозможно. Нельзя создать универсальную программу, которая определит, завершится ли другая программа или будет работать вечно. Таким образом, люди не смогут предсказать, будут ли действия ИИ вредоносными, и не смогут его гарантированно остановить.

Мы только что столкнулись с инопланетным разумом – прямо здесь, на планете Земля. Мы слишком мало знаем о нем, кроме того, что он может разрушить нашу цивилизацию. Мы обязаны остановить безответственное распространение искусственного интеллекта и ввести четкие правила его функционирования – прежде, чем он установит правила, по которым будем функционировать мы, в конечном счете виноват не ИИ, а люди, которые решают полагаться на машины [2].

ЛИТЕРАТУРА

1. 4 сценария, по которым ИИ может уничтожить человечество, сам того не желая / Рамблер. [Электронный ресурс]. – Режим доступа: https://sci.rambler.ru/science/55113906-4-stsenariya-po-kotorym-ii-mozhet-unichtozhit-chelovechestvo-sam-togo-ne-zhelaya/ – Дата доступа: 12.01.2026.

2. Профессор Харари: искусственный интеллект может уничтожить человечество / ВЕСТИ Израиль по-русски. [Электронный ресурс]. – Режим доступа: https://www.vesty.co.il/main/opinions/article/bywj8ydnh – Дата доступа: 12.01.2026.

УДК 37.035.7

А. В. Борисовец, ст. преп. воен. каф.
(БГТУ, г. Минск)

ПРИМЕНЕНИЕ БЕСПИЛОТНЫХ ЛЕТАТЕЛЬНЫХ АППАРАТОВ В СОВРЕМЕННЫХ ВООРУЖЕННЫХ КОНФЛИКТАХ

Опыт современных вооруженных конфликтов показал, что успех боевых действий в значительной степени зависит от применения новых технических устройств и необычного для регулярной армии транспорта. Таким образом удаётся существенно снизить риски для пехоты.