Райков А.Н.

Ловушки безопасности на пути развития сильного искусственного интеллекта

– **Аннотация:** Внедрение сквозных цифровых технологий, особенно искусственного интеллекта (ИИ), направлено на дальнейшее повышение эффективности человеческого труда. Вместе с тем это внедрение уже сейчас заставляет более внимательно относиться к вопросам обеспечения безопасности. Прежде ИИ являлся безобидным помощником человека в рутинных делах, сейчас же он становится опасным конкурентом для работника. Исследования дальнейшего развития ИИ до уровня Сильного ИИ обнаруживают опасные ловушки. Это требует соответствующих упреждающих решений, как в области методологии, так и этики.

– **Ключевые слова:** безопасность, сильный искусственный интеллект, ловушки, этика

Качество функционирования информационных систем обеспечения безопасности (национальной, экономической, информационной и др.) в значительной степени определяется их способностью упреждать и отрабатывать непредвиденные события. Возможности таких систем заметно возрастают при использовании средств искусственного интеллекта (ИИ).

Вместе с тем, феномен цифровой экономики привносит все новые особенности непредвиденных ситуаций, которые ранее не принимались во внимание. Это, прежде всего, их некаузальное (беспричинное) проявление, которое не может быть объяснено с помощью логики любого уровня сложности. Описания событий далеко не всегда имеют под собой вероятностную основу. Поведение наблюдаемого объекта может зависеть от состояния иного объекта, место расположения и функционал которого неизвестны. Само наблюдение за ситуацией, ее измерение, приводит, как в хорошо известном квантовом случае, к ее искажению. Задачи становятся обратными, причем такие задачи приходится решать на неметрическом пространстве, методы и инструментарии для чего пока отсутствуют.

Такие особенности в создании систем поддержки решений и ИИ меняют парадигму развития последнего, заставляют делать его много более сильным, а это, в свою очередь, порождает новые угрозы и риски, природа которых пока не изучена.

ИИ сейчас все больше проникает в социально-гуманитарную и производственную сферу, помогает решать вопросы государственного и муниципального управления. Пока термин ИИ больше связывают с компьютерной обработкой данных, логикой и исчислениями, онтологиями и нейросетями, алгоритмизацией и программированием, аддитивными и ассоциативными схемами, знаками и их семантической интерпретацией, образами и визуальной аналитикой. Вместе с тем, сейчас даже его традиционная версия уже стала опасным конкурентом для любого работника.

ИИ все больше уходит в пространство смыслов, в способы поддержки коллективной творческой деятельности, глубже проникает в тайны субъективного и коллективного бессознательного. А с этим начинают проступать черты следующего поколения ИИ–Искусственного общего (сильного) интеллекта (Artificial General Intelligence, AGI), «который намного умнее лучшего человеческого разума практически во всех областях, включая научное творчество, мудрость и социальные навыки» [1].

С появлением AGI не исключена его опасность для общества. На пути его создания могут встретиться ловушки, попадание в которые способно нанести непоправимый ущерб социуму.

Первая ловушка – это ловушка цифровизации, замены непрерывного природного сигнала цифрой (битами и байтами). И с какой бы точностью компьютер не восстанавливал непрерывный сигнал по его точкам, ошибка моделей накапливается. Это хорошо известно, например, из теоремы Котельникова. Цифровой сигнал имеет ограниченный спектр частот. Такой сигнал неспособен эмулировать всю глубину человеческих эмоций и чувств. Дискретность обработки данных может привести к падению уровня культуры и духовности социума.

Вторая ловушка – ловушка рациональности. Человек, анализируя ситуацию, пытается найти рациональное зерно. Анализ – это деление целого на части. Обратная операция – синтез, который много более сложное действие и носит творческий

характер. При чисто рациональном подходе к синтезу возможности получения хороших решений с помощью ИИ ограничены. Как следствие, увеличивается риск роста ошибок решения важных проблем, особенно стратегических, где решения носят телеологический характер, то есть строятся из будущего.

Третья ловушка – ловушка каузальности, причинности. Традиционный ИИ использует в выводах решений логику, статистику, большие данные, нейронные сети, которые несут груз прошлого опыта. Прогнозы часто делаются на основе накопившегося за определенный период времени опыта. Но жизнь ставит неожиданные проблемы и в новых обстоятельствах, она ведет себя нелогично, чего современный ИИ реализовать неспособен.

Четвертая ловушка – феноменологическая. Человеческий разум усиливается эмоциями, чувствами, медитативными слоями сознания. Эти феномены являются источниками инсайта – мгновенного постижения целого, озарения, прозрения разума. Они характеризуются полной неформализуемостью, нелогичностью, интенсивностью, длительностью, предметностью, тональностью и пр. Традиционный ИИ пока не может эти слои сознания охватить.

Пятая ловушка – тоталитарность. Благое стремление увеличить число диагностических датчиков и собирать большие объемы информации о поведении организма человека, предприятия, поезда, автомобиля, муниципалитета и пр. может иметь обратную сторону. В результате такого тренда создается система, которая может начать диктовать человеку нестандартные для него решения, польза от которых может быть сомнительна.

Шестая ловушка кроется в области генетики. Безобидный, казалось бы, компьютерный генетический алгоритм может оказаться сокрушительным для будущих поколений, если его некорректно использовать в сфере генетической инженерии. Известно, что определенные способности человека передаются по наследству сложными и хаотичными комбинациями большого множества генов. Поэтому имплантация редуцированных генетических схем может составлять для человека угрозу.

Перечень ловушек можно продолжить. Так, попадание в недобросовестные руки элементов AGI («фейковых людей», искусственного эмоционального интеллекта, анализа настроений и

выборных предпочтений и т.д.) может оказать дестабилизирующее воздействие на социальное и политическое развитие любой страны.

Для упреждения нежелательных последствий уже сейчас, например, в области политики следует рассмотреть реализацию следующих мероприятий:

Создание реестра угроз AGI;

Мониторинг практики применения AGI с точки зрения психологического воздействия на любую систему;

Оценка рисков AGI, создаваемых различными субъектами с антисоциальными целями и др.

Становится все более очевидным, что для будущего развития AGI необходима синергия междисциплинарных исследований с охватом, как минимум, порядка 25 дисциплин, более критическое отношение к дискретному представлению данных, погружение информационных моделей в концептуальные пространства (топологии [2], категории, топосы и пр.), снятие противоречий между различными областями физики, обращение к потенциалу космологических и биологических исследование и др.

Возможные опасности, связанные с грядущим появлением AGI, заставляют задуматься об их упреждении. Для этого ученые, государственные и общественные деятели, профессионалы и эксперты вырабатывают этические принципы, которых стоит придерживаться, чтобы развитие AGI продвигалось в безопасном и высоконравственном направлении. Для этого, в частности, уже сформированы известные принципы безопасного развития ИИ [3]. Они, в основном, охватывают сферу традиционного ИИ и его ближайшей перспективы.

Для поддержки указанных принципов, а также учитывая общественную и государственную значимость вопроса возможных рисков развития AGI, которые могут возрасти непредсказуемым и скачкообразным образом, целесообразно предложить органам государственной власти, научному и экспертно-аналитическому сообществу следующий минимальный набор принципов в сфере развития AGI:

1. Важнейшей целью развития AGI должно быть не только повышение эффективности труда, но и освоение глубинных слоев человеческого сознания (эмоции, чувства, медитативные слои

сознания), совершенствование гражданского участия [4], усиление учета социально-гуманитарного фактора.

2. AGI должен быть абсолютно, на 100%, безопасными для людей, включая экологическую и нравственную чистоту, независимо от места его применения: государственное управление, робототехника, реклама, производство, интеллектуальные помощники и пр.

3. Создание комфортной и прозрачной сетевой среды для постоянного роста эффективности виртуального (сетевого) сотрудничества в различных областях деятельности на базе AGI органов власти, ученых, преподавателей, инженеров, студентов, школьников и других слоев общества.

4. При инициации проектов, заключении договоров на создание систем AGI должна определяться персональная ответственность руководителей, ученых, разработчиков и инженеров за возможное причинение вреда и порядок возмещения возможного ущерба производству, человеческому общежитию.

5. Функционирование автономных коллективных и единичных систем AGI не должно противоречить этическим и общечеловеческим нормам и ценностям, сложившимся канонам свободы совести и вероисповедания.

6. Развитие AGI не должно угрожать трудоустройству любого человека. Любое внедрение AGI должно сопровождаться ростом удовлетворенности людей, увеличением числа модифицированных рабочих мест.

7. Человек всегда должен иметь право ответственного выбора: принять решение самостоятельно или поручить это системе AGI, и любая такая система должна проектироваться с учетом того, чтобы человек имел возможность вмешиваться в процесс решения, реализуемый этой системой.

8. Абсолютно все риски развития AGI должны быть контролируемыми и упреждаться соответствующими организационными, научными и инженерными приемами.

9. В развитии систем на основе методов традиционного ИИ особое внимание должно уделяться построению именно AGI (полного, сильного, общего, коллективного, когнитивного и др.). Именно он сможет проявлять сверхчеловеческие возможности

чувственного, эмоционального и трансцендентального слоев сознания.

10. Системы AGI, в том числе коллективные, способные к автономному самоуправляемому поведению, саморазвитию и самовоспроизведению, должны находиться под особо строгим контролем человека.

Таким образом, в условиях, когда торговые войны, фейковые новости, манипулирование политической блогосферой и выборными компаниями, распад региональных союзов, экономические санкции, деградация парадигмы глобализации— становятся реальностью, переломить негативные тенденции, угрожающие возможными масштабными силовыми конфликтами способна цифровая трансформация и контроль за корректным развитием сильного искусственного интеллекта, включая продвижения его этических канонов.

Работа выполнена при поддержке Российского научного фонда, проект № 17-18-01326

Литература:

1. *Bostrom N.* Superintelligence: Paths, Dangers, Strategies. https://en.wikipedia.org/wiki/Superintelligence:_Paths,_Dangers,_Strategies (Дата обращения: 04.11.2019).
2. *Иванов В.К.* Некорректные задач в топологических пространствах // Сибирский математический журнал. — 1969. — № 5. — С. 1065–1074.
3. Asilomar AI principles, https://futureoflife.org/ai-principles/ (Дата обращения: 04.11.2019)
4. *Raikov A.N.* Accelerating technology for self-organising networked democracy. Futures. – V. 103. – 2018, – P. 17-26. https://doi.org/10.1016/j.futures.2018.03.015 (Дата обращения: 04.11.2019).