Возможные угрозы, идущие от искусственного интеллекта

Олег Фиговский и Олег Пенский.

В настоящее время все развитые государства мира устремились в безудержную и беспрецедентную в истории гонку за искусственным интеллектом, его внедрением, а порой, и жестким проталкиванием в жизнь общества. Однако до сих пор, насколько известно авторам настоящей книги, нет единого определения искусственного интеллекта, а поэтому до конца не понятно: – Что же именно внедряют в социум политики и бизнесмены?

«В 2018 году в Праге на площадке Чешского Технического Университета одновременно прошли конференции, посвященные искусственному интеллекту человеческого уровня, общему искусственному интеллекту, биологически-вдохновленным когнитивным архитектурам, а также нейро-символьным технологиям. На конференциях были представлены доклады ведущих специалистов компаний и учреждений, лидирующих в сфере исследовательских разработок в области искусственного интеллекта (ИИ): Microsoft, Facebook, DARPA, MIT, Good AI. В этих докладах были обозначены как текущее состояние разработок в области ИИ, так и стоящие перед обществом нерешенные проблемы, а также угрозы, возникающие в ходе дальнейшего развития этой технологии.

Однако, прежде всего, необходимо уточнить значение некоторых терминов, которые обычно употребляются совместно с ИИ в различных контекстах: *слабый или специализированный ИИ, автономный ИИ* (Autonomous AI), *адаптивный* ИИ (Adaptive AI), *общий ИИ* (Artificial General Intelligence, AGI), *сильный ИИ* (Strong AI), *ИИ человеческого уровня* (Human-Level AI), *ИИ сверхчеловеческого уровня* (Super-human AI).

Слабый или специализированный ИИ представлен всеми без исключения существующими решениями и предполагает способность автоматизации решения одной конкретной задачи, будь то игра в Go или распознавание лиц на видеокамерах. При этом отсутствует возможность самостоятельного обучения другим задачам без перепрограммирования человеком.

Автономный ИИ предполагает возможность системы функционировать долгое время без участия оператора. Например, он позволяет дрону, оборудованному солнечными батареями, совершить многодневное путешествие с Елисейских полей на Красную площадь или в обратном направлении, самостоятельно выбирая как маршрут, так и места для промежуточных посадок для подзарядки аккумуляторов, избегая при этом всевозможные препятствия.

Адаптивный ИИ предполагает способность системы адаптироваться к новым условиям, приобретая знания, не закладываемые при создании. Например, позволить системе поддержания диалогов на русском языке самостоятельно осваивать новые языки и применять их знание в разговоре, попадая в новую языковую среду или на основе изучения учебных материалов для этих языков.

Общий ИИ предполагает настолько высокую адаптивность, что обладающая им система может быть использована в самых различных видах деятельности при соответствующем обучении. Обучение может быть как самостоятельным, так и направленным (с помощью инструктора). В этом же смысле в противопоставление слабому или специализированному ИИ также часто употребляется сильный ИИ.

ИИ человеческого уровня предполагает уровень адаптивности сравнимый с человеческим, то есть система способна осваивать те же самые навыки, что и человек в сопоставимые сроки обучения.

ИИ сверхчеловеческого уровня предполагает еще более высокую адаптивность и скорость обучения. Таким образом, система может обучиться тем знаниям и способностям, которые человеку в принципе не под силу.

Несмотря на множество современных достижений в нейробиологии, на сегодняшний день пока никто точно не знает, как устроен естественный интеллект. Соответственно, точно так же никто не знает, как именно создать искусственный интеллект. Существует ряд известных проблем, требующих решения для его создания и разные мнения по поводу приоритетности достижения тех или иных решений. Например, руководитель международных проектов по созданию искусственного интеллекта с открытым кодом OpenCog и SingularityNET Бен Герцель считает, что все необходимые технологии для создания общего ИИ в принципе уже разработаны, необходимо только соединить их некоторым правильным образом для получения такой синергии, результатом которой станет возникновение общего ИИ. Другие эксперты настроены более скептично, полагая, что необходимо принципиальное решение многих проблем, которые будут перечислены ниже. Также сильно варьируются экспертные оценки срока возникновения сильного ИИ — от десятка до нескольких десятков лет.

Вместе с тем возникновение сильного искусственного интеллекта вполне закономерно в рамках общего эволюционного процесса, как закономерно и возникновение молекул из атомов, клеток из молекул, организмов из клеток, выделение специализированных клеток в центральную нервную систему, возникновение социальных структур, развитие речи, письменности и в конечном итоге — информационных технологий. Закономерность нарастающей сложности информационных структур и способов организации в процессе

эволюции хорошо показана Валентином Турчиным. Если не произойдет гибель человеческой цивилизации, то такая эволюция будет неизбежной, и в самой долгосрочной перспективе это будет спасением человечества, постольку только не биологические формы существования информации смогут пережить неизбежную со временем гибель Солнечной системы и смогут сохранить во Вселенной информационный код нашей цивилизации.

При этом важно осознавать, что для того чтобы построить сильный искусственный интеллект, не обязательно понимать, как устроен естественный, так как необязательно понимать, как летает птица, чтобы сделать ракету. Очевидно, это будет сделано рано или поздно тем или иным способом или, возможно, несколькими способами».

«В качестве принципиальных проблем, решение которых еще предстоит для создания общего или сильного ИИ, большинство экспертов выделяют следующие:

- **Быстрое обучение** (few-shot learning) — необходимость построения систем, обучающихся на небольшом объеме материала в отличие от существующих систем глубокого обучения, требующих большие объемы специально подготовленного обучающего материала.
- **Сильная генерализация** (strong generalisation) — создание технологий распознавания ситуаций, в которых распознаваемые объекты встречаются в условиях, отличных от тех, в которых они встречались в использованном для обучения материале.
- **Генеративные или генерирующие модели обучения** (generative models) — разработка технологий обучения, когда объектом запоминания являются не признаки объекта распознавания, а принципы его формирования, что может позволить отражать более глубокие сущности распознаваемых объектов и осуществлять более быстрое обучение и более сильную генерализацию.
- **Структурированное обучение и предсказание** (structured prediction and learning) — развитие технологий обучения на основе представления объектов обучения в виде многослойных иерархических структур, где более низкоуровневые элементы определяют более высокоуровневые, что может оказаться альтернативным решением проблем быстрого обучения и сильной генерализации.
- Решение проблемы **катастрофического забывания** (catastrophic forgetting) — присущего большинству существующих систем, которые, будучи изначально обучены на одном классе объектов и затем до-обучены

распознаванию на новом классе объектов, теряют способность распознавать объекты первого класса.

- Достижение **возможности инкрементального обучения** (incremental learning), предполагающего способность системы накапливать знания и совершенствовать свои возможности постепенно, не теряя при этом полученные ранее знания, но приобретая новые знания применительно к системам диалогового общения на естественном языке. Идеальным является прохождение «младенческого теста Тьюринга» (Baby Turing Test), в случае которого система должна продемонстрировать возможность постепенного освоения языка от уровня младенца до уровня взрослого человека.
- Решение проблемы **сознания** (consciousness) — предполагает формирование проверенной рабочей модели сознательного поведения, обеспечивающего эффективное прогнозирование и целенаправленное поведение за счет формирования «внутренней картины мира» в рамках которой возможен поиск оптимальных стратегий поведения по достижению поставленных целей без фактического взаимодействия с реальным миром, что существенно повышает безопасность, проверки гипотез, а также повышает скорость и энергетическую эффективность этой проверки, тем самым создавая возможность самообучения живой или искусственной системы в «виртуальном мире» собственного сознания. С прикладной точки зрения проблема сознания имеет две стороны. С одной — создание систем ИИ, обладающих сознанием, позволит резко повысить их эффективность. С другой — появление сознания у таких систем вызывает как дополнительные риски, так и вопросы этического плана, поскольку такие системы в какой-то момент смогут быть по уровню самосознания приравнены к самому человеку с вытекающими из этого последствиями в правовом поле.

Возникновение систем даже просто автономного или адаптивного, а тем более общего или сильного ИИ связывается с несколькими угрозами разного масштаба, актуальными уже сегодня.

Во-первых, угрозу для человека может представлять интеллект не обязательно сильный, общий, человеческого или сверхчеловеческого уровня, так как достаточно иметь автономную систему, оперирующую большими объемами информации с большими скоростями. На её основе могут быть созданы так называемые «автономные системы смертоносных вооружений» — Lethal Autonomous Weapons Systems (LAWS), простейший пример которых — дроны для заказных убийств, печатаемые на 3D-принтерах как в массовых масштабах, так и небольшими партиями в кустарных условиях.

Во-вторых, угрозу для государства может представлять ситуация, когда другое государство (потенциальный противник) получает вооружения с более

адаптивным, автономным и общим искусственным интеллектом с повышенной скоростью реакции и предсказательной способностью.

В-третьих, угрозу для всего мира представляет вытекающая из предыдущей угрозы ситуация, когда государства вступают в новый виток гонки вооружений, совершенствуя уровни интеллекта автономных средств поражения и уничтожения — как это было предсказано Станиславом Лемом несколько десятков лет назад.

В-четвертых, угрозу для любой стороны представляет любая, не обязательно боевая, но и промышленная или бытовая интеллектуальная система с определенной степенью автономности и адаптивности, способная не только к целенаправленному действию, но и к сознательному целеполаганию, при том что автономная постановка целей системы может привести к постановке целей, противоречащих целям человека и людей, а возможностей достижения этих целей у системы будет намного больше, в силу её более высокого быстродействия, большего объема обрабатываемой информации и большей предсказательной способности. К сожалению, масштабы именно этой угрозы сообществом не вполне изучены и осознаны.

В-пятых, угрозу для общества представляет переход к новому уровню развития производственных отношений в тоталитарном обществе, когда более малочисленная часть населения получает возможность контролировать материальное производство, исключая из него подавляющую часть населения за счет еще большей автоматизации, что может приводить к еще большему социальному расслоению, снижению эффективности «социальных лифтов» и увеличению массы «лишних людей» с соответствующими социальными последствиями.

Наконец, угрозу для человечества в целом может представлять автономизация глобальных вычислительных систем обработки данных, распространения информации и принятия решений на основе глобальных сетей, поскольку скорость распространения информации в таких системах и масштаб воздействия может приводить к непредсказуемым с позиций имеющегося опыта и существующих моделей управления социальным явлениям. Например, внедряемая система социального кредита в современном Китае является уникальным экспериментом цивилизационного масштаба с непонятными на сегодняшний день последствиями.

Сложность контроля за системами искусственного интеллекта на сегодняшний день обусловлена, в частности, «закрытостью» существующих прикладных решений на основе «глубоких нейронных сетей», которые не позволяют не только верифицировать правильность принятия решений перед их исполнением, но даже по факту проводить анализ решения, которое было принято машиной. Решению этой проблемы сейчас посвящено как новое

направление «объяснимый искусственный интеллект» (Explainable Artificial Intelligence,EAI), так и новый интерес к интеграции ассоциативного (нейро-сетевого) и символьного (основанного на логике) подходов к проблеме».

«Представляется безусловно необходимым принятие следующих мер для предотвращения катастрофических сценариев развития технологий ИИ и их применения:

- Международный запрет «автономных систем смертоносных вооружений» (LAWS) и разработка и внедрение мер международного контроля за его исполнением.
- Государственная поддержка работ, направленных на решение обозначенных проблем, в особенности «объяснимого искусственного интеллекта» и интеграции различных подходов, а также изучения принципов работы создания механизмов целеполагания с целью получения эффективных средств программирования и контроля интеллектуальных систем, когда средством программирования будут не правила, а ценности, а контролироваться должны не действия, а цели.
- Демократизация доступа к технологиям и методам ИИ, например, за счет реинвестирования доходов от внедрения интеллектуальных систем в массовое обучение вычислительным и когнитивным технологиям, а также создание решений ИИ с открытым кодом и разработка мер стимулирования открытия кодов существующими «закрытыми» системами ИИ. Например, проект Aigents направлен на создание персональных агентов ИИ для массовых пользователей, работающих автономно и не подверженных централизованным манипуляциям.
- Регламентирование на межгосударственном уровне открытости алгоритмов ИИ, протоколов работы распределенных систем обработки данных и принятия решений на их основе с возможностью независимого аудита как международными и государственными органами, так и частными лицами. Одной из инициатив в этом направлении является создание платформы и экосистемы приложений ИИ с открытым кодом SingularityNET».

Мы не будем анализировать или критиковать многочисленные определения и приведенную выше типизацию ИИ, а попробуем дать свое определение: «Искусственный интеллект – это автоматизированный аналог мышления, способный самостоятельно принимать решения».

Анализируя устремления политиков и ученых многих стран, мы можем сделать вывод о том, что они, порой, сами того не понимая, повсеместно внедряя методы искусственного интеллекта, преследуют именно ту цель, которая прописана в приведенном выше определении: «Заставить автоматы

принимать решения за живых людей», тем самым дать возможность машинам определять многие человеческие судьбы.

«Примером реализации такой возможности является КНР. Пока российские суды осваивают онлайн-заседания, видео-конференц-связь и голосование через блокчейн, китайская судебная система ушла далеко вперед. В судах Китая роботы уже помогают судьям принимать решения, составляют сторонам процессуальные документы и ведут печатные онлайн-трансляции заседаний без помощи человека. Причем такие технологии применяются по всей стране в судах разного уровня. На развитие и внедрение новшеств у китайцев ушло примерно пять лет. Искусственный интеллект помогает судьям в рассмотрении простых дел, например, споров по контрактам с низкой стоимостью. Таким образом, судебные органы страны пытаются достичь единообразия в практике по одинаковым делам». «Искусственный интеллект используется в китайских судах с начала 2019 года, пока это использование ограничивается главным образом представлением доказательств, связанных с делом, и помощью в расследованиях» , т.е. искусственный интеллект уже принимает решения о том, включать или не включать имеющиеся у следствия материалы в качестве доказательств преступлений, совершенных человеком.

Конечно, приведенный пример жесткого использования в социуме искусственного интеллекта является пока исключением из правил, но он очень красочно говорит о реальных целях разработчиков алгоритмов этого интеллекта, которые на сегодня самими разработчиками, может быть, пока не осознаются, но которые подтверждают приведенное нами в начале определение ИИ.

Как отмечалось выше, существует несколько типов искусственного интеллекта. Повторим определение одного из них: «**Автономный ИИ** предполагает возможность системы функционировать долгое время без участия оператора. Например, он позволяет дрону, оборудованному солнечными батареями, совершить многодневное путешествие с Елисейских полей на Красную площадь или в обратном направлении, самостоятельно выбирая как маршрут, так и места для промежуточных посадок для подзарядки аккумуляторов, избегая при этом всевозможные препятствия» Мы видим, что в примере для описания автономного ИИ нет ничего опасного, но определение автономного ИИ говорит о том, что в настоящее время создаются приборы, способные работать без вмешательства и контроля со стороны человека – а это уже опасно. Сейчас наука развивается гораздо быстрее, чем осознание разумом человека (пускай, даже самого умного) всех последствий ускоренного развития этой науки.

Алгоритмы и аппарат искусственного интеллекта создал сам человек. Но создавал он этот аппарат не «по образу и подобию своему», а на основе

собственных гипотез о том, как думает и устроен сам, и, называя, порой, интеллектом похожие на человека лишь внешне и схематически, структуры искусственного интеллекта. Именно так родились очень популярные сейчас схемы, названные нейро-сетевыми, которые, на самом деле, имеют не очень много общего с реальными нейронами человеческого мозга. Но именно искусственные (не природные) нейросети открыли большие возможности для сверхбыстрого решения многих задач и приобрели огромную популярность во всем мире.

Экспертные системы, применяемые для получения новых результатов при обработке, например, больших массивов данных, тоже создал человек, но и эти системы, как развитая часть методов искусственного интеллекта, сами являются искусственными и лишь немного похожими на некоторые способы мышления человека. Аналогичными особенностями обладают и другие направления искусственного интеллекта.

Все, что создано человеком в сфере искусственного интеллекта является по самой своей сути искусственными для человека конструкциями и используют лишь аналоги, дополненные неким вымыслом, связанным с, порой, фантастическим искусством математики. Поэтому человек создал интеллект по сути своей отличающийся от психологии человека, а, следовательно, чуждый его природе. Но, благодаря самому человеку, быстро развивающийся искусственный интеллект требует от человека искать безопасные пути сосуществования с чуждым по своей природе интеллектом.

В настоящее время порожденный человеком и находящийся пока в начале своего исторического развития искусственный разум начинает серьезно влиять на психологию самого создателя. Этот эффект, на наш взгляд, может наиболее сильно повлиять на молодое поколение. Ни для кого уже не является секретом, что сейчас роботы активно внедряются в сферу образования. Учителя школ, преподаватели вузов, пусть пока не во всех сферах их деятельности, но уже начинают заменяться искусственным интеллектом. Поэтому возникает риторический вопрос:

- А не станут ли ученики и студенты мыслить алгоритмами и схемами искусственного интеллекта, потеряв при этом свой природный разум?

Приведем небольшой пример влияния на аудиторию средств массовой информации для формирования общественного сознания аудитории в нужном кому-либо направлении.

В работе Олега Фиговского впервые введен и подробно описан термин «социальные нанотехнологии», которому можно дать следующее определение: «Социальные нанотехнологии – это невербальное (на уровне подсознания) влияние субъекта на аудиторию». Это влияние может оказываться многими способами, даже, например, подбором ведущих телевизионных или Интернет-

программ с необходимым заказчику внешним видом, темпераментом, уровнем общей культуры, тембром речи и т.д. Но уже сейчас телеведущие начали заменяться роботами, т.е. искусственным интеллектом. Замена человека роботом-ведущим обусловлена малыми экономическими затратами на его круглосуточное функционирование в эфире. В настоящее время стали широко известны роботы-художники, роботы-музыканты, роботы-композиторы, роботы-актеры и т.д. Эти роботы-творческие личности своей «профессиональной» деятельностью призваны влиять на подсознание живых людей, становясь автоматическими «социальными нанотехнологами», воздействующими своим «неодушевленным» искусством на духовную сферу жизни человека.

– Казалось бы, что может быть плохого во влиянии искусственного искусства на человека?

– Но робототехнический искусственный интеллект сейчас стоит дорого и в развитие его алгоритмов вносят деньги очень состоятельные люди и политики, которые по своей сути являются прагматиками. Поэтому неизбежно проявление психологии прагматизма и у искусственного искусства, которое породит прагматизм у живых людей. А это все и является робототехническими социальными нанотехнологиями.

В монографиях Олега Пенского описываются математические модели цифровых двойников человека. Пока эти модели основаны на математизации «бытовой» психологии отдельного субъекта и групп субъектов. Эта математизация, в числе прочего, позволила объективно описывать пока простейшее поведение систем «робот–человек».

На основе математических моделей была доказана теорема о том, что робот с абсолютной памятью (робот, который в отличие от человека помнит все) опасен для человека, где под опасностью следует понимать психологическое давление более воспитанного (здесь воспитание – это одновременно и психологический, и математический терминпо абсолютной величине робота на менее воспитанного по абсолютной величине человека.

В наших предыдущих работах впервые введены, так называемые коэффициенты мягкого и жесткого влияния одного субъекта на другого, а разработанная компьютерная программа позволяет вычислять эти коэффициенты. Удалось сформулировать следующую гипотезу: «Влияние искусственного интеллекта на воспитание человека с целью полного принятия человеком убеждений робота неизбежно». Доказательство гипотезы может быть основано на следующих очевидных фактах: «При общении робота с человеком коэффициент психологического влияния человека на бездушного робота равен нулю, а коэффициент влияния робота на одухотворенного

человека больше от нуля. Поэтому робот влияет на воспитание человека, а не наоборот».

– Как выходить из этой непростой ситуации?

Ответ на этот вопрос может быть таким:

– Для психологической безопасности человека необходимо создавать эмоциональных роботов с коэффициентами влияния на человека, меньшими, чем коэффициенты влияния человека на роботов.

Искусственный интеллект, синтетическая биология и так называемые неизвестные неизвестные могут уничтожить человечество до 2100 года, считает сооснователь Skype Яан Таллинн. Из трех угроз, которые больше всего беспокоят Таллинна, он сосредоточен на искусственном интеллекте и страны тратит миллионы долларов, пытаясь обеспечить безопасное развитие технологии. Это включает в себя инвестиции на ранних стадиях в лаборатории искусственного интеллекта, такие как DeepMind (отчасти для того, чтобы он мог следить за тем, что они делают) и финансирование исследований безопасности ИИ в таких университетах, как Оксфорд и Кембридж. Ссылаясь на книгу оксфордского профессора Тоби Орда, Таллинн сказал, что вероятность гибели людей в этом веке составляет 1 к 6. Согласно книге, одной из самых больших потенциальных угроз в ближайшем будущем является именно ИИ, а вероятность того, что изменение климата приведет к вымиранию человечества, составляет менее 1%.

Сейчас практически невозможно предугадать, каким будет развитие искусственного интеллекта, насколько умными станут машины в следующие 10, 20 или 100 лет. Попытки предсказать будущее ИИ осложняются тем обстоятельством, что системы ИИ начинают создавать другие системы ИИ уже без участия человека. Об опасности, связанной с выходом из-под контроля ИИ, не раз говорил и основатель SpaseX и Tesla Илон Маск.

По словам Таллинна, если выяснится, что ИИ не очень хорош для создания других ИИ, тогда человечеству не стоит слишком беспокоиться, однако в обратном случае «очень оправданно беспокоиться... о том, что произойдет дальше»».

Зададимся вопросом:

– Может ли искусственный интеллект для создания новых алгоритмов искусственного интеллекта обладать, например, интуицией и озарениями, как наиболее нестандартным проявлением исследовательского интеллекта, присущего, как считают многие современные ученые, только человеку?

В настоящее время нами уже разработаны первые, пока упрощенные, алгоритмы и интуиции и озарений роботов что открывает новые возможности для самореализации саморазвивающегося искусственного интеллекта в

социуме. Отметим то, что в настоящее время вопросы психологической безопасности искусственного интеллекта для человека специалистами по кибербезопасности почти не рассматриваются (за исключением исследований психологов, посвященных компьютерной зависимости человека) и опубликованы лишь единичные работы математиков, посвященные исследованию этой проблемы.

Президент РФ В.В. Путин 4 декабря 2020 года в своем выступлении на конференции по искусственному интеллекту, проходившей в Ново-Огарево (Московская область) , сказал следующее: «Искусственный интеллект - это, безусловно, основа очередного рывка вперед всего человечества… Есть опасение, что машины будут контролировать людей, но люди будут контролировать эти машины». На наш взгляд, для контроля влияния (не только психологического) искусственного интеллекта на социум необходимо привлекать математический аппарат, без которого определять величину опасности искусственного интеллекта для человека невозможно.

Величину опасности ИИ можно определить только тогда, когда известна конкретная цель, обеспечивающая безопасность его внедрения в социум и когда можно с помощью математики вычислить оценку достижения этой цели.