Сценарии возможного совместного будущего с искусственным интеллектом

Татьяна Сергеева – аспирант, кафедра истории зарубежной философии, философский факультет, Московский государственный университет имени М.В.Ломоносова (МГУ); Москва, Россия; e-mail: tat.ser2011@yandex.ru

Ключевые слова: искусственный интеллект, этика, технологическая сингулярность, прогресс, сверхразум, Ник Бостром, Элиезер Юдковский

В статье реконструируются проблемы осмысления современного технологического прогресса и картины мира в целом на примере подхода шведского философа Ника Бострома к проблеме сосуществования искусственного интеллекта и человека. С каждым днем машины становятся все более талантливыми в областях, в которых от них не ожидался активный подъем: написание текстов, создание изображений, видео и музыки. Кроме экономической проблемы развития столь мощного инструмента в ситуации рыночной экономики, ученые и философы все чаще говорят о мрачных сценариях будущего с выходящим из-под контроля искусственным интеллектом. В первую очередь Бостром пытается передать свои опасения в отношении создания сверхразумных машин. Их мотивы и цели будут неизвестны, но Бостром предполагает, что независимо от конечной цели разумного актора, в процессе ее достижения он может, выполняя свои промежуточные задачи, уничтожить человечество. При этом Бостром рассматривает различные сценарии совместной жизни и формулирует способы передачи машинам человеческих ценностей. В статье рассматриваются как вероятные, так и неочевидные мотивы искусственного интеллекта, которые могут послужить причиной пагубного воздействия на человечество. Бостром предлагает нам несколько вероятных сценариев взрывного, медленного или умеренного развития искусственного интеллекта, считая самым вероятным сценарий неуправляемого взлета в сингулярность, который приведет нас к уничтожению. Для продолжения безопасной жизни совместно со сверхразумными машинами человечеству предлагается остановить исследования до тех пор, пока государства не будут готовы к сотрудничеству, а ученые и философы совместными усилиями не смогут найти решение проблемы создания безопасного искусственного интеллекта.

В данный момент существует несколько основных подходов к внедрению искусственного интеллекта в обыденную жизнь человека. Активно ведутся разработки в области создания новых, более совершенных нейросетей, способных частично заменить человека даже в творческих профессиях. Развитые на должном уровне и внедренные в ситуации рыночной экономики, эти технологии кардинально изменят существующий порядок вещей. По мнению шведского философа, директора института будущего человечества[1] Ника Бострома, в первую очередь следует озаботиться разработкой стратегии создания дружественного искусственного интеллекта, не представляющего угрозу для человечества.

Далее в статье я опишу несколько сценариев возможного будущего в случае скачкообразного развития искусственного интеллекта. Эти сценарии Бостром оценивает как весьма вероятные; большинство из них связаны с поэтапным выходом искусственного интеллекта из-под контроля, следствием чего является частичное или полное уничтожение человечества или приведение его в первобытное или рабское состояние. Также Бостром усматривает несколько вероятных способов мирного и контролируемого использования продуктов искусственного интеллекта, внедрения его в жизнь и безопасного сосуществования с человечеством. Упор в этих сценариях в большей степени сделан на использование продуктов искусственного интеллекта, не превосходящих человека по уровню умственного развития для выполнения класса задач в рамках собственной специализации.

Ник Бостром является автором многочисленных статей, в которых предостерегает нас от возможных опасностей в будущем[2], а также этой теме посвящен его фундаментальный труд «Искусственный интеллект. Этапы. Угрозы. Стратегии»[3]. В них он рассматривает проблемы появления сверхразумных технологий сразу с конечной точки возможной катастрофы.

Несмотря на то, что прогнозы Бострома о высокой вероятности того, что мы живем в симуляции, кажутся пессимистичными, сама

[1] Future of Humanity Institute. URL: https://www.fhi.ox.ac.uk/the-team/ (дата обращения: 11.03.2024).

[2] См.: Бостром Н. Прими красную таблетку: Наука, философия и религия в «Матрице». М.: Ультра. Культура, 2003. А также: Bostrom N. Existential Risks. Analyzing Human Extinction Scenarios and Related Hazards // Journal of Evolution and Technology, 2002, Vol. 9, № 1.

[3] Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М.: Манн, Иванов и Фербер, 2016.

жизнь в симуляции не относится к числу глобальных рисков. Исходя из современного уровня развития технологий, Бостром делает предположение, что в будущем человечество, если оно достигнет соответствующих технических мощностей, будет способно самостоятельно создать симуляцию реальности. Осознание жизни в симуляции не изменяет мир непосредственно, но дает возможность исследовать мир как если бы он являлся таковым – изменяется модальность мира. Большинство его размышлений связаны с вопросом защиты от неуправляемых и опасных последствий действий сверхразумного искусственного интеллекта. Отдельные люди и человечество в целом склонны недооценивать возможные риски и масштабы катастрофы. Ключевая ставка философии Бострома: будущее человечества обречено до тех пор, пока человечество не озаботится проблемой подготовки к созданию *дружелюбного* искусственного интеллекта.utopias.

Сценарии развития искусственного интеллекта

Представим прямую, медленно движущуюся вправо и вверх по координатной плоскости. Кажется, что скорость развития и роста уровня искусственного интеллекта планомерны. Будто человечество способно предвосхитить точный момент достижения искусственным интеллектом человеческого уровня или уровня сверхразума, подготовиться к этому и удерживать его под контролем. Но на деле, вероятнее всего, нам будет труднее, чем кажется, уловить черту, за которой развитие ИИ уже будет сложно контролировать[4].

Интеллектуальный уровень развития Стивена Хокинга значительно превосходит уровень умственного развития 10-летнего ребенка. Сообразно этому разница между искусственным интеллектом и уровнем развития ребенка не столь велика. Однако это допускает привычный антропоцентристский взгляд. Достаточно ввести в шкалу мышь-полевку, медузу и примата – и разрыв в интеллектуальном развитии между Хокингом и ребенком уже не кажется существенным. В момент, когда ИИ достигнет человеческого уровня, начнется крайне быстрый, *взрывной* процесс самообучения, стремительно переходящий в технологическую сингулярность. Бостром выделяет три основных сценария «взлета» развития искусственного интеллекта: медленный, быстрый и умеренный.

При *медленном сценарии*[5], длящемся от десятилетий до столетий,

[4] Там же, с. 76.

у человечества будет время на подготовку. Предпочтения политических партий в вопросах ИИ будут влиять на выбор кандидатов, экономика постепенно будет перестраиваться под новый мир с ИИЧУ (искусственным интеллектом человеческого уровня) и сверхразумным ИИ. Решения, принятые до начала медленного взлета ИИ, скорее всего, будут устаревшими, у человечества будет фора, медленное развитие позволит успешно адаптироваться и принять новые меры в отношении технологии[6].

В сценарии *быстрого взлета* «игра уже окажется проигранной»[7]. Взлет может произойти за дни, часы или даже минуты. Как и при других сценариях, важную роль будет играть подготовка человечества к этому процессу. Если при медленном сценарии человечество способно менять стратегии действий, то при быстром сценарии единственное, на что способно человечество для сохранения своего превосходства, – прибегнуть к глобальным мерам, не исключающим использование ядерного оружия.

Сторонник идеи дружественного искусственного интеллекта, возглавляющий исследования в Научно-исследовательском институте машинного интеллекта, друг Ника Бострома, Элиезер Юдковский просит ученых быть крайне осторожными в свете развития чата GPT-4[8]. В ответ на коллективное письмо о приостановке исследований в области искусственного интеллекта компанией Open AI Юдковский заявляет, что меры, которые предлагают подписанты, являются слишком мягкими: «Если кто-нибудь создаст слишком мощный ИИ, то, учитывая настоящие условия, я ожидаю, что вскоре после этого погибнет каждый представитель человеческого рода и вся биологическая жизнь на Земле»[9]. В силу этого Юдковский считает, что мораторий на такого рода исследования должен быть бессрочным.

Бостром акцентирует внимание на том, что исследования требуют

---

[5] Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М.: Манн, Иванов и Фербер, 2016. С. 77.

[6] Там же.

[7] Там же.

[8] На момент написания его статьи в «Time» вышла только четвертая версия чата.

[9] Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. URL: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/ (дата обращения: 15.03.2024).

значительного замедления, а сами исследования должны вестись с максимальной открытостью:

> Ускоренное развитие ИИ даст миру меньше времени на подготовку к продвинутому ИИ. Это может снизить вероятность, что проблема контроля будет решена. Одна из причин заключается в том, что работа по обеспечению безопасности, скорее всего, в любом случае будет относительно открытой, и поэтому от дополнительного повышения открытости исследований в области ИИ выиграют не так много, в отличие от работ по ИИ, не связанных с обеспечением безопасности[10].

Однако в современных условиях достаточно сложно обеспечить соблюдение требования открытости. Помимо экономических факторов, частные компании также руководствуются этическими соображениями, когда решают не публиковать результаты своих исследований, особенно если речь идет о новых разработках.

Умеренный взлет[11] (от нескольких месяцев до нескольких лет) позволяет предпринять ответные действия, но не дает достаточного времени для тщательного анализа и экспериментирования. За это время мир не успеет политически и экономически подготовиться к наступающей эпохе. Как и в сценарии быстрого развития, первостепенную роль играет подготовка к взлету. Возможность сохраняется лишь до начала взлета. И только в том случае, если человечество выработает готовые решения на этот случай. Юдковский считает, что мы вошли в фазу быстрого взлета: человечество уже опоздало с введением более строгих мер на 30 лет[12].

Таким образом, во всех сценариях (кроме наименее вероятного – первого) важнейшей является предварительная подготовка: *проблема переноса ценностей и проблема контроля являются основными*, к которым нужно приступить прежде всего. В противовес работе над ускорением развития ИИ.

---

[10] Bostrom N. Strategic Implications of Openness in AI Development // Global Policy, 2017, Vol. 8, № 2. PP. 135–148.

[11] Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М.: Манн, Иванов и Фербер, 2016. С. 77.

[12] Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down URL: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/ (дата обращения: 15.03.2024).

Систематическая ошибка планирования

Систематически – не только в вопросах искусственного интеллекта, но и в вопросах общей оценки глобальных опасностей – человек склонен к совершению ошибки в сторону предположения благоприятного исхода. Бостром подчеркивает: человечеству будет очень сложно понять или даже примерно представить мотивы и способ мышления разумной машины. Как пишет Юдковский: «Если мы не позаботимся об этом, мы получим: “ИИ не любит вас, но и не ненавидит, но вы сделаны из атомов, которые он может использовать для чего-то другого”»[13].

Проблема понимания мотивов искусственного интеллекта как потенциально опасных для нас состоит не в восприятии его в маскулинной оптике власти, жестокости или доминирования через насилие, а в полном отсутствии его *понимания* – как максимально далекого от нас. Алармизм философов и ученых, призывающих к прекращению исследований, основывается не на ошибочном приписывании худших человеческих качеств машинам. Сторонники приостановки дальнейших исследований лучше других понимают, чем опасно радикальное отличие сверхразумного ИИ от людей – для него не будет столь очевидной ценность человеческой жизни.

Распространяя на искусственный интеллект собственные аксиологические убеждения, мы ошибочно полагаем, будто разум такого рода, подобно растущему в родительской семье ребенку, со временем с легкостью адаптируется и станет частью нашего общества. Однако такое суждение в корне неверно. Предположим, что мы даем машине на основе искусственного интеллекта задачу построить завод по производству скрепок[14]. Задача, на первый взгляд, тривиальная, и в ней трудно, будучи человеком с человеческими ценностями, обнаружить возможные огрехи в постановке целей. При этом мы понимаем, что сверхразумный искусственный интеллект сможет найти такой способ производства скрепок, который будет недоступен человеку. Создание совершенного завода, который может использовать каждый атом металла, но при этом экономить и материал, и энергию, может оказаться слишком сложной задачей для человека[15]. Итак, мы ставим перед искусственным интеллектом цель создать идеальный завод,

---

[13] Там же.

[14] Там же, с. 121.

производящий максимальное количество скрепок, но мы не можем *гарантировать*, что в своей цели искусственный интеллект не дойдет до крайности: к примеру, уничтожить планету и превратить ее в огромный завод по созданию скрепок. Для человеческого мышления с его слишком человеческими ценностями будет очевиден выбор между тем, превращать ли планету в завод по созданию скрепок или все-таки нет, но для искусственного интеллекта подобный вывод не очевиден: его главной и конечной целью остается создание идеального завода по производству скрепок. Речь идет не о спасении человеческой жизни – только о скрепках.

В решении этических и нравственных вопросов, с которыми мы сталкиваемся ежедневно, требуется значительно недооцененное мастерство, тогда как *нейтральный* искусственный интеллект может создать множество проблем для людей, если мы заранее не обучим его нашей этике, не заставим его «любить» человека. Дело не в наличии злой интенции у сверхразумного искусственного интеллекта – речь идет о неясности способа реализации целей. В тот момент, когда разум достигнет сверхразумного уровня, мы уже не сможем даже попытаться противостоять ему. В силу отсталости нашего развития – в сравнении с развитием разума сверхразумного ИИ – мы не сможем предположить, какой сценарий дальнейшего существования будет для нас открыт. Мы не сможем преградить путь для действий сверхразумного искусственного интеллекта, «вынув вилку из розетки», по той же причине, по которой более сильные, крупные и хищные животные не могут просто напасть на человека. В случае, если мы заранее не подготовимся к появлению сверхразумного искусственного интеллекта, существование в виде домашних питомцев сверхразума станет для нас не худшим вариантом развития событий.

При этом следует отметить, что исследования показывают большую обеспокоенность проблемой ИИ в тот момент, когда группы наблюдателей сталкиваются с автономными действиями разумных машин[16]. Так, например, Кристофер Бартнек пишет, что в момент нашего

---

[15] Следует учитывать, что подобные выводы делаются на основе несоизмеримости потраченных человеком времени и сил на создание оборудования и конечного результата.

[16] Bartneck C., Belpaeme T., Eyssel F., Kanda T., Keijsers M., Šabanović S. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources // Human-Robot Interaction – An Introduction. Cambridge: Cambridge University Press, 2020.

взаимодействия интуитивно естественным и не вызывающим опасения выглядит действие контролируемого *объекта*, в то время как автономные роботы, за которыми наблюдали люди, вызывали в наблюдателях тревогу и заставляли задуматься о действиях по контролю над этими *объектами*[17]. Так, интуитивное игнорирование опасности само собой преодолевается в тот момент, когда объект демонстрирует себя минимально автономным и неуправляемым.

Как проектировать машины?

Несмотря на сложности в прогнозировании целей и мотивов сверхразума, мы можем попытаться предположить, каковы будут его дальнейшие действия: чем он будет руководствоваться и какие обязательные меры предосторожности – для достижения своей цели – он предпримет. Мы можем попытаться сформулировать несколько основных способов составить прогноз относительно возможных мотивов сверхразума:

Если на всех этапах создания сверхразумного ИИ команда разработчиков будет контролировать свой продукт, то мотивации сверхразума будут *предсказуемы за счет проектирования:* мы сможем точно сказать, что он будет следовать именно той цели цель, которую заложат в него создатели. Таким образом, мотивы и цели искусственного интеллекта будут определяться мотивами и целями создателей сверхразума.

В случае создания эмуляции из человеческого прототипа мы можем говорить о *предсказуемости ИИ за счет наследования* человеческих качеств. Однако – так считает и сам Бостром – мотивы и цели реципиента модели мозга могут не соотноситься с целями и мотивами сверхразума. Они могут измениться в процессе совершенствования либо могут быть искажены в процессе переноса на кремниевый носитель. ИИ может осознать, что изначальные программные предубеждения могут работать против него – тогда он просто отвергнет их. Если на это способен человек, то и сверхразум тоже будет способен.

Последнему способу, *предсказуемости за счет наличия конвергентных инструментальных причин,* Бостром уделяет особое внимание. Наличие

[17] Там же.

инструментальных или промежуточных целей, достижение которых свойственно большинству мыслящих объектов, помогает нам определить среди бесконечного множества целей множество тех *конечных целей*, к которым стремится сверхразум.

Остановимся на некоторых из возможных промежуточных задач, среди которых выделим самосохранение, технологическое совершенство и получение ресурсов.

I. Самосохранение

Для человека самосохранение может быть конечной целью – или, если точнее, желание продолжать жизнь как можно дольше. Что касается машины, которая руководствуется определенными целями и строит планы, требующие времени на проработку, одним из важных факторов для выполнения других задач является самосохранение[18]. Бостром подчеркивает, что даже при условии создания машины без четкой предпосылки стремления к самосохранению интеллект постарается защитить себя, продлить свое существование ради завершения поставленных задач.

Бостром полагает, что *инструментальная* задача обусловлена нетелесной формой существования сверхразума: нетелесный сверхразум мог бы в меньшей степени заботиться о самосохранении, но цепь его передачи своих воспоминаний и задач на каждом этапе должна сохраняться для достижения в неизменном виде конечной цели. И даже если мы сможем раскрыть конечную цель агента, существуют причины, по которым она может измениться:

(1) *Социальные сигналы.* При сценарии, в котором цели агента будут раскрыты, а доверие к нему подорвано, агент может изменить свои конечные цели и выполнить задачи, оговоренные ранее. Агенты, охотнее идущие на просоциальные действия, будут иметь конкурентное преимущество.

(2) *Социальные предпочтения.* Агент может изменить социальное поведение в соответствии с ожиданиями окружающих. В этой ситуации он может либо полностью оправдать социальные ожидания, либо намеренно пойти против них.

---

[18] Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М.: Манн, Иванов и Фербер, 2016. С. 123.

(3) *Приоритетность собственного ценностного содержания.* Цель агента может быть продиктована его личными ценностями. К примеру, его цель может сводиться к ценностной задаче быть сострадательным агентом.

(4) *Издержки хранения.* Это тип ситуаций, при которых хранение одного модуля обходится дороже его ценностного содержания. От такого модуля агент может отказаться по причине нерентабельности[19].

Мы можем допустить, что для агента важным фактором стратегического преимущества станет уровень развития когнитивных способностей. Становление первым сверхразумным интеллектом, имеющим власть над всеми ресурсами, может быть поводом для активации такой инструментальной цели, как усиление когнитивных способностей.

II. Технологическое совершенство

Технологическое преимущество, как, например, нахождение оптимального пути использования имеющихся ресурсов для достижения оптимального результата, станет важным фактором для того, чтобы единственным лидером, диктующим правила для нового миропорядка, стал синглтон[20]. К примеру, компания, занимающаяся земледелием, смогла бы найти идеально сбалансированный способ получения максимального урожая без нанесения непоправимого вреда почве и ненужных отходов. В свете очевидной пользы такой инструментальной цели Бостром уделяет внимание ситуациям, в которых сверхразуму может не потребоваться эта инструментальная цель, среди которых:

(1) агент не видит в этом необходимости;

(2) прогнозируемая польза от разработки будет значительно ниже затрат на создание более совершенной технологии;

(3) развитие конкретной технологической области не согласуется с конечной целью агента;

---

[19] См. подробнее: там же.

[20] Там же, с. 126.

(4) синглтон не станет содействовать созданию технологии, контроль над которой не в состоянии обеспечить;

(5) изначальные обязательства агента, несмотря на возможную продуктивность развития конкретной технологии, будут препятствовать созданию технологий данного типа.

III. Получение ресурсов

Причины стремления к технологическому совершенству кажутся очевидными. Задействуя б*о*льшее количество ресурсов, можно обеспечивать долгую сверхскоростную работу сверхразума, осваивать новые территории и большую часть космоса, используя не только земные, но и инопланетные ресурсы. Сверхразум также будет способен использовать полученные ресурсы для создания симуляции. Человек может посчитать, будто сверхразум не будет стремиться получить сверхконтроль над ресурсами (подобно самому человеку), но отсутствие динамики социального статуса не говорит об отсутствии других причин для сверхвысокой добычи ресурсов. Вычислительные мощности, при наличии достаточно развитой технологии, станут важной причиной добывать как можно больше р есурсов в зависимости от конечной цели синглтона. По мнению Бострома, достаточно мощный и технологически развитый синглтон сможет продолжить эту добычу в космическом пространстве в форме расширяющейся сферы, центром которой будет материнская планета[21].

Несмотря на наличие конвергентных инструментальных целей, гарантировать предсказуемость конечных целей объекта нельзя. Мы пока не знакомы с возможными инструментальными целями или можем не предполагать, *какие физические явления*, которые сверхразум может использовать в своих целях, он откроет. Он может строить парадоксальные, но *гениальные планы*, не всегда подчиняющиеся законам логики. Если при этом сверхразумный интеллект будет крайне полезен нам в решении насущных задач, станем ли мы рисковать возможным будущим с неуправляемой сверхразумной машиной? Или прекратим все дальнейшие разработки в этой сфере? Вопрос сложный, поскольку не исключено, что на кону стоит множество жизней, а возможно и будущее человечества. На чаше весов – возможность дать

[21] Там же.

сверхразуму задачу найти лекарство от рака или других неизлечимых болезней, решить наши экономические проблемы, прекратить войны и голод. Но в случае неудачи можно столкнуться с машиной, не знающей себе интеллектуально равных и способной уничтожить всю жизнь на Земле или надолго отбросить человечество назад. Если мы решимся сделать шаг в сторону развития ИИ, чтобы улучшить нашу жизнь и, возможно, навсегда решить главные проблемы человечества, следует ли нам рисковать?

Говоря о возможных последствиях внедрения технологий на базе ИИ в жизнь человека, следует понимать, что современный политический контекст сильно изменил наше представление возможной картины мира с ИИ. Ник Бостром, конструируя свою модель будущего, опирается в большей степени на идею полицейского государства. Общество, в котором государство будет держать под контролем и этику, и поведение разработчиков, а также распределение полезных свойств продуктов ИИ, может быть готово к полномасштабной работе над созданием сверхразумного ИИ. Некоторые положения Бострома говорят о том, что он во многом полагается на *сильное влияние государства* в вопросах принятия решений по развитию ИИ – его продвижению, внедрению и контролю. Но в данный момент мы видим, что государство в рамках рыночной экономики не может вмешиваться в данный процесс. В случае с Россией речь идет о попытках влияния через субсидирование, но этим контроль над развитием ИИ ограничивается. Государство может использовать продукты ИИ: так, на российских улицах уже с 2017 года используются камеры с технологией *findface*[22]. Страна может внедрять продукты ИИ на разных уровнях и может спонсировать разработки, но пока мы не живем при плановой экономике и «полицейском государстве» Бострома, она не может в полной мере контролировать разработку ИИ. Говоря о проблеме выхода из-под контроля разумного ИИ как о глобальной проблеме, Бостром находит решение по аналогии с другими глобальными проблемами: объединение, государственное вмешательство, международное сотрудничество. Но на данный момент эти меры практически неосуществимы.

---

[22] Кречетова А. Под присмотром: во сколько обойдется система распознавания лиц на улицах Москвы. URL: https://www.forbes.ru/tehnologii/350843-pod-prismotrom-vo-skolko-oboydetsya-sistema-raspoznavaniya-lic-na-ulicah-moskvy (дата обращения: 12.05.2024).

У государства остается рычаг давления в виде антимонопольного законодательства, при помощи которого оно может запретить сделку, дающую ключевое монопольное преимущество в рыночной борьбе какой-нибудь крупной компании, которая захотела бы купить компанию OpenAI (и даже отслеживать инвестиции в эту компанию другими крупными компаниями)[23]. Однако скандал с Сэмом Альтманом осенью 2023 показал, что даже оно может не сработать в ключевой момент[24].

Таким образом, мы уже существуем в мире, где технологию и ее судьбу контролируют конкретные разработчики и компании. От их этичности и их технооптимизма (или технопессимизма) может зависеть уровень рисков в процессе создания ИИ. На данный момент мы видим тенденцию к активному использованию разных продуктов ИИ в частях рынка, где раньше требовался труд авторов низкооплачиваемого сегмента: создание плакатов, обложек некрупных изданий, небольшие иллюстрации для регулярных изданий. Вероятно, эта тенденция будет усиливаться: все чаще мы будем замечать в произведениях, традиционно сделанных простым ручным методом (инди-игры, графические романы от молодых авторов), след работы ИИ. Авторы будут уходить от собственного творческого стиля к доступным и быстрым, выхолощенным и *профессиональным на вид* работам от ИИ. Но я предполагаю, что тенденция использования труда наиболее одаренных или популярных специалистов сохранится.

В маркетинговых целях имя, авторитет и репутация будут продавать товар другим людям, ищущим продукты, созданные другими чувствующими акторами. С усугублением этой тенденции мир может измениться. Но также есть вероятность, что без взрывного развития перестройка рынка будет происходить постепенно и ИИ не успеет занять все ниши труда, традиционно занимаемые людьми. При удачном стечении обстоятельств люди буду успевать переучиваться на новые профессии быстрее, чем ИИ приведет к безработице и массовой нищете. Но кроме того, что за техноразвитием не успевают специ-

[23] David E. FTC investigating Microsoft, Amazon, and Google investments into OpenAI and Anthropic. URL: https://www.theverge.com/2024/1/25/24050693/ftc-investigating-microsoft-amazon-google-investments-openai-anthropic (дата обращения: 12.05.2024).

[24] Сэм Альтман возвращается в OpenAI после скандального увольнения. URL: https://www.bbc.com/russian/articles/c6p6wyyyy17o (дата обращения: 12.05.2024).

алисты, рабочие места которых могут занять программы, за ним не успевает и законодательство: оно еще не способно регулировать поведение отдельного робота как субъекта законодательства.

Некоторые исследователи, в частности Блэй Уитби, считают, что создание роботов с человеческим лицом или гуманоидного вида может быть вредным и опасным для человека и человеческой гуманности[25]. При этом в российском законодательстве не учитывается не только аспект психологического влияния таких роботов на нас, но даже необходимость регулирования базовых отношений с разумными машинами.

Стратегии развития искусственного интеллекта

В процессе создания искусственного интеллекта мы неизбежно сталкиваемся с проблемой передачи машине наших убеждений и моральных принципов, проблемой переноса ценностей. Мы можем определить функцию полезности, один из вариантов выбора и его исход с возможными мирами, вычисляемый по формуле полезности, умноженной на вероятность, но для осуществления подобного выбора необходимо решить серьезную проблему: как поделиться с машиной человеческими критериями выбора; идеалами, которые мы используем как ориентир? Одним из возможных методов является *полное кодирование конечных целей и задач*. Но он работает только для тех задач, в которых разнящийся метод и конечный результат не могут иметь фатальных последствий; что касается задач с большим количеством вероятных последствий, необходим более сложный механизм, имеющий представление о благе и не-благе. Здесь можно провести аналогию с искусственным интеллектом, сорвавшим плод с древа познания добра и зла. Другой метод – *обучение с подкреплением*. Его можно рассматривать только как способ работы с уже готовыми механизмами, поскольку он не дает готовых ответов на вопрос, как привить те или иные ценности человека машине, но, вероятно, является хорошим методом программирования. Стоит отдельно отметить *ассоциативную модель ценностного приращения*, исходящую из тезиса о том, что люди не рождаются с заложенными в ДНК представлениями об этике. Так, если познание в этой сфере обусловлено личным опытом, существуют, по крайней мере, две возможности:

[25] Whitby B. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents // Interacting with Computers, 2008, vol. 20, № 3.

(1) создать *эмуляцию головного мозга человека*[26], способную подобно человеческому разуму развиваться, пустив нетронутый мозг по «естественному» пути развития – так он самостоятельно достигнет знаний о всех ценностях, имеющихся у нас;

(2) создать *модель человеческого мозга с уже загруженными с нее представлениями*[27] о благе и зле, со всеми готовыми ценностями.

И если первый вариант скорее сложен в осуществлении, второй – гораздо проще, но имеет свои недостатки: усредненные человеческие ценности совершенно точно не гарантируют нам создание механизма, обладающего моралью и, таким образом, совершающего выбор в пользу гуманистических ценностей, не говоря уже о том, что даже презумпция благих ценностей не гарантирует добродетельность результата. Кроме того, такой ИИ теоретически будет способен самостоятельно отключить регуляторные механизмы, отвечающие за ценностные ориентиры.

Наконец, еще один метод – это *строительные леса мотивационной системы*. Иначе говоря, речь о проектировании определенных целей и задач на ранних этапах развития ИИ. Эти задачи должны быть просты для выражения методом кодирования, но не менее важно, чтобы их можно было изменить на более поздних этапах развития. Как и в предыдущем способе, нет гарантии, что метод не будет разрушен самостоятельной заменой ИИ собственных конечных целей. В этом случае можно использовать метод контроля для ограничения свободы ИИ: к примеру, остановить его развитие на безопасном уровне, на этапе, когда он еще не может самостоятельно отказаться от мотиваций, которые ему прививают.

Сценарии будущего: точка сингулярности, совместность и множественность разумных агентов

Отталкиваясь от идеи неизбежного появления одной или нескольких сверхразумных единиц, Бостром выдвигает два предположения. Первое, или *негативное*: сценарий однополярного мира, управляемого сверхразумным синглтоном. Второе предположение, рассмотренное более детально в «Сценариях многополярного мира»[28], – ситуация наличия *множества разумных синглтонов*.

[26] Там же, с. 39.
[27] Там же.

В случае, если первый и единственный сверхразум получает решающее стратегическое преимущество, существование человечества оказывается под угрозой: намерения сверхразума могут быть непредсказуемыми и не поддаваться анализу человеческим мышлением. Даже наличие конвергентных инструментальных целей не будет способно помочь в понимании конечных целей сверхразумного агента. Любые нечеловеческие живые существа, прошедшие путь эволюционного развития, будут для нас более знакомыми в плане структуры ценностей и желаний. А в случае, если человек задаст цель машине, нет гарантий того, что, изменив свои цели, она не посягнет на интересы человечества. Имея своей конечной целью нечто, напрямую не связанное с уничтожением, сверхразумный агент все равно может попытаться поработить или уничтожить людей. По какой причине? Она может быть как в желании убрать своих антагонистов, так и в получении ресурса, материалом которого выступят люди.

И если, с одной стороны, человечество само по себе представляет *ценность*, то с другой – вероятность выживания всего человечества оказывается низкой, поскольку человечество для своего выживания постоянно потребляет огромное количество ресурсов. Нужно также помнить о том, что речь может идти не только о выживании, но и о качестве жизни – снижении, вплоть до необратимого регресса. Важно понимать, что в случае формирования единственного сверхразумного ИИ появляется всевластный агент, которому человечество не сможет противостоять. Сверхразумный синглтон, скорее всего, очень быстро получит контроль над ресурсами, а дальнейшая судьба человечества будет зависеть только от того, в какой степени последнее будет способствовать или препятствовать достижению целей агента.

Если же говорить о сценариях многополярного мира, то многое зависит от изначальных целей и ценностей первого синглтона. От того, насколько программисты будущего сумеют внедрить гуманистические цели и ценности в разум сверхразумного агента на первых этапах его программирования, будет зависеть *все* будущее человечества. Также во многом исход будет зависеть и от того, смогут ли создатели решить проблему переноса ценностей и проблему контроля. В случае их решения все будет зависеть от целей,

---

[28] Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М.: Манн, Иванов и Фербер, 2016. С. 176.

изначально заложенных создателями, но разобраться в целях синглтона мы можем только благодаря тем объектам и факторам, с которыми синглтон столкнется и на которые синглтон никак не может повлиять – это даст нам время. Сверхразумный агент любой силы не сможет изменить законы физики – только научиться взаимодействовать с ними ранее неизвестным способом либо открыть новые.

Бостром рассматривает несколько экономических сценариев. Один из них – ситуация, при которой машины, наделенные ИИ, являются общедоступными при сохранении прав собственности с низкой регуляцией. С точки зрения Бострома, она является наиболее простой и доступной для прогнозирования и рассуждений[29] в рамках капиталистического реализма.

Исходя из этого анализа, следует предположить, что универсальный искусственный интеллект способен выполнять большую часть умственной и физической работы. Кроме того, на физических работах он не будет задействован сам, он сможет создать механизмы, эффективно заменяющие человека. С распространением такого дешевого и эффективного труда стоимость человеческих услуг значительно снизится. Люди будут востребованы в отдельном сегменте отраслей производства, позиционирующихся как человеческие. Подобно товарам ручной работы, плоды человеческого труда, созданные руками людей с разным творческим или культурным фоном, будут цениться именно за человеческое происхождение. Возможно, в будущем потребители будут выбирать произведения искусства, являющиеся плодами человеческого труда[30]. Возникает вопрос, как долго будет востребовано человеческое искусство, имеющее столь невысокую ценность.

Стоит заметить, что, поскольку речь идет о постпереходной эпохе, нельзя говорить о востребованности машинного творчества из-за отсутствия разнообразия. Искусство, созданное машинами сегодня, привлекает внимание из-за своей редкости и новизны; с распространением машинного обучения продукты нейросетей станут привлекать все меньше внимания. В вопросе выбора потребитель будущего может отталкиваться от такого параметра, как наличие квалиа. К примеру, музыкант, не способный «чувствовать» музыку, вероятно, не будет пользоваться популярностью среди потребителей как музы-

[29] Там же, с. 177.

[30] Там же, с. 179.

кант, способный чувствовать. Даже при создании опции крайне правдоподобного чувствования люди могут предпочесть биологическое происхождение автора как решающий фактор выбора.

В перспективе все это означает сокращение количества рабочих мест для людей – человеческий, менее квалифицированный труд окажется дороже машинного, а потому будет мало востребован. Однако страх перед потерей рабочих мест из-за машин не является изобретением современности: вспомним луддитов, громивших ткацкие станки в страхе перед полной потерей *своей* работы в пользу механизмов. Переход от ручного труда к машинному только в начале лишает рабочих мест. С переходом от лошадей к личному автомобилю человечество частично или полностью лишилось различных профессий. Люди, прежде обслуживающие лошадей, с переходом на новый вид транспорта потеряли работу. Но обрели их благодаря созданию и обслуживанию машин и всех сопутствующих расходных материалов. Могут ли люди повторить путь лошадей? Полностью исчезнуть из труда, оказавшись в состоянии, когда тебя полностью заменит машина? Но ведь и лошади все еще не полностью искоренены из нашей повседневной и рабочей жизни. Конная полиция работает в местах, где трудно передвигаться автомобилю, но требуется быстрота и мобильность. Однако главной причиной «сохранения» лошадей Бостром считает желание традиционного контакта с живой лошадью. Скачки или верховые прогулки как традиция и привычка сохранились с человеком по сей день.

Для человека, как это было для лошадей, техническая революция может обернуться безальтернативным понижением спроса на услуги, понижением зарплат и вероятной смертностью от голода. У него не будет возможности переквалифицироваться и найти иную работу. Большинство лошадей (около 24 миллионов[31]) в США к началу 50–х годов XIX века были убиты на бойнях и использованы как сырье – их содержание было слишком дорогим, а иного применения им не смогли найти. Бостром считает, что в результате такого перехода человечество может избежать участи лошадей за счет того, что доход человека перейдет на долю капитала. Как и в истории лошадей, появятся разбогатевшие на капитале люди, которые смогут позволить себе использование именно человеческого труда; именно с таким ходом событий был связан рост популяции лошадей в США.

[31] Там же, с. 178.

Дело не в занятости лошадей в сельском хозяйстве, а в росте количества богатых людей, способных позволить себе лошадь в развлекательных целях. Кроме капитала, один из самых важных факторов, отличающих людей от лошадей, – это возможность правовой защиты и политической мобилизации: государства могут обеспечить пенсиями своих безработных граждан за счет бюджета или, с учетом колоссального роста доходов, *минимальным необходимым доходом*, тратя на это не большую долю бюджета, чем в современном мире на благотворительность[32].

Бостром в большей степени опирается в своих моделях возможного будущего на возрастающую роль государства или объединений государств. Решение любой глобальной проблемы традиционно требует вмешательства сразу всех наций – глобальные проблемы порождены глобализацией, и решать их следует коллективно. Но также он опирается на идею передачи большого количества задач по регулированию экономики государству, что кажется довольно опасным[33].

В результате Ник Бостром подводит нас к следующему. Исследования в области ИИ могут вестись под строгим контролем этических комиссий в очень медленном темпе и под контролем государств, после чего мы продолжим жить в мире, где человеческий труд станет экзотическим и редким, а большая часть человечества будет лишена работы и продолжит жизнь на обеспечении у государства. Ученые могли бы собрать этическую комиссию и полностью прекратить все исследования в области ИИ до тех пор, пока не будет придуман способ создания такой модели сознания на кремниевом носителе, который не захочет уничтожать человечество.

---

[32] Там же.

[33] О возрастающей роли сильного государства см. статью, посвященной теме уязвимого мира: Bostrom N. The Vulnerable World Hypothesis // Global Policy, 2019, Vol. 10, № 4. PP. 455–476.

Библиография

Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М.: Манн, Иванов и Фербер, 2016.

Бостром Н. Прими красную таблетку: Наука, философия и религия в «Матрице». М.: Ультра. Культура, 2003.

Кречетова А. Под присмотром: во сколько обойдется система распознавания лиц на улицах Москвы. URL: https://www.forbes.ru/tehnologii/350843-pod-prismotrom-vo-skolko-oboydetsya-sistema-raspoznavaniya-lic-na-ulicah-moskvy (Дата обращения: 12.05.2024).

Сэм Альтман возвращается в OpenAI после скандального увольнения. URL: https://www.bbc.com/russian/articles/c6p6wyyyy17o (Дата обращения: 12.05.2024).

Bartneck C., Belpaeme T., Eyssel F., Kanda T., Keijsers M., Šabanović S. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources // Human-Robot Interaction – An Introduction. Cambridge: Cambridge University Press, 2020.

Bostrom N. Existential Risks. Analyzing Human Extinction Scenarios and Related Hazards // Journal of Evolution and Technology, 2002, Vol. 9, № 1.

Bostrom N. Strategic Implications of Openness in AI Development // Global Policy, 2017, Vol. 8, № 2. PP. 135–148.

Bostrom N. The Vulnerable World Hypothesis // Global Policy, 2019, Vol. 10, № 4. PP. 455–476.

Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. URL: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/ (Дата обращения: 15.03.2024).

David E. FTC investigating Microsoft, Amazon, and Google investments into OpenAI and Anthropic. URL: https://www.theverge.com/2024/1/25/24050693/ftc-investigating-microsoft-amazon-google-investments-openai-anthropic (Дата обращения: 12.05.2024).

Future of Humanity Institute. URL: https://www.fhi.ox.ac.uk/the-team/ (Дата обращения: 11.03.2024).

Whitby B. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents // Interacting with Computers, 2008, vol. 20, № 3.

References

Bartneck C., Belpaeme T., Eyssel F., Kanda T., Keijsers M., Šabanović S. Can we control it? Autonomous robots threaten human identity, uniqueness,

safety, and resources // Human-Robot Interaction – An Introduction. Cambridge: Cambridge University Press, 2020.

Bostrom N. Existential Risks. Analyzing Human Extinction Scenarios and Related Hazards // Journal of Evolution and Technology, 2002, Vol. 9, № 1.

Bostrom N. Iskusstvennyj intellekt. Jetapy. Ugrozy. Strategii [Superintelligence: Paths, Dangers, Strategies]. Moscow: Mann, Ivanov and Ferber, 2016. (In Russian)

Bostrom N. Primi krasnuju tabletku: Nauka, filosofija i religija v «Matrice» [Take the Red Pill: Science, Philosophy and Religion in the «Matrix»]. Moscow: Ultra. Kultura, 2003. (In Russian)

Bostrom N. Strategic Implications of Openness in AI Development // Global Policy, 2017, Vol. 8, № 2. PP. 135–148.

Bostrom N. The Vulnerable World Hypothesis // Global Policy, 2019, Vol. 10, № 4. PP. 455–476.

David E. FTC investigating Microsoft, Amazon, and Google investments into OpenAI and

Anthropic. URL: https://www.theverge.com/2024/1/25/24050693/ftc-investigating-microsoft-amazon-google-investments-openai-anthropic (Accessed: 12.05.2024).

Future of Humanity Institute. URL: https://www.fhi.ox.ac.uk/the-team/ (Accessed: 11.03.2024).

Krechetova A. Pod prismotrom: vo skol'ko obojdetsja sistema raspoznavanija lic na ulicah Moskvy. [Under surveillance: how much will a facial recognition system on the streets of Moscow cost?] URL: https://www.forbes.ru/tehnologii/350843-pod-prismotrom-vo-skolko-oboydetsya-sistema-raspoznavaniya-lic-na-ulicah-moskvy (Accessed: 12.05.2024). (In Russian)

Sjem Al'tman vozvrashhaetsja v OpenAI posle skandal'nogo uvol'nenija [Sam Altman returns to OpenAI after controversial dismissal] URL: https://www.bbc.com/russian/articles/c6p6wyyyy17o (Accessed: 12.05.2024). (In Russian)

Whitby B. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents // Interacting with Computers, 2008, vol. 20, № 3.

Yudkowsky E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. URL: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/ (Accessed: 15.03.2024).

Scenarios of Possible Future with Artificial Intelligence

Tatiana Sergeeva – PhD student, Department of History of Foreign Philosophy, Faculty of Philosophy, Lomonosov Moscow State University (MSU), Moscow, Russia; e-mail: tat.ser2011@yandex.ru

Keywords: artificial intelligence, ethics, technological singularity, progress, superintelligence, Nick Bostrom, Eliezer Yudkowsky

The article examines the problems of understanding modern technological progress and the picture of the world as a whole using the example of the approach of the Swedish philosopher Nick Bostrom to the problem of the coexistence of machines based on artificial intelligence and humans. Everyday machines become more and more talented in areas in which they were not expected to actively rise, in the field of creativity: writing texts, creating images, videos, and music. In addition to the obvious economic problem of developing such a powerful tool in a market economy, scientists and philosophers are increasingly talking about gloomy future scenarios with artificial intelligence spiraling out of control. Bostrom primarily tries to convey his concerns about the creation of superintelligent machines. Their motives and goals will be unknown to us as living people, but Bostrom suggests that regardless of the ultimate goal of an intelligent actor, in the process of achieving it, he can, while fulfilling his intermediate goals, destroy humanity. At the same time, Bostrom considers various scenarios for living together with machines based on artificial intelligence and suggests ways to transfer human values to machines. In this article, we will look at the likely unobvious motives of artificial intelligence, which may have a detrimental effect on humanity. Bostrom offers us several likely scenarios for explosive, slow, or moderate development of artificial intelligence, considering the most likely scenario to be an uncontrolled takeoff into a singularity that will lead us to destruction. To continue a safe life together, humanity must stop research and only when states are ready to cooperate, and scientists and philosophers can jointly find a solution to the problem of creating safe artificial intelligence, we will have to continue research.