ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: НАДЕЖДЫ И ОПАСЕНИЯ

А.И. Минникес

О том, каким именно будет искусственный интеллект и каким образом его создать, ученые спорят с 1956 года. Тогда искусственный интеллект существовал лишь в футуристических романах фантастов, и время, когда он будет в действительности создан, указывалось как 80-90-е годы двадцатого века. Однако искусственный интеллект не

создан до сих пор, и понимание того, каким он должен и может быть, по-прежнему является объектом научных споров.

Искусственный интеллект, в дальнейшем ИИ, традиционно определяется как попытка симуляции человеческого разума машинными или программными средствами. Обсуждение сущности разума выходит за рамки этой статьи, посему укажем лишь основные границы функциональности подобных программ. Итак, интеллектуальной может считаться программа, которая способна рассуждать, обучаться, общаться, воспринимать, работать с собственным исходным кодом, способна к обобщению и аналогиям. Кроме того, обладающая собственными средствами получения информации от внешнего мира, т.е. органами восприятия: зрением, слухом, возможно, обонянием. Оценив эти критерии «интеллектуальности», мы получаем фактически человека, воссозданного в машине.

В настоящий момент существует три основных подхода к созданию ИИ.

1. «Аппаратная аналогия», также называемая нейрокибернетика.

Данный подход основан на понимании работы человеческого мозга и ее имитации программными и машинными средствами. Детищем этого подхода стали нейрокомпьютеры, нейронные сети. Традиционной критикой считается то, что механизмы работы человеческого мозга до сих пор окончательно не изучены, а следственно, не могут быть доподлинно воспроизведены.

С философской точки зрения, мы можем столкнуться с проблемой «чудовища Франкенштейна». Монстр был собран из деталей человеческого тела и воссоздан в точности. Но жизнь в нем не зародилась до тех пор, пока не появилось электричество. Была ли на молнию возложена функция божественного вдохновения – вопрос спорный и выходящий за рамки данной статьи. Однако чтобы ожила даже точнейшая имитация человеческого мозга, кто-то должен вдохнуть в нее «душу».

2. «Аналогия поведения»: моделирование не мозга, но поведения человека.

Данный подход исходит из того, что не обязательно воссоздавать физическую модель мозга, достаточно лишь точно скопировать поведение человека, а также его мышление, эмоции, мотивации и др.

Данный подход зачастую не выдерживает критики ввиду разнообразия толкований мотивации поступков и мышления. Кроме того, создание подобного ИИ представляет опасность в силу факторов, которые будут рассмотрены позднее в этой статье.

3. «Метафора колеса»: аналогии с человеческим мозгом и поведением не обязательны, нужно создавать что-то принципиально своё.

ИИ должен обладать способностями человеческого мозга, не повторяя принципы его работы. Суть подхода заключается в том, что когда было создано колесо, его не копировали с чего-то существующего, однако, оно все-таки покатилось.

Будет неправомерным утверждать, что успехи человечества по созданию искусственного интеллекта пока ограничиваются тиражами фантастических романов по теме. В разные годы различным программам присваивалось звание ИИ.

• 80-е годы: Экспертные системы (симуляция знаний и аналитических навыков одного или нескольких экспертов)

• 90-е годы: ИИ-подобные программы работали в сфере логистики, медицинского прогнозирования, data mining и высокотехнологичных областях промышленности.[1]

Тем не менее, настоящий искусственный интеллект до сих пор не создан. Имитации ребенка, способность к обучению, являются существенным шагом на пути к этому. И даже на этот небольшой шаг к ИИ у человечества ушло более 50 лет.

Описание технической стороны создания машинного разума не входит в задачи этого исследования, тем более, что технологии различны и многообразны. Гораздо важнее те принципы, которые будут заложены в основу будущего ИИ.

На пути к ИИ ученых подстерегают не только технические, но и философские неудачи, и по последствиям гораздо более пагубны вторые, а не первые. Потенциальные неудачи в попытках создания ИИ делятся на две неформальные категории, техническую ошибку и философскую ошибку.

Техническая ошибка состоит в том, что ИИ не функционирует так, как должен – и никто не может понять, как работает на самом деле созданный код. При более безопасном раскладе код просто не работает.[2]

Философская неудача заключается в попытке построить заведомо неправильную вещь. Иными словами, при создании ИИ допускается критический просчет, и его дальнейшее функционирование становится непредсказуемым. Чаще всего философская неудача истекает из так называемой «ошибки антропоморфизма». [6]

Ошибка антропоморфизма определяется как неосознанное присвоение животным и неодушевленным предметам человеческих качеств, мышления, мотивации, целеполагания и так далее. Такая ошибка характерна, прежде всего, для зоологии, где естественному отбору когда-то приписывалась осознанность. [3], [4] , [5]

Но и в кибернетике риск такой ошибки велик, особенно при создании искусственного интеллекта, который по умолчанию кажется антропоморфным. Надежда на это, не подкрепленная заранее заложенными в ИИ рамками – верный путь к глобальной катастрофе. И даже предварительно заложенные рамки не являются абсолютной панацеей,

поскольку ИИ способен не только самообучаться, но и самостоятельно модифицировать свой код. При этом если человечеству понадобилось 50 лет на небольшой шаг в сторону создания ИИ, самому ИИ понадобится значительно меньше времени. Настолько, что у человечества просто не будет возможности оперативно реагировать на изменения. Это является одним из главных рисков при создании ИИ. Человек неспособен улучшать себя рекурсивно, в то время как для ИИ это нормально.

Другим риском, вытекающим из первого, является возможность самомодификации ИИ. Он может переписать свой код с самого начала, кардинально изменив не только задачи, но и механизм работы. Кроме того, был способен к ликвидации любых рамок, заложенных программистами. И эта возможность может быть использована для любых целей.

Один из спорных вопросов состоит в том, что именно захочет делать созданный ИИ. Спрогнозировать это не представляется возможным, поскольку ИИ не может быть антропоморфным. Соответственно, ему чужды человеческие мотивации. Существует такое понятие, как ошибка «гигантской ватрушки», которую можно истолковать следующим образом: чем больше вычислительные мощности и возможности ИИ, тем больше он «захочет» сделать. Например:

- Достаточно сильный ИИ может преодолеть любое человеческое сопротивление и истребить человечество;
- И ИИ решит сделать это.
- Поэтому мы не должны создавать ИИ.

Или:

- Достаточно сильный ИИ может создать новые медицинские технологии, способные спасти миллионы человеческих жизней.
- И он решит сделать это.
- Поэтому мы должны создать ИИ.

Оба предположения могут быть как верны, так и ошибочны, потому что ИИ в действительности может и не «захотеть» поступать именно таким образом.

Рассмотренные примеры ясно дают понять, что создавать всемогущую машину, которая будет работать совершенно непредсказуемым образом – рискованное занятие для человечества, потому что соперничать с уже созданным ИИ оно не сможет. Это приводит нас к выводу о том, что при создании ИИ необходимо заранее ориентироваться на определенный результат, а не ставить эксперименты с непредсказуемым исходом.

Одним из направлений развития «прогнозируемого» ИИ является создание так называемого NICE AI или Дружественного Искусственного интеллекта.

Дружественный ИИ – это ИИ, созданный со специфической мотивацией. На первый взгляд, попытки создать такой ИИ заранее обречены на ошибку, так как ИИ может самостоятельно модифицировать свой код и разорвать любые наложенные ограничения. В данном случае наблюдается уже рассмотренная выше ошибка «гигантской ватрушки». Любой ИИ, имеющий свободный доступ к своему исходному коду, в принципе, будет обладать способностью изменить его таким образом, что изменится и цель оптимизации. Но это не означает, что ИИ имеет побуждение изменить свои собственные побуждения.

На настоящий момент имеется неограниченное количество расплывчато убедительных аргументов, почему Дружественный ИИ может быть не под силу человеку, и всё же гораздо вероятнее, что проблема разрешима. Но не следует слишком быстро списывать проблему, особенно учитывая масштаб ставок. [7]

Поле исследований ИИ адаптировалось к тому жизненному опыту, через который оно прошло за последние 60 лет, в частности, к модели больших обещаний и следующих за ними публичных провалов. Культура исследований ИИ адаптировалась к следующему условию: имеется табу на разговоры о способностях человеческого уровня. Есть ещё более сильное табу против тех, кто заявляет и предсказывает некие способности, которые они ещё не продемонстрировали на работающем коде. [7]

Складывается впечатление, что каждый, кто заявляет о том, что исследует Дружественный ИИ, косвенным образом заявляет, что его проект ИИ достаточно мощен, чтобы быть Дружественным. Очевидно, что это неверно ни логически, ни философски. Зрелый ИИ, который достаточно мощен для того, чтобы быть Дружественным, и, более того, если, в соответствии с желаемым результатом, этот ИИ действительно является Дружественным, потребует на свое создание годы и годы. Дружественный ИИ – это не модуль, который можно мгновенно изобрести, в точный момент, когда он понадобится, и затем вставить в существующий проект.

Поле исследований ИИ имеет ряд техник, таких как нейронные сети и эволюционное программирование, которые росли маленькими шажками в течение десятилетий. Но нейронные сети непрозрачны. Пользователь не имеет никакого представления о том, как нейронные сети принимают свои решения. Эволюционное программирование является стохастическим, и не сохраняет точно цель оптимизации в сгенерированном коде. Это мощная, зрелая техника, которая по своей природе не подходит для целей Дружественного ИИ, который требует рекурсивных циклов самоулучшения, абсолютно точно сохраняющих цель оптимизации. [6]

В научных кругах существует мнение, что человечество пока недостаточно развито для общения с Дружественным ИИ, а значит, любые попытки создать его заранее

обречены на провал и даже могут привести к созданию лженауки. С другой стороны, необходимые знания существуют, хотя и не объединены в единую науку, а разрозненны по множеству разных: теории решений и эволюционной психологи, теории вероятностей и эволюционной биологии, когнитивной психологии и теории информации и в области знаний, традиционно известной как «Искусственный интеллект». [7]

Таким образом, создание Дружественного ИИ не будет возможно без консолидации имеющихся знаний из различных областей наук в единую науку, а также проведения целенаправленных исследований, четкого представления о конечной цели разработок и заложенной в механизм работы синергии человека и машинного разума.

Список использованных источников и литературы

1. Белозор Р.Ю. Место виртуального в жизнедеятельности человека// Студенческая конференция ТТИ ЮФУ. - 2007
2. Нариньяни А.С. Очень искусственный интеллект. - http://www.ng.ru/science/2006-02-22/14_intellect.html
3. Anissimov M. The Relationship Between AI and MNT. - http://www.acceleratingfuture.com/michael/blog/2007/02/the-relationship-between-ai-and-mnt/
4. Hay N. Anthropomorphic AI.- http://www.singinst.org/blog/2007/10/05/anthropomorphic-ai/
5. Molloy C. Marking Territories. - http://limen.mi2.hr/limen1-2001/clair_molloy.html
6. Vehkavaara T. Natural self-interest, interactive representation, and the emergence of objects. - http://www.uta.fi/~attove/gath2_end.pdf.
7. Yudkowsky E. AI and global risk. - http://yudkowsky.net/singularity/ai-risk