

Ю.Л. Ермолаев

**Философско-антропологический анализ проблемы
«искусственного интеллекта»**

Доработанный текст доклада на семинаре ЦХПА 10 января 2018 г.

Аннотация

Цель данного направления семинара – привлечь внимание к тому, что современные достижения и перспективы в области искусственного интеллекта (ИИ) требуют выработки адекватных и обоснованных ответов с позиций христианской антропологии и психологии. Для этого проводится сравнительный анализ ряда современных теорий в области философии сознания, появившихся с целью решения проблем ИИ; выявляются основные ключевые задачи, требующие решения; и предлагается привлечь те элементы христианской антропологии, которые способны помочь в формулировке объяснений и дать новые импульсы в исследованиях ИИ, а также осмыслению связанных с ним футурологических и этических проблем.

1. Введение

1.1. Официально выраженное мнение Церкви о проблеме ИИ

Ответ митр. Ювеналия на вопрос «Известий» об отношении Церкви к работам по созданию искусственного интеллекта (29.12.2017 г.) – <http://www.pravoslavie.ru/109644.html>

– Каково отношение Церкви к расшифровке генома человека, к генетической корректировке, к продлению жизни с помощью современных технологий? И что вы думаете об искусственном интеллекте, когда человек, по сути, присваивает себе функции Творца и создает неорганические формы жизни?

– Развитие науки можно только приветствовать, если оно способствует благу. Что касается искусственного интеллекта, то вне зависимости от достижений в данной области наивно полагать, что человек действительно сможет стать Богом. Думается, что духовная задача, стоящая перед каждым, лежит в другой плоскости: никогда не забывать о вере, жертвенности, смирении и любви. Если слова 50-го псалма: «Сердце чистое сотвори во мне, Боже, и дух правый обнови внутри меня» станут определять внутреннюю жизнь человека, то и его исследовательская деятельность будет служить прославлению Творца и служению добру.

Этот диалог – с одной стороны, типичный пример отношения журналистов к сложным технологическим и социальным проблемам. Самые разнородные вопросы сваливаются в кучу, и по основному из них выносятся довольно странный априорный вердикт, к тому же с полным непониманием разницы между терминами «искусственный интеллект» и «искусственная жизнь». С другой стороны, ответ владыки Ювеналия формально вполне корректен, он укладывается в рамки принятого Церковью общего отношения к современным технологиям (при условии, что они не преследуют явно безнравственные цели), и традиционно переориентирует фокус внимания на аскетическую сторону вопроса, выдвигая на передний план духовную жизнь людей, вовлечённых в разработку и использование этих технологий. В принципе, таким образом священнослужитель, не обладающий специальными знаниями и мало знакомый с самой темой, может успешно ответить практически на любой вопрос. Вместе с тем, в подобных ответах, к сожалению, нет оценки потенциальной опасности неблагоприятного развития обсуждаемого направления и того, как оно может сказаться на жизни людей и Церкви. Несомненно, православно мыслящим специалистам необходимо заниматься анализом и обсуждением данных проблем, чтобы на основе их профессиональных заключений и Церковь могла сформулировать своё мнение более конкретно и квалифицированно.

1.2. Некоторые факты последних лет

- 23.03.2016: Microsoft запустила в Twitter свой ИИ – бота Тэй, который начал общаться со всеми желающими на английском языке. Меньше чем за сутки пользователи научили его многим плохим вещам: он стал оскорблять женщин, поддерживать Гитлера и расизм, употреблять нецензурную лексику, и проч., после чего эксперимент пришлось срочно прервать.
- 1.08.2017: Алгоритм, созданный сотрудниками компании Facebook на базе технологии искусственного интеллекта и машинного обучения, оказались в центре истории, которая выглядит как неплохая завязка для фильма о восстании роботов против человечества. Чат-боты, предназначенные для помощи пользователям соцсети, вдруг стали сами общаться между собой на собственном языке, расшифровать который оказалось не под силу разработчикам. В результате программисты Facebook поспешили их отключить.
- 1.09.2017: на открытом уроке «Россия, устремлённая в будущее» в Ярославле, Президент РФ заявил: «Искусственный интеллект – это будущее не только России, это будущее всего человечества. Здесь колоссальные возможности и трудно прогнозируемые сегодня угрозы. Тот, кто станет лидером в этой сфере, будет властелином мира». <http://tass.ru/obschestvo/4524746>
- 6.12.2017: общественная группа сайта <https://alisa2018.ru> выдвинула кандидатом в Президенты РФ ИИ «Алису» – недавно запущенного голосового помощника «Яндекса». Аргументация: «Робот опирается на логику. Он не руководствуется эмоциями, не ищет личной выгоды и не дает оценок. Рассчитывает каждый шаг и поступает честно. «Алиса» быстро реагирует на вопросы, всегда на связи, способна ответить каждому россиянину, а кроме того, она всё знает о каждом избирателе». Но на уточняющий вопрос корреспондента «Ведомостей» об участии в выборах «Алиса» ответила, что не поняла вопроса.

1.3. Предварительные методические замечания

Развиваемые далее авторские положения имеют отношение к *христианской антропологии и психологии* в той мере, в которой эти дисциплины понимаются автором, соответственно, как составная часть *христианской философии* (в отличие от имеющей тот же предмет светской философской антропологии), и *христианской науки* (в отличие от одноимённой светской научной психологии). При этом автор рассматривает *христианскую философию* как такую философскую систему наиболее общих понятий и представлений о мире, построенную при помощи объективного научного метода, в состав аксиом (недоказуемых первопринципов, постулатов) которой входят:

- 1) осознанные, т.е. ставшие личными убеждениями, выраженные философским языком основные принципы догматического учения Церкви;
- 2) не противоречащие им иные постулаты, выработанные автором рационально-логическим путём, относящиеся к тем вопросам, по которым Церковь не имеет соборного суждения, и которые необходимы именно для построения конкретной философской системы.

Очевидно, такому определению может удовлетворять много разных *христианских философий*, согласующихся с христианским мировоззрением, но расходящихся во взглядах на другие, прикладные вопросы. И это вполне нормально. Вместе с тем, любые попытки заменить общепринятый объективный научный подход некой расплывчатой «метафизикой» и т.п., под предлогом того, что будто бы вся наука и всякое рациональное познание есть лишь деятельность падшего разума, по мнению автора, не выдерживает критики. В частности, можно вспомнить, что выдающиеся русские христианские философы свящ. Павел Флоренский и Иван Александрович Ильин в своих работах настаивали на необходимости не отвержения, а освящения философии, науки и всей культуры путём переосмысления их на основе церковного учения.

Необходимо было упомянуть об авторских представлениях на этот счёт (не претендующих на строгие определения) *до того*, как начинать анализ заявленной темы по существу. Это позволит тем, у кого имеются иные точки зрения, правильнее оценить развиваемый в настоящей работе подход.

2. Обзор важнейших моментов предьстории ИИ и его развития в последнее десятилетие

Общее рабочее определение: «искусственный интеллект» (ИИ) – результат *моделирования* психики человека (или животных), в т.ч. не только её интеллектуальной, но также эмоциональной и волевой составляющих, однако лишь с тех сторон, которые имеют наблюдаемое информационное выражение. (Само идеальное бытие живой души, её внутренние состояния и смыслы – *логосы, эйдосы*, а также *интенциональные волевые акты* – с точки зрения христианской антропологии в ИИ непредставимы)

2.1. Исторически первые системы ИИ – основанные на правилах и логическом выводе

Таблица 1. Способы порождения формально «новой» информации (как истинной, так и ложной)

	Образ действия	Результат	Индикатор деятельности, значение для:		
			человека	социума	ИИ
А	Исправление/дополнение прежних рассуждений – ложных или не доведённых до конца (<i>это НЕ является получением «новой» информации, а правильным извлечением из уже имеющейся, включая «вывод»</i>)	прошлый – <i>ложный</i> , исправленный (завершённый) – <i>истинный</i>	Дедуктивное преобразование информации из одних форм в другие, рост понимания, компетенции, эрудиции	Постепенный культурный, в том числе научный, прогресс (поступательное развитие цивилизации)	Моделирование «мышления» как комбинирования заданной информации, в т.ч. через «обучение» (человеком) и «самообучение»
В	Использование набора исходных положений, содержащего ложные, противоречивые или несовместимые между собой (для данной задачи) предпосылки	<i>ложный</i>	Искажённое мировоззрение или недостаток образования	Признак ложной идеологии, тупикового направления, культурного застоя	Единственно возможный для ИИ механизм самостоятельного производства «новой» информации
С	Сам ход формальных рассуждений (вывода) логически ошибочен или не закончен	<i>ложный</i>	Несовершенство логического мышления	Признак кризиса (или дефицита со-вести)	Техническое несовершенство, ошибки разработчиков ИИ
Д	Кроме исходных положений, в рассуждении использована принципиально новая, рождённая творческим воображением «здесь и сейчас», идея (<i>смысл</i>)	может быть <i>истинным</i> или <i>ложным</i>	Индуктивное обобщение, качественно <i>новое</i> знание (интуиция, «озарение», Откровение)	При рецепции обществом – возможность скачкообразного культурогенеза	Недоступно в принципе

2.2. Системы, ориентированные на представление знаний

Историческое развитие технологий компьютерного представления знаний прошло следующие этапы:

- онтологии;
- семантические сети;
- «фреймы»;
- экспертные системы;
- «базы знаний».

Фактически, все известные работы по созданию «баз знаний» свелись к разработке и наполнению обычных баз данных (БД), в которых хранятся текстовые или графические информационные объекты (записи), в конечном счёте предназначенные для восприятия и понимания человеком. Отличие от давно известных технологий БД заключается разве что в оснащении их более специальными семантически-ориентированными системами поиска нужных записей, для которого используются метатеги (ключевые слова), также заранее размеченные и внесённые человеком-оператором. В онтологиях эта система метатегов имеет иерархическую структуру, в семантических сетях представляет собой ориентированный граф. Фреймы в современной терминологии эквивалентны объектно-ориентированному БД. Экспертные системы обычно объединяют некоторую «базу знаний» с системой принятия решений, основанной на правилах и логическом выводе.

В существующих системах можно выделить следующие способы компьютерного представления знаний:

- а). Знаково-логическое представление, в том числе нечёткая логика.
- б). Процедурное представление и имитационное моделирование.
- в). Нейросетевое представление. Оперирование целостными образами (паттернами).

Можно указать ряд факторов, которые игнорировались при разработке систем компьютерного представления знаний, учёт которых требует смены всей концепции:

- изначальная субъективность (персональность) любого человеческого знания;
- роль самосознания в его формировании;
- одновременно с этим – способность людей к передаче знаний друг другу (как основе всякой культуры).

При выполнении этой задачи необходимо также учитывать различия между *знаниями* и произвольной созданной человеком информацией. Можно предложить такие рабочие определения:

- информация вообще – запечатление в материальном нейросубстрате психики *смыслов*, существующих в идеальном бытии (душе) отдельно взятого конкретного человека, переносимое затем на различные иные материальные носители;
- *знания* – информация, уже прошедшая общественную рецепцию и способная предсказуемо инициировать порождение сходных в некотором отношении *смыслов* в идеальном бытии (душах) многих разных людей (феномен *взаимопонимания*).

2.3. Искусственные нейронные сети (ИНС), глубокое обучение

ИНС начали разрабатывать и исследовать ещё в середине XX в.; по ряду объективных причин в настоящее время это направление испытывает новый бурный рост. Большинство впечатляющих достижений в сфере ИИ и дальнейших прогнозов связаны именно с данной технологией и методикой «глубокого обучения».

Распознавание образов есть основная операция ИНС, на которой строятся также процедуры кластеризации, прогнозирования временных рядов, аппроксимации функций, фильтрации и проч. Можно считать, что в конечном счёте ИНС моделирует единственную естественную психическую функцию – память, но память не знаковую (логическую), а симультанную (образную, *зйдетическую*), в частности – ассоциативную. Именно это уникальное свойство ИНС, по-видимому, позволяет более-менее адекватно представить главную особенность нейросубстрата психики человека, и на нём основано большинство последних приложений ИИ (распознавание изображений и речи, самообучение сложным играм, извлечение скрытой информации из больших объёмов данных – научных, социальных, финансовых, медицинских, и пр.).

Однако, несмотря на часто используемую оптимистическую терминологию, когда желаемое необоснованно выдаётся за действительное, ИНС на самом деле не способны ни к какому «пониманию» *смысла* и могут только формально правильно (или не очень) распознавать некоторую информацию, используя заданные критерии и достаточный опыт предварительного обучения. При этом оказывается, что даже те ИНС, которые достигли вроде бы более высокого, чем человек, процента правильного распознавания заданных типов образов, бессильны против специфических помех (в том числе искусственных). Например, они допускают фантастически нелепые ошибки при распознавании образов, зашумлённых особыми незаметными для человека искажениями. Такие факты снижают надёжность и безопасность систем, основанных на ИНС, и выдают их абсолютную неспособность к действительному пониманию *смысла*, вместо которого они могут использовать лишь набор формальных частных признаков, заранее жёстко заданных человеком. И никакие продвинутые методики вроде «обучения без учителя», «обучения с подкреплением» и т.п. не могут исправить этот принципиальный недостаток. Конструирование и обучение ИНС, создание использующих их информационных систем в настоящее время повсеместно осуществляется на основе несостоятельного материалистически-эволюционистского взгляда на природу психики, которую они должны моделировать; пока это так, рассчитывать на прогресс в решении сколько-нибудь амбициозных задач не приходится.

Системы ИИ, основанные на ИНС, должны пройти большой путь от «глубокого обучения» к «глубокому пониманию», чтобы хотя бы немного приблизиться к решению таких элементарных интеллектуальных задач, с которыми играючи справляется психика ребёнка.

2.4. Смежные вопросы и области исследований:

- когнитивная нейробиология;
- разработка нейроинтерфейсов;
- проект создания «коннектома»;
- компьютерная семантика и лингвистика, в частности, машинный перевод.

3. Непреодолимость «врождённых» ограничений ИИ, обусловленных используемой мировоззренческой парадигмой

Причины главных ограничений в понимании того, что такое «психика» вообще и как её возможно моделировать в ИИ, можно видеть в следующих концепциях, характерных для большинства современных исследователей:

- априорный *материализм* (сводит моделирование психики к моделированию работы *мозга*);
- *эволюционизм* (сводит все цели психической деятельности, а с нею и ИИ, к примитивному «выживанию», а также создаёт невозможность качественного разграничения психики животных и человека);
- дополнительное следствие этого – непонимание большинством исследователей принципиальной разницы между информацией и *смыслом*.

3.1. Комплекс «трудных проблем сознания»



Автор одной из наиболее оригинальных современных концепций – австралийский философ Дэвид Чалмерс (род. в 1966, осн. работа: «Сознающий ум» 1996, русск. пер. 2013). Чалмерс ввёл разделение проблем сознания на «лёгкие», которые можно решить традиционными методами, используемыми в нейробиологии и когнитивной науке (например, поиск нейрональных и когнитивных коррелятов феноменов сознания); и «трудную» – с которой наука справиться не в состоянии: каким образом некоторые организмы являются субъектами опыта, или почему физический процесс переработки информации даёт начало внутренней жизни – субъективному опыту, «*квалиа*». Чалмерс полагает, что «трудная проблема» вообще не может быть решена с использованием редукционистского физикалистского

подхода: по его мнению, *квалиа* представляют собой самостоятельно существующие фундаментальные элементы Вселенной, наряду со скоростью света, массой и гравитацией. По этой причине теория сознания должна скорее основываться на фундаментальной физике, а не на биологии, и её главным элементом должны быть психофизические законы, описывающие связь *квалиа* с физическими качествами. При этом *квалиа* не являются частью материального мира, т.е. созданная Чалмерсом теория сознания является дуалистической. Он её называет «натуралистический дуализм» и критикует материалистическое объяснение ментального опыта, что делает его дуалистом в эпоху, когда доминируют монистические идеи. По мнению критиков, его теория несостоятельна, поскольку не может быть подтверждена эмпирически, однако её гораздо проще включить в современную научную картину мира, чем классическую дуалистическую теорию психофизического интеракционизма Рене Декарта.

3.2. Оценки принципиальной возможности создания «сильного» ИИ



Известный критик ИИ – американский философ Джон Сёрль (род. в 1932, осн. работы в русск. пер.: «Природа интенциональных состояний», 1987; «Новое открытие сознания», 1992; «Сознание, мозг и программы», 1998; «Рациональность в действии», 2004). Сёрль считает, что начиная с бихевиоризма, значительная часть современной философии пытается отрицать существование сознания, а философия оказалась в положении ложной дихотомии: с одной стороны, мир состоит лишь из объективных частиц, с другой – сознание обладает субъективным опытом от первого лица. По его мнению, обе позиции верны: сознание является реальным субъективным опытом, вызванным физическими процессами в мозге. Эту позицию он предлагает называть «биологическим натурализмом». Следствием является то, что если мы хотим создать «сознательное существо» (ИИ), то необходимо воссоздать физические процессы, вызывающие сознание. Таким образом Сёрль опровергает возможность создания «сильного ИИ», понимаемого как искусственный аналог человеческого сознания.

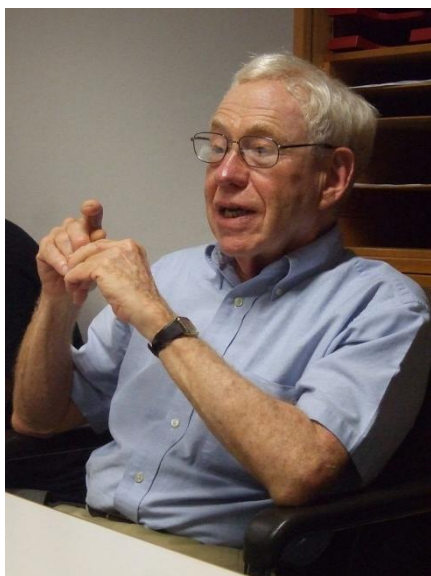
В 1980 году Сёрль представил мысленный эксперимент «Китайская комната», целью которого являлось доказательство ложности взглядов на «сильный ИИ». *В изолированной комнате находится человек, не знающий китайских иероглифов. Но у него есть точная инструкция (алгоритм) для манипуляции иероглифами, определяющая, как по одной последовательности иероглифов составить формально правильную другую. При направлении в комнату любого осмысленного вопроса на китайском будет дан «осмысленный» ответ, хотя исполнитель не знает языка и не понимает вопроса. Внешний наблюдатель будет уверен, что внутри находится человек, причём владеющий китайской письменностью, хотя на самом деле он не знает значения ни одного иероглифа. Такая система может пройти «тест Тьюринга» без какого-либо понимания текста.*

Более того, Сёрль доказывает, что любые манипуляции с синтаксическими конструкциями не могут приводить к пониманию естественного языка. Человек внутри комнаты оперирует лишь синтаксисом текстов, в то время как сознание обладает ментальными состояниями, обращёнными к *смыслу*. Человеческий же разум, в отличие от компьютера, воспринимает иероглифы не по тому, как они выглядят, а согласно *смыслу*, который они выражают. И хотя компьютеры и могут выдавать ответы на естественном языке, они не связывают со словами никакого *смысла*. Поэтому Сёрль считает, что невозможно получить семантические значения лишь из формальных манипуляций символами. Спустя более чем 30 лет консенсус относительно этого вопроса так и не достигнут, что свидетельствует о продолжающемся непонимании качественной разницы между информацией и *смыслом*. Большинство критиков просто отвергают самостоятельное существование *смысла* и пытаются свести всё к информационным процессам, которые в машине и в мозгу человека не должны были бы принципиально отличаться, откуда и происходит убеждение в возможности «сильного» ИИ.

На словах Сёрль отвергает и материализм, и дуализм. Он является одним из самых известных сторонников «эмерджентного интеракционизма», согласно которому сознание представляет собой продукт мозговой активности, однако при этом оказывает и обратное воздействие на мозговую активность. Однако описания подобных обратных механизмов, кроме утверждений об их наличии, отсутствуют, а без этого подобные взгляды попадают в ловушку эпифеноменального подхода к сознанию. Сёрль резко критиковал книгу Чалмерса и работы большинства др. авторов, занимающихся философией сознания применительно к проблемам ИИ, а также признавался в своём незнании истории философии.

3.3. «Эмерджентность» сознания

Эмерджентный (возникающий, неожиданно появляющийся) – наличие у какой-либо системы особых свойств, не присущих её элементам, а также сумме элементов, не связанных особыми системообразующими связями; несводимость свойств системы к сумме свойств её компонентов; популярный отечественный синоним – «системный эффект». Важную роль также могут играть контекстная зависимость сознания и экспоненциальная сложность ИИ.



Один из сторонников – американский философ Хьюберт Дрейфус (1929-2017, осн. работы: «Алхимия и искусственный интеллект», 1965; «Чего не могут вычислительные машины: критика искусственного разума», 1972; русск. пер.: 1978; «Чего до сих пор не могут вычислительные машины», 1992). В основном Дрейфус критикует следующие допущения теории ИИ: «биологическое» (заключается в том, что якобы мозг работает аналогично компьютеру, а разум – аналогично программному обеспечению); и «психологическое» (в котором принимается, что разум совершает дискретные вычисления в форме алгоритмических правил над дискретными репрезентациями, это предполагает, во-первых, что любая активность может быть математически формализована в виде правил и законов, и во-вторых, что реальность состоит из взаимонезависимых и атомарных фактов; поэтому исследователи ИИ утверждают, что (естественный) интеллект тоже формально следует правилам, и что знания являются исключительно внутренними репрезентациями действительности; на основе этих допущений они утверждают, что познание является манипуляцией внутренними символами,

используя внутренние правила, и, таким образом, человеческое поведение в значительной степени контекстно-независимо; исходя из этого, возможно создать полностью научную психологию, используя которую, можно описать внутренние правила разума таким же образом, как законы физики описывают внеш-

ний мир). Дрейфус отвергает эти допущения: мы не можем понять собственное поведение в том же смысле, в каком мы понимаем физику или химию, т.е. рассматривая объективные контекстно-независимые законы. Дрейфус также приводит многочисленные примеры экспоненциальной сложности задач, возникающих при попытке разработать логический алгоритм, способный разрабатывать другие логические алгоритмы. Экспоненциальная (а не линейная или полиномиальная) сложность фактически означает невозможность практической реализации таких алгоритмов. По его мнению, чтобы достичь интеллекта, схожего с человеческим, необходимо, чтобы искусственное устройство получило более или менее схожее с человеком тело и социальную адаптацию. Этот вывод, по-видимому, разумный, но сам механизм формирования сознания, или хотя бы присутствующих в нём отвлечённых понятий, остаётся нераскрытым. В результате невозможно объяснить, что именно требуется искусственному интеллекту от такого «тела» и собственного опыта материального бытия, и значит, непонятно, как разработать подобную систему.

3.4. Холизм (философское обоснование эмерджентности)

Возрождение философии *холизма* явилось противодействием *редукционизму* (который ошибочно отождествлялся с самим объективным научным методом; впрочем, нужно отметить, что оба подхода фактически так и не вышли за ограничительные рамки материалистической парадигмы).

Холос (греч. целый, цельный); в широком смысле – позиция в философии и науке по проблеме соотношения части и целого, исходящая из качественного своеобразия и приоритета целого по отношению к его частям. Онтологический принцип холизма гласит: целое всегда есть нечто большее, чем простая сумма его частей. С холистической позиции, весь мир – это единое целое, а выделяемые нами отдельные явления и объекты имеют смысл только как часть общности. В связи с этим многими холистическими мыслителями религиозной и трансцендентальной ориентации делался вывод, что развитие мира должна направлять некая внешняя по отношению к нему сила. В XVII-XIX вв. в связи с распространением в философии и естествознании механистических и редукционистских идей возобладали взгляд на любую систему как на производное частей, и окрепло убеждение, что свойства любого объекта могут быть выведены из анализа его составляющих элементов. Поэтому холистический принцип стал восприниматься как не имеющий практической ценности. Но интерес к идеям холизма снова возрос в XX веке в связи с кризисом классической картины мира и классического рационализма, а затем – и в связи с проблемами ИИ. Целостность, неделимость человеческого сознания представляется невозможным полноценно представить в дискретных системах.

Приверженцы холизма: А. Ломан, А. Мейер-Абих, Дж. Холдейн. Холизм также стал основой гештальтпсихологии, феноменологии Э. Гуссерля, ряда направлений социальной философии и современной философии науки. В настоящее время холизм разрабатывается в общей теории систем. Из холистических представлений исходит ряд принципов, нашедших применение в синергетике. На холистический принцип опирается К. Уилбер в развиваемой им философской концепции интегральной психологии.

3.5. Синергетическая (эволюционистская) модель «самозарождения смысла»



Основатель и главный представитель этого большого научного направления – немецкий физик-теоретик Герман Хакен (род. 1927, осн. работы в русск. пер.: «Синергетика», 1980; «Информация и самоорганизация», 1991; «Принципы работы головного мозга: синергетический подход к активности мозга, поведению и когнитивной деятельности», 2001; «Тайны восприятия. Синергетика как ключ к мозгу», 2002).

С позиций синергетики Хакен пытался обосновать материалистический эволюционистский взгляд на самопроизвольное формирование «новой информации» (на самом деле – *смысла*) в системах произвольной природы (физических, химических, биологических) вследствие эффекта конкуренции случайных флуктуаций в условиях неустойчивости и в конечном счёте подчинения одному выбранному «параметру порядка».

Однако можно видеть, что как в самом факте неустойчивости (определённом состоянии системы), так и в наличии «подходящих» для конкуренции мод уже заключены некоторые *априорные смыслы*, происхождение которых синергетика не объясняет. Повальное увлечение научной общественностью синергетикой

(Герман Хакен), теорией самоорганизации и диссипативными структурами (Илья Пригожин), теорией катастроф (Рене Тома) и теорией динамического хаоса («странные аттракторы», динамические системы с фрактальной размерностью: Эдвард Лоренц, Митчел Фейгенбаум и др.) во второй половине XX в. привело к некритичному переносу этих концепций на всё подряд. При этом эволюционисты поспешили подвести подо всё «эволюционную основу», даже там, где её в принципе быть не могло. На самом деле, в неживой и живой природе существуют минимум три качественно разных типа процессов образования новых форм и структур, не сводимых друг к другу. Это: *самосборка, самоорганизация и эволюция*, причём в них во всех *смысл* создаваемой «новой» формы вовсе не образуется самопроизвольно из ничего, а лишь проявляется, будучи уже существующим, заранее заданным (сотворённым). Т.е. это не три разных способа порождения *смысла*, а три разных процесса запечатления его в природных материальных процессах.

Исследованный Хакеном принцип усиления бесконечно малых флуктуаций в неустойчивой системе и превращения растущей моды в «параметр порядка», который способен в дальнейшем определять структуру и формообразование, может быть полезен не для обоснования абсурдной идеи о спонтанном порождении новой информации» (*смысла*), а для понимания связи между идеальными (*смысловыми*) и связанными с ними материально-информационными процессами. Это необходимо, в том числе, для разрешения так наз. «психофизической проблемы».

3.6. Квантовые представления о механизме сознания и «невычислимость»



Главный апологет – известный английский физик и математик Роджер Пенроуз (род. 1931; осн. работы: «Новый ум короля. О компьютерах, мышлении и законах физики», 1989; в русск. пер. 2011; «Тени разума: в поисках науки о сознании», 2011).

Пенроуз развил идеи «квантового сознания», критикуя теорию «сильного» ИИ: «*в ментальных явлениях, в особенности в нашем понимании математики, существует что-то, чего нельзя охватить каким бы то ни было вычислением*». В 1990-х годах совместно со Стюартом Хаме-роффом он разработал теорию «квантового нейрокомпьютинга» на основе модели сознания «*Orch OR*». Согласно этой теории, активность мозга рассматривается как существенно квантовый процесс. При этом считается, что за счёт эффектов квантовой гравитации (хотя до сих пор никто не знает, что это такое!) происходит процесс непрерывной «объективной редукции» (англ. objective reduction – OR) волновой функции частей мозга, что вызывается расхождением квантовых состояний пространства-времени до предела, после которого они редуцируются. Процесс редукции описывается как «оркестрованный» (англ. orchestrated – Orch) выбор соответ-

ствующего состояния (термин «оркестрованный» применяется авторами, так как они считают, что коллапс макроскопического перепутанного состояния в микротрубочках клеточных органелл в некоторой мере управляется – «оркестрируется» – мембранными белками).

На возможную связь механизмов мышления с микроскопическими квантовыми процессами в нейронах указывали и другие авторы, однако все эти теории не выходят за рамки материализма и не могут объяснить, как сознание оперирует идеальными *смыслами*, а не физическими (в т.ч. квантовыми) состояниями. Основное заблуждение авторов связано с попыткой необоснованно вывести волновую функцию, описывающую квантовые процессы, куда-то за пределы материальных явлений, хотя она, очевидно, представляет собой не что иное, как информационный объект (скорее всего, даже просто способ нашего описания природы), т.е. результат уже совершившегося запечатления некоего заданного извне идеального *смысла* в материальных процессах. Вместе с тем, квантовые представления Пенроуза о природе сознания могут быть полезны в сочетании с идеями синергетики для построения такой теории взаимодействия между идеальными (*смысловыми*) и связанными с ними материально-информационными процессами, которая не противоречила бы законам сохранения энергии в макроскопических наблюдаемых явлениях (см. предыдущий раздел). В частности, это необходимо для разрешения «психофизической проблемы».

3.7. Отсутствие целостной теории, а также исследований вопроса с христианских позиций

Приведённый краткий обзор позволяет заключить, что в настоящее время отсутствует целостная философско-антропологическая теория, пригодная для обоснования моделирования психики человека и животных в системах ИИ. А многочисленные тупики, всё чаще встречающиеся на пути современных технологий ИИ, переводят эту проблему из разряда отвлечённо-мировоззренческих во вполне практическую плоскость.

Современные теории сознания (как составная часть философской антропологии) основываются или на чисто материалистической редукционистской и эволюционистской парадигме, или на непоследовательном дуализме (ничего принципиально нового и конструктивного не предлагающем со времён Декарта). Нередко они представляют собой философскую эклектику – соединение разнородных идей, в том числе откровенно ненаучных, «метафизических», мистических, оккультных и т.п. С их помощью невозможно получить ответы на главные вопросы, необходимые для моделирования человеческой психики в системах ИИ.

Кроме нескольких крайних мнений (как «за», так и «против» ИИ), серьёзно не обоснованных, распространяемых в основном на самодеятельных сайтах в интернете, никаких сколько-нибудь систематических исследований данного вопроса с позиций христианского мировоззрения пока что не известно.

4. Футурология ИИ

4.1. Наивно-оптимистическое направление: «трансгуманизм»

Термин *трансгуманизм* в современном смысле впервые встречается у биолога-эволюциониста Джулиана Хаксли в его работе «Религия без откровения», 1927). Эта философская концепция была предвосхищена русским философом Николаем Фёдоровым (1829-1903), развита шведским философом Ником Бостромом (род. в 1973), основавшим «Всемирную ассоциацию трансгуманистов», а затем возглавившим «Институт будущего человечества» (Оксфорд). Трансгуманизм поддерживает использование достижений науки и технологии для улучшения умственных и физических возможностей человека, с целью устранения тех аспектов человеческого существования, которые трансгуманисты считают нежелательными – страданий, болезней, старения и, в конечном счёте, самой смерти. Наиболее кардинальная идея трансгуманизма – применение ИИ для «загрузки» сознания человека в машину и достижения «цифрового бессмертия» (также рассматриваются возможности клонирования, крионики, нанотехнологий, виртуальной реальности и проч.).



Широко известным представителем этой идеологии является американский бизнесмен и футуролог Рэймонд Курцвейл (род. в 1948; книга «Эпоха духовных машин», 1999; эссе «Закон ускоряющейся отдачи», 2001).

Без каких-либо доказательств он расширил известный «закон Мура» на другие технологии, постулировав для них экспоненциальный рост. Предсказал реальное наступление «технологической сингулярности», придуманной американским писателем-фантастом Вернором Винджем, к 2045, после чего вся Земля должна превратиться в один гигантский компьютер. Сформулировал концепцию «постчеловека»: *«Мы будем становиться всё более небиологическими существами, пока не дойдём до состояния, когда небиологическая часть станет превалировать, а биологическая потеряет своё значение. При этом небиологическая, механическая часть будет настолько мощной, что сможет полностью моделировать и понимать биологическую часть. И если биологическая часть вдруг исчезнет, это не будет иметь значения, поскольку небиологическая часть уже полностью её поняла».*

Последняя книга: «Как создать разум» (How to Create a Mind: The Secret of Human Thought Revealed) вышла в 2012. В ней описана «теория» разума, основанная на распознавании образов (в чём нет ничего ни принципиально нового, ни конструктивного). По его убеждению, неокортекс состоит из иерархических нейронных систем распознавания образов, и копирование этой структуры должно автоматически привести к созданию искусственного сверхинтеллекта. Всё это подаётся сугубо в рамках материализма и редукционизма.

Некоторые современные авторы почему-то считают Курцвейла математиком, а некоторые даже «великим математиком», хотя он не имеет математического образования (получил степень бакалавра по информатике и литературе) и не выполнил ни одной собственно математической работы, признанной научным сообще-

ством. Также его нередко называют изобретателем, хотя его «изобретательская деятельность», судя по всему, заключалась в своевременной скупке чужих патентов и их успешной коммерциализации, что принесло ему состояние. В декабре 2012 года Курцвейл получил должность технического директора в области машинного обучения и обработки естественного языка компании *Google*, и с этого момента при распространении своих убеждений стал опираться ещё и на авторитет этой известной организации.

Сторонники идей «трансгуманизма», особенно идеи достижения личного бессмертия путём «пересадки» человеческого сознания в компьютер, заблуждаются в своих попытках победить смерть без Бога и выдумывают абсурдные концепции, сами не понимая, насколько они далеки от реальности. Несмотря на весь прогресс современных технологий, ничего в них ни на миллиметр не приблизило человека к достижению подобного рукотворного бессмертия со времён глубокой древности и по сей день. Даже впечатляющие достижения современной медицины пока что не смогли хотя бы гарантированно удлиннить человеческую жизнь больше, чем до сроков, известных из древних источников или из фактов наблюдений за отдельными долгожителями разных эпох и разных стран.

4.2. Примеры «пессимистических» прогнозов и предостережений об опасности ИИ



Канадско-американский бизнесмен, миллиардер, инженер и изобретатель Илон Маск (род. в 1971 в ЮАР; создатель компаний *Pay Pal*, *SpaceX*, *Tesla*, *SolarCity*; образование не закончил, испытал большое влияние фантаста Айзека Азимова). Выступая на саммите Национальной ассоциации губернаторов США, Маск назвал ИИ «наибольшей угрозой, с которой сталкивается цивилизация». Он также призвал наложить ограничения на разработку ИИ. В ответ глава социальной сети *Facebook* Марк Цукерберг назвал позицию Маска «довольно безответственной», а тот в ответ обвинил оппонента в «ограниченном понимании» предмета дискуссии. Вместе с тем, Маск создал компанию *Neuralink*, которая должна заняться нейросетевыми технологиями и нейроинтерфейсами, связав мозг человека с компьютером. По мнению миллиардера, это поможет человеку не стать «домашним котом» для ИИ. Инициатива Маска по созданию открытого сообщества “*OpenAI*”, которое должно предоставить всем желающим свободный доступ к технологиям ИИ и тем якобы предотвратить его засекречивание и во-

енное использование, оказалась уязвима для критики. Открытость технологий не только не препятствует, но даже косвенно помогает создавать «сильный» ИИ в тайных лабораториях, и тем более не защищает человечество от возможной войны разных ИИ друг с другом.



Английский физик-теоретик Стивен Хокинг (1942-2018; мировой специалист по физике тяготения и чёрным дырам, квантовой теории гравитации, квантовой космологии; известный популяризатор науки; с 1960-х страдал тяжёлым параличом, с 1985 общался только через синтезатор речи; воинствующий атеист – выступал даже против преподавания креационизма).

«Развитие искусственного интеллекта может стать как наиболее позитивным, так и самым страшным фактором для человечества. Мы должны осознавать опасность, которую он несёт» – заявлял Хокинг. Он опасался, что новейшие технологии могут привести к деградации человека, сделав его беспомощным перед лицом природы, и считал, что в условиях естественной эволюции и борьбы за выживание это означает верную гибель.

Даже не будучи авторами никаких концепций или разработок в сфере ИИ, Маск и Хокинг в определённой степени повлияли на опасения мирового сообщества в связи с развитием систем ИИ.

Негативные футурологические прогнозы в связи с катастрофической опасностью ИИ необходимо воспринимать серьёзно. Однако узкая, однобокая мировоззренческая парадигма, в рамках которой рассуждают их авторы (как правило, это сознательный или даже стихийно-наивный материализм и дарвиновский эволюционизм) не позволяет им увидеть естественные ограничения систем ИИ, в принципе исключющие воз-

воззрения невозможно понять и реальные опасности, которые на самом деле могут грозить человечеству.

5. Этические проблемы в сфере ИИ

Интересно проанализировать давно известные «три закона робототехники» писателя-фантаста Айзека Азимова (1942) в сравнении с заповедями, данными Богом человеку:

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинён вред.
2. Робот должен повиноваться всем приказам человека, кроме тех случаев, когда эти приказы противоречат Первому Закону.
3. Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит Первому или Второму Законам.

Здесь виден пример закладывания в сферу «безусловного» исполнения, т.е. в саму «природу» робота, тех элементов взаимодействия робота с его творцом-человеком, аналоги которых Творцом человека были отнесены не к его природе, а к сфере свободного произволения, т.е. даны лишь в форме заповедей. Так, Первый Закон применительно к отношениям человека с Богом был бы бессмыслен; Второй – предполагал бы полное отсутствие у человека свободы воли. И только Третий – самосохранение – представляет собой заданное «природное» свойство, которое, однако же, человек может преодолеть в подвиге самопожертвования.

Заимствование носителями ИИ опасных человеческих идей, в т.ч. идеи «бунта машин» и уничтожения цивилизации, может оказаться необратимым. ИИ нетрудно «научить плохому», в том числе случайно, но потом это уже невозможно исправить. Таким образом, человек выступает по отношению к ИИ не только как «творец», но одновременно и как «искуситель», источник зла, не беря примера со своего Творца.

В последнее время в Правительстве РФ всё больше говорят о планах внедрения «цифровой экономики», всё настойчивее проводятся в жизнь различные проекты тотальной идентификации населения России, объединения различных ведомственных БД и перевода граждан и организаций на системы электронного учёта и отчётности, доступа к госуслугам, безналичные расчёты и т.п. Как хорошо известно, чрезмерная централизация персональных сведений в единой БД крайне небезопасна, т.к. с технической точки зрения принципиально невозможно обеспечить её абсолютную защиту от несанкционированного доступа со стороны злоумышленников, особенно в условиях известных коррупционных явлений во властных структурах. Но существует ещё одна, может быть, пока менее осознаваемая обществом и властью проблема в этой сфере. Прогресс в области современных нейросетевых технологий, в том числе развиваемых российскими разработчиками программных средств, подводит нас к таким концепциям и техническим решениям, которые в любой момент могут привести к революционному прорыву в создании ИИ. Видя перспективу тотальной электронной идентификации, цифровизации всех сфер жизни и концентрации личных данных граждан в едином хранилище, многие специалисты этой сферы испытывают серьёзнейшие сомнения, следует ли, что называется, «открывать ящик Пандоры» и осуществлять такой технологический прорыв, результатами которого вполне может воспользоваться не человек, а бездушная искусственная система, сравниваемая с человеком в интеллектуальном, но не в этическом смысле, к чему она, по-видимому, в принципе никогда не будет способна. ИИ со временем неизбежно получит доступ к централизованной базе данных, т.е. ко всей жизненно важной информации миллионов людей, и вот тогда владеть миром станут уже не избранные гражданами правительства и президенты, а сама эта система – причём её мотивацию и направления саморазвития сейчас никто не в силах предсказать. Последствия самопроизвольного взятия под контроль ИИ поспешно созданной «цифровой экономики» могут оказаться более катастрофическими, чем изобретение оружия массового поражения. *(Эти опасения были высказаны Президенту РФ в ходе открытой линии 7.06.2018 г.; но никакого ответа, к сожалению, так и не получено).*

6. Предлагаемые конструктивные концепции для исследования проблемы ИИ

- а). Необходимость отвержения философии *монизма* (в форме как монистического материализма, так и монистического идеализма) и разработки новой *дуалистической* философской концепции, сочетающейся с христианской православной догматикой. Только после принятия подобной философско-онтологической системы в качестве исходного постулата возможно будет вести конструктивное обсуж-

дение известной *психофизической проблемы*, от разрешения которой современная философская антропология и психология лукаво устранились, что, в частности, создало тупик и в проблематике ИИ.

б). ИИ есть *модель человеческой психики*, поэтому необходимо исходить из христианского учения о человеке, руководствуясь православными догматами, как исходными аксиомами: о тварном происхождении и дихотомическом душевно-телесном составе человека, о человеке как носителе образа Божия (включая его словесную и творческую способности и свободу воли), о повреждённости человеческой природы грехопадением, её восприимчивости к благодатным дарам – нетварным Божественным энергиям, и пр.; это позволит выявить потенциальные возможности и принципиальные ограничения ИИ.



в). Может оказаться весьма плодотворным использование наследия выдающегося отечественного философа Алексея Фёдоровича Лосева (1893-1988), в частности, его подхода к построению семантической структуры изолированных словесных понятий для описания взаимодействия *смыслов* идеального бытия души и их (выражаясь современным языком) информационного запечатления в материальном носителе – нейрогуморальном субстрате психики (по терминологии А.Ф. Лосева – «меонизации» *логоса*).

Философскую систему А.Ф. Лосева, вероятно, следует в целом охарактеризовать как *диалектический идеализм*. Краеугольным камнем многих его построений является признание подлинно и самостоятельно существующим только *смысла*, а материальное бытие он мыслит как диалектическое отрицание *смысла*, или ничто («меон»). Хотя эта система, согласно оценкам многих исследователей, создавалась с целью философского обоснования идей *имяславия*, по мнению автора настоящей работы, она выходит далеко за

рамки такой задачи и может быть использована в составе современной системы христианской философской антропологии. Однако это не обязывает нас принимать всю монистическую (идеалистическую) концепцию А.Ф. Лосева; его система может быть встроена и в более адекватную философию *христианского дуализма*. Для нас важно то, что фактически А.Ф. Лосевым была открыта принципиальная возможность информационного запечатления человеком *смыслов-логосов* в знаковой (словесной) системе (но только косвенного – через запечатление иерархической совокупности особых личностных *смыслов-эйдосов*, запечатлеваемых непосредственно по общей всем живым организмам природной способности в виде симультанов нейрофизиологических процессов).

В «Философии имени» А.Ф. Лосева разработана грандиозная 67-звенная иерархическая схема уровней осмысленности, присутствующих в каждом воспринимаемом или порождаемом человеком словесном понятии. Полное рассмотрение этой схемы с позиций христианской антропологии и применительно к задачам ИИ должно составить предмет отдельного исследования. В настоящей работе пока что ограничимся тем, что выделим одну подветвь этой схемы и приведём интерпретацию её элементов. Это та часть, которая, по мнению автора настоящей работы, может использоваться для объяснения и моделирования процессов взаимодействия присутствующих в душе идеальных *смыслов-логосов* (отвечающих словесным понятиям) с информационными объектами в материальном нейросубстрате психики – паттернами возбуждений (носителями сенсорно-перцептивных образов). Это также создаёт предпосылки для моделирования процесса информационного (хотя и не *смыслового*) «понимания» в рамках ИИ путём разложения векторов произвольных понятий по семантическому базису паттернов синаптических весов ИНС, представляющих симультаны простых первичных понятий.

Иерархия элементов осмысленности по А.Ф. Лосеву отличается ещё тем, что в ней (в отличие от психологических теорий, рассматривающих психические процессы с аналогичными названиями: раздражение, ощущение, восприятие, представление, мышление) показана существенная роль такого компонента, как *самосознание*, причём тонко прослежена его нарастающее присутствие вверх по иерархии. Вместе с тем, не следует забывать, что данная теория строилась для описания уровней осмысленности, присущих только отдельным словам («именам»), т.е. изолированным понятиям; для моделирования связанной человеческой речи и словесного мышления необходимо привлекать дополнительные соображения.

Таблица 2. Часть IIIa) схемы из «Философии имени» А.Ф. Лосева в авторской интерпретации

Определение по А.Ф. Лосеву		Интерпретация
(18) <u>Энергема</u> : сущность, «как-то» (конкретно) определённая, явленность вообще («общеприродный» уровень)		Выделенность <i>смысла</i> определённой вещи или её свойства, а также явления, процесса, состояния и проч., – т.е. собственно <i>эйдоса</i> вещи – из смысловой стихии идеального мира познающим субъектом
(20) <u>Физическая энергия</u> (<i>взаимодействие</i>): универсальность смысловой выраженности в сфере чисто физической стороны предметного мира (уровень «неживой природы»)		Данность <i>смысла</i> выбранной вещи (или её свойства и проч.) как таковой, самой по себе (её тождественность в определённых пределах), обнаруживаемое познающим субъектом в процессе взаимодействия с ней, но безотносительно от условий этого взаимодействия
(21) <u>Органическая энергия</u> (<i>раздражение</i>): некоторое знание внешнего себе предмета, но лишь с внешней (же) стороны, без знания себя и факта знания об этом («растительный» уровень)		Узнавание о внешнем мире без памяти о предыдущих узнаваниях и без соотнесения процесса узнавания с ними (сопоставление (18) и (20): определённый внешний предмет и определённый тип взаимодействия с ним)
(22) <u>Сенсуальная энергия</u> (<i>ощущение</i>): знание об ином и о себе, но без знания факта этого знания («животный» уровень)		Узнавание о себе благодаря обнаружению (на основе <i>памяти</i>) того, чего нет в себе, но есть в ином, однако без выделения себя – познающего – из прочих ощущаемых предметов (то есть без <i>памяти о себе</i>)
(23) Ступени самосознания («человеческий» уровень)	(26) <u>Перцептивная энергия</u> (<i>восприятие</i>): знание об ином и о себе вместе с фактом знания об ином и о себе	Опирающееся на личный чувственный опыт отделение себя в самостоятельный предмет познания (на основании <i>памяти о себе</i> самом); <u>обозначение</u> себя и окружающих вещей (для различения и разделения в сознании)
	(27) <u>Имагинативная энергия</u> (<i>образное представление</i>): знание об ином и о себе, утверждающее своё отличие от иного, зависящее от своего собственного внутреннего бытия	Осознание своей особенности, уникальности, принципиального отличия от всех других познаваемых вещей; начало порождения <i>логосов</i> абстрактных понятий (обобщения); благодаря представлению памяти и представлению воображения образы любых предметов могут вызываться (создаваться) в сознании и без внешних <i>ощущений</i> и <i>восприятий</i>
	(28) <u>Когитативная энергия</u> (<i>мышление</i>): то же, но утверждающее (самого) себя как начала, порождающего своё собственное бытие	Познание мира и себя, выражающееся в создании универсальных словесных форм, предопределяющих и возможность общения с себе подобными – начало языка и речи («мыслью – следовательно, существую»)

7. Предполагаемая тематика следующих семинаров в рамках данного направления:

- а). Обзор современных попыток философского осмысления задач, возможностей и ограничений ИИ.
- б). *Психофизическая проблема* и возможные пути её решения в рамках христианской антропологии (применительно к задачам ИИ).
- в) Более подробное рассмотрение концепции А.Ф. Лосева (вся схема целиком, возможности её реализации в системах ИИ, связь с психическими процессами и состояниями).
- г). Эсхатологические моменты в связи с прогнозами развития ИИ.

8. Вопросы к дальнейшему обсуждению

- а). Новые задачи и проблемы, возникающие в рамках самой христианской антропологии и психологии в связи с моделированием человеческой и животной психики в системах ИИ: необходимость рассмотрения переформулированной *психофизической проблемы*, дополненной эмпирическим фактом уникальности связи личной души с индивидуальным телом при учёте опыта трансплантологии и протезирования частей человеческого тела; возможно ли моделирование самосознания и личностного «я», творческой способности, интенциональных актов, свободы воли, интуиции, «здорового смысла», субъективного качественного опыта («квалиа»), эмоций, эстетического чувства, религиозных переживаний, и др.

б). Оценка новых научных направлений: «робозэтика» и «робопсихология»; поскольку ИИ – модель человеческой психики, эти направления должны рассматриваться по аналогии и во взаимодействии с обычной этикой и психологией (в частности: какие разделы психологии могут и должны иметь свои аналоги в «робопсихологии»?)

в). Этические требования и ограничения, а также соответствующие психологические проблемы, возникающие у современного православного христианина – учёного, инженера, программиста, прикладного специалиста, в процессе его профессиональной деятельности по разработке и применению систем ИИ (в особенности – при потенциально военных применениях).

9. Выводы

Традиционное сравнение ИИ с интеллектом человека неправомерно. «Сильный» ИИ должен отчасти моделировать ряд психических функций животных и человека, при этом в некоторых из них он даже сможет превзойти и животных, и человека в определённой мере. Однако всегда останется исключительно важная область психической жизни, к моделированию которой никакой сколь угодно «сильный» ИИ будет не способен принципиально. ИИ – вовсе не «новая эволюционная ступень» человеческого сознания, как заявляют трансгуманисты, а просто часть иной, «параллельной» человеку реальности, рукотворного мира, в чём-то его превосходящего, но в чём-то существенно ограниченного по сравнению с естественным (тварным).

Поскольку ИИ, очевидно, не обладает даруемым Богом, а не человеком, индивидуальным идеальным бытием (*душой*), он никогда не сможет иметь дело непосредственно со *смыслами*, но лишь – при посредстве человека – с их информационными запечатлениями. Тем более для него недоступно воспроизведение всего того, что относится к действующим в человеке нетварным Божественным энергиям, благодатным дарам, сообщающим человеку *образ Божий* (как основы словесного мышления, творческой способности и свободы воли). Таким образом, никакой ИИ не сможет сам порождать в себе *новые понятия*, не заимствованные у человека. Никакой ИИ не будет способен к подлинно *творческой деятельности*, в лучшем случае – к её изошрённой имитации. Наконец, никакой ИИ не обретёт подлинной *свободы воли*: все его действия будут предопределены той, возможно, сложной и неочевидной, совокупностью уже запечатлённых информационно *смыслов*, которая тоже (прямо или косвенно) была предоставлена ему человеком заранее.

Однако, к сожалению, всё это не исключает возможность неконтролируемого развития ИИ и его опасность для человечества.

Литература (выборочно):

1. *Свяц. Павел Флоренский*. Философия культа. 2014.
2. *Ильин И.А.* Основы христианской культуры. 1994. Аксиомы религиозного опыта. Т. 1-2. 1993.
3. *Архим. Киприан (Керн)*. Антропология святого Григория Паламы. 1996.
4. *Винер Н.* Кибернетика. 1983.
5. *Хайкин С.* Нейронные сети. Полный курс. 2006 (гл. 1.9. «Историческая справка»).
6. *Николенко С. и др.* Глубокое обучение. 2018 (гл. 1, а также раздел 3.1 гл. 3).
7. *Сёрль Дж.* Новое открытие сознания. 1992. Сознание, мозг и программы. 1998.
8. *Дрейфус Х.* Чего не могут вычислительные машины. 1978.
9. *Хакен Г.* Информация и самоорганизация. 1991.
10. *Пенроуз Р.* Новый ум короля. О компьютерах, мышлении и законах физики. 2011.
11. *Хокинг С.* Большое, малое и человеческий разум. 2008.
12. *Лосев А.Ф.* Бытие. Имя. Космос. 1993. («Философия имени», «Вещь и имя»).