4.3. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК СМЕРТНЫЙ ГРЕХ ЧЕЛОВЕЧЕСТВА[1]

А.Н. Козырев, д.э.н., к.ф.-м.н.,
руководитель научного направления в Центральном Экономическом институте РАН
Рисунки Елизаветы Вершининой

Развитие искусственного интеллекта сопровождается широким освещением успехов по каналам СМИ. Информационные сообщения подпитывают ожидания на скорое появление сверхмощного искусственного интеллекта. Детальный анализ показывает, что эти ожидания являются завышенными, а связанные с ними страхи едва ли реализуются в ближайшем будущем. Скорее, стоит опасаться косвенного влияния искусственного интеллекта на разрушение отрицательных обратных связей, которые поддерживают человеческое общество. Конкретные примеры таких нарушений указаны Конрадом Лоренцом, к которому и отсылает название данной статьи.

Искусственный интеллект (далее – ИИ) стал важным фактором геополитики благодаря в основном двум важным обстоятельствам. Во-первых, сейчас как никогда раньше реальна возможность использовать ИИ в военной сфере, в том числе, в качестве оружия. Задачи распознавания шумов, производимых винтами подводных лодок, и другие инженерные по сути задачи военного назначения решались и раньше, но сегодня речь идет о создании роботов, которые будут убивать людей. Во-вторых, стремительно растут вложения в ИИ гражданского назначения, прежде всего, вложения частных средств в развитие коммерческих приложений ИИ.

Тенденции развития ИИ и ожидания в оптимистическом ключе рассмотрены в докладе (CD Insights, 2018)[2], подготовленном специалистами CD Insights. Вместе с тем, далеко не все тут однозначно. Влияние вложений в ИИ на экономический рост в настоящее время скорее отрицательно, то есть постоянно возникает несоответствие между ожиданиями и статистикой. Таков главный вывод отчета Национального бюро экономических исследований, озаглавленного «Искусственный интеллект и современной парадокс производительности: столкновение ожиданий и статистики» (NBER, 2017). Еще более ярко разрыв между реальными достижениями ИИ и ожиданиями показан в публикации двух авторов (Marcus & Davis, 2018)[3], один из которых психолог, второй – специалист по информационным технологиям.

Не меньшее разочарование на сегодняшний день вызывает рост затрат вычислительных мощностей на глубокое обучение. В недавнем исследовании (Amodei & Hermander. 2018)[4] приведены цифры, показывающие, что с 2012 года количество вычислений, используемых в крупнейших тренировочных прогонах AI, растет экспоненциально с 3,5-месячным периодом удвоения (для сравнения, закон Мура имел 18-месячный период удвоения). При таком росте затрат вычислительной мощности развитие вычислительной техники не успевает за потребностями.

Настоящая публикация посвящена другим не менее важным и гораздо более тревожным аспектам развития технологий ИИ. Речь идет о влиянии ИИ на людей и нашу цивилизацию в целом. Обращение к понятию смертного греха в названии доклада на мировом форуме и этой публикации – не попытка эпатажа, а всего лишь следование традиции, заложенной Конрадом Лоренцом – величайшим естествоиспытателем и философом 20-века. Название одной из его книг (Lorenz, 1973) – «Восемь смертных грехов цивилизованного человечества» ассоциируется с библейским текстом о семи смертных грехах. Но речь идет не о грехах отдельного человека, а об опасных тенденциях в политике развитых стран и развитии человечества в целом, способных привести человечество к гибели.

Общее у восьми смертных грехов по Лоренцу – то, что все они связаны с нарушением отрицательных обратных связей в природе и обществе как результатом целенаправленных действий людей с благими намерениями. В частности, речь идет об уничтожении отрицательных обратных связей, обеспечивающих гомеостаз в биологической системе, частью которой является человечество, но не только о них. Отрицательные обратные связи обычно воспринимаются как неудобства или даже зло. Обрыв таких связей в каждом отдельном случае легко оправдать, поскольку он вызван либо необходимостью решения каких-то серьезных, в том числе, геополитических проблем, либо соображениями гуманности, либо желанием повысить качество жизни именно в том смысле, как это понимается здесь и сейчас – на момент принятия конкретного решения. Но каждый из таких обрывов

[1] Доклад на панельной сессии: Искусственный интеллект и геополитика в рамках 7-го Всемирного форума за мир в Пекине (7th World Peace Forum (WPF), July 14-15, 2018, at Tsinghua University in Beijing, China).

[2] https://www.cbinsights.com/research/report/top-tech-trends-2018/

[3] https://www.nytimes.com/2018/05/18/opinion/artificial-intelligence-challenges.html

[4] https://goo.gl/CaZCZ8

обратной связи имеет отдаленные последствия, причем не всегда предсказуемые и, возможно, фатальные для человечества в конечном итоге.

Каждому из названных Лоренцем смертных грехов соответствует отдельная глава в его книге. Всего в ней девять глав. Первая глава посвящена гомеостазу, обратным связям и другим общим вопросам, развиваемым в последующих восьми главах применительно к каждому из явно названных восьми грехов. Ниже они перечислены в порядке, выбранном Лоренцем, чтобы потом было удобно ссылаться.

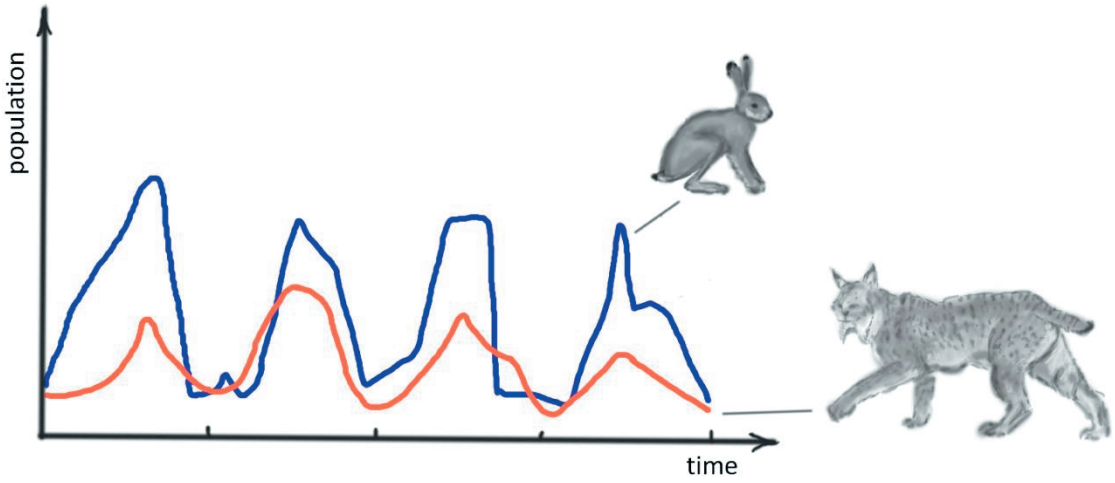

Рисунок 1. Простейший пример гомеостазиса

Глава 1. Структурные свойства и нарушения функций живых систем.
Глава 2. Перенаселение.
Глава 3. Опустошение жизненного пространства.
Глава 4. Бег наперегонки с самим собой.
Глава 5. Тепловая смерть чувства.
Глава 6. Генетическое вырождение.
Глава 7. Разрыв с традицией.
Глава 8. Индоктринируемость.
Глава 9. Ядерное оружие.

Среди перечисленных явно восьми грехов (главы 2-9) не упомянут ИИ. Не упоминается он и в тексте книги. Однако легко заметить, что ИИ явно имеет отношение, как минимум, к трем упомянутым грехам – бег наперегонки с собой (глава 4), разрыв с традицией (глава 7) и, как следует из публикации (RAND, 2018)[5], ядерное оружие (глава 9). Более глубокий анализ показывает, что связей гораздо больше. Возможно, хотя и это спорно, нет связей ИИ с перенаселением Земли (глава 2) и опустошением жизненного пространства (глава 3). В остальном связь просматривается достаточно ясно. Постоянное общение с ботами не может не сказаться на чувствах и способности чувствовать (глава 5), избавление от рутинных операций оборачивается деградацией в том, что не является рутиной (глава 6). Так, постоянное пользование навигатором при вождении автомобиля приводит к неспособности ориентироваться на местности в реальной обстановке. Но самое интригующее – это связь ИИ с тем, что в книге Лоренца получило название Индоктринируемость. В случае с ИИ эта специфическая болезнь современной науки проявляется ярко, как нигде больше, если не считать экономическую науку.

Отдельным смертным грехом могло бы стать создание искусственного интеллекта, превосходящего по своей мощи человеческий интеллект. Возможные проблемы, возникающие или, точнее, ожидаемые в этой связи, активно обсуждаются в некоторых аудиториях и организациях. Например, этому вопросу посвящены публикация (Yudkowsky, 2008) и библиография к ней, а также многие более поздние публикации. Но в основном внимание привлечено к реальным или мнимым успехам в области построения нейросетей и глубокого обучения.

Также надо учесть, что в далеком уже 1972 году, когда Лоренц писал свою книгу, реальные достижения ИИ были еще слишком скромны, чтобы ставить их разрушительную силу в один ряд с ядерным оружием и генетическим вырождением. Например, робот с ИИ мог сложить башенку из 4-х кубиков, используя «руку» – манипулятор. На 5-й кубик интеллекта уже не хватало.

[5] https://www.rand.org/pubs/perspectives/PE296.html

Но в произведениях фантастов еще задолго до того (в 1950-60-х) роботы умели не только служить человеку, но также выходить из-под его контроля и даже любить, рискуя сжечь в любовном порыве все лампы своего электронного «мозга».

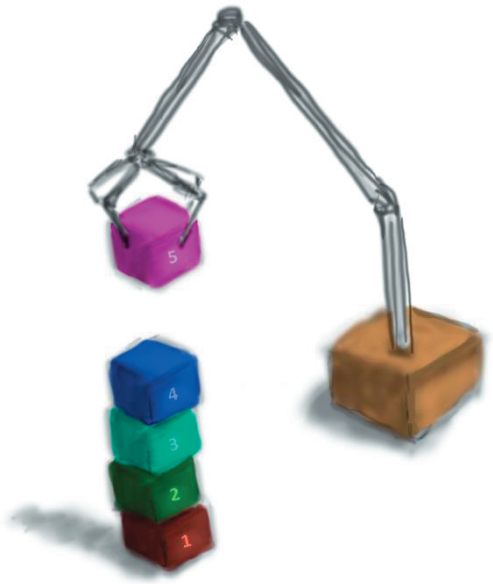

Рисунок 2. Реальность 1972 года

Отчасти эти настроения разделяли и сами исследователи. Как вспоминает в одном интервью С.В. Карелов – один из отечественных исследователей: «Я помню ожидания после рывка 1950–60-х. Казалось, вот-вот — еще 3 года! Еще 5 лет! — и будет создан искусственный интеллект общего назначения, способный себя осознавать, самостоятельно ставить цели, сотрудничать с человеком на равных. А потом стало ясно, что исследователи топчутся на месте, и ничего важного не происходит».

Сейчас ситуация с применением ИИ во многих аспектах отличается от той, что была в шестидесятые и начале семидесятых годов 20-го века. Сегодня ИИ— своеобразный технологический агрегатор, вобравший в себя множество разных направлений, от добычи данных до робототехники. Он привлекает максимальный интерес инвесторов и бизнеса. И все же сходство в некоторых фундаментальных проблемах самого ИИ разительно. Тут самое время вспомнить про Индоктринируемость (глава 8) – одну из болезней современной науки и научного сообщества. В разных странах эта болезнь имела тогда и имеет сегодня свои оттенки, в частности, это касается ИИ.

У нас в 1960-70-х (в СССР) тематика ИИ была окружена особым ореолом, в котором явно присутствовал своего рода декаданс. Среди интеллигенции считалось, что кибернетику у нас «разгромили» по политическим мотивам, снабдив непристойным ярлыком – «Продажная девка империализма». Так ли было на самом деле? Вопрос более, чем спорный, поскольку на связанные с обороной и космосом направления кибернетики тратились очень значительные средства. Велись и вполне мирные исследования. Однако в глазах широкой публики и гуманитарной интеллигенции транслируемая «сарафанным радио» передача про «девку» вполне объясняла причины довольно скромных успехов кибернетики и, прежде всего, ИИ в гражданском секторе, который у всех на виду. Но то же самое было за океаном, только без идеологического оттенка. Кибернетику «громили» представители точных наук, видевшие ее эклектичность и несоответствие достижений обещаниям. В частности, знаменитый доклад математика сэра Джеймса Лайтхилла, подготовленный по заказу парламента в 1973 году (Lighthill, 1973), привел к почти полному демонтажу исследований ИИ в Англии. Примечательно, что доклад обсуждался публично, дискуссия по нему транслировалась ББС, а запись сохранена. Ее можно посмотреть на ЮТубе[6, 7, 8, 9, 10, 11]. Суть доклада, если ее формулировать одной фразой, состояла в том, что не существует такой дисциплины, как ИИ. Все решаемые ИИ реальные задачи могут быть решены в других дисциплинах. А в объединяющей их части реальных достижений фактически нет.

Нечто подобное происходит и сейчас, но все теперь выстроено вокруг денег. Разрыв между реальными достижениями и фантазиями на тему ИИ стал родовой травмой ИИ, более того, он стал злокачественным, когда фантастов заменили маркетологи. Развивается только то, что приносит быстрые деньги, а потому «Социальные сети распознают котиков, «умные дома» — открывают котикам нижние дверки и, отметим, справедливости ради, помогают хозяевам утилизировать мусор. «Немереные деньги вкладываются в системы, способные в реальном времени подменять голоса и мимику политиков и просчитывать поведение избирателей» (С.В. Карелов, там же)[12]. Но все это – частные задачи.

Сенсации при ближайшем рассмотрении оборачиваются не вполне добросовестной подачей материала. Типичный пример – подача в некоторых источниках (Gerbert, 2018)[13] информации о победе AlphaZero над Stockfish в матче из 100 партий как сенсации мирового уровня.

---

[6] https://www.youtube.com/watch?v=yReDbeY7ZMU часть 1
[7] https://www.youtube.com/watch?v=FLnqHzpLPws часть 2
[8] https://www.youtube.com/watch?v=RnZghm0rRII часть 3
[9] https://www.youtube.com/watch?v=pyU9pm1hmYs часть 4
[10] https://www.youtube.com/watch?v=LRgSiKKwFjE часть 5
[11] https://www.youtube.com/watch?v=3GZWFnWOqkA часть 6
[12] https://officelife.media/article/people/sergey-karelov-winter-of-artificial-intelligence-is-near/
[13] https://sloanreview.mit.edu/article/ai-and-the-augmentation-fallacy/

Game	White	Black	Win	Draw	Loss
Chess	AlphaZero	Stockfish	25	25	0
	Stockfish	AlphaZero	3	47	0

Рисунок 3. Результаты матча между AlphaZero и Stockfish

При попытке объективного рассмотрения выясняется, что «для выравнивания условий» у Stockfish были отключены базы данных по дебютам и эндшпилям при том, что именно они у Stockfish были «фишкой», определявшей силу игры. На рисунках 4 и 5 показано, что может означать такое «выравнивание».

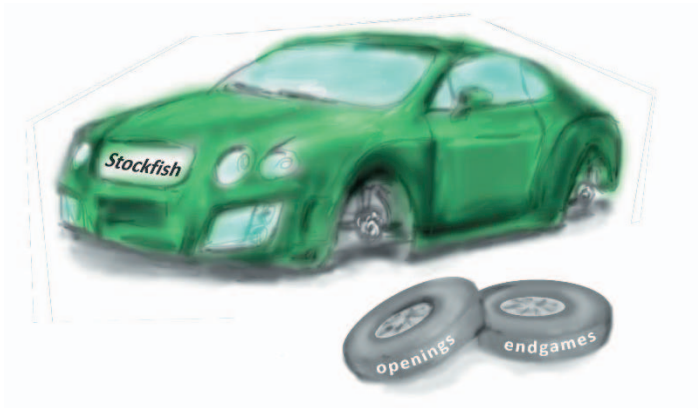

Рисунок 4

Вместе с тем AlphaZero использовал принципиально новый тензорный процессор, а Stockfish – стандартный центральный процессор, что давало AlphaZero дополнительное преимущество. Иначе говоря, здесь скорее можно говорить об успехе в разработке новых процессоров, а не интеллекта.

В разработке нейросетей успех, несомненно, тоже есть, но он отнюдь не сенсационный. И снова речь отнюдь не об универсальном интеллекте, а о новом подходе к решению конкретной задачи.

Тут при желании можно снова сказать, что тематику ИИ «громят» представители точных наук. Типичный пример – лекция известного физика Фримена Дайсона о человеческом мозге (Dyson, 2014)[14]. В частности, он рассказал упоминавшуюся выше поучительную историю с докладом его давнего, но к 2014 уже покойного сэра Джеймса Лайтхилла. Еще тогда в 1973 году все работы по ИИ (в широком смысле) четко делились на три категории, обозначенные в докладе A, B и C. Согласно докладу, категории A и C имеют четко определенные мотивы: каждая из них имеет четко определенное общее направление своих предполагаемых целей, но эти два направления совершенно разные. В обеих этих категориях в течение двадцати пяти лет (начиная со статьи Тьюринга 1947 года "Интеллектуальный механизм" и кончая публикацией доклада в 1973) был достигнут определенный прогресс, хотя ожидания часто не оправдывались. Категорию A составляли чисто прикладные инженерные исследования типа распознавания речи, машинного перевода и некоторые другие практические задачи, решаемые обычно людьми. Категория C - все, что связано с когнитивным, нейроморфным, мозгоподобным компьютингом. Были отмечены некоторые перспективные исследовательские работы в области нейронауки.

Рисунок 5

В течение того же периода проводилась исследования еще одной категории (категория B), где цели и задачи гораздо труднее различить, но которая в значительной степени опирается на идеи как из A, так и из C и, наоборот, стремится влиять на них. Исследования в категории B, если приемлемые аргументы для этого могут быть согласованы, работают на основе его взаимозависимости с исследованиями в категориях A и C, чтобы обеспечить единство и согласованность всей области исследований ИИ. Вместе с тем прогресс в этой промежуточной категории B вызвал еще большее разочарование как в отношении фактически проделанной работы, так и в отношении установления веских причин для такой работы и, таким образом, для создания какой-либо единой дисциплины, охватывающей категории A и C.

[14] https://sureshemre.wordpress.com/2014/11/28/are-brains-analogue-or-digital/

Далее Фримен Дайсон говорит, что спустя еще почти 50 лет, то есть к моменту его лекции, раскладка не изменилась. По-прежнему бурно развивается первое направление: компьютер уже неплохо распознает и переводит. На третьем направлении успехи также заметны: исследователи картировали мозг и стали лучше разбираться в его функциях. А на втором направлении — по-прежнему полный ноль.

В итоге приходим к выводу, что опасаться появления сверхмощного искусственного интеллекта пока не следует. Человечество гораздо быстрее сумеет убить себя, продолжая следовать другим восьми обозначенным Конрадом Лоренцом греховным традициям. А потому опасность ИИ скорее косвенная, связанная с педалированием тех восьми традиций. Это касается и возможности нечаянно спровоцированной ядерной войны, и гонки человечества с самим собой, и утраты чувств, и разрушения традиций. Но главная опасность – Индоктринируемость – болезнь разума цивилизованного человечества.

Литература:

1. Amodei & Hermander (2018) AI and Compute, by Dario Amodei and Danny Hermander https://goo.gl/CaZCZ8
2. CB Insights (2018), 15 Trends Shaping Tech In 2018
3. Dyson, F. (2014) Are brains analogue or digital? Lecture at the University College Dublin
4. Gerbert, Philipp, (2018) AI and the ‘Augmentation’ Fallacy May 16, 2018
5. IBM (2009) The Cat is Out of the Bag: Cortical Simulations with109 Neurons, 1013 Synapses Rajagopal Ananthanarayanan1, Steven K. Esser1 Horst D. Simon2, and Dharmendra S. Modha1
6. Lorenz, Konrad (1973), Die acht Todsünden der zivilisierten Menschheit. R. Piper & Co. Verlag, München, 1973.
7. Lorenz, Konrad (1974), Civilized man's eight deadly sins. “A Helen and Kurt Wolff book”, 1974.
8. Masrcus & Davis (201), A.I. Is Harder Than You Think by Gary Marcus and Ernest Davis. (Mr. Marcus is a professor of psychology and neural science. Mr. Davis is a professor of computer science. May 18, 2018)
9. NBER (2017) ARTIFICIAL INTELLIGENCE AND THE MODERN PRODUCTIVITY PARA-DOX: A CLASH OF EXPECTATIONS AND STATISTICS, by Erik Brynjolfsson, Daniel Rock, Chad Syverson, Working Paper 24001 NATIONAL BUREAU OF ECONOMIC RESEARCH 1050[15] Massachusetts Avenue Cambridge, MA 02138, November 2017
10. RAND (2018) How Might Artificial Intelligence Affect the Risk of Nuclear War? by Edward Geist, Andrew J. Lohn, Perspective EXPERT INSIGHTS ON A TIMELY POLICY ISSUE
11. Yudkowsky, Eliezer (2008), Artificial Intelligence as a Positive and Negative Factor in Global Risk Machine Intelligence Research Institute, Edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

Анатолий Николаевич Козырев (kozyrevant@mail.ru)

Ключевые слова

искусственный интеллект, комбинаторный взрыв, отрицательная обратная связь

Kozyrev A.N. Artificial intelligence as civilized man’s deadly sin.

Keywords

artificial intelligence, combinatorial explosion, negative feedback

Abstract

Success stories from mass-media accompany the development of artificial intelligence. Messages fuel the expectations about rapid creation of superpowerful artificial intelligence. A detailed analysis reveals these expectations to be overstated, and the associated fears are unlikely to realize in the near future. Rather, it is worthwhile to fear the indirect influence of artificial intelligence on the destruction of negative feedback that supports human society. Konrad Lorenz listed the specific examples of such violations and the title of this article refers to his work.

DOI: 10.34706/DE-2018-02-12

[15] http://www.nber.org/papers/w24001