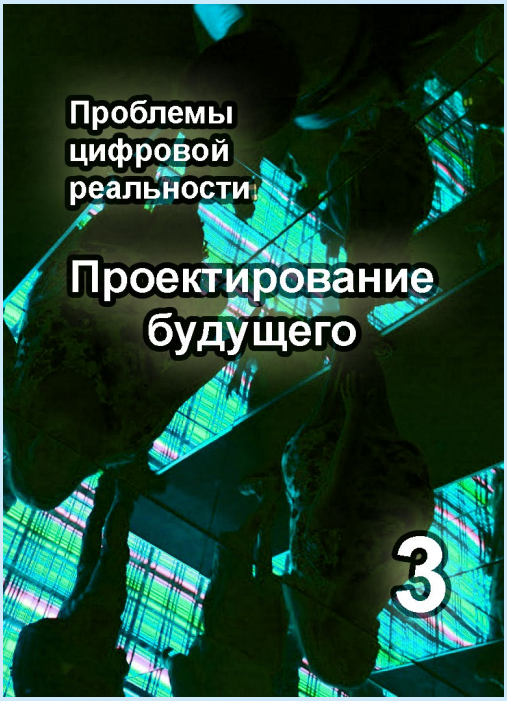
ИПМ им.М.В.Келдыша РАН

Будущее-2020 • Труды конференции



Г.Г. Малинецкий, В.С. Смолин,
О.Ю. Колесниченко, Т.Н. Жилина

Социологическая траектория в становлении ИИ: вызовы неопределенности

Рекомендуемая форма библиографической ссылки

Малинецкий Г.Г. Смолин В.С. Колесниченко О.Ю. Жилина Т.Н. Социологическая траектория в становлении ИИ: вызовы неопределенности // Проектирование будущего. Проблемы цифровой реальности: труды 3-й Международной конференции (6-7 февраля 2020 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2020. — С. 241-251. — https://keldysh.ru/future/2020/22.pdf
https://doi.org/10.20948/future-2020-22

Размещено также видео выступления

Социологическая траектория в становлении ИИ: вызовы неопределенности

Г.Г. Малинецкий[1], В.С. Смолин[1], О.Ю. Колесниченко[2], Т.Н. Жилина[2]

[1]ФИЦ ИПМ им. М.В. Келдыша РАН
[2]ФГАОУ ВО Первый МГМУ им. И.М. Сеченова

Аннотация. Произошедшая в 2010-12 гг. революция ИИ позволяет не только решать всё более широкие классы задач, которые раньше не были доступны человеку, но и более прагматично оценить возможности и цели человеческого мышления. Разработки сильного ИИ идут по разным путям, от имитации работы мозга до создания машины со своими целями и желаниями. Вызовы неопределенности включают внедрение ИИ в ускорение социальной трансформации, порождают риски для жизни и свободы человека, и угрозу техногенных катастроф. Предложенный анализ социологических проблем полезен для понимания как перспектив развития сильного ИИ, так и постановки задач при его внедрении.

Ключевые слова: искусственный интеллект, машинное обучение, искусственные нейронные сети, социум, социология управления, социальная структура

The sociological trajectory in AI drafting: Challenges of uncertainty

G.G. Malinetskiy[1], V.S. Smolin[1], O.Yu. Kolesnichenko[2], T.N. Zhilina[2]

[1]KIAM RAS, Moscow, Russia
[2]I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Abstract. AI revolution of 2010-12 allows not only solving ever wider task classes that were previously accessible only to humans but also more pragmatically assessing the possibilities and goals of human thinking. Developing a strong AI goes in different ways, from the brain work simulating, to creating a machine with its goals and desires. The uncertainty challenges include technological disaster threats, human life and freedom risks, and AI utilization in social transformation acceleration. The proposed analysis of sociological problems is useful for understanding both the prospects for AGI development and setting its deployment goals.

Keywords: artificial intelligence, machine learning, artificial neural networks, society, management sociology, social structure

Введение

Успехи внедрения нейросетевых алгоритмов в устройства и системы, решающие «интеллектуальные» задачи, принято характеризовать как нейросетевую революцию в машинном обучении, произошедшую в 2010-12 гг. Воодушевляющими стали достижения в эффективной автоматической настройке огромного количества латентных (внутренних) параметров нейросетевых алгоритмов. Это сделало доступной обработку сложных сигналов реального мира на качественно более высоком уровне.

Десятилетний опыт использования нейросетевых алгоритмов привёл к созданию целой индустрии со своими технологиями, стандартами, производствами, исследовательскими и учебными институтами. Ретроспективный взгляд на развитие этой отрасли показывает, что практически все идеи и методы, используемые в устройствах и системах с искусственным интеллектом (ИИ), появились в результате развития математических идей и появления новых алгоритмов, а не как заимствование из биологии, нейрофизиологии и психологии. Но опыт применения нейросетевых вычислений для решения широкого круга задач позволяет по-новому взглянуть на обширный массив знаний о человеке, полученных до ИИ.

Формирование представлений об иерархических процессах обработки информации в нейроподобных структурах позволяет глубже понять причины формирования человеческого общества. Можно под другим углом взглянуть на проблемы, возникающие в сложных социальных структурах. С другой стороны, гуманитарный социологический анализ развития цивилизации даёт возможность не только построить прогнозы в отношении взаимодействия социума с ИИ, лучше понять его вклад в гуманитарно-технологическую революцию [1], но и сформулировать требования, следование которым позволит разрабатываемым устройствам и системам ИИ безопасно для человека влиться в социум и послужить дальнейшему прогрессу человеческой цивилизации, а не ее разрушению.

История

С древних времён человек пытается создать устройства, способные самостоятельно осуществлять действия в сложной среде. После доказательства теоремы Тьюринга и появления вычислительных машин на основе схем фон Неймана вопрос перешёл от умозрительной в практическую плоскость. Мы умеем выполнять на ЭВМ любой алгоритм и теперь задаемся вопросом: можем ли мы представить разумную, интеллектуальную деятельность в виде набора алгоритмов?

Считается, что термин «Искусственный Интеллект» (Artificial Intelligence, AI) впервые предложил Джон Маккарти (John McCarthy) на

Дартмутском семинаре, проведённом летом 1956 г. в Дартмутском колледже (США). На семинаре произошло объединение единичных энтузиастов в профессиональное сообщество со своими научными целями и четким самоопределением. Джон Маккарти предположил, что любое свойство интеллекта может быть столь точно описано, что машина сможет его симулировать. «Мы попытаемся понять, как обучить машины использовать естественные языки, формировать абстракции и концепции, решать задачи, сейчас подвластные только людям, и улучшать самих себя», – писал Джон Маккарти в заявке на проведение мероприятия [2]. Однако, надежды, что свойства интеллекта могут быть описаны и реализованы в виде программ для компьютеров, начали всерьёз оправдываться лишь через 55 лет.

После Дартмутского семинара окончательно сформировалось два основных направления в подходах к созданию ИИ: формально-логический и нейросетевой. Сторонники формально-логического подхода называли себя «Neats» (чёткие, клёвые), а нейросетевики – «Scruffies» (нечёткие, неряшливые). В 1960-е гг. лидером «Neats» был Джон Маккарти, утверждавший, что «математически точное мышление = семантические представления логического вывода». Лидером «Scruffies» являлся Марвин Минский (Marvin Minsky), который развивал идеи самоорганизации, машинного обучения, формируемой семантики.

Успехи формально-логических подходов на ранних этапах решения «интеллектуальных» задач были основаны на стремлении имитировать внешние проявления разумной деятельности, используя разработку оптимальных преобразований для реализации требуемых функций. Это более короткий путь к решению отдельных задач, чем разработка универсальных подходов решения широких классов задач, на которые замахивались “Scruffies”. Превзойти уровень «Neats» не удавалось, энтузиазм “Scruffies” сменялся пессимизмом. Две «весны» сменились двумя «зимами» ИИ, прежде чем нейросетевые подходы в 2010-12 гг. начали превосходить формально-логические. Успех сторонников нейросетевых подходов был обусловлен развитием специализированных схем обработки изображений, речи и других модальностей сигналов (динамика конкуренции успешности двух направлений ИИ представлена на рис. 1).

Необходимо отметить, что в 1997 г. в матче с чемпионом мира по шахматам Гарри Каспаровым компьютер Deep Blue одержал победу. В целом, в отдельных задачах (не только в шахматах) уровень человека был превзойдён ещё до 2000 г. формально-логическими методами. С другой стороны, даже сейчас, с использованием развитых нейросетевых подходов, далеко не все «интеллектуальные» задачи удаётся решать хотя бы на уровне человека. Но если до 2015-18 гг. в большинстве «интеллектуальных» задач уровень человека казался недостижимым, то в

этот период по ним начался процесс массового превышения возможностей человека различными техническими устройствами и системами. И процесс активно продолжается в настоящее время.

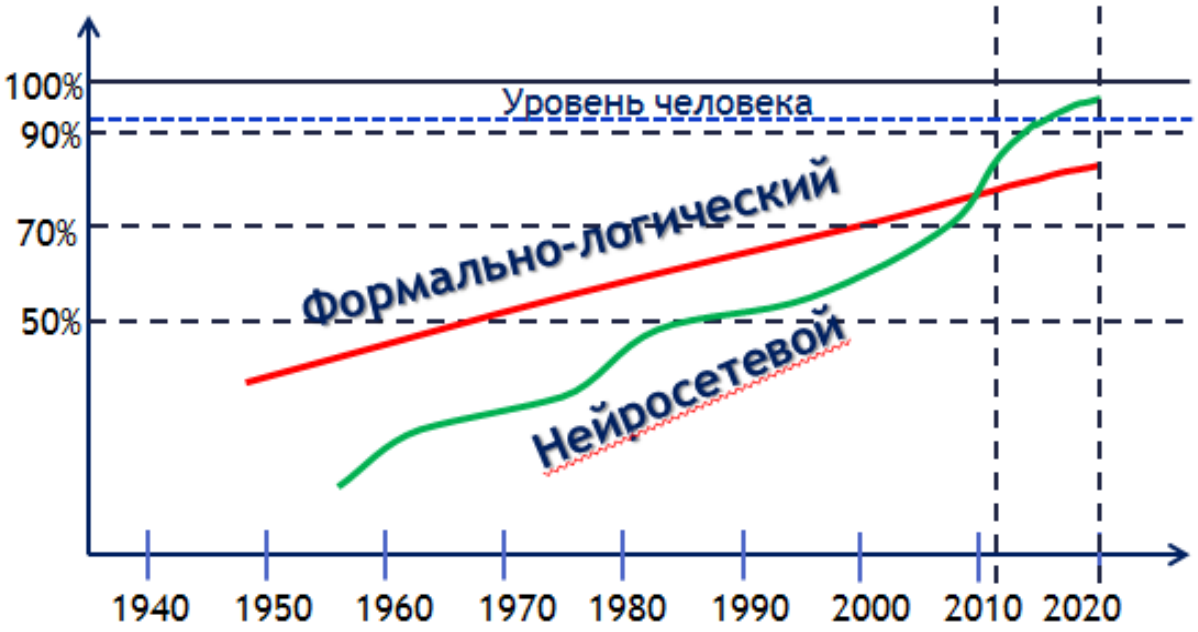


Рис. 1. Процентное возрастание безошибочных решений «интеллектуальных» задач при использовании различных подходов

Современный уровень развития ИИ

Для нейровычислений необходимы мощные компьютеры, большие объёмы данных (которые позволяет собрать сеть Интернет) и теория построения нейросетевых алгоритмов. На сегодняшний день центральным способом решения значительного числа задач ИИ (хотя и не всех) является использование нейросетей глубокого обучения (Deep Neural Networks, DNN). Важнейшим свойством DNN является их способность автоматически настраивать миллиарды своих параметров, применяя метод обратного распространения ошибки (Back Propagation Error, BPE). Успехи теории DNN можно условно разделить на два основных направления: разработку новых схем нейросетевых вычислений и совершенствование методов автоматической настройки параметров (обучения). На рис. 2 показано последовательное усложнение нейросетей, от отдельных элементов до блоков.

Кроме полносвязанных многослойных сетей развитие получило много плодотворных идей: свёрточные сети (CNN – convolutional NN); рекуррентные сети (RNN – recurrent NN); порождающие состязательные сети (GANs – generative adversarial networks); сети глубокого обучения с подкреплением, (DQN – deep Q (quality) networks) и ряд других (рис. 3).

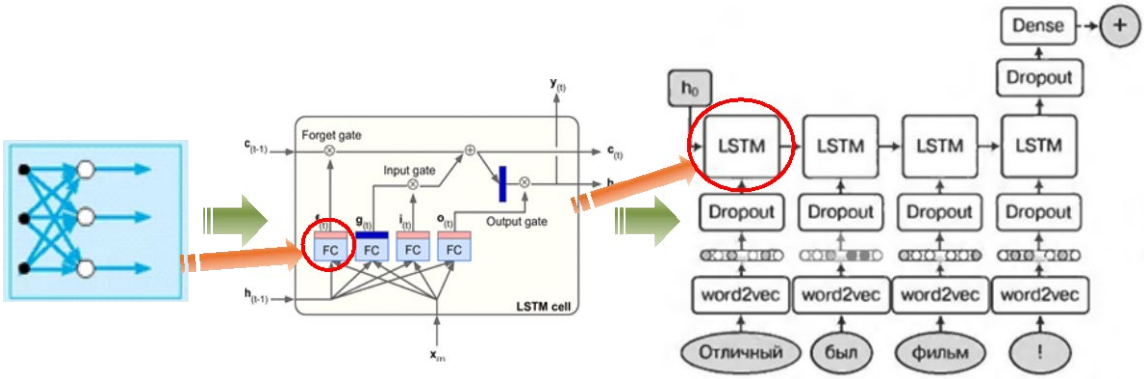


Рис. 2. Этапы усложнения структуры нейросетевых вычислений

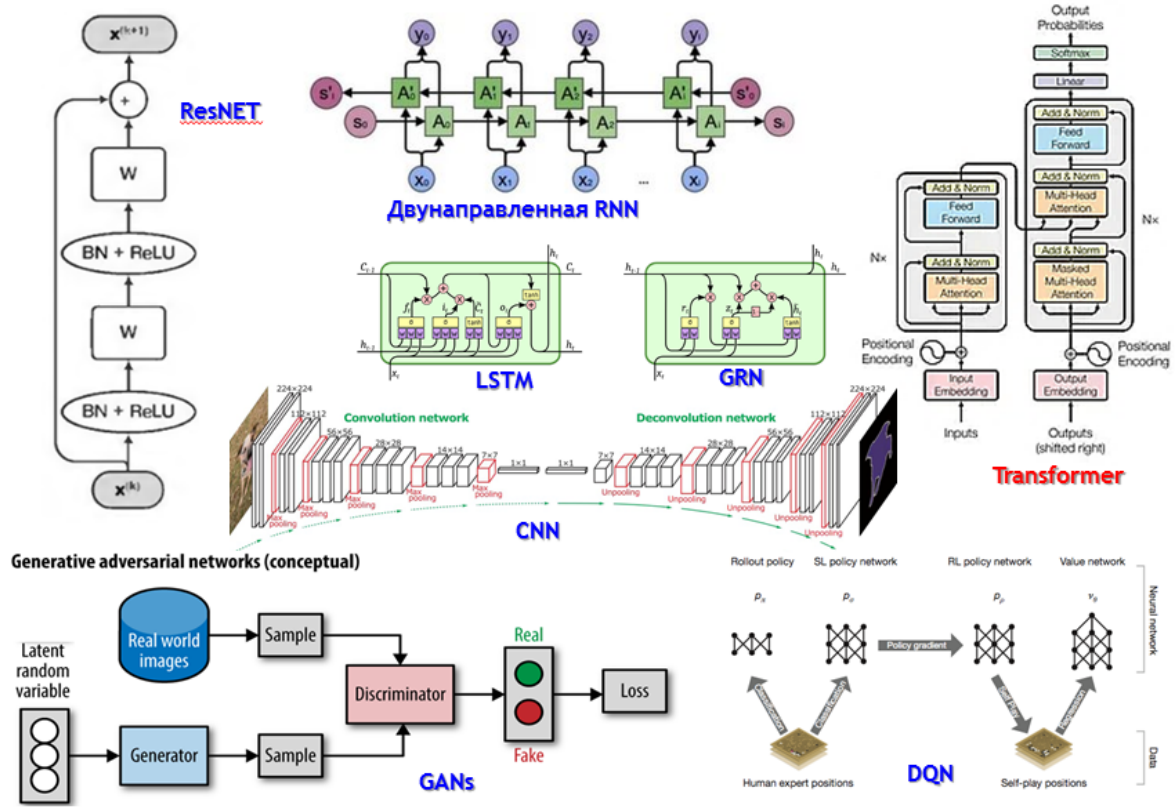


Рис. 3. Схемы сетей формальных нейронов, реализующие основные идеи многослойных (глубоких) нейровычислений

Прогресс в совершенствовании методов обучения не менее значителен: наглядно он вылился в то, что за последние десять лет число слоёв в DNN увеличилось с единиц до сотен, а количество обучаемых параметров с тысяч до миллиардов. Но это «внутренняя кухня», которая конечному пользователю может быть и не заметна.

Впечатление на пользователей оказывают появляющиеся уже каждый месяц новые успехи в обработке сигналов реального мира. Для этого необходимо выразить состояния наблюдаемых объектов и явлений в

виде векторов $\vec{X}$, а затем вектора $\vec{Y}$ использовать для генерации физических явлений – света, звука, давления, движения и др.

В век цифровой техники подобные преобразования производятся во многих устройствах и не вызывают удивления. Но иногда требуется определённая изобретательность, чтобы передать вектором значение слова или выразить настроение в векторной форме. И разработано несколько нейросетевых подходов, например, Wordtovec [3] и Glove, позволяющих формировать векторное описание значений слов. Методы векторизации продолжают совершенствоваться, но уже сейчас есть значительный прогресс в решении широкого класса «интеллектуальных» задач.

Сложность состоит не столько в придании цифровой формы описанию, сколько в выборе (или формировании) удачной метрики. Поскольку, например, похожие по написанию слова могут иметь как близкие, так далёкие значения. Например: «мама» – «мамочка» и «баба» – «бабочка».

Современное развитие схем нейросетевых вычислений позволяет не только создавать аппроксимацию преобразований $\vec{X} \to \vec{Y}$, но и формировать удачные метрики оцифренных векторов (преобразованием $\vec{X} \to \vec{Y}$).

Требования к AGI и высокие цели социума

При всём различии взглядов на проблему создания сильного ИИ (Artificial General Intelligence, AGI), многие сходятся в том, что основным препятствием для создания AGI является сложность описания окружающего мира. Как показали ещё Минский и Пейперт в [4], не все задачи эффективно решаются параллельной обработкой и сводятся к глубокому обучению. В качестве подходов к решению проблемы, например, обладатели премии Тьюринга за 2018 г. Йошуа Бенжио, Ян ЛеКун [4] и Джеффри Хинтон называют развитие таких свойств, как внимание, понимание причинности событий, планирование действий, формирование высокоуровневых концепций. Работы в данных направлениях давно ведутся и почти во всех есть успехи. Проблема состоит в том, что пока не удаётся собрать единую модель, удовлетворяющую одновременно всем желательным свойствам.

Методической сложностью остаётся отсутствие общепризнанного определения AGI и интеллекта в целом. Есть несколько направлений в определении AGI (и того, что же такое интеллект):

Первое направление – *имитировать внешние проявления «интеллектуальной» деятельности*. И до сих пор, успехи ИИ оцениваются в основном по успешности имитации.

Второе направление связано с созданием устройств, *по интеллекту превосходящих уровень человека*. В условиях отсутствия общепринятого

определения интеллекта, сравнение по уровню интеллектуальности не может быть корректным. Сегодня AlphaZero уверенно обыгрывает чемпионов мира по го и шахматам, но никто не считает, что уровень AGI достигнут.

Третьим направлением оценки AGI является *способность к выполнению рациональных или разумных действий.* Разумные действия не всегда рациональны. Главным аргументом критики третьего направления является то, что для рациональных и разумных действий необходима некоторая цель для достижения и оптимизации пути к ней. В процессе эволюции человеческого общества развивались не только орудия труда и иерархические отношения. Формировались, изменялись и проходили отбор нормы этики и морали, а также цели и задачи, которые ставили перед собой как отдельные индивидуумы, так и общества в целом. Даже в мире, пронизанном идеями глобализации, нормы этики и морали имеют заметную вариативность. Это ставит оценку разумности и рациональности действий в зависимость от обстоятельств их выполнения и может обусловить критические социальные проблемы при появлении в перспективе AGI.

Четвёртым направлением является способность *к формированию и выбору целей действий*. Для проявления «интеллектуальности» необходимо наличие свободы выбора значений параметров цели, которая не может быть достигнута, если критерий оптимальности выбора жёстко задан и заложен в систему AGI изначально. Критерии выбора целей AGI должен формировать самостоятельно, на основе собственного опыта взаимодействия со средой и восприятия правил цивилизованного общества, в котором он функционирует.

Пятым направлением определения свойств AGI является *наличие желаний*. Формировать цели, выполнять действия, превосходить человека или другие системы AGI. Если система AGI обладает всеми перечисленными способностями, но не стремится их проявить, то она останется автоматом, выполняющим инструкцию. Наличие желаний реализовать доступные возможности, оценивая при этом на основе своего опыта степень соответствия цивилизационным нормам, позволит AGI взвешенно проявлять свои способности при выборе целей и формировании поведения.

Важный аспект, касающийся социальных процессов, – разделение труда, степень которого увеличивалась в течение всей истории развития человеческого общества. С точки зрения представлений об AGI, декомпозиция выполняемых задач на несколько более простых позволяет достигать лучших результатов обучения. Внедрение ИИ в процесс разделения труда будет оставлять человеку больше времени для осуществления высокоуровневой мыслительной деятельности, не связанной с выполнением рутинной работы, что поменяет ландшафт рынка

труда и образования. Именно мыслительная деятельность человека служит основой для появления новых способов производства, трудовых и социальных отношений и в целом обеспечивает прогресс общества. В этом смысле ИИ будет способствовать ускорению развития всех социальных процессов и бурным социальным преобразованиям.

Исторически сложилась практика принуждения к труду. Можно условно выделить три основных метода постоянного принуждения: угроза применения силы, материальное поощрение, а также создание мифов. Все три метода, как правило, применяются совместно, речь может идти только о вариациях степени вклада каждого из методов в принуждение к труду. Из трёх методов принуждения к выполнению всех работ в рамках разделения труда мифы являются самым эффективным и наиболее широко используемым [6]. Следует ожидать, что в будущем AGI в большей степени будет вовлечен в выполнение задач на идеологическом фронте для управления социальными процессами на рынке труда.

Также стоит отметить, что руководство каждого государства вынуждено искать баланс между собственной независимостью и степенью участия в международном разделении труда. Чем выше уровень импортозамещения, тем меньше зависимость от других государств. Роль развития технологий ИИ в том или ином государстве в мировом разделении труда, очевидно, будет определяющей.

Сложно предсказать последствия создания AGI, способного решать любые доступные человеку задачи, включая управленческие, на социальном уровне. Это не обязательно приведёт к сингулярности (бесконечному росту возможностей AGI и устранению человечества из прогресса цивилизации), но создание даже дружественного (полезного) AGI будет иметь и негативные стороны для верхушки социальной иерархии. Пропадёт моральное обоснование «элитарности» руководящих слоев, и как следствие, понизится привлекательность государственных руководящих должностей, поскольку возможности учета своих личных интересов при приёме управленческих решений сведутся к нулю. Нельзя исключить, что в однополярном мире и AGI тоже будет единой планетарной системой. Более перспективным представляется развитие агентского подхода к AGI, когда сильный ИИ реализован в виде множества независимых устройств AGI, взаимодействующих друг с другом при решении сложных задач. Независимые и менее мощные системы AGI будут не только легче поддаваться контролю, но и смогут осуществлять взаимный контроль. Кроме того, процесс их самосовершенствования тоже проще направлять в сторону развития дружественности и полезности для человеческой цивилизации.

Возможности влияния AGI на политику и экономику

Для AGI в качестве основных обсуждаемых последствий внедрения можно выделить следующее: вытеснение человека из сферы выполнения рутинных работ; плохая объяснимость принимаемых ИИ решений; потенциальная возможность контроля за действиями руководителей; возможность потери контроля за действиями AGI.

Замена человека автоматами, которые без устали выполняют рутинную работу, совершая при этом меньше ошибок и снижая издержки производства, ни у кого возражений не вызывает. Но если массово заменить сотрудников на автоматы, то это создаст мощную волну безработицы и выльется в социальный кризис. Основным путём решения этой проблемы предлагается выплата безусловного базового дохода (ББД или основного дохода, БОД, basic income) с вытеснением населения из экономики. Высокая экономическая эффективность решений на основе ИИ позволяет осуществлять выплату ББД. Критики подхода, в качестве альтернативы ББД, указывают на возможность массового повышения образовательного уровня и привлечения населения к творческому труду. При использовании альтернативного образовательного подхода безликая деградирующая масса превращается из ненужной обузы в мощную производительную силу с ростом естественного интеллектуального потенциала. Данный путь не означает отказа от повышения социальных мер поддержки населения, но даст странам, выбравшим развитие творческого потенциала населения вместо выплаты ББД, заметные преимущества в международной конкуренции.

Проблема объяснимости решений, которые вырабатываются с помощью ИИ, на современном этапе рассматривается как одно из важнейших ограничений, препятствующих его внедрению. В задачах, где ошибочные решения могут приводить к тяжёлым последствиям, таким, например, как начало ядерного конфликта, авария на АЭС, наказание невиновных в суде, смерть пациента при проведении операции – ошибки считаются недопустимыми. Утверждения, что применение ИИ на основе нейронной сети вызывает беспокойство, так как такая система не дает ответа о причинах принятия решения, не соответствуют истине. Современные нейросетевые технологии могут обучиться не только генерировать решение, но и создавать объяснение к этому решению. Но, как и у человека, такое объяснение не обязательно будет правильно описывать причины выбора такого решения.

Технически сохранению сотрудничества между живыми людьми и AGI будет способствовать развитие агентского подхода к AGI. В этом случае отдельные агенты AGI будут вливаться как часть в современные «сверхразумы» – государственные, общественные и коммерческие структуры и смогут органично развивать систему сдержек и противовесов, которая существует в сегодняшнем мире. Конкуренция может быть

распространена на взаимодействие агентов AGI, включая осуществление контроля за их деятельностью, что снизит риски потери контроля.

Достижение сверхчеловеческих возможностей AGI позволит применять его не только в решении технических задач, но и при определении целей развития науки и социума. Отказ от использования AGI в данных вопросах приведёт к проигрышу в конкурентной борьбе. К преимуществам использования AGI можно отнести: участие в экономическом соревновании; более обоснованный выбор экономических и политических решений. К проблемам – зависимость социальных успехов применения ИИ и AGI от типа стратегии управления, выбранной властными структурами. Например, есть альтернатива между выплатой ББД, которая приведёт к вытеснению людей из процесса развития цивилизации с дальнейшим сокращением численности людей и их деградацией, и развитием творческого потенциала населения для участия вместе с AGI в научно-техническом и социальном развитии общества.

Заключение

Можно обозначить основные задачи в развитии технологий ИИ и AGI.

1. Приоритетная поддержка развития ИИ и AGI как технологий, дающих мощный импульс ускорению научно-технического прогресса. Создание условий для поддержки исследователей и секторов, внедряющих пилотные технологии ИИ в социум. Эти идеи получили развитие в ряде национальных программ по развитию ИИ, в том числе в указе президента РФ №490 от 10.10.2019 [7].
2. Формирование взвешенного подхода к использованию ИИ и AGI и ограничения в использовании этих технологий при управлении вооружениями, опасными объектами и в сферах, где ошибки управления могут создать угрозу жизни и свободе человека, что подробно рассматривается, например, в [8].
3. Выработка правил использования ИИ и AGI в конкурентной борьбе. Создание юридической базы для агентов AGI.
4. Интенсификация развития условий для реализации творческого потенциала населения, так как роль творческих профессий будет нарастать по мере создания и внедрения AGI.

Литература

1. *Малинецкий Г.Г.* Цифровая реальность в точке бифуркации и стратегические задачи Союзного государства в контексте гуманитарно-технологической революции // Проектирование будущего. Проблемы цифровой реальности: труды 2-й Международной конференции (7-8 февраля 2019г., Москва). – М.: ИПМ им. М.В.Келдыша, 2019.

2. ru.wikipedia.org/wiki/Дартмутский_семинар#Основные_положения, www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
3. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems, 2013. arXiv:1310.4546
4. *Минский М., Пейперт С.* Персептроны. – М.: Мир, 1971. – 261 с
5. *LeCun Y.* Power & limits of deep learning // 13 Nov. 2017, https://www.youtube.com/watch?v=0tEhw5t6rhc
6. *Харари Ю.Н.* Homo Deus. Краткая история будущего. – М.: Синдбад, 2018. – 320 с.
7. Указ Президента Российской Федерации от 10 октября 2019 г. №490 «О развитии искусственного интеллекта в Российской Федерации»
8. *Тегмарк М.* Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта. – М.: Издательство АСТ, CORPUS, 2019. – 560 с.