Эпистемология и философия науки
2020. Т. 57. № 2. С. 192–207
УДК 167.7

Epistemology & Philosophy of Science
2020, vol. 57, no. 2, pp. 192–207
DOI: https://doi.org/10.5840/eps202057230

Перспективы человека в эпоху технологической сингулярности

Смолин Владимир Сергеевич – научный сотрудник.
Институт прикладной математики им. М.В. Келдыша РАН.
Российская Федерация, 125047, г. Москва, Миусская пл, д. 4;
e-mail: smolin@keldysh.ru

Книга Макса Тегмарка обращает внимание на опасности и преимущества, ожидающие человечество в результате развития технологий искусственного интеллекта (Artificial Intelligence, AI). Космолог и астрофизик Тегмарк, понимая невозможность предсказать последствия развития AI, предлагает захватывающие сценарии вариантов развития цивилизации на десятки, сотни, тысячи, миллионы и миллиарды лет. Анализ противоположных сценариев направлен на формирование мысли, что последствия создания сильного, превосходящего человеческий уровень AI будут значительнее, чем от всех остальных достижений цивилизации. Тегмарк – один из основателей и лидеров движения «За дружественный AI», он излагает результаты обсуждения поднимаемых им вопросов с ведущими специалистами в области AI. Завершает свою книгу Тегмарк призывом к оптимизму: «Своей книгой я убеждаю подумать, какого будущего вы бы *хотели*, а не о том, какое будущее вас *пугает*, только так мы можем найти общие цели, ради которых стоит трудиться».

Ключевые слова: искусственный интеллект (ИИ), сильный ИИ, дружественный ИИ, утопия, антиутопия

The prospects of the mankind in the era of technological singularity

Vladimir S. Smolin – research fellow.
Keldysh Institute of Applied Mathematics, Russian Academy of Sciences.
4 Miusskaya Sq., 125047 Moscow, Russian Federation;
e-mail: smolin@keldysh.ru

The book by Max Tegmark draws attention to the dangers and benefits that await humanity as a result of the Artificial Intelligence (AI) technologies development. Cosmologist and astrophysicist Tegmark, realizing the impossibility to predict the AI development, offers exciting scenarios of civilization development options for tens, thousands, millions and billions of years. The analysis of the opposite scenarios is aimed at the idea formation that the consequences of creating a general AI, superior to the human level, will be more significant than from all other achievements of civilization. Tegmark is one of the founders and leaders of the “Beneficial AGI” movement, he presents the results of the discussion of the issues he raises with leading experts in the field of AI. Tegmark concludes his book with a call to optimism: “My book urge you to think about what future you *would like*, and not what future *scares* you, this way we can find goals for which it’s worth working”.

Keywords: artificial intelligence (AI), general AI, beneficial AGI, utopia, dystopia

Книга Макса Тегмарка[1], шведско-американского космолога и астрофизика, профессора Массачусетского технологического института и члена научной дирекции Института фундаментальных проблем, вышла в 2017 г. Русскоязычная версия появилась двумя годами позже, но, несмотря на стремительный прогресс в области AI, ее актуальность совсем не уменьшилась.

Дело в том, что книга Макса Тегмарка посвящена не современному прикладному AI, решающему узкие практические задачи, достижений которого с каждым месяцем становится все больше. Она рассматривает опасности и преимущества, которые могут ожидать человечество в результате развития технологий AI. Пусть успех AI в прикладных областях, таких как создание «умных» электронных помощников, распознавание и обработка речи и изображений, перевод, управление автомобилями и ряде других, за 2 года значителен. Однако представления о путях развития человечества в эпоху AI все еще находятся в стадии формирования, хотя книга Тегмарка и оказала определенное влияние на эти представления.

В космологии и астрофизике рассматриваются различные сценарии развития вселенной на многие миллиарды лет вперед и назад. В инженерных науках и социологии тоже делаются прогнозы развития, но речи о миллионах и даже тысячах лет в них не идет. Относительно краткосрочные (в сравнении с космологией) 50–100-летние прогнозы в данных областях, как правило, неточны. Ускорение же научно-технического прогресса, особенно с учетом использования искусственного интеллекта (Artificial Intelligence, AI), делает все более сложными для понимания пути развития человеческой цивилизации.

Тем не менее Тегмарк, не споря с невозможностью предсказать последствия развития AI, пытается применить свой опыт космолога и астрофизика в новой для него области технических и социологических прогнозов. Рассматриваемые им сценарии охватывают сотни, тысячи, а в плане расселения человечества по галактикам и контактов с внеземными цивилизациями – миллионы и миллиарды лет. Это обстоятельство позволяет отнести книгу к научно-популярной и научно-фантастической литературе, цель которых не столько показать результаты научных исследований, сколько разжечь интерес читателей к рассматриваемой проблеме.

Вторая цель удалась Тегмарку с блеском: книга читается легко, на одном дыхании. В зависимости от собственных взглядов читателя автор предлагает ему различные маршруты движения по тексту

[1] *Tegmark M.* Life 3.0: Being Human in the Age of Artificial Intelligence. New York: Alfred Knopf Publ., 2017. 364 p. *Тегмарк М.* Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта / Пер. с англ. Д. Баюка. М.: АСТ: CORPUS, 2019. 560 с. ISBN 978–5–17–983126–6.

книги, что позволяет различным группам читателей лучше преодолевать барьеры понимания проблем эпохи AI.

Основной опасностью от технологий AI принято считать последствия создания сильного искусственного интеллекта (Artificial General Intelligence, AGI), и Тегмарк придерживается данных взглядов. Само название – Жизнь 3.0 – отражает направленность книги на изучение вопросов, связанных с AGI. Каждый живой организм Тегмарк разделяет на две составляющие: хард – тело, состоящее из атомов, и софт – информация, состоящая из бит, которые, впрочем, тоже хранятся в структурах из атомов. В книге вводится классификация форм жизни по трем стадиям в зависимости от ее способности к самодизайну:

1. Жизнь 1.0 (биологическая стадия): эволюция «харда» и «софта».

2. Жизнь 2.0 (культурная стадия): эволюция «харда» и дизайн большей части «софта».

3. Жизнь 3.0 (технологическая стадия): дизайн и «софта», и «харда».

Промежуточные варианты, вроде Жизнь 1.3 или Жизнь 2.5, в книге не рассматриваются (хотя варианты 1.1 и 2.1 – упоминаются). Более того, анализируются последствия не столько создания, сколько «правильного» программирования только определенного типа «харда» – средств обработки информации, превосходящих интеллектуальные возможности человека при решении любых задач.

Именно такие средства обработки информации являются принципиальным элементом, отсутствие которого в настоящее время не позволяет перейти к Жизни 3.0. Способность превосходить интеллектуальные возможности человека при решении любых задач является одним из наиболее распространенных толкований понятия AGI. Предполагается, что появление AGI даст мощный толчок ускорения познавательным и творческим процессам нашей цивилизации, поскольку будут развиваться не только схемы познания и творчества, но и «хард», на котором эти схемы реализуется. Причем каждая последующая версия «харда» будет создаваться все быстрее. Поскольку в отличие от ситуации, когда новые версии техники создаются людьми с одинаковыми способностями, AGI будет использовать всё лучшие версии «харда» собственной реализации для создания еще более совершенных версий.

Теория резкого, лавинообразного ускорения научно-технического прогресса после появления AGI, имеющего лучшие способности к самосовершенствованию, чем возможности дальнейшего усовершенствования AI коллективом инженеров и разработчиков, его создавшим, носит название «технологическая сингулярность». Положительная обратная связь в цепной реакции самосовершенствования AGI, согласно данной теории, приводит к бесконечному росту скорости развития AGI. Это, по разным оценкам, за несколько лет, месяцев или даже часов делает недоступными для понимания действия и планы AGI как отдельным человеком, так и человечеством в целом.

Соответственно, наступление технологической сингулярности принято рассматривать как горизонт событий, до которого прогнозы имеют некоторое осмысленное значение, а после него – нет.

Далеко не все ученые придерживаются теории технологической сингулярности даже среди тех, кто считает в принципе возможным создание AGI. Справедливо отмечается, что в природе пока не удалось наблюдать неограниченно возрастающих величин, и у процессов познания и самосовершенствования должны проявиться какие-нибудь свойства, которые не позволят им развиваться бесконечно быстро.

Макс Тегмарк тоже скептически относится к теории технологической сингулярности, хотя и без явной критики упоминает ее в своей книге. Это можно утверждать, поскольку, во-первых, он указывает, что возможности самосовершенствования ограничены физическими законами. Во-вторых, большинство рассмотренных в книге сценариев относится ко времени после создания AGI, т.е. автор неявно отрицает возможность быстрого достижения технологической сингулярности и связанного с ней горизонта событий. И, в-третьих, Тегмарк высказывает компромиссную идею, что после создания AGI резкое ускорение прогресса познания случится, но результатом такого ускорения будет не бесконечный рост знаний, а быстрое достижение практически полных знаний о физических законах в нашей вселенной и способах их оптимального использования. В частности, он предполагает, что при встрече двух различных цивилизаций с развитым AGI (т.е. Жизнью 3.0) их уровни развития будут примерно равны, поскольку обе будут близки к верхнему пределу познания.

В качестве введения Тегмарк предлагает «Сказание о команде “Омега”» – исследовательской группе, написавшей программную систему, обладающую возможностями AGI и пытающуюся использовать преимущества AGI при сохранении контроля над ним. Данная история неявно предполагает, что: а) возможностей вычислительных средств, которые могут быть доступны одной корпорации (сейчас или в обозримом будущем), достаточно для создания AGI, надо только написать правильные программы; б) доступной в Интернете информации достаточно для обучения AGI и, наконец, в) первый же вариант программной реализации AGI будет настолько успешным, что не потребует никаких доработок в процессе его контролируемой эксплуатации.

Впрочем, по современным представлениям, AGI должен обладать способностью решать любую задачу лучше человека. И если поручить AGI совершенствовать свой код, то именно успешность в решении данной задачи может служить критерием появления реально работающего AGI, а не его пробных отладочных версий. Отслеживать же успешность и направленность усовершенствований кода даже коллективу разработчиков будет сложно, поскольку предполагается, что

AGI решает задачи лучше человека не на доли процента, а на несколько десятичных порядков.

Тегмарк в своем «Сказании о команде “Омега”» решает эту проблему, помещая AGI в «песочницу», не имеющую прямого выхода в Интернет для совершенствования своего кода (который, после улучшений, команда «Омега» не может воспринять и оценить). Все запросы и получаемая AGI из Интернета информация контролируются командой «Омега». Интрига заключается в том, что и «Омега», и AGI понимают, что в отсутствие контроля не только развитие AGI пойдет значительно быстрее, но и прикладное использование AGI будет эффективнее. Но нет никаких гарантий, что после снятия контроля команда «Омега» сможет продолжать извлекать выгоду от использования созданного ими детища.

Сложность же сохранения контроля состоит в том, что по мере развития AGI он становится умнее не только отдельных членов, но и всей команды «Омега» в целом. Тем не менее, согласно «Сказанию» команде «Омега» это удается и, планомерно наращивая использование AGI в различных областях деятельности, она скрытно захватывает управление сперва мировой экономикой, а затем и политикой. Тегмарк заключает рассказ словами: «“Омега” завершала наиболее драматическое преобразование в человеческой истории. Впервые на всей планете устанавливалась единая власть, мощь которой многократно усиливалась интеллектом столь могучим, что он мог бы обеспечить процветание жизни на Земле и в окружающем нас космосе на миллиарды лет. Но в этом ли состоял ее план?».

«Сказание о команде “Омега”» – это только один из огромного множества сценариев развития человеческой истории и далеко не самый вероятный. Тегмарк задает читателю вопросы: а как вы представляете себе пути и цели развития человечества в эпоху AGI? Эти риторические вопросы позволяют ему перейти к рассмотрению достаточно большого числа различных сценариев. Естественно, рассматриваемые сценарии не охватывают и малой части всех возможных сценариев, но позволяют высказать несколько интересных идей и обсудить некоторые из распространенных бытовых представлений и мифов об AGI.

Задача оценки, какие из рассматриваемых сценариев более вероятны, в целом не ставится, хотя Тегмарк и понимает важность данного вопроса. Так, например, в книге приводится расчет вероятности ядерной войны на взаимное уничтожение. На основе предположения, что вероятность такой войны в течение одного года составляет 0,001, по элементарной формуле вычисляется вероятность того, что ядерная война случится в ближайшие 10 000 лет (ответ – 99,95%). Но к остальным рассматриваемым в книге сценариям элементарные вероятностные формулы, как правило, неприменимы.

Кроме разбиения своей книги на 8 глав, Тегмарк разделяет ее на 2 части – История разума и История смысла – и оценивает степень

спекулятивности рассмотренных в книге тем. Все оценки Тегмарка варьируются от «не очень спекулятивно» до «исключительно спекулятивно», причем в Истории разума больше оценок «исключительно спекулятивно», а в Истории смысла больше «не очень спекулятивно».

Как настоящий космолог, Тегмарк не может не начать Историю разума с момента Большого взрыва 13,8 млрд лет назад. Тем не менее он делает интересное для астрофизика заявление, что до появления жизни процессы эволюции вселенной были бессмысленными, и если жизнь исчезнет, то рефлексия, красота, надежда, цели и смыслы снова пропадут.

Также Тегмарк определяет в качестве живого любой объект, способный к размножению и самосовершенствованию. Совершенствование как харда, так и софта живого организма может быть случайным, эволюционным, а может и целенаправленным, что, по Тегмарку, определяет принадлежность организмов к стадиям Жизнь 1.0–3.0. Наиболее эффективной является Жизнь 3.0, которая позволяет производить осмысленные изменения как софта, так и харда. Жизнь 3.0 пока не создана, но мы находимся на пути к ее созданию. Единого мнения, когда это произойдет и каковы будут последствия, пока не сформировалось.

Тегмарк выделяет несколько групп специалистов в области AI по их характерным ответам на два вопроса:

1. Когда искусственный интеллект превзойдет человеческий уровень? (ответы: через несколько лет; от 20 до 100 лет; от 100 лет до никогда).
2. Если сверхчеловеческий искусственный интеллект возникнет, будет ли это хорошо? (ответы: плохо; непредсказуемо; хорошо).

Большую группу специалистов Тегмарк относит к «техноскептикам», которые считают, что AGI будет создан очень не скоро (более чем через 100 лет или вообще никогда) и обсуждать, плохо это или хорошо, не имеет смысла.

В прогнозный период от 20 до 100 лет Тегмарк поместил 3 группы специалистов: «луддиты», «движение за дружественный AI» и «техно-утописты». Если луддиты считают AGI абсолютным злом, а техно-утописты – абсолютным добром, то представители движения за дружественный AI придерживаются взвешенной позиции: все зависит от того, какой AGI будет создан, а также кто и как будет его контролировать. Сам Тегмарк не просто относит себя к движению за дружественный AI, но и проводит активную деятельность по организации и руководству данным движением. Во временной градации отрезку «через несколько лет» сопоставлена категория «практически никто». Это одно из редких мест книги, которое успело за 2 года устареть. Если в 2016 г. считалась смелой оценка Рэя Курцвейла – 2045 г., то сейчас многие исследователи помещают прогнозную дату

создания AGI в отрезок между 2025 и 2035 гг., а некоторые – еще раньше, так что категория Тегмарка «Практически никто» активно заселяется. Тем не менее Тегмарк правильно отмечает, что все эти оценки (оптимистические и пессимистические) строятся не на строгих научных прогнозах, а на основе приверженности авторов прогнозов к различным, распространенным в области AI мифам. Автор приводит с десяток таких мифов и анализирует, почему содержащиеся в них утверждения нельзя считать достоверными.

Другой проблемой, связанной с AGI, является отсутствие общепринятых определений таких терминов, как «жизнь», «разум» и «сознание». Тегмарк находит изящное решение данной проблемы: приводит таблицу определений такого рода понятий, которая относительно ясно описывает, в каком смысле эти понятия используются в данной книге.

Суммируя все сложности с мифами и определениями в области AGI, Тегмарк приходит к выводу, что, хотя абсолютно уверенно утверждать о неизбежности создания AGI в обозримом будущем нельзя, вероятность такого события достаточно высока. И поскольку оно повлияет на судьбы человечества даже больше, чем промышленная революция, необходимо заранее пытаться найти ответ на вопрос: на что будет похожа жизнь человечества в эпоху AGI? Ведь это не нашествие инопланетян, а продукт деятельности человека. Понимание того, какие свойства AGI желательны для человечества, а какие нет, может позволить создать дружественный AI, что является центральной идеей развиваемого Тегмарком одноименного движения. И основной объем книги посвящен как раз обсуждению данных вопросов.

Но прежде чем приступить к их рассмотрению, Тегмарк проводит достаточно подробный ликбез по вопросу: как грубая физическая материя может породить нечто представляющееся настолько эфемерным, абстрактным и бестелесным, как разум? Основное направление ответа Тегмарка сводится к тому, что информационные процессы протекают на устройствах, созданных из материи, но являются субстрат-независимыми, т.е. одинаковые информационные процессы могут выполняться на разных физических носителях. Следовательно, разумными могут быть не только люди, но и машины. Несколько высокопарно Тегмарк называет этот ликбез «Основы теории разума», привлекая к обсуждению темы некоторые данные из теорий алгоритмов и нейросетевой обработки информации.

Для построения далеких прогнозов на будущее AGI полезно представлять сегодняшние успехи в области узкого прикладного AI и сформировавшиеся тенденции. Читателю предлагается достаточно широкая подборка материалов по применению AI в различных областях, таких как обработка естественных языков, космические исследования, финансы, производство, транспорт, энергетика, здравоохранение и связь. Отдельное внимание обращается на применение AI

для военных целей и в судебной практике. В отличие от вышеперечисленных областей, где AI используется исключительно «во благо», военная и судебная системы имеют целью наказание «нехороших» людей, и ошибки в их определении могут стоить особенно дорого.

Значительное внимание уделяется анализу успехов программ AlphaGo и AlphaZero фирмы DeepMind (основные успехи AlphaZero состоялись в 2018 г., но в русскоязычное издание книги Тегмарка, вышедшей в 2017 г., вошли). Тегмарк указывает (и это – распространенное мнение), что победы данных программ продемонстрировали не только мощь современных нейросетевых методов машинного обучения, но и возможность превзойти человека при решении сложных интеллектуальных задач. При этом побеждены были не только ведущие игроки в го и другие интеллектуальные игры, но и все игровые программы, написанные традиционными методами с использованием человеческих знаний.

Активно расширяющееся поле применения AI не должно скрывать связанных с этим проблем. Это не только негативные последствия возможных ошибок в тексте программ и злонамеренное использование AI, но и вытеснение людей из ряда областей трудовой деятельности, а также другие, возникающие по мере внедрения AI этические вопросы. Тегмарк призывает не ждать возникновения негативных последствий и переходить от реактивного поведения к проактивному, т.е. как минимум пытаться предвидеть возможные опасности и заранее принимать меры к их недопущению. В качестве одной из превентивных мер Тегмарк указывает на необходимость заключения международных соглашений о запрещении разработки и производства автономного летального оружия, чтобы избежать появления убивающих автоматических устройств, легкодоступных для всех, у кого есть немного денег и достаточно ненависти.

И это всё – уже реалии сегодняшнего дня, когда принципиальный успех – создание AGI – еще не достигнут. Что же будет после этого? Для ответа на этот вопрос Тегмарк возвращается к содержащемуся во введении «Сказанию о команде “Омега”». Он анализирует возможные варианты развития легенды, такие как захват “Омеги” правительственными или преступными силами, утрату контроля над AGI и ряд других. Особенно подробно Тегмарк рассматривает причины, почему AGI будет склонен выйти из-под человеческого контроля. И почему, имея интеллектуальное превосходство, он сможет это сделать.

Важным является вопрос о сохранении многополярности мира. Общая тенденция развития цивилизации приводит к уменьшению количества центров власти – на месте раздробленных княжеств появляются крупные государства и империи. Это связано с развитием технологий. Повышение производительности труда и развитие средств связи создают условия для управления все более крупными

структурами. Создание AGI будет очередным технологическим скачком, который может привести к формированию единого центра власти и однополярного мира. Но возможны и сценарии с сохранением многополярности. Это зависит, кроме прочих причин, и от скорости создания и развития AGI. При высокой скорости больше шансов получить однополярный мир, а при низкой скорости у первого появившегося AGI до полного захвата им экономической и политической власти могут успеть возникнуть несколько конкурирующих AGI. В свою очередь, на скорость создания и развития AGI влияет наличие сопротивления и оптимизационной силы. После создания AGI оптимизационная сила должна заметно возрасти. Но произойдет это не сразу, а после того, как стоимость применения AGI для решения прикладных задач окажется ниже использования человеческого труда, что мы уже сейчас наблюдаем в отношении AI.

Тегмарк предполагает, что в эпоху AGI во избежание «овощного» существования люди должны будут превращаться в «киборгов» или «заливки». Киборг – это человек, сращенный с высокотехнологичными устройствами, причем это не только физические, но и информационные гаджеты. Смартфон с мобильным Интернетом в вашем кармане – начало пути к киборгам.

«Заливки» же пока целиком относятся к области фантастики. Идея состоит в том, что информационные процессы, проходящие в вашей центральной нервной системе, можно прочитать и воспроизвести на технических устройствах другой физической природы со значительно более высокими быстродействием и объемом памяти и возможностями обмена информацией.

Но ни киборги, ни «заливки» не представляются магистральным путем развития AGI. История развития техники показывает, что снятие ограничений, связанных с кодом ДНК, позволяет получать более простые, эффективные и дешевые технические решения. AGI будет развиваться по своим, пока не понятным нам законам. Мы же можем рассматривать только те варианты организации общества в эпоху AGI, которые доступны нашему разуму.

Тегмарк предлагает список из 12 вариантов социального устройства. Он начинает с «либертарианской утопии», когда люди, киборги, «заливки» и AGI мирно сосуществуют, и «шлюзовой обороны», когда единственный AGI следит только за тем, чтобы второй AGI не был создан, и прогресс фактически остановлен. Последние варианты – «победивший AGI», когда человечество быстро вымирает за ненужностью, и «самоуничтожение», когда AGI никогда не был создан по причине гибели человечества от другой катастрофы (скорее всего, техногенной). В целом, если технический прогресс не будет остановлен, то AGI обязательно создадут и с проблемой AGI человечество столкнется. И Тегмарк продвигает идею, что если большинство проблем

сосуществования с AGI не будет решено до его создания, то судьба человечества печальна.

В описании различных вариантов социального устройства встречаются фразы типа «Разумные машины во всем приходят на смену людям», которые можно воспринять так, что AGI – это не единственное интегрированное суперустройство, а как минимум несколько относительно независимых «разумных машин». Но по контексту понятно, что это люди, киборги и «заливки» – множественны и могут иметь конкурирующие потребности, а сверхразум – один, и даже если аппаратно он реализован в виде нескольких независимых устройств, то, по Тегмарку, никаких разногласий между данными устройствами возникать не должно. Более того, даже если в космосе встретятся две или более Жизни 3.0, то их интересы и цели должны оказаться непротиворечивыми. Поскольку их понимание вселенной будет близко к абсолютному знанию и, следовательно, представления о добре и зле тоже должны совпасть.

В числе 12 рассматриваемых Тегмарком вариантов социального устройства будущего человеческого или потомственного ему машинного общества только 8 предполагают существование сверхразума. Но и эти 8 сценариев достаточно сильно различаются между собой. Даже Тегмарк называет одни сценарии утопическими, а другие – антиутопическими. К сожалению, Тегмарк не анализирует, каким образом могут совпасть интересы и цели в случае встречи во вселенной двух цивилизаций с Жизнью 3.0, имеющих сильно отличающиеся социальные устройства.

Рассмотрению вопроса, какие цели может ставить себе AGI, в книге посвящена отдельная глава. Но еще до этой главы космолог Тегмарк как безусловную рассматривает цель сверхразума расселиться в как можно большей части вселенной. Если в области познания Тегмарк считает, что AGI способен сколь угодно близко подойти к пределам познания, ограниченным физическими законами, то в области космологии, где он является специалистом, находит ряд ограничений, которые не позволят никакому AGI даже приблизиться к оккупации всей вселенной. Тем не менее предполагаемые Тегмарком возможности Жизни 3.0 по расселению во вселенной заметно превосходят (как и всё в космологии…) представления читателей, далеких от тематики. Тегмарк рассматривает заселение не только других звезд, но и галактик и более крупных образований во вселенной.

Безусловность стремления Жизни 3.0 к захвату максимально большей части вселенной Тегмарк выводит из того, что для достижения любой цели необходимо обладать ресурсами, которые в ограниченной части вселенной не могут быть бесконечными. Хотя, как космолог, Тегмарк демонстрирует невозможность оккупации всей вселенной (даже в случае, если она не окажется бесконечной), но увеличение освоенной части приводит к росту доступных

ресурсов и позволяет ставить и достигать больше разнообразных целей.

Именно захват максимально большего объема вселенной обеспечит сверхразуму вещество для харда и энергию для обработки информации, которые необходимы для увеличения главной ценности Жизни 3.0 – знания. Рост познаний позволит осуществлять не только экстенсивное (заселение новых миров), но и интенсивное (повышение эффективности процессов) развитие Жизни 3.0. В данном направлении Тегмарк тоже находит физические ограничения: повышение эффективности возможно, но для извлечения энергии – «только» в 10^{10} раз, плотности хранения информации – в 10^{12}–10^{18} раз и быстродействия ее обработки – в 10^{31}–10^{41} раз. И это только нижние оценки – рост знаний должен позволить увеличить приведенные оценки. Хотя нельзя исключить и обратного – познавательный процесс приведет к выявлению новых, неизвестных сегодня физических ограничений (как ранее случилось, например, со скоростью света), которые могут понизить приведенные Тегмарком оценки.

Впрочем, многие вопросы, рассматриваемые в «космологической» главе книги Тегмарка, хочется отнести к риторическим. На фоне сделанной им же оценки, что вероятность самоуничтожения человеческой или потомственной ей цивилизации в ближайшие десять тысяч лет составляет 99,995%, значительно более актуальным кажется анализ возможностей избежать самоуничтожения в ближайшие 1000, 100 и особенно – 10 лет. Это обстоятельство делает абстрактное качественное рассмотрение Тегмарком проблемы постановки целей более практически полезным, чем количественные оценки космологических проблем AGI.

При всей своей любви к космологии Тегмарк пишет, что, «если бы мне надо было одним словом выразить, в чем сложность AI-спора, я бы выбрал слово “цели”». От того, какие цели будут ставить перед собой не только AGI, но и человечество до его создания, в значительной степени зависит, в каком направлении будет развиваться наша цивилизация.

Анализ истории Тегмарк начинает с физических законов. В элементарной физике законы выражаются формулами эмпирически выявленных зависимостей между различными величинами. В теоретической физике эти же самые законы получаются как результат решения оптимизационной задачи. Как правило, минимизируется энергия и максимизируется энтропия. Тегмарк интерпретирует используемый в теоретической физике формализм как наличие у неодушевленной материи целей достичь состояний с минимумом энергии и максимумом энтропии.

Непонятно, как это соотносится с ранее приведенными утверждениями Тегмарка, что «до пробуждения Вселенной (после появления разумной жизни) никакой красоты не было. Если ей суждено

когда-то навсегда вернуться в свою дремоту, в силу какого-то космического бедствия или же нашего саморазрушительного безумия, она, увы, снова лишится всякого смысла». Или: «Если в далеком будущем в нашем космосе останется только один бесчувственный сверхразумный зомби, элегантность Вселенной станет неважной, никто не сможет ни наблюдать ее, ни переживать ее элегантности – космос станет огромным и бессмысленным пропащим местом»? До появления жизни смысла не было, а цели были? Слова «смысл» и «цель» имеют различные значения, но в одном из значений эти слова – синонимы. И глава «Цели» входит в раздел книги «История смысла». В том же разделе, в главе «Сознание» содержится параграф «Смысл». Но даже там вопросы взаимосвязи понятий «смысл» и «цель» Тегмарк не анализирует.

Не слишком удачным надо признать и определение Тегмарком интеллекта как способности достигать сложных целей. Что такое сложные цели? Установление энергетически минимального состояния галактики или теплового равновесия вселенной вряд можно назвать простыми целями, а неживая материя вполне способна их достичь. Следует ли из этого, что она обладает интеллектом? В книге только делается попытка установить, как связаны между собой понятия «смысл», «сознание», «интеллект», «цель» и другие, но удовлетворительных ответов на такие вопросы автор не находит.

Зато Тегмарк определяет, чем отличаются цели живой материи от неживой и разумной Жизни 2.0 от Жизни 1.0. Целью неживой материи является максимальная диссипация (достижение «тепловой смерти» вселенной). Жизнь 1.0 имеет целью репликацию (воспроизведение себе подобных), которая ускоряет диссипацию. Жизнь 2.0 преследует цель удовлетворения своих желаний, которые сформировались еще в бытность ее Жизнью 1.0, чтобы помочь осуществлять успешную репликацию. Машины, механизмы и прочие гаджеты прокладывают путь к Жизни 3.0 и создаются с целью лучшего удовлетворения желаний Жизни 2.0. Можно было бы предположить, что Жизнь 3.0 продолжит эту тенденцию, если бы человечество (Жизнь 2.0) не очень выборочно подходило бы к помощи различным формам Жизни 1.0 в вопросе репликации…

Тегмарк упоминает теорию конвергенции, которая предполагает, что чем сильнее развит разум, тем легче ему ставить цели и находить решения, которые могут удовлетворить всех. И если следовать данной теории, то с развитием AGI все меньше конфликтных целей должно им выдвигаться. Но при этом Тегмарк высказывает сомнение в состоятельности теории конвергенции, ссылаясь на Ника Бострёма, который в своей книге *Superintelligence* выдвигает тезис *ортогональности*, что уровень интеллекта и степень конфликтности целей – независимы. То есть можно быть умным и добрым, а можно – умным и злым.

На основе тезиса ортогональности Тегмарк обнаруживает конвергенцию не в будущем, а в прошлом, когда вся Жизнь 1.0 была нацелена на репликацию (хотя надо отметить, что уже тогда, согласно теории Дарвина, некоторые конфликты целей обеспечивали естественный отбор). Надежды же на конвергенцию целей человечества и AGI в будущем, по мнению Тегмарка, должны быть основаны не на вере, что конвергенция случится сама по себе, а на специально разработанных методах ее достижения. Они, к сожалению, пока нам не известны.

Для достижения конвергенции необходимо, чтобы AGI смог понять, принять и в дальнейшем придерживаться целей, согласованных с человеческими. Некоторые считают, что такими целями могло бы стать соблюдение человеческих этических норм. Но Тегмарк отмечает, что этические нормы работают только внутри человеческого общества, а на другие виды (животных) не распространяются. И не наблюдается серьезных оснований считать, что даже если AGI и будет придерживаться человеческих норм этики, то он в течение длительного времени будет сохранять их применение и на людей.

Тегмарк считает возможным заложить в AGI практически любую цель, но для ее достижения AGI обязательно породит вторичные цели, такие, например, как самосохранение, захват ресурсов и любопытство. Как бы хорошо основная цель AGI ни была согласована с человеческими целями, две первые из перечисленных вторичных целей потенциально могут создать проблемы человечеству, а третья – изменить основную цель.

Автор заканчивает главу, посвященную обсуждению целей, высказыванием: «Не понятно, как вдохновить сверхразумный AI целью, которая не приводила бы к истреблению человечества». Данная цитата показывает оценку автором степени решенности проблемы целей для AGI.

В последней главе своей книги Тегмарк пытается рассмотреть проблему сознания. Он отмечает, что «споры на эту тему порождают больше жара, чем света», а «содержание статьи о сознании в психологическом словаре Макмиллана 1989 года издания ограничивается фразой, что “ничего стоящего прочтения по этой теме не было написано”». Вроде бы каждый индивидуум обладает сознанием, но содержательно описать, что это такое, даже специалистам не удается. Известный нейробиолог Кристоф Кох рассказывал Тегмарку, как его отговаривали заниматься проблемами сознания, пока у него нет постоянной профессорской позиции. И не советовал Коху заниматься сознанием, не имея солидного веса в научном сообществе, нобелевский лауреат Фрэнсис Крик. Но к 2017 г. Макс Тегмарк уже являлся достаточно известным ученым, чтобы к его высказываниям по любой теме прислушивались, включая даже такую, как сознание.

Состояние проблемы сознания в вопросах построения AGI значительно хуже постановки целей для AGI. В книге предполагается, что мы не знаем только, как ставить правильные цели для AGI, но зато нам известно, что такое цели и зачем они нужны Жизни 3.0. В отношении же сознания есть целых 3 аспекта незнания: а) не знаем, как построить AGI, обладающий сознанием; по Тегмарку, нам также не известно; б) что такое сознание и в) зачем вообще сознание нужно разумной жизни.

Тем не менее Тегмарк считает, что после 1989 г. некоторый прогресс в данной области есть. В частности, он пишет: «Мало кто посвятил этому вопросу (сознания) больше усилий, чем Дэвид Чалмерс, известный австралийский философ, со своей неизменной улыбкой и кожаным пиджаком, который так нравится моей жене». По Чалмерсу, построение AGI столкнется с двумя основными проблемами: 1) простой, связанной с задачами построения различных алгоритмов преобразования информации, и 2) трудной, заключающейся в нашем непонимании природы сознания.

Вторую проблему Чалмерс подразделяет на: 2а) «довольно трудную проблему» (pretty hard problem, или PHP), какие физические особенности отличают обладающие сознанием системы от бессознательных?; 2б) «еще более трудную проблему» (even harder problem, или EHP), какие физические свойства определяют субъективные ощущения (квалиа)? и 2в) «по-настоящему трудную проблему» (really hard problem, или RHP), откуда берется сознание? Как настоящий философ, Чалмерс не пытается анализировать возможные пути решения выделенных им проблем, а только пытается доказать невозможность их решения, основываясь на своей вере, что мир зомби – представи́м (возможен мир животных и AGI, не обладающих сознанием, а действующих по заложенным в них правилам).

Тегмарк дает и свое определение: сознание = субъективные переживания. Если бы мы знали (или Тегмарк бы нам рассказал…), что такое субъективные переживания, то, несомненно, мы бы легко поняли и что такое сознание. Так, если мы знаем, что такое 5 кг, то формула $x = 5$ *кг* хорошо описывает нам свойства и значение величины x. Но определение Тегмарка больше соответствует формуле $x = y$, где и x, и y – неизвестные величины. Тогда формула выражает только утверждение, что величины x и y обладают одинаковыми свойствами. В случае определения сознания формула Тегмарка указывает на то, что все аспекты незнания, которые мы связываем с понятием сознания, могут быть отнесены и к субъективным переживаниям.

Тегмарк гордится тем, что его определение не содержит упоминания ни на поведение, ни на восприятие, ни на самоощущение, ни на эмоции или внимание. И, вроде бы, может быть применено как к биологическим, так и техническим системам. Но это – только игра слов. Сам же Тегмарк ниже пишет, что «определение оставляет

открытой (т.е. неопределенной) возможность обладать сознанием для будущих систем с искусственным интеллектом». Поскольку к субъективным переживаниям относятся и 3 аспекта незнания, и 3 уровня трудной проблемы Чалмерса, связанных с сознанием. Зато, при спокойном анализе, определение Тегмарка позволяет соотнести указанные выше аспекты и уровни.

Важнейшим отличием позиции Тегмарка от Чалмерса является вера Тегмарка в возможность решения перечисленных проблем научными методами (в чем Чалмерс высказывает серьезные сомнения). По Тегмарку, сознание только ощущается как нефизическое, поскольку «оказывается дважды независимым от субстрата». Не только представление обрабатываемой информации не зависит от объектов, которые она описывает, но и процесс обработки информации не зависит от физической природы используемых информационных средств. Но, как и всякий субстрат-независимый информационный процесс, сознание всегда должно базироваться на физических носителях, которые могут быть исследованы научными методами.

Следствием данных представлений является вывод, что сознание определяется особыми свойствами информационных процессов, а не наличием специальных полей или частиц. При этом, вне зависимости от физической природы, «информационная система в целом может быть автономна, но каждая ее часть в отдельности – нет».

Завершает главу о сознании Тегмарк следующими выводами:

- «Поскольку без сознания невозможен никакой смысл, то не наша Вселенная дает смысл сознающим существам, а сознающие существа дают смысл нашей Вселенной».
- «Так как людям предстоит уступить место умнейших существ во Вселенной машинам, мы должны свыкнуться с новой ролью – Homo sentiens (человек, субъективно ощущающий), а не Homo sapiens (человек разумный)».

В целом Тегмарку в большей степени удается заинтересовать читателя рассматриваемыми им проблемами, чем найти или хотя бы предположить пути их решения. Книга и хороша как раз тем, что она скорее ставит очень важные вопросы, чем дает на них ответы. Было бы удивительно, если бы человек, в зрелом возрасте сменивший область деятельности, за несколько лет нашел бы решение большого числа проблем в новой для себя тематике. Следует отметить, что книга Макса Тегмарка является не только результатом его размышлений над проблемами развития AI, но также отражает его обсуждения поднимаемых в книге вопросов с ведущими специалистами в области AI. Тегмарк также занимается организационной деятельностью: вошел в число пяти основателей Института будущего жизни (The Future of Life Institute (FLI)). На момент выхода книги в 2017 г. с командой сотрудников FLI провел две конференции по безопасному развитию AI с привлечением ведущих специалистов и ньюсмейкеров

в области AI и продолжает активно участвовать в поддержании и расширении деятельности FLI (https://futureoflife.org/) и после выхода книги.

В заключение следует отметить, что в качестве основной темы в книге рассматриваются вопросы возможных глубоких изменений в истории жизни после создания AGI. Катастрофические или экзистенциальные риски должны регулироваться соразмерными усилиями по планированию и смягчению их последствий. Возможности к самосовершенствованию AGI должны подвергаться самому строгому контролю в плане безопасности и управляемости. И, главное, это общее благо: «Сверхразум может создаваться только для служения всеми разделяемым этическим идеалам и во благо всего человечества, а не какого-то одного государства или какой-то одной организации».

Завершает свою книгу Тегмарк призывом к оптимизму: «На протяжении всей книги я убеждал вас прежде всего подумать о том, какого именно будущего вы бы *хотели*, а не о том, какое будущее вас *пугает*, потому что только так мы можем найти общие цели, ради которых стоит планировать и трудиться».