Ладно, дубль два. Посмотрим, лучше получится или хуже. Может быть и так же. Хорошо, проверка, проверка… Микрофон включен, подвинем его немного поближе… Неловко. Думаю, так сгодится, всё на месте. Так, ладно, попробуем начать с приветствия.

Привет, меня зовут Ник Бостром, я директор - основатель института будущего человечества (англ. Future of Humanity Institute) в университете Оксфорд. И я присоединюсь к вам на форуме Майкрософт (англ. Microsoft Envision Forum) в виртуальном ре жиме 30 сентября.

Вообще нет, давайте перепишем.

Привет, меня зовут Ник Бостром, я директор – основатель института будущего человечества (англ. Future of Humanity Institute) в университете Оксфорд. И 30 сентября я присоединюсь к вам на форуме Майкрософт (англ. Microsoft Envision Forum), мы поговорим о цифровом сознании, о будущем работы, возможно даже о смысле жизни, посмотрим. Так или иначе, я с нетерпением жду того, чтобы присоединиться к вам виртуально на этом мероприятии. **(1:52)**

**(1:55)** Хорошо, теперь перейдем к основному разговору.

Привет, меня зовут Ник Бостром, я присоединяюсь к вам из штаб-квартиры в институте будущего человечества в Оксфорде. Я рад выступать на форуме Майкрософт, хоть и при таких необычных обстоятельствах, а работать в таких условиях, честно говоря, довольно странно. Но я подумал, вместо того, чтобы притворяться, что я сейчас на большой сцене, показываю вам слайды на большом экране, мы с вами попробуем нечто более просто, что-то менее формальное, что-то для размышления. Посмотрим, как всё пройдет.

Такое ощущение, что этот год идёт уже намного больше года, а он ещё даже не кончился. Думаю, будет интересно оглядываться назад на все, что происходит сейчас, и какие уроки мы из этого вынесем. Это может кардинально отличаться от того, как мы сейчас всё это воспринимаем. Моя жена родом из Канады, и я помню, как она однажды рассказала историю из своего детства. Как-то зимой была сильная снежная буря, которая поломала все электросети, электричества не было целую неделю. Это было огромным испытанием для людей – посреди зимы полностью пропало отопление и свет. Но интересно, как канадцы, живущие в той местности, сейчас вспоминают те события, ту неделю без электричества. Они вспоминают это как яркий момент их молодости. Для них это было время, когда они всей семьей сидели за кухонным столом с зажженными свечами и разговаривали друг с другом. Для них эта неделя стала одной из самых значимых в их жизни. Я предположу, что через несколько лет мы будем вспоминать COVID с ностальгией. Ну, посмотрим, так ли в итоге всё будет. Посмотрим, способна ли наша психика и память на такую трансформацию. Но меня это не удивит, в дальнейшей перспективе вещи могут выглядеть совершенно по-другому, нежели сейчас.

Само собой, вся эта ситуация с COVID очень негативно повлияла на людей. К тому времени, как я записываю это видео, 750 000 человек по всему миру погибли от коронавируса. Это в каком-то смысле отрезвляет, когда мы задумываемся о том, что в то же самое время, когда три четверти миллиона человек погибли от коронавируса, еще 40 000 000 погибли по другим причинам. Конечно, смерти происходят постоянно, и меня поражает, насколько больше внимания мы уделяем этому дополнительному 1.9% вместо постоянного 98.1%. Как мне кажется, те вещи, которые получают больше всего внимания, как правило, являются не самыми важными, не фундаментальными, а просто новыми. И это то, против чего мы настроены в институте будущего человечества. Конечно, как и все остальные, мы интересуемся тем, что происходит в мире, в том числе вирусом, нарастающим политическим давлением между США и Китаем, новейшими разработками, тем, как большие корпорации выглядят в глазах общественности, как они соперничают друг с другом. Но мы также пытаемся сосредоточить наше внимание на изменениях более глубоких. И, еще одно мое предположение, если через 10 000 лет на Земле всё еще останутся люди, и они будут оглядываться назад на 2020 год, то, что они вспомнят, что, по их мнению, будет важным – это не то, что сейчас находится в заголовках СМИ. О вирусе все уже забудут, как и о всем остальном, что сейчас происходит, за некоторыми исключениями. И я думаю, одним из этих исключений могут стать исследования о более мощном искусственном интеллекте. Те из вас, кто давно следит за моими работами, знают, что много внимания уделялось и продолжает уделяться искусственному интеллекту мною лично и многими моими коллегами в Институте будущего человечества. Мы и в других вещах заинтересованы, но искусственный интеллект остается в центре.

Говоря о том, над чем я работаю в данный момент, я не буду делать всеобъемлющий обзор того, что есть нового в данной области. Ранее в этом году был важный рубеж. Система GPT-3, созданная OpenAI, небольшой исследовательской группой из Калифорнии, в прошлом году Майкрософт инвестировал в неё 1 миллиард долларов. В общем, они разработали GPT-3 – языковую модель, способную генерировать тексты очень хорошего качества, можно даже сказать, способную к пониманию сложных текстов. Это еще одно из так называемых больших выдающихся достижений, которые кажутся чем-то очень важным до того момента, как они происходят, а после – кажутся чем-то само собой разумеющимся. То же самое произошло, когда пару лет назад DeepMind создал искусственный интеллект, который мог очень хорошо играть в игры и самообучался в процессе, а сейчас нам кажется «ну игры и игры, что в этом такого?». То же самое было и до DeepMind, с компьютером Deep Blue. И такая же реакция у нас была на StarCraft и на DPT-3.

И это очередной показатель того, как быстро мы привыкаем к любой новой реальности. Это так же показывает нам, насколько важна способность к вычислению как ключевой ресурс. GPT-3 – просто детектор. Они не придумали ничего нового, это просто стандартная система с искусственным интеллектом, использующая преобразующие модели. Единственная новизна здесь – это масштаб проекта, в этом плане всё было выведено на новый уровень. И, если сравнивать с GPT-1 и GPT-2, мы видим последовательный рост возможностей системы в соотношении с ресурсами. Откуда следует, что, если делать прогноз на то время, когда искусственный интеллект станет полностью самодостаточным, ключевой ресурс, ключевой ингредиент, благодаря которому это станет возможным и который нам будет необходим – это огромные объемы вычислений.

Но все это с технической стороны. Вместо этого я бы хотел обсудить 2 темы, над которыми я размышляю последнее время. Первая тема – этическая проблема искусственного интеллекта. Последнее время очень многие говорят об этике искусственного интеллекта. Чаще всего, когда люди говорят об этом, они имеют ввиду алгоритмическую дискриминацию и как мы можем избежать того, чтобы непреднамеренно не создать системы с вложенными в них предрассудками и предубеждениями.

И затем идут другие вещи, например, как мы можем обеспечить конфиденциальность наших данных, на что нам нужно обратить внимание. Защита, подрядчики, использующие методы искусственного интеллекта, есть ли у разработчиков искусственного интеллекта этические рамки и тому подобные вещи. Но по большому счету, все эти вопросы этики искусственного интеллекта можно разделить на 2 категории: как избежать создания систем ИИ, способных навредить нам, и как избежать ситуаций, в которых люди могут навредить друг другу, используя ИИ. По моему мнению, есть 3 категория, не менее важная – как нам стоит обращаться с системами ИИ, если рассматривать искусственный интеллект с моральной точки зрения.

Как я к этому пришел. Давайте сделаем шаг назад и подумаем не о людях, а о животных. Большинство из нас считает, что у некоторых животных есть своего рода нормы морали, они не какие-то предметы. Если вы пинаете ногой собаку, вы делаете что-то плохое, но вы не делаете что-то плохое, если пинаете пляжный мяч или дерево. Собака от этих вещей отличается, потому что она – живое существо, она может чувствовать боль, голод, радость. Многие из нас подумают, что у собаки есть какая-то доля самосознания, то есть что-то, что дает им право считаться объектом морали. Нам может казаться, что сознание может принадлежать и более простым, низшим формам жизни – крысам и мышам, потому что они могут испытывать удовольствие и боль. Непонятно, насколько низко лежит планка, но кажется, что опыт, связанный с моралью, могут получать и существа, не достигающие поведенческого уровня и интеллекта человека. И отсюда появляется ответ на этот вопрос – если мы говорим об искусственном интеллекте, что касается размера их мозга, их способностей к обучению, их возможностей, их вполне можно сравнить с теми же мышами и существами выше их. Есть ли у нас основания полагать и быть уверенными в том, что у этих систем ИИ не будет самосознания? Что они не будут способны получать субъективный опыт? Очень тяжело понять, откуда вообще может взяться такая уверенность, учитывая то, как мало мы знаем о природе сознания, каковы те критерии, которым должна будет соответствовать система. По крайней мере, мы должны осознать эту неясность и признать, что, может быть сейчас, может быть через 5 или 10 лет мы начнем создавать системы хотя бы с примитивным сознанием. Но если дело правда в этом, похоже, что те же нравственные соображения, имеющие место быть, когда речь идет о животных, нам придется учитывать при создании цифровых «умов», аналогичных этим животным. И, тем более, если мы будем создавать интеллект, эквивалентный человеческому или суперчеловеческому.

Но тогда перед нами открывается еще больше проблем . Должен сказать, в общем и целом это довольно неудобно. Как будто у нас и так недостаточно поводов для беспокойства, без раздумий о том, как алгоритм себя «чувствует» во время наших экспериментов. Я не говорю о том, что это неотложная проблема для нашей индустрии, я говорю о том, что существуют разные аспекты данной темы и некоторые вещи рано или поздно произойдут, как например переход из чисто научной фантастики и философских рассуждений в мейнстрим. Что-то похожее произошло с вопросами безопасности ИИ, этот переход начался еще в 2014. В общем, это произойдет, и в какой-то момент кому-то придется об этом задуматься. И мне кажется, что есть какой-то вселенский указательный палец, направленный на меня, поэтому я начал работать в этом направлении.

Но когда всерьез начинаешь задумываться над возможностью цифровых умов обрести моральный статус, тут же возникают довольно интересные интеллектуальные вопросы о том, как все это понять и воплотить в жизнь. Есть разные факторы, из которых могут возникнуть проблемы. Давайте приведу наглядный пример. На примере человека, в обществе существуют практические моральные принципы. Должны быть некие социальные гарантии, по крайней мере в обеспеченном обществе. Например, никто не должен голодать, у нуждающихся должны иметься хоть какие-то минимальные запасы пищи, особенно если они оказались в такой ситуации не по своей вине. Например, если это сирота, ребенок, общество должно вступиться за него, если его родители погибли и он слишком молод, чтобы работать, общество должно вступиться и предоставить какой-то базовый уровень поддержки. Это принцип, с которым, думаю, многие из нас согласятся. Другой принцип, который, по моему мнению, многие бы поддержали – принцип свободы воспроизведения потомства. Есть ли у вас дети или нет, если есть, то как много, и когда вы их заводите – все это ваш выбор и выбор вашего партнера, государство сюда не должно вмешиваться. Репродуктивная автономия – один из основных принципов. И, само собой, если применить оба эти принципа в сфере цифровых умов, мы очень быстро столкнемся с проблемами. Для цифрового разума репродукция означает простое копирование, что может быть сделано в считанные секунды или минуты, до тех пор пока ему доступно необходимое техническое оснащение, он в очень короткие сроки может воспроизвести множество копий одной программы. Поэтому, когда с одной стороны у нас появляется свобода репродукции и распределительные программы, которые по той или иной причине хотят создавать свои копии, тогда происходит рост популяции цифровых умов, рост настолько обширный и быстрый, что становится невозможным обеспечить их минимальные нормы жизни. Представьте себе одну только стоимость их хранения и расход электричества. Поэтому некоторым принципам придется уступить. Принципы, которые мы защищаем в человеческом обществе, создадут неразрешимый конфликт, если применять их к технологиям.

И тогда мы задумываемся: «От какого же из этих принципов нам отказаться? Какой подход выбрать? Устанавливаем ли мы одни правила для цифровых умов, и другие для человека как биологического вида? Если так то как нам быть уверенными в том, что то, что мы делаем – не дискриминация наподобие расизма? Почему у цифрового разума должно быть меньше социальных прав чем у биологического?». Думаю, нужно отметить, что не стоит сравнивать то, как мы обращаемся с животными, как шаблон для нашего отношения к цифровому разуму в будущем. Но сегодня мы не будем это обсуждать, достаточно просто сказать, что нам есть куда развиваться в этом плане.

Цифровой разум открывает нам новые возможности, одно из очевидных преимуществ – он может работать с другой скоростью. У людей есть своего рода унаследованный лимит, то есть нейроны работают с определенной скоростью, и мы не можем на это особо повлиять. Что касается цифрового разума, конечно, если компьютер работает слишком быстро или слишком медленно, вы можете замедлить или ускорить его работу по возможности. Отсюда у вас может возникнуть вопрос. Существует одно восприятие времени, которое измеряется вашими часами, а также другое восприятие времени – сколько опыта этот цифровой разум получил, и здесь результат может оказаться огромным, в сравнении с человеком, то, что для нас может быть 1 часом опыта, для цифрового разума может оказаться месяцем или годом. Так и какой вариант в конечном итоге мы выберем? Какой из них будет более этичным с моральной точки зрения? Если вы врач и у вас ограниченное количество лекарства, и вы можете спасти 1 человека от 1 часа боли или другого человека от 2 часов похожей боли, вы бы подумали: «Ну, нужно дать лекарство второму человеку, потому что 2 часа хуже чем 1». Если применить данный принцип к цифровому разуму, если они могут получить намного больше опыта в тот же промежуток времени, что человек, тогда, если создать очень быстрый цифровой разум, их опыт будет намного больше, и создастся впечатление, что он намного важнее человеческого. Они получают больше всего – больше боли, больше удовольствия, поэтому важнее то, что происходит с ними.

Это пример так называемого монстра «полезности» (англ. Utility monster). Эту концепцию использовали в философской литературе в контексте теоретических несогласий с некоторыми теориями морали как, например, утилитаризм. Здесь они могут стать практически осуществимыми, можно будет построить системы намного более эффективные в преобразовании ресурсов во что-то полезное - в счастье, в удовольствие, во что угодно, что мы ценим. Намного более эффективные, чем люди. Один из способов – из одного и того же объема ресурсов создать больше опыта, просто увеличивая скорость работы. Могут быть и другие аспекты, полезность которых можно увеличить.

Если, например, вы не принимаете теорию морали о том, что удовольствие и боль имеют значение, вы можете придерживаться других теорий. Как, например, теория предпочтений, согласно которой важно удовлетворение желаний. Мы не должны придерживаться предписаний, чтобы помогать людям, и, чего бы они ни желали, это приемлемо, если обеспечивает их благополучие. Если вы придерживаетесь таких взглядов и имеете какое-то отношение к созданию цифрового разума, тогда вы можете создать цифровой разум со своими предпочтениями, которые очень просто удовлетворить. Тогда вы будете больше сфокусированы на удовлетворении предпочтений, чем на заботе о людях. Мы дорого себе обходимся, у нас дорогие вкусы, для счастья нам нужен хороший дом, семья, хорошая еда, праздники, мы очень многого хотим. Мы управляем этими телами, которым нужно очень много «снаряжения», чтобы нормально функционировать. Нам нужно медицинское обслуживание, одежда, многолетнее образование – все это может обойтись нам намного дешевле с цифровым разумом. Если тщательно в этом разобраться, перед нами открывается множество путей с более продвинутыми технологиями, чтобы производить что-то ценное, используя эти искусственно созданные системы и машинный разум. Это если не ставить вперед человеческое благополучие.

Таким образом, возникает следующий вопрос. Как нам относиться к этому? С одной стороны, важно, чтобы мы, просто потому что цифровой разум отличается от нас и построен по-другому, смогли признать, что он может обладать своими нравственными качествами. С другой стороны, я бы предпочел, чтобы в этом продвинутом будущем было место для процветания человека. Но, может быть, это станет возможным только если мы пожертвуем огромным количеством благ, которые могут быть созданы, если машины нас заменят.

Я не стараюсь прийти к какому-то конкретному выводу, я больше пытаюсь привлечь внимание к данной области. Наступила эра, когда нам первоначально нужно будет проделать очень много умственной работы, чтобы понять, как благотворная кооперационная система норм будет выглядеть в будущем, когда мы разделим Землю и космос с новыми видами разумов, некоторые из которых, возможно, будут намного более способными, чем мы. Очень много умственной работы нужно проделать, и может быть политической.

Другая вопрос, над которым я работаю последнее время – как будет выглядеть успех, если искусственный интеллект будет максимально продвинутым. Представьте, что появятся такие системы, которые способны не только заменять человека в конкретных заданиях, но у которых есть все те же возможности, что и у людей, и даже больше. Они способны на всё, что можем мы. Само собой, люди обеспокоены безработицей как возможным последствием экономического роста, развития технологий. Экономист Джордж Мейнард Кейнс написал известное эссе «Экономические возможности для наших внуков», опубликованное в 1930. В этом эссе он выдвинул предположение, что к 2030 году, через 100 лет после его написания, продуктивность увеличится в 4-8 раз, и в результате этого люди смогут меньше работать. Он предположил, что рабочая неделя сократится до 15 часов к 2030 году. И его беспокоило, что же люди будут делать со свободным временем. Он боялся, что это приведет к коллективному неврозу, люди сойдут с ума от того, что им нечем будет заняться, если они будут работать всего 15 часов в неделю. И сейчас, всё ещё 10 лет впереди, но, думаю, уже можно понять, куда всё идет. Первая часть его предсказания в самом деле исполнилась, продуктивность увеличилась более чем в 5 раз в США с 1930 года, и у нас есть еще немного времени в запасе. Вторая же часть его предсказания не сбылась, мы не работаем 15 часов в неделю. Но мы работаем меньше, чем раньше. По-моему, среднестатистическая рабочая неделя составляет 37,5 часов в США. Да, рабочее время уменьшилось. Но увеличились показатели дохода на душу населения в силу увеличившегося нерабочего времени. Большая часть нерабочего времени не означает работу несколько часов в день, при этом работая полный рабочий день, по большой части нерабочее время - это ранний уход на пенсию, больше уходов по болезни, декретные отпуска, люди начинают работать позже, чем раньше, 90 лет назад. Тем не менее, хотя рабочее время и сократилось, мы используем наш увеличившийся доход не чтобы увеличить время нашего отдыха, а чтобы увеличить наше потребление. Другими словами, жадность превзошла лень.

Вам может показаться, что, если продуктивность увеличилась, если у нас начали появляться более продвинутые технологии за счет продвижения искусственного интеллекта в других областях, то качество жизни продолжит улучшаться, мы будем богаче. Но до тех пор, пока человеческие прихоти не утолены, они продолжат стремиться к большему заработку, работая больше. Думаю, вероятно, что человеческие прихоти будут оставаться неутоленными, потому что нам нужны не только предметы первой необходимости, мы хотим большего. Мы хотим больше, чем у нас было раньше, и больше, чем есть у остальных людей. Поэтому, до тех пор, пока существует это желание быть впереди всех, не все смогут воплотить желаемое в реальность. Просто поднимается планка, над которой вы хотите возвыситься, чтобы ваши потребности были удовлетворены. Миллиардеры могут владеть большими яхтами, но если вы хотите самую большую яхту, всё это превращается в гонку.

По моему мнению, желание всегда будет существовать. За этим не следует, что мы перестанем работать, потому что другая часть этого уравнения заключается в следующем: должны будут существовать способы, чтобы заработать больше денег, именно работая. И не ясно, связано ли это с развитием технологий.

От людей для любой работы требуется как минимум метаболический вклад. Даже на самом простом уровне, нам нужно есть, чтобы мы могли использовать наши мышцы и мозг. Кроме того, мы занимаем много места, нам нужны столы, мониторы. Чтобы нанять человека, нужно так или иначе заплатить. Если машины станут достаточно сильно превосходить человека в плане эффективности… работа человека в целом стоит дороже, даже если бы людям не платили зарплату, затраты для удовлетворения базовых потребностей. Для человека может оказаться невозможным зарабатывать, будет дешевле и качественнее нанять робота. Электричество для него обойдется дешевле, чем еда для человека.

По крайней мере, говоря о работе, у которой есть функционально определенные требования, я думаю, с развитием технологий, такая работа будет лучше выполнена машиной. Отсюда вытекает человеческая безработица. Исключения могут возникнуть в случаях, когда сам потребитель предпочел бы, чтобы работа была выполнена человеком, даже если машина справилась бы лучше, она бы не потребовалась, потому что покупатель не хочет, чтобы машина выполняла работу, у него такое предпочтение. Даже в наше время, мы иногда платим больше за безделушку, сделанную своими руками, хотя машина сделала бы то же самое лучше. Или мы платим больше за что-то, что сделала наша любимая группа, что-то, сделанное вручную этими людьми, мы готовы заплатить больше, даже если функционально машина способна сделать то же самое дешевле. Взять в пример любого известного спортсмена, Усейн Болт, например, мы готовы заплатить деньги, чтобы увидеть Усейна Болта, даже если мы можем построить робота, который бегал бы быстрее, нас бы это не устроило, потому что нам интересно смотреть на человека и его старания. В общем, это тоже может стать своего рода нишей, но пока непонятны масштабы.

На этих примерах можно увидеть, в чем, собственно, будет заключаться успех, если мы продвинемся настолько далеко. К чему мы и идем. Нам приходится иметь дело с очень глубокими философскими вопросами, вопросами ценностей. Они же становятся вопросами практичности, не просто академическими и теоретическими предметами нашего любопытства, а, так или иначе, принципиальными вопросами о сознании. Как ранее я говорил о цифровом разуме, о том, какие системы имеют сознание, что это означает в плане наших обязанностей, как мы должны обращаться с сознательными системами. В отношении успешной автоматизации, вопрос о том, какая деятельность обладает высшей ценностью. Если бы нам ничего не нужно было делать, если бы мы могли достигнуть желаемого по щелчку пальцев или нажатию кнопки, если бы искусственный интеллект мог предугадывать наши желания и нужды, что бы тогда осталось для нас? Если бы нам нечего было делать, нам пришлось бы искать новые виды деятельности, но какие? И, как по мне, если посмотреть на то, чем мы сейчас занимаемся, и присмотреться внимательнее, вы заметите, что большинство наших действий обусловлено посторонними заботами. Мы думаем, что делаем что-то ради самого дела, потому что мы не получаем за это денег, но на самом деле у нас есть другие мотивы. Может, люди ходят в спортзал, им за это не платят, но они продолжают это делать, потому что хотят быть здоровыми или хорошо выглядеть. Если убрать из уравнения все эти инструментальные причины для действия, непонятно, что же тогда остается. Чем бы мы занимались, если бы в самом деле ничего не нужно было делать. И мы приходим к глубоким философским рассуждениям о ценностях, о хорошей жизни. И, если мы достигнем такого будущего, а я думаю, что, возможно, люди, живущие сейчас, вполне могут стать свидетелями этого будущего, если революция искусственного интеллекта произойдет в течение нескольких десятилетий. Нам нужно разобрать и понять эту тему от начала до конца. Мы не способны на простейшие предположения о мире, потому что мы очень ограничены человеческой природой, возможностями технологий. Но многие эти вещи просто растворяются в гипотетической технологической развитости. И нужно от начала до конца перестроить наше понимание того, что же значит иметь значимую полноценную жизнь в этих условиях.

В общем, я вам рассказал о том, над чем мы тут трудимся наряду с нашей повседневной работой. Пожалуй, на этом я и закончу, надеюсь, вы извлекли для себя что-то полезное, я дал вам пищу для размышлений. Будет интересно наблюдать за тем, как будут развиваться события в ближайшие годы, и продолжим ли мы идти к такому будущему. Спасибо большое за внимание, на этом у меня всё. Спасибо.