DOI: 10.31857/S023620070026101-8

©2023 И.Ю. ЛАРИОНОВ, Н.В. ПЕРОВА

«МАШИНА БОГА» ДЖ. САВУЛЕСКУ КАК МОРАЛЬНЫЙ АГЕНТ И ПРОБЛЕМА ОТВЕТСТВЕННОСТИ

Ларионов Игорь Юрьевич — кандидат философских наук, доцент.
Санкт-Петербургский государственный университет, Институт философии.
Российская Федерация, 199034 Санкт-Петербург, Университетская набережная, д. 7–9.
ORCID 0000-0002-0180-0304
i.larionov@spbu.ru

Перова Нина Вадимовна — лаборант.
Санкт-Петербургский государственный университет, Институт философии.
Российская Федерация, 199034 Санкт-Петербург, Университетская набережная, д. 7–9.
ORCID 0000-0002-1505-5376
nino4kaperova@gmail.com

Аннотация. Данная статья посвящена проблематике моральной ответственности в связи с технологиями искусственного интеллекта. В последние годы искусственный интеллект активно развивается в сторону все большей автономности, что делает крайне актуальной философскую аналитику искусственной моральной агентности (moral agency).

При поддержке гранта РНФ «Трансформации морального агентства: этико-философский анализ», № 22-28-00379.

УДК 17

Актуальность работы также определяется все большим распространением искусственного интеллекта в профессиональных сферах, в том числе связанных с принятием ответственных управленческих, финансовых и тому подобных решений. Основным объектом исследования в статье выступает умозрительный эксперимент — проект так называемой «Машины Бога» западного философа Дж. Савулеску, который позволяет обратиться к такому важному аспекту, как возможность влияния искусственного интеллекта на развитие нравственного сознания человека, поскольку современные технологии, наряду с проектами биологического нравственного улучшения человека, претендуют на способность создания искусственных моральных агентов. Авторы дают всесторонний анализ этого проекта, соотнося его с другими концепциями современной аналитической философии сознания, в том числе Г. Франкфурта. Теоретический фокус исследования направлен на концепты ответственности и свободы как ключевые в формулировке критериев морального агентства. При этом в статье рассмотрены особенности реализации критериев моральной агентности в отношении искусственного интеллекта. В статье оценивается состоятельность проекта Дж. Савулеску, в частности: насколько искусственный интеллект наподобие Машины Бога может рассматриваться как моральный агент, сможет ли человек при помощи такого искусственного посредника обрести характеристики нравственной, ответственной и свободной личности и насколько подобные проекты способствуют нравственному совершенствованию человечества. Указываются риски патерналистского вмешательства в развитие морали. Отдельно рассматривается феномен моральных дилемм в связи с проблематикой встраивания искусственного интеллекта в жизнь современного общества, а также логические аспекты принятия решения в условии моральных конфликтов. Формулируется вывод о ключевом значении категории ответственности для моделирования взаимоотношений человека и искусственного интеллекта в машинной этике.

Ключевые слова: моральная ответственность, моральное агентство, искусственный интеллект, биологическое нравственное улучшение, свобода, моральные дилеммы, Дж. Савулеску, Г. Франкфурт.

Ссылка для цитирования: Ларионов И.Ю., Перова Н.В. «Машина Бога» Дж. Савулеску как моральный агент и проблема ответственности // Человек. 2023. Т. 34, № 3. С. 24–40. DOI: 10.31857/S023620070026101-8

В наши дни проблема моральной ответственности напрямую связана с развитием технологий, в том числе искусственного интеллекта (ИИ) и биологического нравственного улучшения. Актуальность работы определяется все большим распространением ИИ не только в областях повседневной жизни, но и в профессиональных сферах, в том числе связанных с принятием ответственных управленческих, финансовых и тому

подобных решений. Философия все больше обращается к вопросу, как распространение ИИ влияет и может повлиять на моральную составляющую жизни человека и общества в перспективе развития высоких технологий. Доверие к ИИ вызывает к жизни теоретические проекты его использования для достижения благих целей совершенствования человека и общества, среди которых заметным направлением мысли начала 20-х годов XXI века следует признать теории так называемого нравственного биологического улучшения человека (human moral bioenhancement), то есть искусственного расширения и развития ряда человеческих способностей. Техники такого биоулучшения могут быть более и менее «инвазивными», однако все они ориентируются на достижение ценностно значимых результатов [Ларионов, Марков, 2022].

Современные технологии претендуют на способность создания искусственных моральных агентов. К таким технологиям относятся биологическое нравственное улучшение и ИИ. Своеобразным синтезом этих исследований выступает так называемая «Машина Бога» (The God Machine) — умозрительный эксперимент Дж. Савулеску, который можно рассматривать как максимизацию футуристических идей в этих областях. Однако претензия на создание искусственного морального агента связана с рядом этических проблем, которые, как будет показано в статье, не в последнюю очередь связаны с понятием моральной ответственности.

Проект Дж. Савулеску состоит в следующем: в каком-то обозримом будущем (Савулеску называет 2050 год) исследования морали оказались близки к завершению и в рамках некого Великого нравственного проекта (Great Moral Project) был создан сильнейший суперкомпьютер, который способен отслеживать мысли, убеждения, желания и намерения каждого человека. Эта Машина Бога может за наносекунды влиять на них и изменять, причем так, что человек не осознает факта внешнего изменения. Машина запрограммирована таким образом, что вмешивается в деятельность человека только для предотвращения существенного вреда или несправедливости — в терминах Савулеску, «крайне аморального поведения» (grossly immoral behavior). Машина Бога фактически устроена так, что подключенные к ней люди не смогут совершить такие крайне аморальные поступки, она им просто не позволит.

Частью Великого нравственного проекта была не только Машина Бога. В его рамках было реализовано полномасштабное внедрение технологий нравственного биоулучшения. Важным результатом, согласно Савулеску, стало общее повышение уровня альтруизма и чувства справедливости, благодаря чему люди почти перестали совершать безнравственные поступки. Все это

фактически позволяет говорить о создании совершенных моральных агентов [Savulescu, Persson, 2012].

Разумеется, проект Савулеску вызывает множество вопросов и возражений. В первую очередь сама по себе идея нравственного улучшения человека посредством Машины Бога, как и идея биологического нравственного улучшения человека, ставит вопрос о критериях статуса человека как морального агента.

Феномены нравственной жизни, наиболее ценимые на протяжении всей истории человечества, традиционно описываются так, будто они предполагают в качестве необходимого условия так или иначе понимаемую свободу индивидуума. Особое значение свобода имеет для концепта ответственности, причем данная категориальная связь характерна не только для моральной, но, например, и для правовой ответственности. Важное место в определении моральной агентности играет моральная ответственность. Если при этом мы исходим из того, что ключевым критерием моральной ответственности является свобода, то можно говорить, что Машина Бога, по меньшей мере, ставит этот статус под вопрос.

В первую очередь следует обратить внимание на то обстоятельство, что воздержание, уклонение от крайне аморального поведения в случае вмешательства Машины Бога не может рассматриваться как результат свободного выбора агента (тем более, как результат внутренней борьбы и непростых размышлений). Можем ли мы в данном случае нравственно одобрить самого человека за такой поступок? Правильно ли будет связывать его личность с позитивными последствиями поступка и/или вменить ответственность за них?

Проект Дж. Савулеску сталкивается с компатибилистской проблематикой, получившей значительное развитие в этике XX века, — попытками продемонстрировать, что детерминизм не противоречит свободе и совместим с возможностью вменять ответственность за совершенные деяния. Обращает на себя внимание схожесть образа Машины Бога и логики рассуждения американского философа Г. Франкфурта. Франкфурт оспаривает представление о том, что человека следует считать морально ответственным за поступок только в случае, если у него была возможность поступить иначе. Такое убеждение философ именуют «принципом альтернативных возможностей» и считает минимальным понятием о свободе. Франкфурт приводит гипотетическую ситуацию: представим, что агент совершает выбор между действиями *A* и *B*, в то время как существует некая причина *C*, вследствие которой возможность действия *B* на самом деле отсутствует, хотя агент об этом не знает. Франкфурт предлагает представить, что имеется крайне суровая угроза жестокого и непреодолимого

принуждения совершить именно *A*, скрытая от агента до того момента, пока он не решит избрать *B*. Но философ упоминает и возможность прямого контроля деятельности мозга агента. В таком случае, исходно вознамерившись избрать и избрав *A*, агент будет нести за это полную меру ответственности, хотя возможности поступить иначе у него и не было [Франкфурт, 2017]. Основное возражение данному аргументу Г. Франкфурта состоит в том, что можно не согласиться с самим определением свободы как наличия альтернативных возможностей. Во-вторых, описание моральной личности здесь выглядит упрощенным. Это возражение можно высказать и по отношению к Машине Бога Савулеску: едва ли все разнообразие нравственной жизни человека можно свести к способности самостоятельно воздерживаться от крайне аморального поведения.

Отчасти компенсировала недостатки компатибилизма Франкфурта его теория личности (person), которую он определяет как сущность, располагающую так называемым «волением /способностью воления второго уровня» («second-order volitions»), то есть способностью хотеть наличия у себя определенного типа хотения («первого уровня») и стремиться к таковому. Например, стремиться (желать) не иметь (и подавлять) желания совершить крайне аморальный поступок. Таким образом, ориентированная осознанная мотивация и является, по Франкфурту, подлинным содержанием свободы воли. Она согласуется с детерминизмом и вообще не требует ни наличия альтернативного выбора, ни отсутствия вмешательства извне. Однако разум, замечает Франкфурт, необходим для формирования «волений второго уровня». Машина Бога, говоря языком Г. Франкфурта, не отнимет у агента ни статуса личности, ни свободы воли, если вмешается в ситуацию путем формирования у рационального агента осознанного отвращения к собственному желанию совершить глубоко аморальный поступок [Frankfurt, 1971].

Сопоставляя проекты вмешательства в автономные действия агента у Г. Франкфурта и Дж. Савулеску, мы видим, что для последовательной их концептуализации требуется согласовать определение свободы с теоретическим конструированием понятия об агенте как личности, обладающей сложной структурой способностей (или даже тождественной этой структуре). Выявляется также еще один существенный аспект: философская традиция на протяжении веков исходила из предположения, что именно сознательное мышление руководит моральным выбором человека, при том что сама человеческая свобода основана на рациональности как способности видеть альтернативы, принимать решения и реализовывать их в действии. Благодаря силе разума человек

освобождается от страстей и внешнего контроля, что составляет основу его ответственности.

В отличие от Г. Франкфурта, работавшего на «метафизическом» уровне проблематики (вопрос о свободе воли вообще), Дж. Савулеску обращается именно к этическому ее измерению. В первую очередь он отмечает, что речь идет о деятельности, которую человек никогда и не был «свободен» делать. Возможно, здесь происходит подмена понятий: на английском языке «free to act» может быть понято и как «свободен действовать», и как «дозволено действовать», что все-таки разные вещи. Действительно, если речь идет о запрещенных действиях (тем более запрещенных юридически, например воровстве или убийстве), человеку вне зависимости от подключения к Машине не «дозволено» их совершать. Однако эти запреты на совершение действия никоим образом не связаны с самой возможностью выполнить его. В этом смысле, хотя Машина Бога, конечно, не изменит уровень «дозволенности», она все же повлияет на уровень «свободы» выполнить действие.

Другой аргумент Савулеску в защиту Машины Бога связан с лишением человека «свободы ошибаться» (freedom to fall): человек будет совершенно свободен, пока не будет на грани совершения крайне аморального поступка. То есть человек сможет иметь аморальные мотивы и желания и даже строить планы по их реализации, и единственная ситуация, в которой Машина Бога вмешается, — прямо перед совершением крайне аморального поступка. Более того, вмешательство Машины заключается в том, что она поменяет мнение человека. Человеку будет казаться, что это было самостоятельное решение, основанное на личных нравственных установках. Для Савулеску это значит, что, хотя сам по себе глубоко аморальный акт человек никогда не сможет совершить, он все равно будет свободен аморально мыслить и развиваться как моральная личность, исходя из аморальных идей и намерений. Отдельный акцент Савулеску делает на то, что в случае с максимально моральным человеком, который никогда даже не помышляет о совершении аморального действия, Машина Бога не будет вмешиваться ни при каких обстоятельствах, а, значит, свобода такого человека вообще никогда не будет нарушена.

Отметим, что речь идет пусть не о полном лишении свободы, но о таком, которое, может быть, незначительно по количеству, но качественно составляет крайне важное условие именно нравственной жизни. Это вызывает серьезные опасения о сохранении свободы. Кроме того, устраняемая здесь возможность моральной ошибки обычно способствует формированию ответственности, по крайней мере психологически.

Вопрос о свободе в рамках Машины Бога ставится в контексте свободы использования. В описании эксперимента указано, что подключение к Машине Бога будет добровольным, хотя и с некоторыми оговорками. Так, все дети подключены к Машине. По достижении совершеннолетия им будет даваться выбор — остаться подключенным или нет. Также взрослый человек в любой момент может отключиться и подключиться по желанию — исключение составляют преступники, которые вместо тюремного заключения будут приговорены к принудительному подключению. Все это, согласно автору, не будет нарушать свободу людей [Savulescu, Persson, 2012].

Дж. Савулеску, утверждающий нравственное биоулучшение как утилитаризм, приводит в подтверждение своих слов цитату из трактата Дж. Милля «О свободе»: «Каждый член цивилизованного общества только в таком случае может быть справедливо подвергнут какому-нибудь принуждению, если это нужно для того, чтобы предупредить с его стороны такие действия, которые вредны для других людей». Савулеску трактует это как дозволение пожертвовать автономией ради всеобщего блага. Однако здесь он упускает важный момент: Машина Бога предполагает жертвовать автономией даже тех, кто не совершает и не совершит никакого глубоко аморального поступка. Лишенным свободы оказывается даже тот, кто не приносит и не будет приносить никакого вреда другим. И в этом он противоречит самим идеям Милля: «Только в том случае дозволительно подобное вмешательство, если действия индивидуума причиняют вред кому-либо» (здесь и выше цит. по: [Милль,1993: 10–11]). При этом, если мы соглашаемся с Савулеску, что подобное нарушение свободы не является угрозой для статуса морального агента, то мы оказываемся перед вопросом иного рода: что является необходимым условием для морального агента?

В анализе этих проблем мы предлагаем обратиться к такому аспекту нравственности, как необходимость взаимодействия с другим моральным агентом, которая приобретает особенную важность именно в процессе нравственного воспитания и совершенствования человека. Тем самым вопрос состоит в том, рассматривать ли подключение к Машине Бога как метод подлинного нравственного воспитания, изменения, развития и совершенствования человека или это всего лишь технический инструмент решения части существующих проблем, связанных с агрессией, но не меняющий ничего по сути.

В этом отношении важно определить, может ли Машина Бога, являющаяся системой ИИ, сама считаться моральным агентом, и будет ли ее таким образом воспринимать сам взаимодействующий

с ней и подключенный к ней человек. Можно ли контакт с ней считать разновидностью взаимодействия с другим моральным агентом, в том числе как важное условие своего нравственного совершенствования? Также стоит определить, можно ли считать, что подключенный к Машине Бога человек (особенно подключенный в сознательном возрасте), будет нести какого-либо рода ответственность перед Машиной Бога или его взаимоотношения с ней будут иного характера?

Один из ключевых вопросов современной машинной этики как раз состоит в том, действительно ли та или иная гипотетическая или реальная система является искусственным агентом или только кажется таковым. Такая постановка вопроса была заложена еще А. Тьюрингом, чей знаменитый тест состоял в распознавании того, взаимодействуем мы с ИИ или с другим человеком [Turing, 1950]. В ответ на работу А. Тьюринга философ Дж. Сёрл предложил мыслительный эксперимент «китайская комната», в котором стремился показать, что даже если мы в тесте Тьюринга перестанем отличать, человек с нами общается или «машина», из этого вовсе не будет следовать, что на другом конце коммуникации имеется личность в собственном смысле слова [Сёрл, 1998: 400]. Современные исследователи М. и С.Л. Андерсон, продолжая рассуждения Сёрла, последовательно проводят мысль, что искусственный моральный агент должен быть таковым, а не казаться [Anderson, Anderson, 2007: 16–17; Разин, 2019].

Исследователи единодушны в том, что обязательным условием для искусственного морального агента должна быть автономия — способность полноценно действовать в отрыве от инженеров, программистов, владельцев и т.п., пусть даже изначально они и были разработаны и запрограммированы определенным извне образом. Машина Бога у Савулеску, по всей вероятности, создана для автономного функционирования [Anderson, Anderson, 2007].

В случае Машины Бога мы, очевидно, имеем дело с системой, обладающей немалым массивом знаний в области морали. В то же время, если в ней и есть модуль морального рассуждения, то используется он крайне узко — для определения степени опасности действия, совершить которое намерен подключенный к ней человек. По всей видимости, мы не можем говорить о наличии рационального выбора у Машины Бога. В любом случае, она не принимает решения в отношении самой себя. По видимости, Машина Бога в узкой сфере своей «работы» не может совершить ошибки. Но будет ли подлинно нравственным такое безупречное существо?

Неотъемлемым требованием для полноценных искусственных моральных агентов ряд авторов считает интенциональность,

преднамеренность действия (intentionality), которое на уровне теории выводится из самого понятия «агентства». Данный подход был поддержан Д. Деннетом: моральная ответственность может быть вменена роботу, если он действовал на основании некой интенции и даже «higher order intentionality», то есть интенции относительно другой интенции (например, намерения не нарушать правило непричинения человеку вреда). По Деннету, этого достаточно также для того, чтобы считать робота личностью. Параллелизм рассуждений Д. Деннета и Г. Франкфурта здесь не случаен: оба предполагают, что подобного рода явления сознания второго порядка могут быть достаточным критерием для личности [Dennett, 1998].

О Машине Бога мы едва ли можем утверждать, что она действует из некоего намерения в отношении к подключенному к ней человеку, в том числе из намерения способствовать нравственному улучшению такового.

Практически во всех исследованиях искусственное моральное агентство в первую очередь связываются с ответственностью. По всей видимости, это следствие того, что источником актуальности проблематики искусственного морального агентства с самого начала была возможность причинения роботом вреда человеку. При этом вопрос о том, кому вменить ответственность за деяния робота, является и этическим, и правовым. Ответ же прямо связан с признанием возможности искусственного морального агентства или отрицанием такового, сведением роботов к инструментам (принцип «blame the user, not the tool») и перенесением ответственности на человека [Floridi, Sanders, 2001]. Умеренная позиция по данному вопросу может быть сформулирована так: если представляется, что робот действует, исходя из определенных обязанностей к кому-либо, то его можно рассматривать и как морального агента [Sullins, 2006].

Говорить о Машине Бога как об ответственном агенте можно только в очень узком смысле. В первую очередь в своем функционировании она практически не причиняет вреда, ибо ее цель — сократить меру вреда, причиняемого людям друг другу. В то же время она должна оставаться беспристрастным участником ситуации, поэтому нельзя сказать, что она в том или ином роде берет на себя ответственность за свои действия. Более того, запрограммированное пресечение крайне аморального деяния не выглядит как самостоятельный выбор Машины Бога.

Отдельно следует учесть возможность ситуаций, когда Машина Бога оказывается перед лицом моральной дилеммы, от решения которой зависит ее реакция на возможность крайне аморального поступка. В исследовании М. и С. Андерсон рассматривали возможность использования при программировании «моральных»

роботов принципов утилитаризма, этики добродетели, деонтологии, норм профессиональной этики, специфику принципов prima facie и т.п. Авторы пришли к выводу, что говорить об искусственном моральном агентстве можно только при наличии у системы алгоритма принятия решения в сложных в моральном отношении ситуациях [Anderson, Anderson, 2007].

Разумеется, Машина Бога с необходимостью должна быть сложнейшим, саморазвивающимся, самообучающимся ИИ. Однако Машина Бога не может быть просто сводом моральных норм. Даже если представить, что Великий нравственный проект дал нам совершенную единую систему этических норм, все равно речь будет идти о системе, соответствующей определенному обществу в конкретный период времени. Даже включение в нее различных синхронно существующих культурных и национальных особенностей уже приведет к тому, что одни и те же действия могут быть расценены как моральные и как аморальные. Кроме того, профессиональные этики могут включать в себя нормы, противоречащие общепринятым, например, военная этика или политическая этика [Гарвардт, Перова, 2022].

Тем самым Машина Бога должна иметь максимально сложные настройки, включающие возможность определения одного и того же действия как морального и аморального исходя из обстоятельств, притом что, учитывая трансформацию морали вследствие социальных изменений и развития технологий, эти настройки должны меняться постоянно. Процесс изменения морали Машины Бога должен быть самостоятельным, поскольку, если будут выделены человек или группа людей, которые будут следить за работой Машины и при необходимости корректировать ее, то мы получим общество, в котором малая группа будет решать, какое именно поведение может быть отнесено к крайне аморальному. То есть будет построено общество, основанное на нравственном патернализме, что также недопустимо.

Одна из сложностей, связанная с попыткой авторов Великого нравственного проекта «поиграть в Бога», состоит в том, что особенностью нравственной жизни и нравственного совершенствования является способность осмысленного принятия решения в ситуации сложных, неоднозначных дилемм. В некотором отношении критерием подлинного («сильного») ИИ может быть способность решать сложные познавательные и творческие задачи с учетом всего разнообразия контекста социального бытия и поведения человека. ИИ вроде Машины Бога, работающий с чем-то настолько сложным, как мораль, обязательно столкнется с противоречиями.

В качестве примера, способного вызвать такую сложность, возьмем убийство из самозащиты. Представим случай, в котором

человек попадает в такую ситуацию случайно (а Машина Бога способна различить, действует ли человек преднамеренно и со скрытыми мотивами или нет), события развиваются крайне быстро, так что субъект находится на пороге совершения действия в состоянии аффекта. Машина Бога здесь выбирает между допущением гибели одного из участников ситуации. Ее положение похоже на позицию беспристрастного наблюдателя в дилемме вагонетки — выбор между неизбежной смертью пяти невинных людей и возможностью их спасти, направив движущуюся вагонетку в сторону еще одного невинного человека. Так или иначе особенность таких моральных дилемм в том, что они требуют от принимающего решение не только когнитивных, но и нравственных усилий для обоснования своего выбора.

Вполне реальной технической разработкой, сталкивающейся с этой областью проблем, являются автономно действующие средства вооружения, а также автомобили-беспилотники [Ларионов, 2022]. Беспилотный автомобиль может оказаться перед подлинно моральным выбором: дорожная обстановка чревата трагическими ситуациями, в которых невозможно сохранить здоровье (а иногда и жизнь) всех ее участников. В подобных условиях человек обнаруживает понимание возможности пожертвовать жизнью одного человека ради спасения другого, принятие на себя ответственности, при том что все решения будут случаями крайне аморального поступка.

Есть основания думать, что и гипотетическая Машина Бога столкнется с похожими ситуациями и с необходимостью принимать такие решения. Онлайн-проект так называемой Моральной машины («Moral Machine») 2016–2018 годов на базе Массачусетского технологического института (MIT), в рамках которого собиралась и анализировалась информация о том, какие решения склонны принимать люди в сложных дорожных ситуациях, показал крайнюю неопределенность и многообразие вариантов, предложенных участниками [Awad et al., 2018]. Почти полувековое обсуждение дилеммы вагонетки в теоретической литературе также не пришло к однозначному результату [Mason, 1996]. Данное обстоятельство позволяет поставить под вопрос возможность самого завершения Великого нравственного проекта.

Моральные дилеммы с точки зрения логики могут быть формализованы как противоречия. Современные логические исследования содержат теории о том, как именно ИИ будет работать после выявления противоречия. В классической логике, которую можно считать наиболее непредвзятой, противоречия приводят к взрыву, что означает прекращение процедуры вывода, то есть ИИ перестает выдавать результаты. На практике это будет означать следующее:

после выявления противоречия Машина больше не сможет принимать решения о нравственности или безнравственности конкретного поведения. Для преодоления взрыва можно обратиться к неклассическим логикам. Однако здесь и появляется значительная проблема для требования о непредвзятости. Своевременное распознание дилеммы требует обучения ИИ модальным логикам. Дальнейшее выявление и устранение взрыва невозможно без обращения к параконсистентной логике [Альчуррон, Герденфорс, Макинсон, 2013].

Проблема в том, что неклассические логики не могут быть непредвзятыми. Для того чтобы ИИ мог должным образом реализовывать все возможные пути выявления и преодоления противоречий, необходимо глубоко погрузиться в частные условия конкретных ситуаций. Каждый случай дилеммы с необходимостью будет содержать уникальные факты, которые будут определять необходимость внимательного изучения ситуации со стороны ИИ и в конечном итоге предвзятое решение о дальнейшей деятельности. ИИ будет иметь возможные, объективно равносильные варианты продолжения деятельности, выбор между которыми возможен только при предвзятом включении ИИ в ситуацию.

Особая коллизия связана с проблемой так называемой моральной неудачи, моральных ошибок. Подлинная моральная дилемма характеризуется, в частности, тем, что обе возможности представляют собой морально недопустимое действие, «moral failure» (ср. [Harris, 2011]). Моральные ошибки и неудачи нередко рассматриваются как часть нравственного становления и совершенствования личности (см. выше о «свободе ошибаться»).

Таким образом, даже если Машина Бога будет эффективно устранять ситуации моральных конфликтов, возникают сомнения, будет ли это способствовать именно нравственному совершенствованию.

* * *

Проведенный в статье анализ позволяет, во-первых, говорить о том, что в рамках влияния Машины Бога нельзя говорить о свободе человека и, тем более, о свободном нравственном совершенствовании. В то же время данный мыслительный эксперимент ставит под вопрос сами основы понимания нравственной свободы, а также содержательные характеристики категорий свободы, которые могли бы быть применены для описания нравственного поступка. Современные исследования в области моральной философии также предполагают возможность совместимости нравственной жизни и детерминизма. Само нравственное совершенствование, учитывая

неотъемлемый нормативный характер морали, вовсе не обязательно предполагает полноценную реализацию свободы индивида[1]. Хотя свободу нельзя считать достаточным условием для моральной агентности, она тем не менее выступает необходимым ее условием. Самое существенное: она выступает основанием для моральной ответственности, которую и следует признать тем ключевым критерием морального агентства, который ставит под вопрос Машина Бога.

Во-вторых, наиболее весомым аргументом против того, что подключение к Машине Бога сможет хоть как-то способствовать нравственному совершенствованию человека, будет отсутствие у такового именно ответственности. Если готовность быть подключенным к Машине Бога может рассматриваться как часть общего стремления агента к нравственному совершенствованию, то постоянное осознание возможности вмешательства Машины не будет способствовать формированию у него навыка воздержания от аморального поступка. Напомним, что мыслительный эксперимент Г. Франкфурта относится к прямо противоположной ситуации — ответственность агента сохраняется, когда его действия не были пресечены извне.

Исключение возможности ошибки, особенно в ситуации сложной моральной дилеммы, может возыметь приемлемый с психологической точки зрения результат, однако устранит важный мотив действовать с учетом всех вероятных последствий. Подключенность к Машине приведет к тому, что человек станет в определенном отношении нравственно безвольным. Тем самым Машина Бога будет, скорее, лишать подключенного к ней человека статуса морального агента, чем способствовать его нравственному совершенствованию.

В-третьих, важный компонент формирования нравственной ответственности — взаимодействие с другими моральными агентами. Именно такой характер имеет нравственное воспитание детей, а также развитие нравственной личности взрослого. Как мы видели, в описании работы Машины Бога дается во многом упрощенное представление о характере нравственного сознания, а также социального аспекта морали, требующей от агента развитых способностей ориентироваться в сложных, неоднозначных и конфликтных ситуациях. Мораль как сферу жизни общества характеризует то, что люди

[1] К схожим выводам относительно возможностей использовать ИИ для нравственного совершенствования человека приходит исследователь Ф. Лара. Пренебрежение личной автономией при прямом вмешательстве технологий не может быть более эффективным для нравственного совершенствования, чем традиционное воспитание человека человеком. В качестве альтернативы автором предлагается ИИ как помощник, который с помощью диалога и технологий виртуальной реальности сможет научить пользователей самостоятельно принимать более эффективные моральные решения [Lara, 2021].

вступают в сложные взаимоотношения, рассматривая самих себя, а также эти отношения как осознанную ценность, имеющую нормативную силу, видоизменяющую (в том числе и ограничивающую) и направляющую поступки [Философия ответственности, 2014: 172].

Нравственное улучшение непосредственно зависит от ответственности, понятой как интеракция с другим моральным агентом. Однако может ли Машина Бога считаться полноценным моральным агентом? Исходя из рассмотренных выше критериев, можно сделать вывод, что Машина Бога (а также существующие системы ИИ вообще) не может выступать моральным агентом. Она не является полностью свободной, так как имеет базовую программу, определенную извне. Какой бы проработанной ни была изначальная этическая программа, она не сможет соответствовать реальности нестатичности морали. Это означает, что вся самообучаемость Машины должна подвергаться внешнему контролю. Мы также не можем говорить о непредвзятом ИИ: изначальная предвзятость алгоритма сосуществует с необходимой предвзятостью решений, принимаемых при выявлении любых противоречий. Поскольку деятельность людей не соответствует требованию непротиворечивости — мы не можем исключить из жизни спорные случаи и моральные дилеммы, — поведение ИИ должно это учитывать. Это значит, что для непрерывной работы Машина Бога должна быть активно включена в частные случаи деятельности людей, что исключает возможность непредвзятости. Наконец, ИИ не может быть моральным агентом с точки зрения моральной ответственности. Общество не может возлагать на ИИ ответственность за совершенные действия и принятые решения. Также в отношении ИИ не может быть ожидания осознания моральной ответственности как осознания последствий за совершенные поступки.

Подводя итог сказанному, в рассматриваемом проекте Дж. Савулеску мы видим неустранимое противоречие — Машина Бога не сможет воплотить в жизнь гипотетические достижения Великого нравственного проекта, поскольку она устраняет два важнейших условия бытия морального агента: ответственность и свободу.

J. Savulescu's "The God Machine" as a Moral Agent and the Problem of Responsibility

Igor Yu. Larionov
PhD in Philosophy, Docent.
Institute of Philosophy, Saint-Petersburg State University.
7-9 Universitetskaya nab., Saint Petersburg 199034, Russian Federation.
ORCID 0000-0002-0180-0304
i.larionov@spbu.ru

ФИЛОСОФИЯ ЧЕЛОВЕКА

Nina V. Perova
Researcher. Institute of Philosophy, Saint-Petersburg State University.
7-9 Universitetskaya nab., Saint Petersburg 199034, Russian Federation.
ORCID 0000-0002-1505-5376
ninio4kaperova@gmail.com

Abstract. This article is devoted to the issue of moral responsibility in connection with artificial intelligence technologies. In recent years, artificial intelligence has been actively developing towards greater autonomy, which makes the philosophical analysis of artificial moral agency extremely relevant. The relevance of the work is also determined by the increasing spread of artificial intelligence in professional areas, including those related to the adoption of responsible managerial, financial, etc. solutions. The main object of research in the article is a thought experiment — the project of the so-called "The God Machine" by the Western philosopher J. Savulescu, which allows us to turn to such an important aspect as the possibility of the influence of artificial intelligence on the development of human moral consciousness, since modern technologies, along with projects of biological moral human enhancement, claim the ability to create artificial moral agents. The authors give a comprehensive analysis of this project, correlating it with other concepts of modern analytical philosophy of consciousness, incl. H. Frankfurt. The theoretical focus of the study is directed to the concepts of responsibility and freedom as key in formulating the criteria of moral agency. At the same time, the article considers the features of the implementation of the criteria of moral agency in relation to artificial intelligence. The article assesses the viability of J. Savulescu's project, in particular: to what extent artificial intelligence like the The God Machine can be considered as a moral agent, whether a person can acquire the characteristics of a moral, responsible and free person with the help of such an artificial mediator, and to what extent such projects contribute to the moral enhancement of mankind. It also points to the risks of paternalistic interference in the development of morality. Separately, the phenomenon of moral dilemmas is considered in connection with the problems of embedding artificial intelligence in the life of modern society, as well as the logical aspects of decision-making in the context of moral conflicts. The conclusion is formulated about the key importance of the category of responsibility for modeling the relationship between man and artificial intelligence in machine ethics.

Keywords: moral responsibility, moral agency, artificial intelligence, biological moral enhancement, freedom, moral dilemmas, J. Savulescu, H. Frankfurt.

For citation: Larionov I.Yu., Perova N.V. J. Savulescu's "The God Machine" as a Moral Agent and the Problem of Responsibility // Chelovek. 2023. Vol. 34, N 3. P. 24–40. DOI: 10.31857/S023620070026101-8

Литература/References

Альчуррон К.Э., Герденфорс П., Макинсон Д. Логика теории изменения: функции ревизии и сокращения через частичное пересечение // «Нормативные

системы» и другие работы по философии права и логике норм. СПб.: Издательский дом СПбГУ, 2013. С. 318–343.

Alchurron K.E., Gerdenfors P., Makinson D. Logika teorii izmeneniya: funkcii revizii i sokrasheniya cherez chastichnoe peresechenie [The Logic of the Theory of Change: Revision Functions and Reduction through Partial Intersection]. *"Normativnye sistemy" i drugie raboty po filosofii prava i logike norm.* St. Petersburg: Izdatelskij Dom SPbGU Publ., 2013. P. 318–343.

Гарвардт М.А., Перова Н.В. Универсализм и релятивизм морального агента биологического нравственного улучшения (на примере политической деятельности) // Дискурсы этики. 2022. № 1(13). С. 69–84.

Garvardt M., Perova N. Universalism and Relativism of the Moral Agency of Biological Moral Enhancement (on the Example of Political Activity)]. *Discourses of Ethics*. 2022. N 1(13). P. 69–84.

Ларионов И.Ю. Кто несет ответственность за применение оружия с искусственным интеллектом? // Интернет и современное общество: сборник тезисов докладов [Электронный ресурс]. Труды XXV Международной объединенной научной конференции «Интернет и современное общество» (IMS-2022), Санкт-Петербург, 23–24 июня 2022 г. СПб.: Университет ИТМО, 2022. URL: http://ojs.itmo.ru/index.php/IMS/issue/view/78, свободный.

Larionov I.Yu. Kto neset otvetstvennost za primenenie oruzhiya s iskusstvennym intellektom? [Who is Responsible for the Use of Weapons with Artificial Intelligence?] // *Internet i sovremennoe obshestvo: sbornik tezisov dokladov* [Elektronnyj resurs]. Trudy XXV Mezhdunarodnoj obedinennoj nauchnoj konferencii "Internet i sovremennoe obshestvo" (IMS-2022), St. Petersburg, 23–24 iyunya 2022 g. SPb.: Universitet ITMO, 2022. URL: http://ojs.itmo.ru/index.php/IMS/issue/view/78, svobodnyj.

Ларионов И.Ю., Марков Б.В. Морально-антропологические риски технологий биологического совершенствования человека // Дискурсы этики. 2022. № 4(16). С. 25–46.

Larionov I., Markov B. Moral and Anthropological Risks of Human Bioenhancement Technologies // *Discourses of Ethics.* 2022. N 4(16). P. 25–46.

Милль Дж. О свободе // Наука и жизнь. 1993. № 11. С. 10–11.

Mill J. O svobode [*On Liberty*]. *Nauka i zhizn*. 1993. N 11. P. 10–11.

Разин А.В. Этика искусственного интеллекта // Философия и общество. 2019. № 1. С. 57–73.

Razin A.V. Etika iskusstvennogo intellekta [The Ethics of Artificial Intelligence]. *Filosofiya i obshestvo*. 2019. N 1. P. 57–73.

Сёрл Дж. Сознание, мозг и программы // Аналитическая философия: становление и развитие. М.: Дом интеллектуальной книги: Прогресс-Традиция, 1998.

Searle J. Soznanie, mozg i programmy [Minds, Brains, and Programs]. *Analiticheskaya filosofiya: stanovlenie i razvitie*. Moscow: Dom intellektual'noj knigi: Progress-Tradiciya Publ., 1998.

Философия ответственности / под ред. Е.Н. Лисанюк, В.Ю. Перова. СПб.: Наука, 2014.

Filosofiya otvetstvennosti [Philosophy of Responsibility]. Ed. by E.N. Lisanyuk, V.Yu. Perova. St. Petersburg: Nauka Publ., 2014.

Франкфурт Г. Альтернативные возможности и моральная ответственность // Философия. Журнал Высшей школы экономики. 2017. Т. I, № 4. С. 129–140.

Frankfurt H. Alternativnye vozmozhnosti i moralnaya otvetstvennost [Alternate Possibilities and Moral Responsibility]. *Filosofiya. Zhurnal Vysshej shkoly ekonomiki*. 2017. Vol. I, N 4. P. 129–140.

Allen C., Varner G., Zinser J. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*. 2000. Vol. 12, N 3. P. 251–261.

Anderson M., Anderson S.L. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine.* 2007. N 4(28). P. 15–26.

Awad E., Dsouza S., Kim R. et al. The Moral Machine Experiment. 2018. *Nature.* Vol. 563. P. 59–64.

Dennett D. When HAL Kills. Who's to Blame? Computer Ethics // Stork, David, HAL's Legacy: 2001's Computer as Dream and Reality. MIT Press, 1998.

Floridi L., Sanders J. Artificial Evil and the Foundation of Computer Ethics. *Ethics and Information Technology.* 2001. N 3(1). P. 56–66.

Frankfurt H. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy.* 1971. Vol. 68, N 1. P. 5–20.

Harris J. Moral Enhancement and Freedom. *Bioethics.* 2011. Vol. 25. P. 102–111.

Himma K.E. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent? *Ethics and Information Technology*. 2009. N 11. P. 19–29.

Lara F. Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates*? Science and Engineering Ethics*. 2021. N 42. P. 27.

Moral Dilemmas and Moral Theory / ed. by H.E. Mason. N.Y.: Oxford University Press, 1996.

Nadeau J.E. *Only Androids Can Be Ethical* // Thinking about Android Epistemology / ed. by K. Ford, C. Glymour. MIT Press, 2006. P. 241–248.

Savulescu J., Persson I. Moral Enhancement, Freedom and the God Machine. *The Monist*. 2012. N 95(3). P. 399–421.

Sullins J.P. When Is a Robot a Moral Agent. *International Review of Information Ethics.* 2006. Vol. 6(12). P. 23–30.

Thomson J.J. The Trolley Problem. *The Yale Law Journal*. 1985. Vol. 94, N 6. P. 1395–1415.

Turing A. *Computing Machinery and Intelligence* // Mind. 1950. LIX (236). P. 433–460.