Железнов А. С. Идея нравственности в экспериментальной моральной философии // Философия. Журнал Высшей школы экономики. — 2022. — Т. 6, № 3. — С. 181–207.

Андрей Железнов*

Идея нравственности в экспериментальной моральной философии**

Получено: 06.10.2021. **Рецензировано:** 29.11.2021. **Принято:** 22.07.2022.

Аннотация: Экспериментальная моральная философия показывает нравственность как стремление к благу, превосходящему любой данный опыт. Для того чтобы продемонстрировать это, мы будем обсуждать ряд исследований, которые разделяют методологическую парадигму, предполагающую использование методов естественных и социальных наук для изучения содержания и механизмов функционирования морали. При этом мы будем концентрироваться не на дискуссии о прямых результатах экспериментальных исследований, а на обсуждении отношения к морали, которое проявляется в подобных исследованиях косвенным образом, — так нам удастся выделить два основных аспекта нравственности. Во-первых, определяя моральный поступок, исследователи фактически всегда показывают, что он совершается ради целей, противостоящих очевидным интересам самого агента. В этом смысле не принципиально, как именно обосновывается приоритет моральных целей, но сама способность действовать морально связывается со способностью видеть нечто более важное, чем очевидные собственные интересы. Во-вторых, в отношении результатов экспериментального исследования предполагается наличие некоторой позиции или процедуры оценки правильности норм и механизмов морали, открытых в результате экспериментальных исследований. Возможность этой позиции или процедуры не следует из фактов, но допускает убежденность или веру в существование некоторого специального морального «блага», превосходящего любой данный опыт. Эта трактовка нравственности довольно формальна, но в то же время она имеет вполне конкретные последствия для этики, а именно отказ от любых попыток обосновать объективную необходимость соблюдения моральных норм.

Ключевые слова: этика, нравственность, экспериментальная моральная философия, моральные дилеммы, предвзятость, методологическая парадигма.

DOI: 10.17323/2587–8719–2022–3–181–207.

Ценность разговора об экспериментальной моральной философии

Говорить об экспериментальной моральной философии сейчас имеет смысл по нескольким причинам. Первой из них может быть отсутствие

*Железнов Андрей Сергеевич, к. филос. н., независимый исследователь (Москва), itsnomoredancing@gmail.com, ORCID: 0000–0001–9516–2392.

в русскоязычной литературе обсуждений данного направления: в феврале 2021 г. поиск по запросу «экспериментальная моральная философия» в базе Google Scholar, РИНЦ показал всего одну работу (Дроздова, 2017). Правда, наряду с отсутствием прямого обсуждения экспериментальной моральной философии в русскоязычном поле присутствует довольно активное обсуждение «нейроэтики»: например, в 2020 г. вышел номер журнала «Философия. Журнал Высшей школы экономики», посвященный этой теме. По своей проблематике нейроэтика действительно близка к экспериментальной моральной философии, но составляет (как мы постараемся уточнить ниже) только один из вариантов реализации ее методологической парадигмы. Учитывая, что экспериментальная философия вообще и экспериментальная моральная философия в частности представляют собой заметное явление, более широкое знакомство русскоязычного читателя с этой методологией может представлять самостоятельную цель. Данная статья, однако, преследует такую цель лишь косвенно — наш интерес имеет другую природу.

Для нас обсуждение экспериментальной моральной философии — это способ поговорить об идее нравственности как таковой или показать границы экспериментального, эмпирического подхода к морали, поэтому мы не планируем осуществлять лобовую критику результатов и метода экспериментальной философии, как если бы мы подвергли разбору и обсуждению способ получения выводов. Вместо этого мы постараемся подсветить установки, разделяемые экспериментальными исследователями до какой бы то ни было рефлексии.

Здесь следует сделать пару терминологических отступлений и объяснить, как мы будем употреблять понятия морали и нравственности, а также определить, что же такое экспериментальная моральная философия.

Термины «мораль» и «нравственность» в этом тексте не выступают синонимами. Слово «мораль» используется нами для описания норм, правил или принципов: это могут быть социально установленные принципы и правила, механизмы функционирования нервной системы или же принципы, предложенные кем-то из философов. Поскольку экспериментальные исследования зачастую нацелены на описание моральных норм или механизмов суждения, разделяемых обычными людьми, то их результатом как раз и является получение описания морали. Термином «нравственность» мы называем способность или склонность индивида принимать всерьез моральные принципы и стремиться к их

соблюдению. В этом смысле мы говорим, что экспериментальная философия исследует мораль, но позволяет нам косвенным образом судить и о нравственности.

Хотя мы явно отличаем мораль от нравственности, на уровне высказываний относительно поступков или поведения в тексте ниже практически не разграничиваются моральный поступок (моральное поведение) и нравственный поступок (нравственное поведение). Дело в том, что мы не называем моральным поступок, который признан соответствующим морали постфактум, без знания его мотивов. А называем моральным мы такой поступок, цель которого — реализация моральной нормы, то есть этот поступок направляется нравственной мотивацией. Возможно, было бы правильнее говорить здесь о «нравственных поступках», сохранив для «моральных» именно аспект внешней оценки. Однако, сохраняя близость к англоязычному «moral», в этом тесте мы будем употреблять выражение «моральные поступки» как синоним выражения «нравственные поступки».

В определении экспериментальной моральной философии нам кажется уместным использовать то, как ее понимают сами исследователи, разделяющие данный подход. Например, Джошуа Нобе и Марк Альфано — авторы статей Стэнфордской философской энциклопедии, посвященных экспериментальной философии вообще и экспериментальной моральной философии в частности. Использование самоопределения скорее отсылает нас не к объективной оценке результатов и места данной методологии, а к декларируемым целям и ожиданиям, что может быть не до конца корректно с точки зрения историко-философской оценки, но вполне достаточно, исходя из целей нашей работы.

Экспериментальная философия вообще определяется как междисциплинарный подход, объединяющий попытки разрешить философские вопросы с помощью методов психологии и когнитивных наук (Knobe & Nichols, 2017). Экспериментальный подход в данном случае противопоставляется спекулятивной или кабинетной философии, он предполагает возможность прояснения философских концептов за счет обращения к представлениям и суждениям «простых людей» («folks»). А экспериментальная моральная философия составляет подраздел экспериментальной философии, имеющий своим предметом моральные интуиции, суждения и поведение.

Как и другие формы экспериментальной философии, она предполагает применение экспериментальных методов для получения данных, которые исполь-

> зуются для обоснования, опровержения или пересмотра философских теорий. В данном случае рассматриваемые теории касаются природы морального рассуждения и суждения; объема и источников моральных обязательств; природы хорошего человека и хорошей жизни; даже масштаба и характера самой теории морали (Alfano et al., 2018).

Собственно, экспериментальная философия вообще и экспериментальная моральная философия в частности активно используют методологии широкого круга эмпирических дисциплин: нейрофизиологии, психологии и социальных наук. Нейроэтика, то есть «область нейробиологических исследований когнитивных процессов, обеспечивающих моральные реакции и решения» (Апресян, 2020: 13) или исследование когнитивных и нейронных коррелятов морального суждения (Федорова, 2020: 176), может рассматриваться в качестве реализации методологической парадигмы экспериментальной философии с использованием методов нейрофизиологии. В сущности, довольно обширное пересечение исследований, которые относятся как к нейроэтике, так и к экспериментальной философии, также говорит в пользу такого соотношения двух дисциплин. Среди тех работ, которые мы будем обсуждать в данной статье, к нейроэтике и экспериментальной моральной философии относятся, в частности, исследования Грина (Greene, 2017), Кушмана и Янг (Cushman & Young, 2009), Кристенсен и Гомилы (Christensen et al., 2014; Christensen & Gomila, 2012).

Однако к экспериментальной моральной философии относятся и исследования практических вопросов применения искусственного интеллекта, например поведения самоуправляемого автомобиля в ситуации, когда он вынужден принести одного из участников ДТП в жертву («such as how a self-driving car should respond to sacrificial dilemmas» (Alfano et al., 2018)). Мы будем достаточно много говорить про такие исследования, так как, разделяя общую эмпирическую установку и методологию, они зачастую более явно выводят на первый план позицию самого исследователя.

Говоря об экспериментальной моральной философии в общем, нам также кажется важным заметить, что она не занимается выработкой собственной методологии, а просто применяет практики, сложившиеся в рамках психологии и других наук (Anstey & Vanzo, 2016: 98). Самым распространенным подходом в границах экспериментальной моральной философии является изучение «интуиций», т. е. склонности

простых людей выносить те или иные суждения в отношении философских проблем и категорий (Knobe & Nichols, 2017; Alfano et al., 2018). Предполагается, что выявление изначального смысла понятий, активно используемых философами, поможет отказаться от излишнего усложнения и надуманных споров, а установление корреляции между склонностью к тем или иным суждениям и социальными или психологическими качествами респондентов позволит свести философские и моральные споры к некоторой «реальной» основе. В таком духе рассматривается связь между моральной оценкой действия и суждением о его преднамеренности (Knobe, 2003), связь экстраверсии со склонностью утверждать свободу воли (Feltz & Cokely, 2008) или естественная распространенность компатибилистской установки (Sommers, 2010).

нравственность как преодоление данных интересов

Первый аспект экспериментальных исследований, который мы хотим обсудить, — это определение моральных поступков через их противопоставление очевидным интересам самого агента. Обоснование необходимости предпочесть общие или высшие интересы частным в данном случае можно вынести за скобки: вне зависимости от того, как это объясняется (и объясняется ли), в представление о морали заложено противопоставление морального поведения наличным интересам. Обсуждая исследования атрибуции морального намерения, связи склонности к моральному поступку со свободой воли и уровнем осознанности эмоций, а также связи склонности к суждению в духе одной из этических теорий (деонтологии или утилитаризма) с формулировкой моральной дилеммы, мы покажем, что безотносительно согласия с результатами таких исследований нравственность фактически понимается здесь как способность превзойти очевидные интересы ради некоторого высшего блага.

Такое представление о нравственности присутствует в одном из самых известных экспериментальных исследований морали — работе Джошуа Нобе о «побочных эффектах» (Knobe, 2003). Она посвящена выявлению связи между моральной оценкой поступка и атрибуцией намерений агенту, который его совершает. В эксперименте респонденты должны были оценить, являются побочные эффекты решения персонажа намеренными или же случайными. Для этого респондентам предлагались две истории: одна — о решении председателя совета директоров запустить новую экономическую стратегию, вторая — о решении лейтенанта послать подразделение на штурм высоты. У каждого из этих решений был отрицательный или положительный побочный эффект (влияние

экономической программы на окружающую среду и возможная гибель солдат), и в обоих случаях агент заявлял о своем безразличии к тому, проявится ли этот эффект. Одной части респондентов давалось описание ситуации, в которой побочный эффект был положительным (вред от производства для окружающей среды снижался, солдаты устраняли причину опасности), а другой — где этот эффект оказывался отрицательным. В результате

> Нобе обнаруживает, что участники эксперимента были намного более склонны считать, что главный герой истории намеренно вызывал побочный эффект, когда такой эффект был отрицательным (причинение вреда окружающей среде), чем в ситуации, когда побочный эффект был положительным (помощь окружающей среде). Этот психологический эффект воспроизводился десятки раз, а его область применения существенно увеличилась, включая факты приписывания намерения после нарушения как собственно моральных, так и иных норм (Alfano et al., 2018).

Однако прежде чем соглашаться с этими выводами, давайте посмотрим на то, по какому принципу в этом кейсе противопоставляются моральное и аморальное поведение. Формулировки из описания кейса содержат асимметричные подсказки о намерениях председателя: раз он говорит, что ему не важны последствия для окружающей среды, то сделать из этого выводы о его заботе о ней достаточно сложно, а вот осудить его безразличие вполне логично — это и делают респонденты. Более того, осуждение действий председателя происходит из-за вполне конкретных моральных ожиданий и ценностей (Wagner, 2014: 70):

> моральные ожидания от заинтересованного в прибыли председателя совета директоров заключаются в том, что он не будет начинать программу, чтобы избежать экологического ущерба, так как важность сохранения окружающей среды намного выше, чем получение прибыли.

Поэтому ответы агента оцениваются как решение пренебречь моральными нормами ради собственных интересов.

Респонденты имеют два варианта трактовки ситуации: либо агент видит самоцелью намерение причинить вред окружающей среде, либо он намеревается действовать вопреки (моральному) требованию заботы о ней, что приводит к негативным экологическим последствиям. В обоих вариантах мы имеем дело с намерениями, но первый из них не выглядит убедительным в контексте самой истории, а вот второй соответствует словам агента о самом себе: он открытым текстом говорит, что ему безразличны последствия. Одобрять случайно полученный позитивный

результат поэтому нет никакого смысла (Wagner, 2014: 70–71), осудить же пренебрежение общими моральными принципами очень даже логично. И в этом смысле никакой асимметрии приписывания намерения тут нет: респонденты осуждают намерение действовать вопреки моральным требованиям, а не намерение получить негативный (или позитивный) побочный эффект сам по себе. Методологически (и мы еще будем говорить об этой проблеме) у нас сложилась ситуация, когда исследователь получил ответ не на тот вопрос, который задавал. Вместо того чтобы отвечать, были ли намеренными побочные эффекты решения агента, респонденты выражали свое осуждение его намерения поступиться общественной пользой.

Исходя из определения нравственности, важно обратить внимание на противопоставление морального поведения личным интересам: аморальным поведением называется следование личным интересам вопреки общим. И наоборот, моральное поведение предполагает способность отказаться от личной заинтересованности ради общего блага. Этот смысл нравственности имманентно заложен в постановке исследования, он не подвергается рефлексии или изучению, но сам определяет их координаты.

Похожее противопоставление мы находим в исследовании связи представления о свободе воли и склонности к нравственному поступку (Vohs & Schooler, 2008). Его смысл заключается в выявлении корреляции между признанием детерминизма и склонностью к аморальным поступкам. Сценарий эксперимента сводился к тому, чтобы оценить готовность респондентов жульничать после прочтения ими текста, утверждающего (или отрицающего) детерминизм. (Как несложно догадаться, респонденты, прочитавшие текст о предопределенности, были более склонны к жульничеству.)

Жульничество, которое совершали испытуемые, заключалось в том, что вместо самостоятельного выполнения сложных вычислений в уме они могли подсматривать ответы или просто пропускать вопросы. Респонденты упрощали себе жизнь, обманывая другого (в данном случае экспериментатора), их личные интересы были противопоставлены абстрактному принципу честности. В данном примере слабой выгоде противопоставляется слабое моральное требование: возможность избежать умственного напряжения не до такой степени важна, как возможность разбогатеть, а жульничество с тестом (подобно списыванию на экзамене) не приводит к явным страданиям, например, ученого-экспериментатора. Вне зависимости от способа обоснования морального требования мы

рассматриваем конфликт, в котором способность к моральному поступку определяется как способность предпочесть собственным интересам нечто более важное.

В других исследованиях связи идеи свободы и готовности к моральным поступкам это противопоставление проявляется аналогичным образом.

> Вера в свободную волю может быть решающей в мотивации людей контролировать их бессознательный порывы ради форм поведения, учитывающих интересы общества. [...] Детерминисткие убеждения предполагают, что человек не мог действовать иначе — это напоминает стандартную форму оправдания («я ничего не мог поделать»), — и, таким образом, могут побуждать людей действовать недальновидно, импульсивно, эгоистично (Baumeister et al., 2009: 261).

Моральное поведение предполагает заботу о других или «просоциальное» поведение. Ему противопоставляется эгоизм с присущими ему импульсивностью и недальновидностью — неспособностью видеть цели более важные, чем очевидные в моменте (ibid.: 267). Нравственность тождественна способности вести себя в соответствии с принципом, который выше (важнее) конкретных интересов.

Связь морали и способности действовать вопреки неосознанным порывам и реакциям может быть также раскрыта как связь морали с уровнем осознанности эмоций. Основная гипотеза и вывод проведенного на эту тему исследования звучат достаточно очевидно (Cameron et al., 2013: 723):

> ...люди могут развивать «моральную экспертизу» и формировать более взвешенные решения, когда они лучше осознают собственные эмоции. Когда люди умеют четко определять свои эмоции, в процессе формирования морального суждения они могут распознавать неадекватные эмоции и снижать их значимость. А эмоциональная подготовка к моральному решению зависит от той ясности, с которой люди осознают свой эмоциональный опыт.

То есть более обдуманные решения принимаются, когда их автор учитывает влияние на него эмоций.

Стоит обратить внимание на выражение «more informed moral judgments» — здесь дается оценка качеству морального суждения, которое оно получает вне зависимости и даже в некотором смысле вопреки эмоциям. Суждения, вызванные эмоциями, рискуют быть случайными, как говорят об этом авторы ранее (ibid.: 720):

> Мы предполагаем, что те люди, которые умеют более точно различать собственные переживания, будут выносить моральные суждения, менее подверженные эмоциональному влиянию, ведь они лучше понимают источник своих переживаний.

Моральное суждение тут противопоставляется не эмоциям как таковым, а именно случайности, неосознанности. Осознание эмоций предполагает не исключение их из процесса принятия решения, а уверенность в наличии соображений более важных, нежели очевидный эмоциональный порыв. То есть структура морального суждения допускает возможность перехода от очевидности данных через эмоции интересов и целей к не столь очевидному, но более важному содержанию.

Представление о моральном решении как о преодолении очевидно данных интересов проявляется в исследованиях корреляции моральных суждений не только со внутренними реакциями респондента, но и с описанием ситуации, в которой принимается такое решение. Это показывают исследования влияния формулировок моральных дилемм на решение респондента, проведенные Кристенсен и Гомилой (Christensen et al., 2014; Christensen & Gomila, 2012).

Вообще, для экспериментальной моральной философии характерно использование дилемм в качестве инструмента, который должен выявить механизмы работы моральной интуиции или привычного способа рассуждения. Моральная дилемма — это

> короткая история, содержащая моральный конфликт. Здесь моральный конфликт представляет собой ситуацию, в которой противоречащие друг другу моральные принципы направляют субъекта к разному поведению. А это приводит к осознанию несовместимости двух способов поведения и их последствий (ibid.: 1251).

Рассуждения респондента, решающего дилемму, должны показать его предрасположенность к утилитаристскому (здесь это понятие тождественно термину «консеквенциалистский») или деонтологическому способу определения должного.

Рассматривать противопоставление консеквенциализма и деонтологии в качестве ключевого также характерно для современной моральной теории. Но именно то, каким образом это противопоставление происходит в экспериментальных исследованиях, косвенно говорит нам о восприятии нравственности. Выбор, совершаемый при решении дилеммы, — это выбор между причинением вреда другому («жертве») ради выгоды (собственной или всеобщей) и воздержанием от этого, исходя из

моральных принципов. Согласие на причинение вреда рассматривается как склонность к утилитаристскому взгляду на мораль, а воздержание от него — как склонность к деонтологическому. Иными словами, деонтология понимается как воздержание от причинения вреда другому в ситуации очевидной выгоды от такого вреда.

Но давайте снова обратим внимание на формулировки, которые используются в этом противопоставлении (Christensen et al., 2014: 3):

> Решение причинить вред рассматривается как утилитарное моральное суждение, потому что принимается в результате взвешивания цены и выгод, тогда как решение воздержаться от причинения вреда деонтологическое, так как оно основывается на значимости требования «не убий».

Авторы исследования не называют причинение другому вреда (принесение его в жертву) морально правильным (как сделал бы последовательный утилитарист), но такое действие всегда рассматривается в качестве моральной трансгрессии, нарушения. В этом смысле получается, что противопоставляются не два принципа построения морали, а ее соблюдение и нарушение норм. Соблюдение моральной нормы тождественно воздержанию от причинения вреда другому, и оно противоречит получению выгоды самим агентом или сообществом. Понятно, что это противопоставление не абсолютно: речь не идет о том, что моральное поведение всегда противоречит личным интересам. Однако дилемма, ситуация морального выбора, — это именно конфликт между выгодой и соблюдением моральной нормы.

Противопоставление нравственной мотивации, т. е. стремления выполнить моральную норму, выгоде для агента проявляется не только в трактовке деонтологии, но и в конечных выводах о том, как формулировка дилемм влияет на принимаемое решение. Чтобы оценить эту корреляцию, Кристенсен с коллегами классифицирует дилеммы по следующим четырем факторам: по персональности вклада (Personal Force), получателю выгоды (Benefit Recipient), возможности избежать вреда (Evitability) и намеренности (Intentionality) (ibid.: 20). Поведение респондента — будет ли он стремиться любой ценой соблюсти моральное требование и воздержаться от причинения вреда другому (деонтология) или же станет действовать исходя из рационального расчета и принесет разумную жертву (утилитаризм) — коррелирует с особенностями формулировки дилеммы.

Люди склонны нарушать моральные требования, т. е. следовать утилитаризму, в ситуациях, когда это нарушение спасает лично их (а не

других или общество в целом), а также тогда, когда вред для жертвы не является неизбежным и непосредственным результатом их действий (Christensen et al., 2014: 29). С другой стороны, участники эксперимента демонстрировали деонтологический стиль рассуждений (т. е. воздерживались от причинения вреда), когда вред описывался как результат их персонального вклада, выгоду получали другие (или общество в целом), при удачном стечении обстоятельств жертва могла не пострадать, а причинение вреда было намеренным (а не побочным эффектом действий агента).

> Другими словами, когда вред другому был абстрактным, далеким и отчуждаемым от намерений агента, у участников эксперимента наблюдалась бóльшая склонность нарушить моральные нормы, нежели в ситуации, когда вред для другого описывался как практически непосредственный и оставлял ощущение «крови на руках» (ibid.).

Эти выводы могут показаться банальными: люди склонны согласиться на причинение вреда другим, если речь идет о весьма вероятной для них выгоде и не такой очевидной ответственности, тогда как если дело касается далеких абстрактных выгод и очевидной ответственности за причиненный вред, то они предпочитают соблюдать моральные нормы. При этом в данном прочтении утилитаризма и деонтологии фактически противопоставляются не две разных этических системы или два разных принципа определения моральной нормы, а забота о моральной норме (в лице деонтологии) и забота о выгоде (собственной или общей, сопровождающейся высокой личной ответственностью или нет).

Тут можно отдельно обстоятельно порассуждать о том, что это абсолютно не соответствует идеям консеквенциализма и утилитаризма как таковым. С точки зрения последовательного утилитариста, причинение ограниченного вреда ради большей выгоды — это морально правильное решение. А вот воздержание от такого поступка ради соблюдения общепринятой (или заложенной биологически) нормы следует считать морально вредным. Но Кристенсен совершенно игнорирует такое понимание утилитаризма. Для нее нарушение нормы и причинение вреда другому — это всегда «моральная трансгрессия». То есть поступать морально — значит действовать вопреки выгодам (личным или общим). Моральная норма тут очевидно противопоставляется любой выгоде, а стремление заботиться о ее соблюдении, т. е. нравственность, рассматривается как способность превзойти имеющиеся у индивида или сообщества интересы ради высшего блага.

неизбежность оценки фактически данной морали

Представление о нравственности как о способности превосходить наличные интересы ради высшего блага относится, однако, не только к простым личным интересам (вроде денежного вознаграждения или лени), оно применяется и по отношению к самим моральным нормам. Иными словами, если моральные нормы воспринимаются в качестве фактов (социальных или биологических), нравственное требование заключается в том, чтобы найти лучшую версию этих норм или превзойти их. Этот аспект экспериментальных исследований морали проявляется в обсуждении их результатов. Следует отметить, что мы говорим тут не столько о критике экспериментальной моральной философии (которую отчасти также затронем), сколько о том, как формируется пространство обсуждения результатов самими авторами исследований. Исследователь, получивший описание функционирования механизмов морали или ее актуальное содержание, как правило, переходит к вопросу о том, насколько приемлема эта мораль, не должна ли она быть улучшена. Эта дискуссия о «правильности» фактически данной морали происходит за рамками эмпирического анализа в априорном ключе. Необходимость априорно оценивать фактически данную мораль также показывает, что нравственность соотносится не с реальным, фактическим данным, а с возможным (или невозможным).

Чтобы продемонстрировать это, мы обсудим примеры исследований моральной интуиции в ее связи с принципом принятия решения (опять же, деонтология или утилитаризм), поговорим об оценке результатов обобщения реальных моральных норм в контексте создания или обучения искусственного интеллекта, затронем влияние определения понятий на моральные оценки и придем в итоге к критике экспериментальной философии. Обсуждая эти сюжеты, мы будем концентрироваться на том, что ожидает исследователь от «истинной морали» и какова может быть сама та позиция, с которой происходит оценка фактической морали.

Один из примеров перехода от констатации фактического устройства морали к ее оценке можно найти в работах Джошуа Грина, посвященных моральным интуициям. Грин исследует корреляцию разных типов нервной активности с утилитаристскими (консеквенциалистскими) и деонтологическими принципами принятия решения. Активность частей мозга, отвечающих за интуицию, фиксировалась в момент принятия «деонтологических» решений, а активность частей, отвечающих за рациональное рассуждение, соответствовала консеквенциалистским

суждениям (Cushman & Young, 2009: 11). Грин не останавливается на выявлении этой корреляции, а, переходя к теоретическому осмыслению результатов исследования, делает попытку редуцировать различия механизмов к единому принципу: «деонтологическая природа» интуиции объясняется закреплением прошлого опыта, так же как это происходит при выработке условного рефлекса. Например, интуитивная истинность запрета убивать другого человека закрепляет результаты длительного социально-эволюционного опыта, в котором убийство почти всегда приводит к дурным последствиям. Вышеизложенное позволяет утверждать, что все моральные решения являются консеквенциалистскими, просто в некоторых случаях рассуждение подменяется выработанным на опыте правилом, действующим как рефлекс. Отсюда следует, что противоречия, возникающие между интуитивным деонтологическим («нельзя убить толстяка») и рациональным консеквенциалистским суждениями («лучше убить одного толстяка, но спасти трех школьниц») при решении моральных дилемм, объясняются не фундаментально иной природой интуиции, а спецификой дилеммы, которая создает непривычный контекст, запутывающий нашу интуицию (Greene, 2017: 10).

Немного отвлекшись, стоит заметить, что в такой трактовке моральные интуиции приобретают легкий оттенок манипуляции, так как предполагают следование определенным интересам и логике, которые индивид не осознает и не принимает на рациональном уровне: он подчиняется интуиции, а не собственному решению. Эта логика очень похожа на классический подход, в котором необходимость следования интересам общества или сочувствие к другому рассматриваются как искаженная форма заботы о себе, причем настолько искаженная, что между моральным требованием и очевидностью собственной пользы находится разрыв, нуждающийся в осознанном или волевом преодолении.

Однако нас в большей степени интересует, как Грин обсуждает полученные результаты — он буквально задает вопрос: что если «интуиции, хорошие с биологической точки зрения, плохи с точки зрения моральной?» (ibid.: 6) (О том, что называется «плохим с моральной точки зрения», можно прочитать в сноске к этому пассажу: морально плохи корысть и трибализм.) Вместе с тем постановка этого вопроса необъяснима исходя из логики эмпирического исследования: если механизмы принятия моральных решений — результат эволюционного процесса, то они логически не могут быть «плохими с моральной точки зрения». Получается, что «моральная оценка» должна быть априорной, происходить

с позиции, лежащей вне опыта. В сам*ой* попытке такой оценки имплицитно содержится идея высшего блага (или принципа), с точки зрения которого можно и нужно оценивать любую фактически данную мораль. Уверенность в наличии или вера в существование превосходящей любой данный опыт позиции, с которой должна оцениваться мораль, сложившаяся, например, в результате эволюции, определяет нравственность в качестве стремления к высшему благу, не данному ни в каком опыте.

Переход к априорным суждениям ради оценки фактически данной морали особенно ярко проявляется в исследованиях, имеющих практическую направленность, — в частности, тех, цель которых — формирование правил поведения искусственного агента.

Самое известное исследование такого плана — проект MIT «Moral Machine» — представляет собой глобальный сбор мнений опрашиваемых о том, какой выбор должен совершить беспилотный автомобиль в ситуации, когда причинение вреда кому-то из участников движения неизбежно. Участникам эксперимента предлагалось выбрать жертву из альтернатив в духе «пожилая женщина или молодой мужчина», «собака или девушка» и т. п. Идея исследования заключалась в следующем: изучая, как индивиды и страны различаются по этическим предпочтениям, можно прийти к созданию «универсальной машинной этики» (Awad et al., 2018: 59).

Задачи исследования предполагали, что придется столкнуться с проблемой отличий моральных кодексов разных культур: одним из результатов исследования стала карта глобальных различий по вопросу о выборе «предпочтительной» жертвы (молодой/ая или старый/ая, мужчина или женщина и т. д.). Но эти различия не кажутся авторам исследования критическими (ibid.: 61), напротив, утверждается, что они показывают возможность общего морального консенсуса. Здесь интересно снова обратить внимание на то, как безмятежно авторы эмпирического исследования поднимаются над полученными фактами (обходят их): они очень легко предполагают наличие такого консенсуса, который окажется важнее и действеннее, чем моральные принципы, закрепленные в разных культурах. Реальный опыт моральных суждений оказывается нефатальным — он менее важен, чем возможность общего консенсуса. Исследователи фактически рассуждают так, будто некоторая способность или стремление получить единую мораль выше и важнее фактов действительных моральных норм.

С аналогичной установкой мы имеем дело и в случаях проектирования систем, использующих машинное обучение для выдачи рекомендаций.

Предвзятость (bias) таких систем, очевидно, требует моральной коррекции и моральной оценки, причем эта оценка происходит с позиции, превосходящей фактически данные моральные нормы.

Примеры предвзятости искусственного интеллекта довольно распространены. Мы можем привести три из них: алгоритм найма сотрудников[1] предпочитал мужчин женщинам, а системы оценки необходимости дополнительного медицинского ухода за пациентами (Obermeyer et al., 2019) и оценки риска ареста за совершение правонарушения (Hao & Stray, 2013) дискриминировали чернокожих граждан. В каждом из этих случаев создатели искусственного интеллекта были вынуждены признать, что поведение их творений требует моральной коррекции.

Предвзятость искусственного интеллекта связана с техническими аспектами его создания, в частности с парадигмой машинного обучения. Дело в том, что логика принятия решений искусственным интеллектом формируется в результате его обучения на основании большого массива данных о решениях, которые в аналогичном случае принимались людьми. И, оказывается, эти решения далеко не всегда были такими, как нам хотелось бы. Проблема не только в том, что искусственный интеллект воспроизводит реальные установки рекрутеров или специалистов по страхованию, — она еще глубже. В частности, система оценки потребности в дополнительном уходе обучалась на основании данных о реальных затратах без учета того, что чернокожие граждане фактически тратят на лечение меньше не потому, что меньше нуждаются, а потому, что меньше могут себе позволить. Подобным образом чернокожих граждан чаще подвергали аресту, в том числе и по причине предвзятости правоохранительных органов. Концептуально мы имеем дело с проблемой, когда искусственный интеллект корректно воспроизводит общественную мораль «как она есть», но это не устраивает его создателя. В случае с прогнозом правонарушений чернокожими

> они также в среднем подвержены большему риску, и большая их доля будет отмечена как находящиеся в зоне риска, что одновременно корректно и некорректно (ibid.).

Создатель искусственного интеллекта желает изобрести агента, который, можно сказать, действует исходя из должного, а не сущего.

[1] См. материал Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

В результате во всевозможных гайдах и правилах создания искусственного интеллекта прописывается необходимость оценивать нравственность реальных решений и вычищать из данных неотрефлексированные предрассудки[2].

Это требование, однако, труднодостижимо не только из-за качества данных и сложности их обработки, но и потому, что моральная позиция самого́ создателя — это тоже фактически данная моральная установка, влияющая на результат. Позиция, с которой проводится ревизия моральных фактов и установок, не может быть одной из таких установок (личной или культурной), не может быть одним из таких фактов. Подобная нравственная оценка или попытка занять позицию, с которой будет происходить нравственная оценка, совершенно выходит за пределы эмпирической установки — она предполагает наличие некоторой априорной нравственной позиции, принципиально не определяемой данным в опыте. И дело тут не в том, что исследователь берет на себя право судить других людей, — его собственная мораль также несовершенна и может быть пересмотрена. Однако предполагать, что такой пересмотр вообще возможен или тем более необходим, можно, только разделяя веру в то, что нравственное определяется и обнаруживается за рамками данного в опыте.

Стремление к априорной оценке относится не только к конкретным моральным нормам, но и к этическим понятиям и категориям. Критика экспериментальной моральной философии зачастую строится также именно на указании необходимости предварительного априорного прояснения моральных понятий. Так, можно говорить о том, что сравнение суждений различных респондентов не имеет смысла без предварительного достижения консенсуса относительно философских и моральных категорий, а попытка прояснить эти категории через эмпирическое исследование запускает логический круг.

Факт влияния опыта априорного анализа моральных категорий на моральные суждения выявился в эксперименте Кевина Тобиа. Его суть заключалась в том, что двум группам людей — одну из них составляли профессиональные философы (т. е. научные работники) — были предложены гипотетические ситуации, в которых следовало решить,

[2]См. Guidance for Regulation of Artificial Intelligence Applications: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf. А также позицию руководства Google: https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04.

должен ли агент принести случайного человека в жертву для спасения других. Одной части респондентов давалось безличное описание этой ситуации (принять решение должен был некий «Джек»), а другой — персонифицированное (принятие решения возлагалось лично на респондента). В итоге устанавливалась корреляция между персональностью ответственности и готовностью принести жертву. Корреляции в этих двух группах были различными: философы оказались более склонными к тому, чтобы в ситуации персональной ответственности принимать морально сложное решение, — они чаще отвечали: «Да, я обязан убить туземца» или «Да, допустимо убить пятерых» (Tobia et al., 2013: 631–634).

Интерпретируя полученные результаты, можно говорить о том, что философы реже убегали от ответственности, т. е. «возможно, философы более готовы к тому, чтобы оценивать ситуацию с моральной точки зрения и брать на себя моральные обязательства» (ibid.: 636). Но вполне вероятно также, что они были лучше подготовлены к оцениванию моральной составляющей ситуации рационально — вопреки собственным первоначальным интуициям. В любом случае этот эксперимент показывает, что разный уровень подготовки в работе с этическими категориями существенно влияет на ответы респондентов. И речь тут не о том, что смещение акцентов в понимании категорий может изменить моральное решение, потому что они сложны (sophisticated), а о том, что их прояснение не может происходить через экспериментальные процедуры или за счет них. Сами результаты экспериментальных процедур зависят от тонких аспектов понимания респондентами моральных и философских концептов.

Эта зависимость может быть продемонстрирована на примере эксперимента, устанавливающего корреляцию между представлением о детерминизме и идеей свободы воли: разница в нюансах понимания детерминизма сказывается на готовности респондентов признать наличие или отсутствие свободы воли. Участникам эксперимента задавали вопрос о моральной ответственности выдуманных жителей другой планеты, предопределенность поведения которых научно доказана. Разным группам испытуемых предлагались истории, содержащие разные описания детерминизма: поведение жителей определялось либо химическими реакциями и устройством нервной системы (механистически), либо мыслями, желаниями и планами (психологически) (Nahmias et al., 2007: 224). Участники, прочитавшие механистическое определение, были склонны

отрицать возможность свободы воли, тогда как прочитавшие психологическую трактовку не видели проблемы совместимости детерминизма со свободой воли (Nahmias et al., 2007: 232).

Здесь важны нюансы понимания детерминизма. Изначально он определяется как «тезис о том, что связь (возможно, причинно-следственная) между состоянием системы в разное время управляется законами природы» (ibid.: 215), и это в общем-то вполне корректно и строго. Однако в сюжете, который предлагался участникам описанного в предыдущем абзаце эксперимента, эта строгость теряется. Психологический вариант выглядит так (при удалении механистического варианта из текста) (ibid.: 224):

> Психологи также открыли, что мысли, чувства, желания и планы полностью определяются текущей ситуацией в жизни эртанина (выдуманного жителя другой планеты. — А. Ж.) и предыдущими событиями из его или ее жизни. Эти предыдущие события также полностью определяются более ранними событиями, в итоге восходя к событиям, которые произошли до его или ее рождения.

И хотя здесь сказано, что мысли и желания определяются событиями прошлого, остается существенный пробел в понимании участия в них самого агента. Мысли и планы вполне могут восприниматься как «продукт» собственной деятельности человека, как нечто, что он может менять. Мысли и планы — мои, они могут принадлежать мне в том смысле, что я ими распоряжаюсь. И это их принципиальное отличие от химических реакций, которые уж точно не мои. В этом смысле высказывание «Ваше поведение полностью определяется вашими желаниями» звучит тождественно тому, что «Ваше поведение полностью определяется вашим выбором желать одно или другое». Поэтому психологическая формулировка содержит возможность подрыва детерминизма: респонденты, прочитавшие эту версию, могли вообще не понять, что в данном случае речь идет о нем.

Для нас очевидно, что как между разными респондентами, так и между респондентами и исследователем существует заметная разница в понимании используемых в эксперименте терминов, и это — основа для серьезной критики экспериментальной моральной философии. Хотя первоначальная цель экспериментальных исследований звучит как прояснение философских понятий через обращение ко мнениям простых людей, в итоге оказывается, что моральные категории вовсе не проясняются в результате изучения мнений — напротив, они только запутываются.

Об этом много и, на наш взгляд, довольно обоснованно говорит Каупинен, формулируя собственную критику экспериментальной философии.

В качестве примера приведем разбор (мнимого) противоречия между выводами двух эмпирических исследований, посвященных связи морального суждения и мотивации (Kauppinen, 2007: 98–99). В первом исследовании утверждается, что способность морального суждения, очевидно, связана с личной мотивацией, — вывод делается на основании согласия респондентов с тем, что некто (Джордж) не может судить о том, что война — это хорошо, если сам он ее избегает. Во втором эксперименте выводится обратное, и это обосновывается согласием респондентов с тем, что убийца может понять аморальность собственных действий, даже если у него отсутствует способность испытывать сострадание к своим жертвам. Противоречие между выводами определяется использованием разных концепций нравственной мотивации: мотивация делать что-то не тождественна эмоциональной реакции (эмоциональным мотивам). Это наблюдение за разным употреблением понятий позволяет Каупинену предложить в качестве ключевого поинта своей критики экспериментальной философии «аргумент о несогласии»: проблема экспериментальных исследований философских концептов заключается в различной трактовке необходимых для эксперимента понятий исследователем и респондентами, а также различными исследователями. Это приводит к тому, что результаты экспериментов оказываются несравнимыми между собой.

Получается, что исследования народных интуиций не имеют смысла без прояснения исследуемых концептов. Это прояснение, о котором мы говорим, относится не только к специальным терминам и понятиям, таким как «детерминизм», «компатибилизм» и другие измы, но и к общим, привычным, вроде понятия «моральная ответственность». Экспериментальная моральная философия сама показывает, что невозможно определить основополагающие моральные (и философские вообще) понятия эмпирическим путем, на опыте. Напротив, на интерпретацию экспериментальных данных влияет определение этих понятий, которое должно быть сделано до опыта или в не-опытном поле, скажем, в пространстве диалога, как предлагает, например, Каупинен (ibid.: 110–111). И здесь может быть задан практически кантовский вопрос о том, как можно неэмпирически провести определение нравственных категорий. На основании чего и с какой позиции такое определение может быть получено? И что это скажет нам о природе нравственности?

Следует заметить, что схожая проблема обнаруживается в обсуждении моральных принципов искусственного интеллекта. Например, в свежей работе Ясона Габриэля обосновывается необходимость регулирования искусственного интеллекта с опорой именно на моральные ценности, а не на прямые инструкции или обязанности заботиться о благополучии людей. При этом сами такие ценности в итоге оказываются лишены конкретного или позитивного содержания. Они вовсе не даны в качестве ясных, четких установок, они не являются фактами общественного мнения, механизмов психики, нервной системы, равно как и консенсусом в отношении определения. Но искомые моральные ценности функционируют в роли указателей на некоторую цель или некоторое содержание, которое они пытаются выразить. Поэтому, как замечает сам Габриэль, моральные ценности — это предмет веры, а «искусственный интеллект должен быть согласован с верой в ценности, а не с ценностями как таковыми» (Gabriel, 2020). Например, справедливость — это не конкретное представление о правильном устройстве общества, а смутный идеал, который конструируется, а не констатируется. Это вера в то, что есть нечто называемое справедливостью или нечто, что выражается в справедливости. Рассматриваемая вера в ценности концептуально очень близка к выводу, который мы сделали, наблюдая отношение исследователей к фактической морали: ожидание лучшей, более правильной морали не основано на фактах.

В начале этой части статьи мы утверждали, что любое описание фактической морали не удовлетворяет самих исследователей. Полученная в результате их исследований мораль не принимается в качестве той, которую мы будем теперь использовать. Вместо этого исследователь ищет способ оценить ее, сделать лучше. Следствием такого положения оказывается неудовлетворенность в понимании основных этических концептов — это понимание также не может быть принято как факт. Оно требует прояснения этих концептов или выяснения того смысла и значения, которое они должны иметь, тем более что такое значение непосредственно влияет на понимание моральных суждений. И это прояснение также не имеет свойства эмпирической констатации фактов или связей. Просто выяснить, на какое явление (по мнению большинства) эти слова указывают, не представляется возможным — нет реальных объектов, на которые они бы указывали, а достижение договоренности относительно их значения тождественно ведению моральной дискуссии.

Таким образом, мы можем констатировать, что логика эмпирического, экспериментального исследования морали приводит к запуску априорного обсуждения морали, которое предполагает принципиальную недостаточность любых наблюдений за фактическим положением вещей.

Вкупе с замечанием о том, что моральное действие всегда превосходит известные интересы и мотивы, этот тезис предполагает необходимость признать направленность нравственности к благу, находящемуся за пределами любого возможного опыта. Ниже мы набросаем перспективу того, что это может значить для этики.

к идее нравственности

Для того чтобы обозначить перспективу, создаваемую вышеизложенными наблюдениями, имеет смысл кратко вспомнить логику, в которой разворачивалось наше обсуждение экспериментальной моральной философии. Наш подход можно назвать косвенным в том смысле, что, не вступая в прямую дискуссию относительно метода или выводов, мы пытались заметить установки и идеи, присущие исследователям-экспериментаторам и формирующие координаты их понимания нравственности. Что же мы обнаруживаем таким косвенным образом?

Прежде всего, исследователи-экспериментаторы идентифицируют моральные поступки как совершаемые ради целей, превосходящих наличные интересы агента. Абстрагируясь сейчас от объяснения причин такого поведения (это объяснение может быть вполне традиционным: будущие выгоды предпочитаются сиюминутным), мы должны признать, что форма морального действия, остающаяся после вынесения за скобки теоретических рассуждений о природе нравственной мотивации, ее содержании или источнике, заключается именно в том, чтобы действовать ради высшего блага. Это утверждение о моральном действии не является результатом эмпирической проверки, но скорее представляет собой рамку экспериментальных исследований — мы обнаруживаем его косвенно, — но оно же и указывает на содержание того, что мы называем нравственностью как таковой.

Вторая черта, присущая эмпирическим исследованиям, — стремление (или необходимость) исследователей оценивать полученную в результате эксперимента мораль. Эта черта отличает экспериментальную моральную философию не только от физики, но и от близкой к ней психологии. Грубо говоря, никто не пытается оценивать фактическое устройство психики, например когнитивные искажения, — оно принимается как

норма, которой можно определенным образом оперировать (в том числе избегать ее или преодолевать). С моралью такого не происходит: представления обычных людей о должном подвергаются оценке и обсуждению; предполагается, что людям следует иметь иные, лучшие моральные представления. Добавим сюда, что оценка описанной морали связана также с обсуждением этических понятий, которые тоже не получается определить эмпирически. Это порождает философскую дискуссию, в которой понятия скорее конструируются, нежели оказываются констатированы. Здесь прослеживается тесная связь с первым замечанием: так как благо, ради которого совершается моральный поступок, не может быть дано в опыте, то и истинная мораль не может быть сведена к констатации фактического положения вещей, интересов или правил.

Вышеизложенные замечания можно тезисно обобщить: экспериментальная моральная философия показывает нравственность в качестве способности действовать ради такого блага, которое превосходит данное нам в опыте. Поступок считается нравственным, если он совершается не ради очевидных интересов, но ради такого высшего, которое определяется исключительно априорно.

Этот тезис на первый взгляд может показаться поверхностным и лишенным практического смысла: действительно важный вопрос о нравственности — как определить эти поистине высшие интересы. И в самом деле, авторы экспериментальных исследований остаются во многом на уровне обыденных представлений о морали, они чаще всего не делают сложных философских выводов и тем более не исходят из фундаментальных этических теорий. Однако, во-первых, экспериментальная моральная философия и не ставит себе такой цели — напротив, ее цель обратна: она заключается в том, чтобы избегать излишней терминологии. И поэтому нет ничего удивительного, что собственные установки исследователей выглядят обыденными. Философская рефлексия над этими установками, обнаружение их в контексте истории этики и этических доктрин могли бы иметь отдельный смысл и стать целью отдельной работы, которая не вмещается в наш текст по объему. Во-вторых мы полагаем, что не стоит упрощать и недооценивать потенциал наших наблюдений о смысле нравственности: их последовательное раскрытие может привести к двум довольно неожиданным выводам.

Первый вывод такой: если нравственность заключается в том, чтобы превосходить любое благо, данное в опыте, это значит, что попытки обосновать мораль, свести ее к любому закону теряют смысл. Так, классическая попытка обнаружить высшие интересы, ради которых

следует принести в жертву личную выгоду, чаще всего предполагает существование данности или реальности, которой нравственность подчинена. И сами экспериментальные исследования часто содержат в себе эту установку, т. е. предполагают, что нравственность выражает природную заботу о себе или стремление размножаться. Но, как мы показывали выше, само обсуждение полученной в результате исследований морали подрывает этот ход рассуждений. Если наши действия определяются эволюционно-биологической логикой, то нравственно будет научиться ее превосходить. Если мораль можно зафиксировать в качестве социальных договоренностей о правильном, то нравственным станет поиск и достижение лучших договоренностей. Если основной мотив — это стремление к благополучию, то способность пренебрегать сиюминутным благополучием ради будущего окажется тождественной способности пренебречь благополучием в принципе. Таким образом, подход к обсуждению экспериментальной моральной философии может быть для нас моделью критики классической попытки обосновать мораль некоторым общим законом: этот подход показывает, что нравственность как таковая предполагает превосхождение любой обоснованной моральной доктрины.

Вот второй вывод: признание, что идея нравственности заключается в стремлении к высшему благу, означает, что нет и не может быть никакого позитивного критерия или данного в опыте закона, соблюдение которого или подчинение которому гарантирует моральное поведение, — любой такой данный критерий или закон должен быть превзойден. Также это означает, что сам внутренний критерий нравственности — это только само по себе стремление к высшему благу. Мыслить нравственность трансцендентально — значит перестать искать позитивные основания или мотивы действовать в соответствии с моральными нормами. Единственным мотивом быть нравственным может служить только разделение агентами веры в возможность высшего блага. Нравственность, не определенная никаким данным опытом, никаким устройством мира, принимается свободно только в качестве некоторой возможности. Следствием такого понимания нравственности становится утрата ею дисциплинарного смысла: она более не может быть обязательной для всех. Вера в возможность такого блага делает нравственное стремление разумным и необходимым, но сам шаг веры остается при этом совершенно свободным.

Литература

Апресян Р. Г. Нейроэтика : вызовы и недосмотры // Философия. Журнал Высшей школы экономики. — 2020. — Т. 4, № 1. — С. 13—23.

Федорова М. В. Нейроэтика «тогда и сейчас» // Философия. Журнал Высшей школы экономики. — 2020. — Т. 4, № 1. — С. 171—199.

Alfano M., *Loeb D.*, *Plakias A.* Experimental Moral Philosophy / The Stanford Encyclopedia of Philosophy ; ed. by E. N. Zalta. — 2018. — URL: https://plato.stanford.edu/archives/win2018/entries/experimental-moral/ (visited on Oct. 10, 2021).

Anstey P. R., *Vanzo A.* Experimental Philosophy // In A Companion to Experimental Philosophy / ed. by J. Sytsma, W. Buckwalter. — Oxford : Blackwell, 2016. — P. 87–102.

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

Baumeister R. F., *Masicampo E. J.*, *DeWall C. N.* Prosocial Benefits of Feeling Free : Disbelief in Free Will Increases Aggression and Reduces Helpfulness // Personality and Social Psychology Bulletin. — 2009. — Vol. 35, no. 2. — P. 250–268.

Cameron C. D., *Payne B. K.*, *Doris J. M.* Morality in High Definition : Emotion Differentiation Calibrates the Influence of Incidental Disgust on Moral Judgments // Journal of Experimental Social Psychology. — 2013. — Vol. 49, no. 4. — P. 719–725.

Christensen J., *Flexas A.*, *Calabrese M.* Moral Judgment Reloaded : A Moral Dilemma Validation Study // Frontiers in Psychology. — 2014. — Vol. 5. — P. 607.

Christensen J. F., *Gomila A.* Moral Dilemmas in Cognitive Neuroscience of Moral Decision-Making : A Principled Review // Neuroscience and Biobehavioral Reviews. — 2012. — No. 36. — P. 1249–1264.

Cushman F., *Young L.* The Psychology of Dilemmas and the Philosophy of Morality // Ethical Theory and Moral Practice. — 2009. — Vol. 12, no. 1. — P. 9–24.

Feltz A., *Cokely E. T.* Do Judgments About Freedom and Responsibility Depend on Who You Are? Personality Differences in Intuitions About Compatibilism and Incompatibilism // Consciousness and Cognition. — 2008. — Vol. 18, no. 1. — P. 342–350.

Gabriel I. Artificial Intelligence, Values, and Alignment // Minds & Machines. — 2020. — Vol. 30, no. 3. — P. 411–437.

Greene J. D. The Rat-a-gorical Imperative : Moral Intuition and the Limits of Affective Learning // Cognition. — 2017. — Vol. 167. — P. 66–77.

Hao K., *Stray J.* Can You Make AI Fairer Than a Judge? / MIT Technology Review Play Our Courtroom Algorithm Game. — 2013. — URL: https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/ (visited on Oct. 10, 2021).

Kauppinen A. The Rise and Fall of Experimental Philosophy // Philosophical Explorations. — 2007. — Vol. 10, no. 2. — P. 95–118.

Knobe J. Intentional Action and Side Effects in Ordinary Language // Analysis. — 2003. — Vol. 63, no. 3. — P. 190–194.

Knobe J., Nichols S. Experimental Philosophy / The Stanford Encyclopedia of Philosophy ; ed. by E. N. Zalta. — 2017. — URL: https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/ (visited on Oct. 10, 2021).

Nahmias E., Coates J., Kvaran T. Free Will, Moral Responsibility, and Mechanism : Experiments on Folk Intuitions // Midwest Studies in Philosophy. — 2007. — Vol. 31. — P. 214–242.

Obermeyer Z., Powers B., Vogeli C. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations // Science. — 2019. — Vol. 366, no. 6464. — P. 447–453.

Sommers T. Experimental Philosophy and Free Will // Philosophy Compass. — 2010. — Vol. 5, no. 2. — P. 199–212.

Tobia K., Buckwalter W., Stich S. Moral Intuitions : Are Philosophers Experts? // Philosophical Psychology. — 2013. — Vol. 26, no. 5. — P. 629–638.

Vohs K. D., Schooler J. W. The Value of Believing in Free Will : Encouraging a Belief in Determinism Increases Cheating Behavior // Psychological Science. — 2008. — Vol. 19, no. 1. — P. 49–54.

Wagner V. Explaining the Knobe Effect // Experimental Ethics / ed. by C. Luetge, H. Rusch, M. Uhl. — London : Palgrave Macmillan, 2014.

Дроздова Д. Н. Экспериментальная философия 2.0 : новое вино в старых мехах? // Философия и наука : проблемы соотнесения. Алёшинские чтения 2016 / ed. by Т. А. Шияна. — М. : РГГУ, 2017. — P. 119–128.

Zheleznov, A. S. 2022. "Ideya nravstvennosti v eksperimental'noy moral'noy filosofii [The Idea of Morality in Experimental Moral Philosophy]" [in Russian]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 6 (3), 181–207.

Andrey Zheleznov
PhD in Philosophy; Independent Researcher (Moscow, Russia); orcid: 0000-0001-9516-2392

The Idea of Morality in Experimental Moral Philosophy

Submitted: Oct. 06, 2021. **Reviewed:** Nov. 29, 2021. **Accepted:** July 22, 2022.

Abstract: This paper aimed to show a concept of the moral that is indirectly contained in the experimental moral philosophy. This concept can also contradict the very idea of experimental moral philosophy. A significant number of studies are called experimental moral philosophy, as they use the methodological paradigm of applying the methods of the natural and social sciences to study the mechanisms of the functioning of morality. When discussing such studies, the focus will be on their direct results. This paper aimed to show a concept of the moral that is indirectly contained in the experimental moral philosophy. Experimental moral philosophy

is a methodological paradigm that uses methods of natural and social science to explore how morality works. In the paper we will not discourse directly on the results of such research, but pay attention to the expectations from morality, which can be shown indirectly. Thus, we can find that experimental moral philosophy shows morality as the pursuit of a good that transcends any given experience. There are two main aspects of morality, which are revealed in the discussion of experimental research. Firstly, moral driving action is performed for the sake of goals that oppose the apparent interests of the agent themselves. No matter how these goals were justified, moral action itself is made possible by the ability to see something more important than apparent self-interests. Secondly, a particular position or procedure for assessing the correctness of discovered norms and mechanisms is assumed concerning the discussion of the empirical results of the experimental research. The possibility of this position/procedure does not result from the facts but still presupposes an a priori way of thinking.

Keywords: Ethics, Morality, Experimental Moral Philosophy, Moral Dilemmas, Bias, Methodological Paradigm.

DOI: 10.17323/2587–8719–2022–3–181–207.

REFERENCES

Alfano, M., D. Loeb, and A. Plakias. 2018. "Experimental Moral Philosophy." Ed. by E. N. Zalta. The Stanford Encyclopedia of Philosophy. Accessed Oct. 10, 2021. https://plato.stanford.edu/archives/win2018/entries/experimental-moral/.

Anstey, P. R., and A. Vanzo. 2016. "Experimental Philosophy." In *In A Companion to Experimental Philosophy*, ed. by J. Sytsma and W. Buckwalter, 87–102. Oxford: Blackwell.

Apresyan, R. G. 2020. "Neyroetika [Neuroethics]: vyzovy i nedosmotry [Challenges and Omissions]" [in Russian]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philisophy. Journal of the Higher School of Economics]* 4 (1): 13–23.

Awad, E., S. Dsouza, and R. Kim. 2018. "The Moral Machine Experiment." *Nature* 563:59–64.

Baumeister, R. F., E. J. Masicampo, and C. N. DeWall. 2009. "Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness." *Personality and Social Psychology Bulletin* 35 (2): 250–268.

Cameron, C. D., B. K. Payne, and J. M. Doris. 2013. "Morality in High Definition: Emotion Differentiation Calibrates the Influence of Incidental Disgust on Moral Judgments." *Journal of Experimental Social Psychology* 49 (4): 719–725.

Christensen, J., A. Flexas, and M. Calabrese. 2014. "Moral Judgment Reloaded: A Moral Dilemma Validation Study." *Frontiers in Psychology* 5:607.

Christensen, J. F., and A. Gomila. 2012. "Moral Dilemmas in Cognitive Neuroscience of Moral Decision-Making: A Principled Review." *Neuroscience and Biobehavioral Reviews*, no. 36, 1249–1264.

Cushman, F., and L. Young. 2009. "The Psychology of Dilemmas and the Philosophy of Morality." *Ethical Theory and Moral Practice* 12 (1): 9–24.

Drozdova, D. N. 2017. "Eksperimental'naya filosofiya 2.0 [Experimental Philosophy 2.0]: novoye vino v starykh mekhakh? [New Wine in the Old Skins]." In *Filosofiya i nauka [Philosophy and Science] : problemy sootneseniya. Alëshinskiye chteniya 2016 [Problem of Correlation. Aljoshinskie Chtenija 2016]*, ed. by T. A. Shiyan, 119–128. Moskva [Moscow]: RGGU.

Fedorova, M. V. 2020. "Neyroetika 'togda i seychas' [Neuroethics 'Then and Now']" [in Russian]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philisophy. Journal of the Higher School of Economics]* 4 (1): 171–199.

Feltz, A., and E. T. Cokely. 2008. "Do Judgments About Freedom and Responsibility Depend on Who You Are? Personality Differences in Intuitions About Compatibilism and Incompatibilism." *Consciousness and Cognition* 18 (1): 342–350.

Gabriel, I. 2020. "Artificial Intelligence, Values, and Alignment." *Minds & Machines* 30 (3): 411–437.

Greene, J. D. 2017. "The Rat-a-gorical Imperative: Moral Intuition and the Limits of Affective Learning." *Cognition* 167:66–77.

Hao, K., and J. Stray. 2013. "Can You Make AI Fairer Than a Judge?" MIT Technology Review Play Our Courtroom Algorithm Game. Accessed Oct. 10, 2021. https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

Kauppinen, A. 2007. "The Rise and Fall of Experimental Philosophy." *Philosophical Explorations* 10 (2): 95–118.

Knobe, J. 2003. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63 (3): 190–194.

Knobe, J., and S. Nichols. 2017. "Experimental Philosophy." Ed. by E. N. Zalta. The Stanford Encyclopedia of Philosophy. Accessed Oct. 10, 2021. https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/.

Nahmias, E., J. Coates, and T. Kvaran. 2007. "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions." *Midwest Studies in Philosophy* 31:214–242.

Obermeyer, Z., B. Powers, and C. Vogeli. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–453.

Sommers, T. 2010. "Experimental Philosophy and Free Will." *Philosophy Compass* 5 (2): 199–212.

Tobia, K., W. Buckwalter, and S. Stich. 2013. "Moral Intuitions: Are Philosophers Experts?" *Philosophical Psychology* 26 (5): 629–638.

Vohs, K. D., and J. W. Schooler. 2008. "The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating Behavior." *Psychological Science* 19 (1): 49–54.

Wagner, V. 2014. "Explaining the Knobe Effect." In *Experimental Ethics*, ed. by C. Luetge, H. Rusch, and M. Uhl. London: Palgrave Macmillan.