УДК: 141.3
DOI: 10.15372/PS20210408

И.К. Ставровский

РАВНОВЕСИЕ ВИНЕРА И ТЕХНИЧЕСКИЙ ПРОГРЕСС

Технический прогресс может быть описан в виде игры под названием дилеммы изобретателя, которая является аналогом дилеммы заключенного. Два субъекта, которые независимо друг от друга разрабатывают потенциально опасную технологию, скорее всего, продолжат заниматься ее созданием, даже если осознают опасность и смогут договориться. Сама ситуация, в которой они находятся, подталкивает их продолжить работу. Таким образом, появление потенциально опасных технологий практически неизбежно. Это значит, что вместо того чтобы пытаться предотвратить появление таких технологий, мы должны сосредоточиться на поиске способа нивелировать или компенсировать негативные последствия их появления.

Ключевые слова: равновесие Нэша; равновесие Винера; дилемма заключенного; дилемма изобретателя; теория игр; технический прогресс; искусственный интеллект; философия техники

I.K. Stavrovsky

WIENER EQUILIBRIUM AND TECHNOLOGICAL PROGRESS

Technological progress can be described as a game called the Inventor's Dilemma which is analogous to the Prisoner's Dilemma. Two actors who independently work on creating a potentially dangerous technology are likely to continue to develop it, even if they recognize the danger and come to an agreement. The very situation, in which they are, prompts them to continue working. Thus, it is virtually inevitable that potentially dangerous technologies will appear. This means that instead of trying to prevent the emergence of such technologies we should focus on finding a way to neutralize or compensate for the negative consequences of their emergence.

Keywords: Nash equilibrium; Wiener equilibrium; Prisoner's Dilemma; Inventor's Dilemma; game theory; technological progress; artificial intelligence; philosophy of technology

В своей книге «Кибернетика, или Управление и связь в животном и машине» Н. Винер, рассуждая о возможности сдержать развитие технологий, писал следующее: «Мы даже не имеем возможности задержать новые технические достижения. Они носятся в воздухе, и самое большее, чего добился бы кто-либо из нас своим отказом от исследований по кибернетике, был бы переход всего дела в руки самых без-

ответственных и самых корыстных из наших инженеров» [2, с. 78]. Это означает, что даже если разрабатываемая технология является потенциально опасной, то отказ от работы над ней может стать не лучшим решением. Объяснение того, почему дела обстоят именно так, и является целью настоящей статьи. В наших рассуждениях мы используем некоторые модели из теории игр для большей наглядности, не обращаясь к математическому аппарату, поскольку в данном случае он избыточен. Однако мы не исключаем возможность формализации излагаемых в статье концепций.

Итак, для иллюстрации мы возьмем равновесие Нэша, а затем модифицируем его под наши цели. Равновесием Нэша принято называть ситуацию в конечной игре, когда ни один из участников не может улучшить свое положение, изменив стратегию, при условии что и другие участники не меняют свою стратегию [5, с. 741]. Классической иллюстрацией равновесия Нэша является *дилемма заключенного*.

Допустим, что полиции удалось схватить двух преступников А и Б во время совершения ими мелкой кражи. При этом полиции известно, что ранее А и Б ограбили банк, но доказательств недостаточно, чтобы посадить их в тюрьму за это преступление. Однако у полиции появляется хитрый план. Преступников изолируют друг от друга и каждому предлагают сделку: если один свидетельствует против другого, а другой хранит молчание, то первый сразу же выходит на свободу, но второй отправляется в тюрьму на 10 лет. Если оба молчат, то каждый получает полгода тюрьмы за мелкую кражу. Наконец, если оба дают показания друг против друга, то каждый получает по 2 года. Важно помнить, что преступники никак не могут договориться. Более того, у них нет причин для уверенности в преданности друг друга. Так мы получаем игру, у которой может быть четыре исхода:

1) оба преступника отказываются давать показания и получают полгода каждый;

2) А дает показания, Б молчит. А выходит на свободу, Б отправляется в тюрьму на 10 лет;

3) Б дает показания, А молчит. Б выходит на свободу, А отправляется в тюрьму на 10 лет;

4) оба преступника дают показания и получают по 2 года.

В данной игре равновесием Нэша является четвертый вариант, так как даже при условии возможности отказаться от своих показаний ни

один из заключенных не может улучшить свое положение, если этого не сделает другой.

По аналогии *равновесием Винера* мы будем называть подобную ситуацию в похожей игре. Для большей схожести назовем ее *дилеммой изобретателя*.

Допустим, у нас есть два изобретателя А и Б, которые независимо разрабатывают системы искусственного интеллекта (ИИ) для военных целей. Волей случая оба изобретателя независимо друг от друга придумывают способ, как создать систему ИИ настолько эффективную, что ее использование гарантирует полную и безоговорочную победу в любом военном конфликте. Поэтому тот, кто создаст данную систему ИИ первым, не только получит признание и деньги, но также гарантирует себе или своему спонсору колоссальную власть. Очевидно, что такая перспектива хотя и может показаться соблазнительной, вместе с тем несет очевидную опасность. Не исключено, что даже для того, кто первым создаст столь мощный искусственный интеллект. Потому изобретатели могут решить, что лучше полностью отказаться от подобных разработок. Хорошая новость в том, что, в отличие от заключенных, изобретатели хотя бы гипотетически могут договориться. Но, как и у заключенных, у них нет причин быть полностью уверенными в честности друг друга. В конце концов, никто не отменял профессиональное честолюбие, и даже изобретатели не лишены жажды власти, славы и богатства. Таким образом, мы получаем еще одну игру, у которой может быть четыре исхода:

1) оба изобретателя отказываются разрабатывать потенциально опасный ИИ;

2) изобретатель А отказывается продолжать работу, изобретатель Б продолжает разработку потенциально опасного ИИ;

3) изобретатель Б отказывается продолжать работу, изобретатель А продолжает разработку потенциально опасного ИИ;

4) оба изобретателя продолжают разрабатывать потенциально опасный ИИ.

Равновесие Винера наступает при четвертом исходе. То есть равновесием Винера мы будем называть ситуацию, когда оба изобретателя (неважно, идет ли речь об отдельных людях, компаниях или даже государствах) принимают решение продолжить разработку технологии, несмотря на потенциальную угрозу.

Дилемма изобретателя очень похожа на дилемму заключенного, однако есть и отличия. Главное из них относится к тому, что в данной игре не очевидно, принесет ли продолжение работы над проектом позитивные, негативные или вообще хоть какие-то результаты в принципе. Возможно, изобретатели ошиблись, а потому на первый взгляд гениальная идея оказалась провальной и не принесет никакого результата. Возможно, разработка займет слишком много времени и потому не получит финансирования. Возможно, изобретателю не хватит ресурсов, для того чтобы довести работу до конца. Однако эти уточнения изменяют ситуацию несущественно, ведь хотя продолжение разработок не гарантирует успеха, отказ от работы практически гарантирует его отсутствие, исключая откровенно фантастические сценарии.

Важно оговориться, что искусственный интеллект был выбран исключительно для иллюстрации. Вместо ИИ можно было бы поставить другую технологию. Главное, чтобы она хотя бы гипотетически могла принести серьезную опасность и эта опасность осознавалась. Отказываться от разработки технологий, которые не представляют опасности, или технологий, опасность которых не осознается, можно либо из-за их низкой рентабельности, либо в силу различного рода внешних ограничений (финансовых, технических, временных), либо по идеологическим причинам. То есть неопасные технологии не представляют большого интереса в рамках обсуждаемой проблемы.

Количество конкурирующих изобретателей также не имеет существенного значения. Увеличение числа игроков лишь снизит вероятность того, что им удастся договориться, так как каждый должен доверять всем прочим участникам. При этом увеличивается вероятность того, что как минимум один из игроков захочет продолжить разработки даже вопреки достигнутым договоренностям. Но что интересно, даже если мы говорим только об одном изобретателе, это почти не повлияет на игру. Единственное принципиальное отличие будет в том, что на его решение в меньшей степени будет влиять опасение, что кто-то разработает технологию раньше. С другой стороны, поскольку риск того, что кто-то создаст технологию раньше, явно меньше (так как другим ее еще надо придумать), то и поводов отказываться от дальнейших разработок тоже меньше. Потому в конечном итоге суть игры меняется не так принципиально, как это было бы, если бы в дилемме заключенного мы оставили только одного преступника.

Следует отметить, что при анализе различных социальных и политических процессов часто забывают, что дилемма заключенного яв-

ляется не единственным типом игры в теории игр. Игнорирование данного факта может склонять к тому, чтобы рассматривать конфликт как оптимальный способ решения различных проблем даже тогда, когда предпочтительно сотрудничество [3, с. 128–129]. Действительно, если отношения между государствами, компаниями, сообществами или индивидами рассматриваются через призму дилеммы заключенного, то может сложиться впечатление, что стратегия предательства является предпочтительной. Однако в реальной жизни мы гораздо чаще сталкиваемся с ситуацией, которую Д. Хофштадтер иллюстрирует с помощью сюжета с обменом закрытыми сумками [5, с. 741–760]:

Продавец А и покупатель Б изо дня в день встречаются в условленном месте и обмениваются закрытыми сумками. Один из участников обмена кладет в свою сумку товар, а второй – деньги. Каждый из участников обмена может обмануть другого или сам оказаться обманутым, так как сумку нельзя проверить на месте. Таким образом, обмануть второго участника обмена можно безнаказанно. Как и в двух предыдущих примерах, в данной игре может быть четыре исхода:

1) А и Б не пытаются друг друга обмануть, обмен продолжается;
2) А обманывает Б, обмен прекращается;
3) Б обманывает А, обмен прекращается;
4) А и Б обманывают друг друга, обмен прекращается.

На первый взгляд кажется, что описанная игра незначительно отличается от дилеммы заключенного. Однако одно отличие этой ситуации в корне меняет игру. В дилемме заключенного мы сталкиваемся с единичной игрой, которая заканчивается, когда все участники сделали свой ход. В случае обмена закрытыми сумками это не так. Игра может продолжаться потенциально бесконечно, если результатом каждого обмена сумками будет вариант, где никто никого не обманывает. Обмануть же можно лишь один раз, после чего игра закончится. И хотя в этот раз есть вероятность максимизировать выгоду, в долгосрочной перспективе это плохое решение, выгода от которого, скорее всего, не перевешивает потерю возможности продолжать обмен.

Такая модель гораздо лучше отражает то, как обычно строятся отношения в обществе. Например, хотя преступник, решивший нарушить закон, может получить сиюминутную выгоду, в долгосрочной перспективе он рискует лишиться репутации, денег, свободы и даже жизни. По этой причине быть преступником обычно является проиг-

рышной стратегией. Аналогично при взаимодействии двух сверхдержав с ядерным оружием стратегия сотрудничества или хотя бы отказа от уничтожения друг друга также выглядит более выгодной. Напротив, рассматривать такое взаимодействие через призму дилеммы заключенного просто опасно, так как данная модель не учитывает долгосрочные последствия.

Это заставляет задаться вопросом: не стоит ли пересмотреть дилемму изобретателя в схожем ключе? На первый взгляд это может показаться хорошим решением, ведь при сотрудничестве изобретателей разработка технологии будет идти быстрее. Однако не следует забывать, что речь идет о создании потенциально опасной технологии, поэтому именно перспектива появления такой технологии оказывается сомнительной. При сотрудничестве вместо двух конкурирующих изобретателей мы получаем одного более эффективного изобретателя. А как уже было сказано ранее, для дилеммы изобретателя не имеет большого значения, о скольких изобретателях идет речь, ведь опасность сохраняется, даже если он один. Если же мы предполагаем, что изобретатели договорятся не разрабатывать потенциально опасную технологию, то мы тем самым и воспроизводим ситуацию дилеммы изобретателя. Таким образом, игра с обменом закрытыми сумками не подходит под описываемую ситуацию, а потому нам не нужно учитывать игру этого типа в дальнейших рассуждениях.

Итак, если наиболее адекватной моделью ситуации оказывается дилемма изобретателя, то какие следствия это имеет? Очевидно, что для описываемой ситуации существует сразу три сценария развития, при которых работа над потенциально опасной технологией продолжается, несмотря на все предостережения и риски.

Это можно резюмировать в следующем тезисе: *если технология X может быть создана, то следует ожидать, что она будет создана*. Под словосочетанием «следует ожидать» подразумевается, что при планировании такая перспектива должна рассматриваться как реальная, а не как всего лишь гипотетическая. Данное утверждение может показаться слишком радикальным, потому важно сделать несколько уточнений.

1. Технология *X* должна быть реализуемой на практике. Иначе любые суждения о будущем подобной технологии будут относиться к научной фантастике.

2. Должно существовать понимание того, как можно создать технологию *X*.

3. Без этого сохраняется вероятность, что никто никогда не сможет придумать способ ее воплощения на практике.

4. Технология *X* должна быть действительно ценной. В противном случае нет причин заниматься разработкой именно этой технологии.

Таким образом, наш тезис не предполагает, что любая технология, которую только можно помыслить, обязательно будет создана. Это потребовало бы веры в неограниченный творческий потенциал человека, позволяющий достигнуть результата вопреки любым объективным ограничениям, и в то же время пришлось бы поверить, что технический прогресс подобен закону природы, который никак не зависит от желаний людей. Такая вера безосновательна.

Однако неверно было бы утверждать и то, что технический прогресс полностью зависит от желаний человека. М. Деланда в работе «Война в эпоху разумных машин» приводит множество примеров того, что технический прогресс обладает своей движущей силой. Причем для этого нам вовсе не нужно предполагать существование некой метафизической сущности. Речь идет скорее о том, что одно, даже на первый взгляд незначительное, событие может иметь целый ряд последствий, которые людям придется принять как внешний и независящий от них факт. Например, появление конусовидной пули значительно увеличило точность стрельбы из огнестрельного оружия. Это вынудило военачальников отказаться от полного контроля над армией, давая солдатам больше автономии, а детально продуманные планы сражений пришлось заменить на более гибкие тактики, где небольшим отрядам солдат дается лишь общая цель и уже командир подразделения на месте принимает решение об оптимальном способе ее достижения [3, с. 10]. Создатель конусовидной пули хотел лишь увеличить точность оружия и едва ли предполагал, что его изобретение со временем изменит сам облик войны. Но такое изменение происходит независимо от воли людей. Ситуация, сложившаяся в результате большой сети причинно-следственных связей, сама собой подталкивает людей к изменению своего поведения. Аналогично дилемма изобретателя лишь создает условия, когда ситуация как бы подталкивается в сторону определенного исхода.

Если наш тезис верен, то не стоит надеяться на то, что нам удастся сдержать появление опасных технологий в будущем. Поэтому если перспектива разработки подобной технологии является реальной, то нет большого смысла искать способы предотвратить ее появление. Скорее

всего, мы сможем достигнуть лишь того, что технологию создаст кто-то другой. И поскольку этот субъект стал разрабатывать технологию, несмотря на явную опасность, мы вправе предположить, что речь идет о циничном или безответственном субъекте, т.е. о том, кто представляет наибольшую угрозу. А ведь именно от подобного субъекта мы бы хотели себя обезопасить. Именно ему мы меньше всего захотели бы доверить опасную технологию, особенно ту, которая делает нас беззащитными перед ним. Таким образом, своим отказом разрабатывать потенциально опасную технологию мы сделали ситуацию еще хуже.

С другой стороны, наличие конкуренции часто еще больше стимулирует развитие технологий. Это особенно очевидно в ситуации, аналогичной гонке вооружений. Появление новых технологий у одной противоборствующей стороны часто является ответом на последние разработки другой. В качестве примера можно привести стелс-технологии, снижающие заметность боевой машины для радара. Они были бы бессмысленными при отсутствии технологий обнаружения. Поэтому вне условий конкуренции технический прогресс замедлился бы в ряде сфер, так как не давал бы заметных преимуществ.

В условиях жесткой конкуренции, особенно когда силы всех сторон примерно равны, ценность новых технологий возрастает. Следовательно, это должно приводить к ускорению технического прогресса в тех сферах, где он может дать хотя бы небольшое преимущество. И хотя есть технологии, которые дают преимущества даже тогда, когда конкуренция отсутствует, в условиях конкуренции скорость разработки даже таких технологий возрастает по указанным ранее причинам. Например, различные способы сокращения расходов на производство будут полезны в любом случае, но в условиях конкуренции они не просто позволяют сэкономить, а зачастую определяют, какая компания останется на рынке.

Кроме того, прогресс в условиях конкуренции ускоряется за счет взаимного шпионажа и использования открытых источников. Простой факт наличия технологии у конкурента доказывает, что она возможна. Это повышает готовность инвестировать средства в аналогичные разработки. И даже страх отставания, который может быть безосновательным, оказывается достаточным стимулом для ускорения работы. Это легко проиллюстрировать реальным примером. В 1945 г. США удалось разработать и испытать первую ядерную бомбу. Советский Союз был вынужден реагировать на этот факт, и уже через четыре года у СССР появляется своя ядерная бомба [1, с. 136–137].

В подобных условиях попытки конвенционально ограничить развитие технологий не только не будут пользоваться популярностью, но едва ли смогут дать значимые результаты. Это особенно заметно в военной сфере [3, с. 76], где технологии всегда были самыми инновационными и быстро развивающимися. И дело не только в том, что любые конвенции можно проигнорировать. Даже при их соблюдении наличие ограничений становится дополнительным стимулом для развития новых технологий, которые не регулируются конвенциями. А поскольку иметь преимущество в этой сфере выгодно, то ограничения лишь подталкивают к тому, чтобы искать способы их обходить.

Однако подобное поведение наблюдается не только в военной сфере. Например, в США долгое время было запрещено тестировать технологию генного редактирования CRISPR на людях по этическим причинам. Лишь несколько лет назад запрет был снят. Но даже сейчас проведение подобных экспериментов сильно затруднено из-за необходимости соблюдать ряд правил. Например, команде ученых из Университета Пенсильвании потребовалось около двух лет, чтобы выполнить все требования контролирующих органов. В то же время в Китае ограничения практически отсутствуют, поэтому одобрение подобного проекта иногда занимает всего один день. И хотя мы можем осудить это с этической точки зрения, именно такой подход к проблеме помогает быстрее добиваться результатов в исследованиях. В перспективе это может сделать Китай лидером в CRISPR [4]. На этом примере мы видим, как попытки заблокировать или замедлить развитие технологии лишь создают пространства для новых возможностей в другом месте.

Из всего вышесказанного можно сделать следующий вывод: поскольку попытки сдерживать технический прогресс не слишком эффективны, а иногда даже контрпродуктивны, то нам следует отказаться от подобного подхода и принять как факт, что даже опасные технологии, скорее всего, продолжат появляться. С этой точки зрения мы и должны подходить к решению проблемы, т.е. мы должны думать не о том, как предотвратить появление потенциально опасных технологий, а о том, как нивелировать или компенсировать негативные последствия их появления. Эти решения никогда не будут совершенными, но альтернатива может быть еще хуже. Такое положение дел, судя по всему, является неотъемлемой обратной стороной технического прогресса, своеобразной платой за блага, которые он несет.

Литература

1. *Бостром Н.* Искусственный интеллект: Этапы. Угрозы. Стратегии. – М.: Манн, Иванов и Фербер, 2016. – 496 с.

2. *Винер Н.* Кибернетика, или Управление и связь в животном и машине. – 2-е изд. – М.: Советское радио, 1968. – 328 с.

3. *Деланда М.* Война в эпоху разумных машин. – М.: Кабинетный ученый; Ин-т общегуманитарных исследований, 2014. – 338 с.

4. *Отказ* от этических правил делает Китай лидером CRISPR-технологий. Хайтек. – URL: https://hightech.fm/2018/01/23/china-crispr (дата обращения: 02.09.2021).

5. *Hofstadter D.* Metamagical Themas: Questing for the Essence of Mind and Pattern. – N.Y.: Basic Books, 1985. – 852 p.

References

1. *Bostrom, N.* (2016). Iskusstvennyy intellect: Etapy. Ugrozy. Strategii [Superintelligence: Paths, Dangers, Strategies], Moscow, Mann, Ivanov and Ferber Publ., 496. (In Russ.).

2. *Wiener, N.* (1968). Kibernetika, ili Upravlenie i svyaz v zhivotnom i mashine [Cybernetics: Or Control and Communication in the Animal and the Machine], 2nd ed. Moscow, Sovetskoe Radio Publ., 328. (In Russ.).

3. *DeLanda, M.* (2014). Voyna v epokhu razumnykh mashin [War in the Age of Intelligent Machines], Moscow, Kabinetnyy Uchenyy Publ. and Institut Obshchegumanitarnykh Issledovaniy Publ., 338. (In Russ.).

4. *Otkaz* ot eticheskikh pravil delaet Kitay liderom CRISPR-tekhnologiy [Abandonment of Ethical Rules Makes China a Leader in CRISPR technologies]. Available at: https://hightech.fm/2018/01/23/china-crispr (date of access: 02.09.2021).

5. *Hofstadter, D.* (1985). Metamagical Themas: Questing for the Essence of Mind and Pattern. New York, Basic Books, 852.

Информация об авторе

Ставровский Игорь Константинович – Центр философско-методологических и междисциплинарных исследований Института философии НАН Беларуси (Республика Беларусь, 220072, Минск, Сурганова, 1/2).

tutoriks@gmail.com

Information about the author

Stavrovsky, Igor Konstantinovich – the Center for Philosophical-Methodological and Interdisciplinary Studies, Institute of Philosophy, National Academy of Sciences of Belarus (1, Bldg 2, Surganova st., 220072, Minsk, Belarus)

tutoriks@gmail.com

Дата поступления 11.09.2021