А.Ю. Алексеев

КОГНИТОТЕХНОЛОГИЧЕСКИЕ ПРОЕКТЫ ИСКУССТВЕННОЙ ЛИЧНОСТИ

В современных исследованиях искусственного интеллекта (ИИ) особое значение придается проекту *искусственной личности* (ИЛ). Проект изучает способы компьютерной реализации систем, которым наблюдатель (человек, группа людей, коллектив экспертов) атрибутирует (приписывает) сознание, самосознание, свободу воли и другие персонологические параметры, например способность к моральному вменению. Проект занимает промежуточное положение между: а) проектом искусственной жизни как неотъемлемой части исследований нано- и биотехнологий, обеспечивающих компьютерную реализацию феномена биологической жизни; б) проектом искусственных обществ – компьютерного фундамента информационных и социокультурных технологий, в которых изучается общественная жизнь, начиная от материально-производственной сферы и завершая религиозными верованиями.

Проект искусственной личности начинает широко обсуждаться с середины 1990-х годов на симпозиумах, конференциях, в статьях, книгах, научно-исследовательских работах. Возникновение этого междисциплинарного направления во многом объяснимо факторами общественной жизни. Сегодня сложные компьютерные системы невозможно рассматривать сугубо в техническом плане. Они приобрели социокультурное, «человекомерное» измерение, т.е. погружены в контекст социальных ценностей, мировоззренческих ориентиров, моральных императивов, правовых норм, эстетических канонов и иных составляющих духовной сферы. Явный

практический смысл приобрели общие теоретические требования постнеклассической переориентации методологии изучения, построения и развития сложных систем, предложенной в отечественной философской науке (В.С. Стёпин)[1]. Однако в большей мере возникновение проекта искусственной личности обусловлено внутренними технологическими факторами, т.е. достижениями в области искусственного интеллекта как метатехнологии комплекса НБИКС (нано-, био-, инфо-, когни-, социотехнологий)[2]. Отметим то, что совсем недавно комплекс расширился новой составляющей – методологической. Комплекс МНБИКС более правдоподобен, так как непонятно, какая составляющая, помимо методологической, способна отвечать за конвергентное развитие НБИКС.

Дефиниции искусственной личности

В зависимости от степени онтологических притязаний к «способу компьютерной реализации» персонологических феноменов выделим ряд последовательных дефиниций «искусственной личности»: 1) имитация; 2) модель; 3) репродукция естественной человеческой личности; 4) креация – создание «сверхличности». В этих определениях в силу принятия постнеклассической установки важен выбор исследовательской позиции относительно «личностности» когнитивно-компьютерной системы. При создании классификации исследователей относительно возможностей компьютерной реализации персонологических феноменов предлагается подход, включающий контекст изучения проблематики фи-

[1] *Стёпин В.С.* Научная рациональность в гуманистическом измерении // О человеческом в человеке / Под общ. ред. И.Т. Фролова. – М.: Политиздат, 1991. – 384 с. См. также: *Алексеев А.Ю.* Гуманизм, персонализм и информатика (к общетеоретическим основам моделирования искусственной личности) // Здравый смысл. – М., 1998. – № 3/7. – С. 52–56.

[2] По поводу представления технологии искусственного интеллекта как метатехнологии технологий НБИКС см.: *Алексеев А.Ю.* Четыре лика электронной культуры // Электронная культура: Феномен неопросветительства / Под ред. А.Ю. Алексеева, С.Ю. Карпук. – М.: МГУКИ, 2010. – С. 50–68.

лософских зомби[1]. Во-первых, философские зомби являются концептуальными двойниками искусственных личностей. Во-вторых, данная классификация более адекватна, нежели широко известное сёрловское деление исследователей на сторонников слабого, умеренного, сильного искусственного интеллекта. В нашем подходе конкретизируется предметная область – изучение не «интеллекта», а более сложных понятий – «сознания», «самосознания», «самости», «я», «личности», «другого».

Азомбисты игнорируют тематику сознания в когнитивно-компьютерных исследованиях (А. Тьюринг). Антизомбисты предполагают то, что, воспроизводя сложную кодовую нейрофизиологическую зависимость на немозговом материальном субстрате, мы с необходимостью репродуцируем сознательные феномены (Д. Дубровский). Нейтральные зомбисты предполагают создание компьютерных моделей «квазисознания», которые имеют некоторые эссенциалистские аналогии с человеческим сознанием (Дж. Маккарти). Зомбисты полагают, что мы способны создать своих поведенческих и функциональных компьютерных двойников, не обладающих сознанием (Д. Чалмерс). Неумеренные зомбисты полагают, что все люди – зомби, поэтому стать сознательными мы можем лишь посредством психоделического «расширения» сознания (Ч. Тард) либо тогда, когда заработает система глобального искусственного интеллекта (Д. Деннетт).

Искусственная личность – имитация естественной личности. Такая когнитивно-компьютерная система обеспечивает полную либо частичную *имитацию* человеческой личности. Компьютерная имитация функционально, поведенчески и в некоторых вариантах не только физически, но даже нанотехнологически *неотличима* от человеческой личности. У наблюдателя, атрибутирующего х-системе персонологические параметры, отсутствуют какие-либо эссенциалистские предубеждения. Он не знает, да ему и не интересно предполагать природу системы, с которой он имеет дело: с естественной личностью, компьютером, Солярисом или марсианином. Показателем «неотличимости» естественной и ис-

[1] *Алексеев А.Ю.* Проблема сознания в электронной культуре // Полигнозис. – М., 2010. – № 3(39). – С. 129–141; *Алексеев А.Ю.* Понятие зомби и проблемы сознания // Проблемы сознания в философии и науке / Под ред. проф. Д.И. Дубровского. – М.: Канон+: РООИ «Реабилитация», 2009. – С. 195–214.

кусственной личностей выступает способность системы пройти комплексный тест Тьюринга (Алексеев А.Ю.)[1] или, по крайней мере, *полный тест Тьюринга (тест Харнада)*, который будет рассмотрен ниже. В этих тьюринговых тестах, помимо вербально-коммуникативной, перцептивно-моторной, анатомо-физиологической и даже микрофизической неотличимости искусственной и естественной систем, всесторонне учитываются поддающиеся внешнему наблюдению показатели духовной сферы личности: «смысл», «свобода», «любовь», «ответственность», «право», «творчество», «красота» и др. Правдоподобие данной дефиниции проблематично. Здесь не ясна связь между внутренним миром человека и внешними его проявлениями. С такой проблемой почти век назад столкнулся методологический бихевиоризм, в последующем – логический бихевиоризм. Методологическая парадигма бихевиоризма воспроизведена в идее игры в имитации А. Тьюринга. Логическая парадигма в 1960 г. четко сформулирована Х. Патнэмом в концепции «машинного функционализма»: состояния сознания – это логические состояния машины Тьюринга. Несмотря на заслуги бихевиоризма в антиэссенционалистском устранении метафизики ментальных сущностей из концептуальных моделей компьютерных систем, очевидна ущербность такого подхода, выявленная в результате широко известной критики правомочности метода оценки сознательной деятельности исходя из наблюдений за актуальным поведением и / или из гипотез о поведенческих диспозициях.

Дефиниции искусственной личности как имитации естественной личности придерживаются *азомбисты*. Они напрочь игнорируют онтологическую проблематику реализации «сознания». Не важно, обладает или не обладает персонологическими феноменами компьютерная система фактически, в действительности. Главное, чтобы была создана правдоподобная имитация человека, т.е. такая компьютерная система, которая способна пройти комплексный тест Тьюринга.

Искусственная личность – модель естественной личности. В рамках этого когнитивно-модульного подхода компьютерная система должна включать в свой состав блок «псевдосознания». Это – компьютерная модель, которую исследователи считают

[1] *Алексеев А.Ю.* Комплексный тест Тьюринга (философско-методологические и социокультурные аспекты). – М.: ИИнтеЛЛ, 2013. – 300 с.

функциональным подобием (аналогом) человеческого сознания, самосознания и пр. (Дж. Маккарти). Такого подхода придерживаются нейтральные зомбисты. Главное для них – руководствоваться психологическими, логическими, лингвистическими и др. моделями для решения инженерной задачи: построения систем, способствующих интеллектуальной деятельности человека. Блок «псевдосознания» предназначен для эффективного манипулирования «данными» и «знаниями» интеллектуальной системы, по аналогии с тем, как, например, осознание боли у человека является механизмом организации эффективной сигнализации о том, что с целостностью его что-то не в порядке. Формальные персонологические зависимости, образуемые в отношениях между когнитивными и компьютерными компонентами, задаются посредством вычислительных элементов, операций и функций. Каузальные зависимости этих отношений принципиально представимы в способах функционирования комплексной машины Корсакова–Тьюринга[1]. Компьютинг понимается расширенно. Объединяются два подхода: репрезентативный, моделирующий данные и знания предметной области, и коннекционистский, моделирующий динамику нейтральной системы. Полученные при этом интегральные репрезентативно-коннекционистские кодовые структуры (этот способ иногда называется «двойным кодированием») операционализируются средствами квазиалгоритмической обработки. В результате многошаговой редукции получаем дефиницию искусственной личности: *личностное (персональное) суть квазиалгоритмическое вычисление*.

Степень адекватности компьютерных моделей личности, однако, сомнительна из-за многочисленных шагов редукции: персональное редуцируется к ментальному; ментальное – к интеллектуальному; интеллектуальное – к структурно-функциональной архитектуре компьютера. Следует отметить практический ориентир такого определения. Однако иногда исследователи из-за сложности модельных

[1] По поводу комплексной машины Корсакова–Тьюринга см.: *Алексеев А.Ю.* Комплексная, коннекционистско-репрезентативная интерпретация машины Корсакова // Материалы Шестой международной научно-практической конференции «Актуальные проблемы современной когнитивной науки». – Иваново: ОАО «Издательство Иваново», 2013. – 330 с. – С. 53–64; *Алексеев А.Ю.* Протонейрокомпьютер Корсакова // Нейрокомпьютеры: Разработка и применение. – М., 2013. – № 7. – С. 6–17.

отношений начинают заблуждаться по поводу онтологического статуса компьютерной системы. Например, нейрокомпьютерная система представляется им фактически интеллектуальной, ведь она реализует нейрофизиологическую модель интеллектуальной деятельности, а нейрофизиологи ошибиться не могут, так как они – специалисты в данной предметной области (Л.Н. Ясницкий). Очевидно, что это совершенно некорректная позиция.

Искусственная личность – репродукция естественной личности. Согласно данному определению, компьютерная система фактически воспроизводит общие и единичные феномены сознания посредством реализации сложной функциональной зависимости нейрофизиологических кодов субъективной реальности на субстрате, инвариантном относительно физиологического строения человеческого мозга (Д.И. Дубровский). Данное определение четко поддерживается антизомбистами.

Искусственная личность – креация естественной личности. Проект искусственной личности креацирует феномены, которые не имеют натуральных персонологических аналогов у человеческого индивида (Д. Деннетт). В предельном варианте данная дефиниция подводит к понятию глобального искусственного интеллекта, т.е. компьютерной системы как «сверхличности», своеобразного варианта понятия «ноосферы».

Репродукционное и креационное определения автор кратко обозначил, так как не считает их значимыми для развития теории и практики построения искусственной личности. В философском отношении они, возможно, интересны, так как раскрывают предельные технологические перспективы. Однако к научному дискурсу имеют достаточно отдаленное отношение. Принципиально отрицательная позиция к таким подходам будет высказана в заключение данной работы. Здесь же подчеркнем то, что современные проекты искусственной личности развиваются в условиях лингвистической разобщенности, т.е. в условиях многообразия дефиниций искусственной личности. Причем, заметим, разнообразие интерпретаций возникает не только в силу различий в определении «человеческой личности» (здесь выработка четкого понятия принципиально невозможна), но по причине многообразия представлений о путях компьютерной реализации персонологических феноменов.

Многообразие проектов искусственной личности. По сути, история философии искусственного интеллекта – это история обсу-

ждения возможностей компьютерной реализации персонологических параметров. Достаточно вспомнить полемический стандарт А. Тьюринга по поводу построения мыслящих машин (1950) как совокупность аргументов и контраргументов в решении «основного вопроса» философии искусственного интеллекта: «Может ли машина мыслить?»[1] В «стандарт» вошли аргументы и контраргументы, не имеющие прямого отношения к «интеллекту»: теологический, антисциентистский, креационистский, «от первого лица», «от другого сознания» и даже экстрасенсорный.

С начала 1990-х годов в широкой печати выделяются следующие крупные проекты ИЛ: 1) OSCAR Дж. Поллока, сформулированный в рамках «универсальной теории рациональности» с ее приложениями для построения искусственных рациональных агентов («артилектов»)[2]; 2) проект «человекоподобных агентов» А. Сломана, призванный реализовать широкий спектр персонологических параметров, например, «любовь», «свободу»[3]; 3) проект гуманоидных роботов КОГ, в котором Д. Деннетт усматривает апробацию собственной теории «множественных набросков», где персональное возникает из сложного сочетания безграничной серии нарративов, а личность и социум – это субстанциональные системы бесчисленных роботов, в которых «ментальное» представляется компонентой функциональной самоорганизации[4].

---

[1] Серию аргументов и контраргументов, приводимую А. Тьюрингом для дискуссии по поводу мыслящих машин, мы предложили назвать Полемическим стандартом Тьюринга, так как она придала форму и отчасти содержание всем современным дискуссиям в области философии искусственного интеллекта (см.: *Алексеев А.Ю.* Возможности искусственного интеллекта: Можно ли пройти тесты Тьюринга // Искусственный интеллект: Междисциплинарный подход / Под ред. Д.И. Дубровского и В.А. Лекторского. – М.: ИИнтеЛЛ, 2006. – С. 223–243.

[2] *Серёдкина Е.В.* Общая теория рациональности и артилекты в проекте OSCAR Дж. Поллока // Философско-методологические проблемы искусственного интеллекта: Материалы постоянно действующего теоретического междисциплинарного семинара / Под ред. Е.В. Серёдкиной. – Пермь: Изд-во Перм. гос. техн. ун-та, 2007. – С. 108–122.

[3] *Sloman A.* What sorts of machines can love? Architectural requirements for human-like agents both natural and artificial. – Mode of access: http://www.sbc.org.uk/literate.htm

[4] *Деннетт Д.С.* Виды психики: На пути к пониманию сознания / Пер. с англ. А. Веретенникова; Под общ. ред. Л.Б. Макеевой. – М.: Идея-Пресс, 2004. – 184 с.

В этих проектах предлагаются концептуальные, логико-математические, программные решения. Рассмотрим одно из них.

Типовая когнитивно-компьютерная архитектура искусственной личности. Как правило, рассматривается трехуровневая архитектура когнитивно-компьютерной системы, претендующей на проект ИЛ: 1) уровень коннекционистских образов (паттернов), осуществляющий перцептивную обработку данных; 2) уровень первичных репрезентаций, переводящий восприятия в дискретные представления и суждения; 3) уровень вторичных репрезентаций, на котором осуществляется представление представлений (моделирование других моделей представления знаний и моделирование собственной модели). Особо показателен подход А. Сломана. Он задает конкретный вопрос: «Какие машины могут любить?» – и предлагает архитектуру «любящих» компьютеров, состоящую из: 1) реактивного; 2) обдумывающего; 3) рефлективного (метауправляющего) уровней[1].

Помимо внутренней архитектуры системы искусственной личности, немаловажным является *вопрос о внешнем, физическом подобии искусственной и естественной систем.* Показательными представляются два подхода: 1) Дугласа Лената, автора широко известной программы «Автоматический математик» (1976), который считает, что следует подражать психологическим, социокультурным, лингвистическим, интеллектуальным и др. особенностям личности, а физическое подобие системы – второстепенный, несущественный фактор; 2) подход Родни Брукса, ученика Д. Лената, автора вышеупомянутого робота КОГ и знаменитой версии Kismet, имитирующей мимику человеческого лица. Р. Брукс полагает, что физическая антропоморфность системы (робота) – первичное и необходимое качество его персонологического подобия, так как все социокультурные понятия, на основании которых будет функционировать робот, предопределены физическими особенностями. Например, «безногому» роботу трудно будет уловить «смысл» выражения «перевернуть с головы на ноги».

Подход Д. Лената и Р. Брукса можно обозначить, соответственно, как *неантропоморфный и антропоморфный* подходы к построению ИЛ.

---

[1] Ряд технических характерстик трехуровневой архитектуры представлен в работе: *Алексеев А.Ю.* Трудности и проблемы проекта искусственной личности // Полигнозис. – М., 2008. – № 1. – С. 20–44.

Чтобы четче обозначить различия между этими подходами, рассмотрим иерархию тестов Тьюринга (ТТ), предложенную С. Харнадом (тест Харнада)[1].

ТХ0. Это – уровень «игрушечных» ТТ – не полноправных тестов, но лишь некоторых фрагментов, ограниченных как по длине, так и по содержанию. Такие тесты не отвечают исходному замыслу А. Тьюринга (частично они реализованы в играх на премию Лойбнера). Однако все попытки моделирования интеллекта, известные на сегодняшний день, выше данного уровня не поднялись.

ТХ1. Это – уровень оригинального теста Тьюринга с учетом ограничений на длину теста, объем компьютерной памяти и скорость вычислений.

ТХ2 – общепринятое понимание теста Тьюринга, который часто называется «друг по переписке». Длина теста равна протяженности человеческой жизни.

ТХ3 – «роботизированная» версия ТХ2. Имеется возможность манипуляции предметами внешнего мира. Процедура идентификации систем, которые проходят данный тест, требует принципа «Вскрытие покажет» (принцип Н. Блока), т.е. анатомическое исследование системы. Такой прием часто используется в фантастических фильмах для отличения роботов от людей. При поражении тела робота из него вытекает жидкость зеленого, белого или иного некрасного цвета.

ТХ4 – это компьютерные системы, неотличимые как в плане ТХ3-неотличимости, так и в плане микрофизической организации системы. Здесь имеет место «тотальная неотличимость» компьютерной системы от человека, включая мельчайшие внутренние нюансы телесного строения. Теперь из поврежденного робота будет вытекать как бы настоящая кровь. Перспективы теста связываются с нанотехнологиями.

Как мы видим, экспертный подход должен пройти уровень ТХ2, робототехнический – уровень ТХ3. Неантропоморфные версии – это уровни ТХ0-ТХ3, антропоморфная – ТХ4.

В англо-американской философии ИИ в основном преобладает антропоморфный подход, согласно которому ИЛ – это обла-

[1] *Harnad S.* Minds, machines and turing: The indistinguishability of indistinguishables // J. of logic, language, and information. – Mode of access: http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/harnad00.turing.html

дающий квазисознанием автономный робот. Экспертная методология не признается такими авторитетными зарубежными философами ИИ, как Дж. Маккарти[1], А. Сломан[2]. В отечественной же науке, напротив, сложился экспертный подход: ИЛ рассматривается в социально-эпистемологическом контексте междисциплинарных взаимодействий специалистов, в ходе которых формируются «данные», «знания», «смыслы» социокультурного и персонологического содержания. Рассмотрим подробно эти подходы.

Искусственная личность – робототехническая система. Искусственная личность – это робот, наделенный квазисознанием и благодаря этому реализующий некоторые персонологические способности и качества человеческой личности. Показательна дискуссия по поводу книги Селмера Брингсйорда «Чем могут и не могут быть роботы» (1992, 1994)[3]. В работе позитивные утверждения относительно проекта ИЛ сопровождаются обстоятельной критикой. Основной девиз работы звучит так: «В будущем робот будет делать все то, что делаем мы, но не будет одним из нас», т.е. он не будет сознательным.

С. Брингсйорд доказывает, что: а) когнитивно-компьютерная технология будет производить машины со способностями проходить все более и более сильные версии ТТ, однако: б) проект ИЛ по созданию машины-личности будет неминуемо проваливаться. В защиту (а) предлагается индуктивный вывод при построении версий тестов: тьюринговая игра в имитацию $\rightarrow$ наблюдение внешнего вида игроков $\rightarrow$ изучение их сенсомоторного поведения $\rightarrow$ сканирование мозга и пр. (т.е., по сути, прохождение рассмотренной выше иерархии тестов Харнада). В основе доказательства (б) – несостоятельности проекта ИЛ – лежит modus tollens из суждений: (1) «Проект ИЛ» $\rightarrow$ «Личность – это автомат»; (2) $\neg$ «Личность – это автомат»; $\Rightarrow$ (3) $\neg$ «Проект ИЛ». Посылка (2) интуитивно апеллирует к наличию у человека и отсутствию у автомата ряда

[1] *McCarthy J.* Artificial intelligence and philosophy. – Mode of access: http://cogprints.ecs.soton.ac.uk/archive/00000420

[2] *Sloman A.* What sorts of machines can love? Architectural requirements for human-like agents both natural and artificial. – Mode of access: http://www.sbc.org.uk/literate.htm

[3] *Bringsjord S.* What robots can and can't be, psycoloquy: 5,#59 Robot Consciousness (1); – Mode of access: http://psycprints.ecs.soton.ac.uk/archive/00000418/#html

машинно не воспроизводимых персонологических параметров F, к которым относятся свободная воля, способность к интроспекции, внутренний проективный опыт «каково быть» («what it's like to be») и пр. К F относится и возможность компьютерного воспроизводства творческих способностей, которую автор подробно анализирует в ряде предыдущих работ на основе рассмотрения программ генерации художественных и философских произведений. Убедительно доказывается: «компьютер творить не может!»[1] Общее заключение следующее: (4) «Личность обладает F»; (5) «Автомат не обладает F»; $\Rightarrow$(6) «Личность не может быть автоматом». С. Брингсйорд достаточно подробно обосновывает невычислимость F, апеллируя к аргументу Гёделя, «Китайской комнате» Дж. Сёрля, аргументу произвольной множественной реализации Н. Блока и др. Итог таков: «Роботы будут многое делать, однако они не будут личностями».

В полемике по поводу работы С. Брингсйорда, продолженной на страницах сайта http://psycprints.ecs.soton.ac.uk (рубрика «Сознание робота»), наблюдается четыре направления критики и поддержки.

I. Проект ИЛ *абсурден*, так как: 1) метафизические понятия «личность», «свобода воли», «интроспекция» не могут быть предметом эмпирического анализа; 2) нечеткое определение понятия личности, образная иррациональность проекта ИЛ – все это не заслуживает научного внимания; 3) персонологические параметры логически невыразимы – попытки их логической экспликации непременно сопровождаются ошибками.

II. Проект ИЛ *неточен*, так как: 1) апелляция к интуитивным аргументам разрушает логическую строгость дедуктивных аргументов; 2) нецелесообразно применять дедуктивные умозаключения в качестве метода аргументации (дедукция – не метод, а схема суждений); 3) необходимо четко формулировать конкретную разновидность функционализма как методологического базиса проекта ИЛ – нет функционализма «вообще», а есть, к примеру, низкоуровневый функционализм, биологический функционализм и пр.

[1] Подход С. Брингсйорда к критике компьютерного творчества явно выражен в аргументе Лавлейс. См.: *Брингсйорд С.*, *Беллоу П.*, *Феруччи Д.* Творчество, тест Тьюринга и улучшенный тест Лавлейс / Пер. с англ. А. Ласточкина // Тест Тьюринга. Роботы. Зомби / Пер. с англ.; Под ред. А.Ю. Алексеева – М.: МИЭМ, 2006. – С. 62–83.

III. Проект ИЛ *реализуем*, так как: 1) человеческая личность – это тоже компьютер, но неклассического типа, т.е. представимый не машиной Тьюринга, а иными, квазиалгоритмическими правилами обработки вычислительных элементов (мы предложили использовать машину Корсакова–Тьюринга); 2) компьютер вовсе не автомат, и кто придерживается противоположного мнения – тому не место в проблематике компьютерного «сознания»; 3) компьютерная система должна тестироваться природой, а не человеком.

IV. Проект ИЛ *значим* вне зависимости от возможностей его реализации: 1) будет ли создана ИЛ либо не будет создана – не принципиально, главное, что проекты ИЛ выявляют механистические аспекты человеческого сознания; 2) проектирование ИЛ проясняет роль и функции сознания в жизни человека путем постановки вопроса о зомби – о созданиях, не обладающих сознанием, но имеющих все поведенческие способности человека.

Добавим еще один пункт.

V. Исследования естественной личности нерепрезентативны, если они не учитывают исследования проекта искусственной личности. Человек оказался «заброшенным» в компьютеризованный мир, поэтому каноны и стереотипы компьютинговой рациональности более прочно и оперативно, нежели в традиционной культуре, фиксируют и закрепляют у него «искусственное». Для выявления «человеческого собственно человеческого» и предназначен проект искусственной личности – кем он, естественный человек, не должен быть.

Имитация и моделирование естественной личности должны быть соразмерными способами реализации проекта ИЛ. Не надо бросать все силы на построение совершенной имитации, проходящей комплексный тест Тьюринга (как это принято на лойбнеровских состязаниях компьютеров и людей). Но так же не надо доверять онтологической полноте построенной компьютерной модели некоторого частного персонологического аспекта. И совершенно не следует навязывать результаты моделирования личности – пусть очень полного и обоснованного – оригиналу этой модели.

Что касается значимости параметра антропоморфности робота – здесь нет жестких ограничений. Зачем роботу, работающему в космосе или в глубине океана, иметь человеческий облик? Однако роботу, который, скажем, подстригает волосы и бороду

(такие роботы уже работают в экспериментальных парикмахерских), предпочтительно все же иметь приятное дружелюбное лицо.

Искусственная личность – экспертная система

Примерно в те же годы, когда проходила обозначенная выше дискуссия, т.е. в середине 1990-х годов, в нашей стране совершенно независимо от зарубежной науки возникла иная идея проекта искусственной личности. Автор данной работы был в эпицентре инициации работ по этому проекту[1]. По заказу Министерства обороны РФ в 27 ЦНИИ МО под руководством проф. В.В. Деева были организованы исследования по программированию и внедрению в системы принятия решений так называемых «нетрадиционных информационных технологий». Необычность этих технологий – в изучении и применении неординарных способностей экстрасенсов для эффективного управления войсками. Проект искусственной личности был призван интегрировать «знания», которыми владели как экстрасенсы (сенситивы), так и традиционные специалисты. Семантическая неопределенность и концептуальный разброс для решения ряда задач отличают данный проект от модных на сегодняшний день проектов «коллективного принятия решений» в системах «управления знаниями».

Исследования базировались на концепции моделирования «смысла». Проект ИЛ рассматривался как этап эволюции интеллектуальных систем. Если традиционная компьютерная технология характеризуется формулой «данные+алгоритм», технология искусственного интеллекта – «знания+эвристика», то проект ИЛ задается формулой «смысл+понимание». Учитывая крупные наработки отечественных специалистов в области построения экспертных систем, в частности в семиотическом моделировании способов представления «знаний», в основу рассматриваемого проекта ИЛ была положена экспертная методология. Специальные программы моделировали процессы «понимания», обеспечивали условия экспликации «смысла», «сущностных» оснований прини-

[1] *Алексеев А.Ю.* Искусственная личность // Сб. трудов конференции «Междисциплинарный подход к изучению человека». – М.: Институт микроэкономики, 1998. – С. 28–31.

маемых решений. Модель «смысла» фиксировала эти «траектории» в определенным образом кодифицированных массивах информации.

Проблеме моделирования «смысла» посвящено много работ. Среди них в отечественной литературе «дисциплинарно» выделяются следующие модели: культурологическая, психологическая, лингвистическая, герменевтическая, риторическая, дискурсивная, семиотическая, поэтическая, иконографическая, эзотерическая, математическая, инженерная, синергетическая, неклассическая и пр.[1] Наиболее продуктивным для компьютерной реализации представляется *контекстуальный подход*, согласно которому «смысл» – это обладающий параметрами системного единства контекст формализованных «знаний» экспертов. Общетеоретической основой этого подхода являются классические модели «смысла» Г. Фреге (смысл – это выражаемый знаком способ означивания значения) и модификации этой модели Б. Расселом и Л. Витгенштейном. Наряду с этим использовались идеи теории репрезентации М. Вартовского, которая в систематическом единстве рассматривает модель, способ ее построения и способ ее интерпретации. Такие связи обеспечивают динамику наполнения «смыслового объема» возможными «смысловыми траекториями» в ходе согласования метода и предмета репрезентации. Эксперт как бы «погружается» в квазиалгоритм построения «модели модели», создает «репрезентацию репрезентации».

Технически модель «смысла» выполнена в виде программно-информационной оболочки над системой представления «знаний». «Смыслообразующие» концепты выражались кодами «личность», «значение», «смысл», «ценность», «понятие», «действие», «роль», «норма», «умение» и рядом др. Конкретизация и экземплификация этих кодов осуществлялись в рамках парадигмы теоретико-деятельностного подхода. Эксперт должен соотносить полученные «смысловые траектории» с конкретными формализованными «знаниями», непосредственно задействованными в модели принятия решения. Название проекта – «искусственная личность» – оправ-

[1] Социокультурные аспекты моделирования «смысла» / *Алексеев А.Ю.*, *Артюхов А.А.*, *Крючков В.Л.*, *Маликова Я.С.*, *Розов М.А.* // Философия искусственного интеллекта / Под ред. В.А. Лекторского и Д.И. Дубровского – М.: ИФ РАН, 2005. – С. 134–138.

дывается тем, что «знания» о принимаемых решениях индексируются персонологическими параметрами, например план решения соотносится с параметром, характеризующим добровольное / принудительное его принятие (это – очевидный аспект морального поведения). Также предполагается, что экспертная система в целом – это социальный институт, интегрирующий персонологические компетенции людей. То есть некоторый человек вносит свои «знания» в систему, другой – свои. Результат интеграции разнородных и гетерогенных «знаний» – это, собственно, факт «деятельности» программ искусственной личности, недостижимый и по большей части невозможный, в условиях общения наблюдателя с конкретными людьми.

Апробация экспертного подхода к проектированию ИЛ, осуществленная в ходе опытной эксплуатации макета, большая часть которого осталась «на бумаге», выявила ряд неудач: методологических, организационных, теоретических и реализационных. Главная – эпистемологическая неудача: модели «смыслов» – это все-таки модели наших «знаний о смыслах», но не «смыслах как таковых».

Несмотря на неудачи с экспертным проектом ИЛ, был достигнут *социокультурный эффект:* предложенный инструментарий вынуждает специалистов осуществлять рефлексию над «знаниями», эксплицировать «смыслы» собственных решений и выставлять их для интерсубъективного обсуждения.

Расширение поля междисциплинарных исследований в проекте искусственной личности. Очевидно, что робототехнический и экспертный проекты ИЛ дополняют друг друга в плане расширения возможностей интеллектуальных информационных технологий. Если первый декларирует: «Искусственная *личность – это робот, наделенный квазисознанием*», то второй утверждает: «Искусственная *личность – это экспертная система, оборудованная механизмами 'смысла'*». На наш взгляд, экспертный вариант проекта ИЛ по порядку и по значимости должен предшествовать робототехническому. Экспертная система, отвечающая на запросы самого различного содержания, обеспечивает построение вербально-коммуникативного теста ТХ2. Лишь базируясь на нем, возможно построение и более сложных тестов ТХ3 и ТХ4, необходимых для робототехнического подхода. До изучения того, как робот автономно либо в окружении колонии роботов способен, например, проду-

цировать «коммуникацию», «волеизъявление», «моральные императивы», «религиозные верования» и пр., следует определиться с операциональными определениями этих понятий. А это достигается в рамках экспертной методологии.

Проекты повышают роль и усиливают значимость междисциплинарного подхода. Собственно, и проект искусственного интеллекта уже немыслим вне поля междисциплинарных исследований психологов, логиков, математиков, лингвистов, нейрофизиологов и др. Но проект ИЛ не ограничивается перечисленными традиционными специалистами ИИ. Он вовлекает в исследования социологов, политологов, экономистов, искусствоведов, правоведов и др. В отечественном проекте ИЛ, как мы отметили выше, участвовали даже представители паранауки, хотя от них толку было мало. Приоритетными в проекте ИЛ становятся отнюдь не специалисты в области естественных и технических наук. Исключительно важна роль экспертов в области общественных и гуманитарных наук. Именно они вводят в предмет конструирования «смыслы», «ценности», «нормы», «идеалы» и пр. составляющие духовной жизни.

Однако представим утопический сценарий: все лучшие умы брошены на реализацию проекта ИЛ, выполнены все необходимые методологические работы по коммуникации, координации, интеграции научного сообщества. Более того, со стороны государства получена безграничная финансово-экономическая поддержка («За искусственную личность платить надо!» – подчеркивал Д. Деннетт[1]). Будет ли реализован проект искусственной личности? Вряд ли. Для этого надо преодолеть ряд методологических трудностей, связанных со сложностью предмета исследований. Эти трудности на сегодняшний день представляются не только практически, но и теоретически непреодолимыми. Проблемы инспирированы смежными проблемами философии сознания и философии ИИ, к числу которых относится *проблема построения теста искусственной личности (теста Тьюринга на персональное)*. Причем данный тест следует рассматривать, начиная с уровня ТХ0. Пока не надо претендовать на антропоморфную неотличимость естественной / искусственной личности.

[1] *Dennett D.C.* The practical requirements for making a conscious robot. – Mode of access: http://www.ecs.soton.ac.uk/~harnad/Papers/Py104/dennett.rob.html

Тест на личностное распадается на два вида: 1) *общий тест* – по каким вопросам к неизвестной х-системе можно определить, личность перед вами или не личность (компьютер); 2) *частный тест*, тест на сознание – по каким вопросам определить, обладает ли система сознанием или не обладает.

Частный тест Тьюринга на личность включает в качестве оцениваемых параметров сознательные способности и свойства. К этим способностям относят: 1) «ощущение» (sentience), т.е. восприятие мира и реагирование на него в соответствии с различными видами перцептивно-эффекторных возможностей; 2) «бодрствование» (wakefulness), т.е. способность фактически реализовывать некоторые возможности, а не только обладать диспозициями к их осуществлению; 3) «самосознание» (self-consciousness), т.е. не только осведомленность о чем-то, но и осведомленность о своей осведомленности; 4) «позиционность» (what it is like), т.е. способность воспринимать мир в соотнесении с собственной позицией «бытия» или, иначе, восприятие среды в соответствии с ответом на вопрос «каково быть?» (например, каково быть летучей мышью, роботом, экспертной системой?).

Помимо демонстрации вышеперечисленных способностей, х-систему можно считать квазисознательной, если наблюдатель приписывает ей то, что она: 1) пребывает в состоянии «осведомленности о»; 2) обладает квалиа – владеет качественными свойствами конкретной разновидности (ощущает красное, чувствует запах кофе и пр.); 3) пребывает в феноменальном состоянии, которое более структурировано, нежели квалиа, и обладает определенной пространственной, темпоральной и концептуальной организацией опыта окружающей среды и опыта самого себя; 4) обладает акцессным сознанием (access consciousness), т.е. доступом к внутриментальным составляющим при управлении речью и действиями (Н. Блок); 5) обладает нарративным сознанием (narrative consciousness), т.е. выдает серию лингвистических выражений, «высказываемых» с перспективы актуальной или виртуальной самости (Д. Деннет); 6) обладает интенциональностью – направленностью своего внимания на предмет сознания, активностью опредмечивания окружающей действительности (Д. Сёрль).

Следует подчеркнуть, что вышеприведенные способности и характеристики «квазисознательности» необходимы, но недостаточны. Несомненно, можно привести длинный список иных харак-

теристик квазисознательного поведения. Проблема полноты сознательных способностей и свойств иначе называется «конъюнктивной проблемой» и ей придается большое значение в современных дискуссиях наряду с так называемой «дизъюнктивной проблемой» – проблемой множественной реализации сознательных феноменов на различных субстратах.

Решение задачи проектирования ИЛ требует специального анализа принципа инвариантности информации (Д.И. Дубровский), в частности, сопоставления этого принципа с аргументом множественной реализации. Для этого надо всесторонне изучить и редукционистские и антиредукционистские парадигмы реализации ментальных феноменов на различных физических системах – на мозге человека, на мозге других приматов и животных, на «тине» в «черепе» гипотетических марсиан, на кодовых структурах искусственной личности.

В завершение данной работы вновь подчеркнем, что проект искусственной личности является одним из наиболее перспективных проектов искусственного интеллекта. Он располагается между проектом *искусственной жизни*, направленным на воспроизводство биологически-эквивалентных форм квазиорганической жизни, и проектом *искусственного общества*, призванного реализовать весь спектр социальной жизни искусственных агентов. Для первого проекта он задает ориентир создания наиболее высокоорганизованной формы искусственной жизни. Для второго – создает базис «общественной жизни» искусственных систем, так как социальное вырастает из взаимодействия персонологически активных искусственных агентов.

На наш взгляд, проект ИЛ следует развивать в форматах компьютерной имитации и моделирования персонологических феноменов. Что касается компьютерной репродукции и креации, здесь имеется очень много проблем – трудных, сложных, самых сложных и пр., о которых автор неоднократно упоминал в статьях и докладах[1].

[1] *Алексеев А.Ю.*, *Кураева Т.А.* Проблема зомби и перспективы проекта искусственной личности // Философия искусственного интеллекта: Материалы Всероссийской междисциплинарной конференции. – М.: МИЭМ, 17–19 января 2005 г. Под ред. В.А. Лекторского и Д.И. Дубровского. – М.: ИФ РАН, 2005. – С. 5–9; *Алексеев А.Ю.* Трудности и проблемы проекта искусственной личности // Полигнозис. – М., 2008. – № 1. – С. 20–44.

Хотя такие проблемы как бы не замечают сторонники общественного движения «Россия-2045», в развитии которого принимают участие широкие слои населения. Ажиотаж вокруг этого движения, поддерживаемый рядом ведущих философов страны, объясняется очевидными преимуществами, которые получает концептуальный проект от рекламного бренда. «2045» – это своеобразное неофёдоровское движение «философии общего дела». Однако имеется кардинальное различие. Первая волна фёдоровцев в начале XX в. стремилась применить «электронные технологии» в альтруистических целях – воскрешения предков. Новая волна, в начале XXI в., инспирирована буржуазным эгоизмом приобретения личного бессмертия. О какой постнеклассической рациональности здесь может идти речь? Что касается технологических аспектов реализации проекта, то налицо некритическое восприятие возможностей компьютерных технологий, прямо скажем, паранормальное отношение к исследованиям искусственного интеллекта[1].

Вряд ли компьютерные технологии «2045», даже конвергентно объединенные в некоторую будущую целостную систему НБИКС, способны решить основополагающие для любого человека вопросы – о Я, смысле жизни, бессмертии, свободе, общественном идеале и пр. Собственно, сюжеты такого рода проектов ярко раскрыты в многочисленных фантастических B-moves. Кстати, известный фильм «Суррогаты» явился непосредственным источником идеи «2045». Однако слишком далек и тернист путь, который надо проделать от художественного образа создания искусственных личностей до фактически реализуемого проекта. Особых отличий мы не обнаруживаем в ветхозаветном «проекте» лепки Богом людей из глины и в современных проектах «2045» создания искусственных личностей посредством компьютерных систем. Метафора создания себе подобного остается той же. Тайна личного Я сохраняется.

[1] *Алексеев А.Ю.* Паранормализация искусственного интеллекта // Поругание разума: Экспансия шарлатанства и паранормальных верований в российскую культуру XXI века: Тезисы к международному симпозиуму «Наука, антинаука и паранормальные верования». Москва, 3–7 октября 2001 г. – М.: Российское гуманистическое общество, 2001. – С. 8–11.