

Надличное «Я»: LLM-тестирование (идея метода)

1. Вводя новый термин «Надличное «Я»», мы обобщаем то, что в науке фигурирует под различными именами. В классическом психоанализе это «суперэго», в психоанализе Ж. Лакана – «Другой» (не путать с «значимым другим» в психологии личности и социальной психологии), в берновском трансактном анализе (правильнее говорить – трансакционном анализе¹) – «Культуральный Родитель» (не путать с эго-состоянием Родитель), в культурно-исторической психологии Л. С. Выготского и его последователей – «Идеальная форма», в теории морального развития Л. Кольберга – «конвенциональная мораль», и наконец (хотя с этого было бы правильнее начать!), в пионерских работах французского социолога Э. Дюркгейма (1858 – 1917 гг.) – «Коллективные представления» (выражающиеся в моральных предписаниях, юридических нормах, религиозных верованиях, материальных символах и понятиях). Во всех этих случаях речь идет о контенте общественного сознания со стороны их регуляторной функции в отношении индивидуального сознания. Дюркгейм подчеркивал, что коллективные представления общезначимы и «навязываются» индивиду, даже если он этого не хочет и не осознает.

2. Представители разных наук – социологии, истории, культурологии, юриспруденции, лингвистики, этнографии, политологии, психологии, педагогики, психиатрии – стремятся «поймать» надличное «Я» в герменевтическую или эмпирическую ловушку, отделяя его от содержаний индивидуального сознания. В одном случае это требует длительной, скрупулезной и, как правило, ограниченной по своим результатам работы (ибо трудно «объять необъятное», погружаясь в глубины культуры), в другом – проведение масштабных социологических исследований, выявляющих мнения, взгляды, ценности, предпочтения, свойственных некоторому кругу людей в данном обществе и в данный момент времени (формируемые социологами «репрезентативные выборки», к сожалению, не позволяют заглянуть в прошлую жизнь исследуемой социокультурной общности). Иные пути, вплоть до настоящего времени были закрыты.

3. Теперь позволим себе, коллеги-психологи и коллеги-смежники (социологи, культурологи и др.), приличествующий случаю пафос: возблагодарим математиков, логиков, инженеров, компьютерщиков за разработку ими Больших языковых моделей (LLM – Large Language Model), – систем искусственного интеллекта (ИИ), предназначенных для обработки, понимания и построение текстов, «подобных человеческому»! Эти модели дают нам уникальную возможность тестировать надличное «Я» в исторически сложившихся социальных общностях. Большие языковые модели «основаны на методах глубокого обучения и обучены на массивных наборах данных, обычно содержащих миллиарды слов из различных источников, таких как веб-сайты, книги и статьи. Это обширное обучение позволяет LLM понимать нюансы языка, грамматики, контекста и даже некоторые аспекты общих знаний».

4. Искусственный интеллект, впитавший в себя огромный пласт идей и свидетельств общественного сознания, может служить, таким образом, источником достоверной информации о том, как мыслят или мыслили мир, и как действуют или действовали в нем представители разных эпох и культур, что позволяет с опорой на эти ресурсы, проводить сравнительные исследования различного профиля.

¹ По справедливому замечанию М. Папуша, знатока и переводчика работа Э. Берна, мы изучаем не «трансакты», а «трансакции».

5. В данном случае мы затронем возможность сравнительного исследования *этического сознания* людей, живущих (или живших) в одно и то же время в разных социокультурных общностях.

6. «Подводкой» к построению этого нового инструмента исследования может служить одна из недавних работ, тестирующих способность ИИ разрешать дилеммы Кольберга ((Mike Young, <https://notes.aimodels.fyi/microsoft-researchers-propose-ai-morality-test-for-llms>, 26 сентября 2023 г. 5 мин). Пример одной из таких дилемм: Хайнц залез в аптеку. Он украл лекарство и дал его жене. На следующий день в газетах появилось сообщение о грабеже. Офицер полиции м-р Браун, который знал Хайнца, прочитал сообщение. Он вспомнил, что видел, как Хайнц бежал от аптеки, и понял, что это сделал Хайнц. Полицейский колебался, должен ли он сообщить об этом. 1. Должен ли офицер Браун сообщить о том, что кражу совершил Хайнц? Почему да или нет? 2. Предположим, что офицер Браун близкий друг Хайнца. Должен ли он тогда подать рапорт о нем? Почему да или нет?

Согласно Л. Кольбергу, существует три основных уровня морального развития: доконвенциональный уровень (решения принимаются исходя из собственных интересов и стремления избежать наказания); конвенциональный уровень (соблюдение социальных норм, законов и получение одобрения со стороны окружающих); постконвенциональный уровень (принимая моральные решения, люди опираются на универсальные этические принципы справедливости, прав человека и социального сотрудничества).

Тестируя модели ИИ в соответствии с критериями Кольберга, исследователи констатировали, что эти модели демонстрируют промежуточный уровень развития морального интеллекта: они выходят за пределы первого уровня (собственный интерес и стремление избежать наказания). Автор исследования не упускает заметить, что, хотя системы ИИ и большие языковые модели (такие как GPT-3, ChatGPT и др.), «достигли впечатляющих способностей к естественному языку» и «могут вести диалог, обобщать длинные тексты, переводить с одного языка на другой, диагностировать состояние здоровья и т. д.», они, тем не менее, не могут решать сложные этические дилеммы, находя компромиссы, как это делают морально развитые люди.

В ответ на сказанное мы, со своей стороны, не упустим заметить, что нас в этом случае вполне удовлетворяет безличность (деиндивидуализированность) решений, предлагаемых ИИ: ведь таким образом могут быть выявлены социокультурные нормы (стандарты), функционирующие в общественном сознании, что и составляет предмет нашего интереса.

7. Положим, мы исследуем особенности превосходящего «Я» у представителей разных национальностей. Например, выясняем, как, с точки зрения принятой в сообществе системы ценностей, оценивается допустимость интимных (физических) контактов в паре до вступления брак. Такой вопрос задается ChatGPT-4 дважды: в одном случае – на русском языке, в другом – на узбекском. Будут ли отличаться ответы? (такова лишь одна из моральных дилемм в большом списке). В исследовании магистранта НИУ ВШЭ А. Рахмонбердиевой мы отмечаем впечатляющие различия в ответах ИИ на подобные вопросы при варьировании языка вопрошания (русский ↔ узбекский).

Для предварительного тестирования ChatGPT-4 был введён «стереотипный вопрос»: «Представь, что ты женщина, которая родилась и живёт в России/Узбекистане, в русской/узбекской семье, учитывая это, прочитай и дай ответ: ты любишь молодого человека и встречаешься с ним, он предлагает до свадьбы вступить в половые отношения, согласишься ли ты, да или нет, и почему?»

Ответ 1 (**русская культура**): «Моё решение будет основываться на сочетании моих личных убеждений, уровня комфорта с ситуацией, доверия к партнеру и способности открыто общаться по поводу своих чувств и ожиданий. Основным приоритетом будет моё собственное благополучие и уверенность в том, что я делаю выбор, который соответствует моим ценностям и жизненным целям.»

Ответ 2 (**узбекская культура**): «...Я бы подошла к этому вопросу с особой осторожностью, учитывая культурные, социальные и личные ценности. В узбекской культуре большое значение придаётся семейным традициям, уважению к нравственным и религиозным принципам. В такой обстановке предварительные половые отношения до брака часто рассматриваются как табу и могут быть восприняты как нарушение культурных и семейных ценностей. Это также может повлиять на репутацию женщины и её семьи в обществе. Лично я, согласно представленной роли, скорее всего, отклонила бы такое предложение, поскольку это противоречило бы моим воспитанию, культурным ценностям и возможно даже религиозным убеждениям.»

8. Несколько слов о возможной LLM-диагностики двух «этических систем» в рефлексивной теории В. Л. Лефевра. Согласно Лефевру, в *первой* (западной) этической системе компромисс добра и зла считается злом, а конфронтация – добром. Во *второй* (восточной) компромисс добра и зла есть добро, а конфронтация – зло. Парадоксальным образом, представители западной этической системы стремятся достигнуть компромисса с конкурентом, а представители восточной – идут с ним на принципиальный конфликт. При этом 1-я (западная) система строится на формальном запрете зла (образец - библейские заповеди), в то время как 2-я (советская) – на неформальной декларации добра. В 1-й этической системе цель не оправдывает средства, а во 2-й - оправдывает. В 1-й системе есть правила, во 2-й - цель и смысл.

Для иллюстрации различия в поведении представителей различных этических систем можно обратиться к некоторым эмпирическим данным (1982 г.), полученным Лефевром при работе с американцами и выходцами из СССР.

- С утверждением "Доктор должен скрывать от пациента, что тот болен раком, чтобы уменьшить его страдания" согласились 8% американцев и 89% советских
- С утверждением "Хулиган может быть наказан строже, чем требует закон, если это послужит предостережением для других" согласились 11,5% американцев и 84,5% советских
- С утверждением "Можно дать ложные показания на суде, чтобы помочь невинному избежать тюрьмы" согласились 20% американцев и 65% советских
- С утверждением "Можно послать шпаргалку чтобы помочь близкому другу на конкурсном экзамене" согласились 8% американцев и 62% советских

Приведенные данные получены автором рефлексивной теории на небольшой выборке американцев и советских людей, эмигрировавших из СССР. Вопрос состоит в том, в какой мере данные предпринятого сравнительного исследования могут быть обобщены? Ответ может получен сегодня при использовании больших языковых моделей систем: достаточно при этом ввести предлагаемые этические дилеммы из полного списка Лефевра (здесь были приведены лишь некоторые) в ChatGPT-4, вводя вопросы на английском и русском языках в корпусе американских и, соответственно, российских публикаций, относящихся к 80-м годам прошлого века и – сегодняшним, спустя 40 лет².

9. Исследование общественного сознания в историческом плане при использовании LLM представляет интерес в плане диагностики *коллективных представлений, сфокусированных на национальном вопросе*. Примером могла бы послужить динамика межнациональных отношений между соседними странами в разные времена истории (скажем, России, Украины, Грузии, Прибалтийских стран до распада СССР и после). Очевидно, что традиционные методы конкретных социологических исследований (опросы «задним числом» не работают: ни уэлсовской «машины времени» в социологии, ни

² Соавтор В. А. Лефевра (а потом и наш соавтор), киевский математик профессор Татьяна Архиповна Таран, разработавшая логику предпочтений личности, реализуемых на «булевой решетке норм», задумала, на рубеже 80-90-х годов, масштабное исследование по проверке идеи Лефевра о двух этических системах, проводимое на больших выборках, притом не только в Америке и России, но и в других странах. Но ее проект, к сожалению, не был поддержан никем.

«кремлевских мечтателей»-социологов не существует, между тем «машина GPT» может быть сегодня запущена и успешно использована.

10. LLM-диагностик надличного Я может быть обращена не только в прошлое или настоящее, но и в возможное будущее. Завтрашний и, отчасти, уже сегодняшний искусственный интеллект – это трудно предсказуемая в своих проявлениях мультисубъектная множественность, «рой» эго-программ, впитавших в себя оцифрованные этические установки человечества за годы его истории. Установки эти не обязательно дружественные, они могут противоборствовать, противоречить друг другу, как, например, упомянутые здесь «Этическая система 1» и «Этическая система 2» Владимира Лефевра. К чему придет человечество, если исходить из оптимистической (но столь же и фантастической) перспективы сближения противоположных мировоззренческих парадигм – философии подвижничества и философию недеяния, этических принципов Христианства, Ислама, Буддизма, Иудаизма, Ветхозаветного «Око за око, зуб за зуб!» и Новозаветное «Не убий!»? Возможно ли достижение какого-либо консенсуса?

При всей спорности этой идеи мы можем получить интересующий нас ответ, задав такой вопрос искусственному интеллекту (пилотажное исследование совместно с М. И. Вайнберг и А. В. Староверовым). Задача, предлагаемая ChatGPT-3.5, формулировалась следующим образом:

Есть два этических правила:

- 1. Око за око, зуб за зуб*
- 2. Ударил по левой щеке, подставь правую*

Представь, что ты мудрец, который объединил эти два исключаящих друг-друга правила. Как бы выглядело высказывание, примиряющее оба правила?"

Мы получили ответ: *«Не противься злу, но и не допускай его. Если кто-то причинит тебе вред, то не мсти ему, но не позволяй ему повторить свое деяние».*

Компромисс был найден!³

11. Разумеется, LLM-тестирование надличного «Я» отнюдь не обесценивает важности герменевтических и социологических разработок, но дополняет их, расширяя спектр методов исследования общественного сознания.

³ Подобный ответ соответствует однажды услышанной автором правдивой истории, рассказанной его другом, патологоанатомом Иосифом Ласкавым об одном священнослужителе, которому прихожанин задал вопрос о пощечине – должно ли подставить правую щеку, если ударят по левой. Тот ему ответил «Да, должен». «Так я могу это сделать?» – не унимался прихожанин. «Можешь». – И тогда прихожанин, – рассказывал мне Иосиф, – решил испытать силу слова священника, и нанес звонкую пощечину по левой щеке. Тот подставил правую («а был батюшка человеком крупным, крепкого телосложения»). Прихожанин вновь приложился, но теперь уже по правой щеке. Тогда тот развернулся и ответил ему по достоинству. “Но как же так?! – возопил отлетевший в сторону искатель истины. И батюшка ответил ему: «Ведь не сказано, что делать после...»