



АВТОРСКАЯ АНАЛИТИЧЕСКАЯ СТАТЬЯ

Редакционная дата: 27 октября 2026 года

Слова, которых нет в интернете

От Финляндии до Саудовской Аравии: почему величайший национальный проект в области искусственного интеллекта может начаться с голоса бабушки, тетради школьника и памяти целого общества?

Летом 2020 года Финляндия обратилась к своим гражданам с необычной просьбой: говорить.

Их не просили читать официальные тексты, произносить тщательно подготовленные в лаборатории фразы или говорить на идеальном финском, которому учат в школе.

Их попросили говорить так, как они говорят в жизни.

О спорте, еде и воспоминаниях, о том, что они любят и о чём спорят. На своих естественных диалектах, с заминками, случайными ошибками, паузами, смехом и выражениями, которых не знает словарь.

Инициатива получила название «Подари свою речь» (Lahjoita puhetta — Donate Speech); её запустила Финская телерадиовещательная компания (Yle) вместе с Банком финского языка (Kielipankki) и партнёрами из науки и технологической отрасли.

Целью было собрать десять тысяч часов живой финской речи, чтобы создавать технологии, способные понимать людей такими, как они говорят на самом деле.

Когда кампания завершилась 6 марта 2024 года, Банк финского языка сообщил, что было собрано более трёх тысяч часов речи. Исследование, опубликованное в 2022 году, описывает более раннюю версию корпуса 2021 года: около 3600 часов и более двадцати тысяч предполагаемых говорящих. Эта оценка исходит из того, что каждый идентификатор клиента приложения соответствует одному говорящему, а не из проверки личности участников.

Инициатива не достигла первоначальной количественной цели, но создала нечто более важное, чем собрание аудиозаписей.

Она превратила обычного гражданина в участника создания инфраструктуры знаний, которая понадобится технологиям будущего.

Исследование *Lahjoita puhetta: A Large-Scale Corpus of Spoken Finnish* показывает, как повседневная речь стала материалом для изучения автоматического распознавания речи (Automatic Speech Recognition — ASR), понимания диалектов и анализа языковых и социальных особенностей.

Но самое прекрасное в этой истории — финны поняли одну предельно простую вещь: каким бы талантливым ни был инженер, он не может сидеть в лаборатории и изобретать то, как разговаривают миллионы людей.

Он может спроектировать систему, которая будет у них учиться, но материал для обучения ему должны дать люди.

Эта финская история приводит меня к вопросу, который, на мой взгляд, заслуживает беспрецедентной саудовской национальной инициативы:

Что, если каждый дом в Саудовской Аравии, каждая школа и каждый университет, каждая профессия и каждое учреждение смогут участвовать в создании крупнейшей инфраструктуры подлинных человеческих данных, представляющей наш язык, культуру и знания?

Не собирая в интернете то, что другие написали о нас, и не генерируя миллионы текстов, похожих на продукцию нынешних моделей.

А начав с самого человека.

Со слов, которые он произносит, смыслов, которые он понимает, документов, которые умеет читать, и деталей, которым машина не может научиться, потому что они никогда не были записаны в доступном ей источнике.

Полвека интернета — а жизнь всё ещё больше него

Представьте, что вы попросили языковую модель объяснить смысл выражения «يا بعد حيي» («я баад хайи»).

Она, возможно, даст приемлемое словарное толкование и объяснит, что это выражение любви и признательности.

Но знать значение выражения — одно, а понимать его место в человеческих отношениях — совсем другое.

Так может сказать пожилой мужчина внуку, женщина — сестре, друг — другу в минуту благодарности. Интонация и смысл выражения будут меняться в зависимости от ситуации, места и отношений между собеседниками.

А теперь представьте инженера на промышленном объекте: он употребляет принятый среди коллег термин для обозначения повторяющейся неисправности, которой под этим названием нет в техническом руководстве.

Или женщину из бедуинской среды, которая умеет назвать едва различимые перемены ветра и направления облаков, тогда как общая энциклопедия даёт им лишь приблизительное описание.

Или медсестру, которая знает, какими словами пациенты описывают боль и чем обиходное описание отличается от того, что врач вносит в медицинскую карту.

Эти знания отсутствуют не потому, что они ничего не стоят.

Во многих случаях их нет потому, что они не попали в цифровую среду, на которой учатся машины.

И здесь открывается проблема глубже, чем нехватка арабских текстов.

В статистике существует проблема, известная как смещение выборки (Sampling Bias). Если данные собраны у группы, которая не представляет всё общество, модель может превосходно понимать эту группу и гораздо хуже — тех, кто в выборку не попал.

С языковыми моделями проблема становится многослойной.

Интернет — не сбалансированная копия общества.

Одни люди пишут больше других; одни профессии оставляют после себя обилие цифровых текстов, а другие группы — лишь слабые следы; некоторые диалекты встречаются только в частных разговорах, которые нельзя собирать без разрешения.

Порой люди, владеющие самыми ценными местными знаниями, представлены в цифровых данных меньше всех.

Поэтому модель может много знать о мире и мало — о человеческом опыте, который совсем рядом с её пользователями.

Она знает опубликованное, но не обязательно известное людям.

В этом и состоит возможность: Саудовской Аравии нужно не столько собирать ещё больше того, что машины уже знают, сколько дать людям возможность научить их тому, до чего прежде они не могли добраться.

Когда народы стали партнёрами в обучении машин

Финляндия не была единственным исключением.

В Индии, где существует огромное множество языков и диалектов, правительство запустило инициативу BHASHINI, в том числе программу языкового донорства Bhasha Daan, которая позволяет гражданам предоставлять тексты, аудиозаписи и переводы.

В марте 2026 года Министерство электроники и информационных технологий Индии сообщило, что к платформе Bhashini Samudaye присоединились более десяти тысяч участников. Она объединяет частных лиц, университеты, компании и лаборатории для создания языковых ресурсов и разметки данных (Data Annotation).

Смысл не в том, чтобы правительство лишь собирало голоса граждан в архив.

Оно создаёт ресурс, в котором участвуют университеты, исследователи, организации и компании и которым могут пользоваться самые разные технологии.

Возможно, именно этим короткая кампания по сбору данных отличается от постоянной национальной инфраструктуры.

В Южной Корее проект принял более широкий институциональный характер.

Национальное агентство информационного общества (National Information Society Agency — NIA) создало платформу AI-Hub, которая, по данным самого агентства, с 2017 года открыла доступ более чем к девятистам наборам данных для обучения искусственного интеллекта.

Среди них — данные о корейском языке, речи, изображениях, медицине, промышленности, сельском хозяйстве и многом другом.

Но особенно важно, что Корея не стала относиться к разметке данных как к закрытой технической работе.

В рамках государственных программ, создававших обучающие данные для размещения на платформе, появились рабочие места и оплачиваемая краудсорсинговая работа по сбору, обработке и проверке качества данных. Агентство также документировало возможность того, что с опытом некоторые работники переходили от разметки к проверке и управлению проектами.

Здесь возникает важная экономическая перспектива: **обучение машин может стать сферой подготовки людей и приобретения ими новых профессиональных навыков.**

В Сингапуре выбрали другой, не менее увлекательный путь.

Организация AI Singapore вместе с Google запустила платформу Aquarium в составе экосистемы SEA-LION, чтобы сообществам и исследователям было проще участвовать в создании наборов данных на языках Юго-Восточной Азии.

Проект задуман не только для сбора материалов на малайском, тайском или индонезийском языках, но и для облегчения доступа к ресурсам, отражающим языковую и культурную реальность региона, а также для разработки инструментов участия, оценки и мотивации.

Возможно, само название — «Аквариум» — несёт красивую метафору: язык — не одна большая рыба, которую можно поместить в резервуар, а живая система разнообразия и связей.

Эстония выбрала иной, образовательный подход.

Она запустила инициативу AI Leap, чтобы ввести искусственный интеллект в общее образование и помочь учителям и ученикам осваивать навыки мышления и критического обучения.

В 2025–2026 учебном году в программе участвовали 154 из 156 старших школ страны, около пяти тысяч учителей и двадцати тысяч учеников.

Эта инициатива не является проектом сбора национальных обучающих данных в том смысле, о котором я говорю здесь.

Но она доказывает, что государство способно превратить школы в широкую национальную сеть обучения и участия в развитии новых технологических практик.

В Соединённых Штатах мы находим прекрасный пример в неожиданном месте — Библиотеке Конгресса (Library of Congress).

Библиотека запустила проект *By the People*, который приглашает добровольцев расшифровывать рукописи и исторические документы, написанные от руки, проверять чужие расшифровки и классифицировать материалы.

Добровольцем может быть школьник, пенсионер или историк-любитель.

С каждой прочитанной строкой исторический документ, трудный для понимания компьютером, превращается в текст, по которому можно вести поиск, который можно изучать и использовать.

Для искусственного интеллекта такая работа показывает ценность преобразования неструктурированных архивов (Unstructured Archives) в пригодное для использования цифровое знание.

Пять примеров, пять разных путей, но все они сходятся в одной мысли:

Данные, которые нужны искусственному интеллекту, не всегда можно купить в готовом виде. Иногда это знания, рассеянные среди миллионов людей, и новаторство заключается в том, чтобы создать способ собрать их, защитить права их носителей и проверить качество.

«Диван»: национальный проект, который пишут люди

Я представляю себе саудовскую инициативу с предлагаемым названием: **Саудовский диван человеческих знаний**.

Я вижу её не приложением, которое людей попросят скачать, а потом забудут, и не сезонным конкурсом, который завершится раздачей призов.

Я вижу постоянную национальную инфраструктуру общественного участия (National Participatory Data Infrastructure), к которой предлагается привлечь школы и университеты, частный сектор, некоммерческие объединения, исследовательские центры, а также граждан и жителей страны, желающих внести свой вклад.

Её целью станет создание крупнейшего в Саудовской Аравии документированного ресурса подлинных и размеченных человеческих данных с продуманным языковым, культурным, профессиональным и социальным разнообразием.

В центре «Дивана» могла бы находиться цифровая платформа, позволяющая участвовать на разных уровнях.

Школьник заходит на неё и находит задание, соответствующее его возрасту и знаниям: объяснить местное выражение, описать сцену, прочитать предложение вслух или проверить текст, автоматически извлечённый из лицензированного документа.

Учительница помогает классифицировать учебные тексты, определить уровень их сложности и объяснить неоднозначные места.

Врач или медсестра проверяет медицинские термины и выражения в примерах, специально подготовленных для обучения, или в учебных данных, анонимизированных по требованиям Исполнительного регламента к Закону о защите персональных данных после оценки риска повторной идентификации, а не в раскрытых записях пациентов.

Инженер объясняет профессиональную лексику, классифицирует изображения оборудования, доступные на законных основаниях, или описывает этапы типового производственного процесса.

Женщина, интересующаяся социальной историей, с собственного согласия записывает объяснение народной поговорки, местного обычая или названия старинного предмета домашнего обихода.

Житель пустыни описывает слова, связанные с местностью, погодой и дикой природой, а горожанин — выражения, возникшие в его современной среде.

Ни у одного диалекта нет заранее заданного преимущества перед другим, как нет его и у одного общественного стиля перед другим.

Задача не в том, чтобы навязать модели единственный образ жителя Саудовской Аравии, а в том, чтобы она увидела разнообразие общества, у которого учится.

Жители страны, не являющиеся её гражданами, тоже смогут участвовать, внося знания своих языков и свой опыт, особенно в областях повседневного общения, труда, услуг и межкультурного перевода.

Так появится возможность создавать модели, которые шире понимают социальный и профессиональный контекст Королевства.

Главное преимущество в том, что участнику не нужно разбираться в программировании или машинном обучении.

Достаточно знать что-то настоящее, чем он может поделиться с другими.

Бабушка может не знать, что такое архитектура трансформера (Transformer Architecture), но она знает истории и выражения, которых инженер, создающий трансформеры, не изобретёт при всей доступной ему вычислительной мощности.

Так инициатива становится больше, чем источник данных.

Она становится общественным признанием того, что знания принадлежат не только людям с учёными степенями, а повседневный опыт сам по себе может быть ценным материалом для развития технологий.

День, когда исправить слово станет делом национального значения

Представим интерфейс «Дивана».

Доброволец видит изображение старинного арабского документа, разрешение на использование которого дала владеющая им организация.

Система оптического распознавания символов (Optical Character Recognition — OCR) пытается извлечь текст, но из-за особенностей почерка или повреждения бумаги неверно читает слово.

Доброволец замечает слово и исправляет его.

Затем исправление показывают второму участнику. Если прочтения расходятся, слово передают более опытному рецензенту.

За несколько минут участники выполняют работу, которая может помочь создать набор данных для развития систем чтения арабских документов.

В другом задании появляется лицензированное изображение предмета традиционного ремесла или сельскохозяйственного орудия.

Платформа просит участников описать его, указать принятое в их регионе название, рассказать о назначении и, возможно, привести другое название из иного диалекта.

Изображение связывается с подходящими описаниями и становится материалом для обучения зрительно-языковых моделей (Vision-Language Learning).

В третьем задании доброволец слушает запись, которую человек предоставил с собственного согласия, а затем проверяет расшифровку системы автоматического распознавания речи.

С технической точки зрения это не отвлечённая идея.

Платформа [Google Crowdsourcing](#) позволяет пользователям проверять метки изображений (Image Label Verification), распознавать рукописный текст (Handwriting Recognition), проверять переводы, описывать изображения, подтверждать аудиозаписи и выполнять другие задания.

Платформа начисляет разное количество баллов в зависимости от задания и его сложности, а за некоторые материалы, подтверждённые другими участниками, даёт дополнительные баллы.

Проект Common Voice фонда Mozilla (Mozilla Foundation) позволяет людям записывать свой голос, а другим добровольцам — проверять записи; так возникают многоязычные аудиокорпуса, которые можно использовать для развития речевых технологий.

Идея уже доказала свою осуществимость.

Но саудовская инициатива могла бы пойти дальше простого копирования.

Она могла бы связать участие с областями знания, создать траектории для продвинутых уровней, разработать сертификаты опыта в языковой, аудио- и визуальной разметке и сделать участие путём к специализации в науке о данных.

Тот, кто начнёт с исправления одного слова в рукописи, через несколько лет может стать исследователем, создающим систему, способную читать тысячи рукописей.

Может ли машина учуять запах машины?

Здесь возникает проблема, от которой не уйти.

Если открыть участие широкому кругу людей, некоторые могут попытаться сгенерировать готовые тексты с помощью таких моделей, как GPT, Claude или DeepSeek, а затем представить их как подлинный человеческий вклад.

И эта угроза вовсе не второстепенна.

Если мы хотим создать ресурс знаний, исходящих от людей, а он наполнится текстами коммерческих моделей, мы рискуем вернуть в него знания, уже содержащиеся в прежних моделях, вместо того чтобы расширить представленность человеческого опыта.

Поэтому я предлагаю включить в платформу инициативы специальную исследовательскую систему для проверки происхождения и качества материалов.

Такая система могла бы использовать анализ последовательностей слов (N-gram Analysis), статистические отпечатки стиля (Stylometric Fingerprints), перплексия, или степень языковой неожиданности (Perplexity), а также поиск повторов и почти одинаковых текстов (Near-Duplicate Detection).

Например, она могла бы искать необычно частое повторение отдельных конструкций, длинные последовательности, совпадающие с опубликованными материалами, или паттерны, которые намного чаще встречаются в машинных текстах, чем в контрольных корпусах человеческого письма.

Но здесь возникает научный парадокс, который нельзя оставлять без внимания.

Нельзя доказать, что текст написан человеком, лишь потому, что он не похож на машинный; нельзя и обвинить человека в использовании машины лишь потому, что его стиль последователен.

Исследование, опубликованное Вэйсином Ляном (Weixin Liang) и его коллегами из Стэнфорда в журнале *Patterns* в 2023 году, показало на изученных выборках и семи проверенных детекторах, что средства обнаружения сгенерированного текста могут быть предвзяты к текстам людей, для которых английский не является родным, и принимать человеческий текст за машинный.

Другое исследование под руководством Деборы Вебер-Вульф (Debora Weber-Wulff), опубликованное в *International Journal for Educational Integrity*, выявило ограниченную надёжность ряда распространённых инструментов обнаружения и в некоторых случаях высокий уровень ложноположительных и ложноотрицательных результатов.

Поэтому предлагаемая система проверки контента не должна становиться судьёй, который запрещает участие на основании одного вероятностного показателя.

Она должна быть частью более широкой архитектуры подтверждения происхождения данных (Data Provenance).

Эта архитектура начинается с самого вклада: ясного заявления о способе его создания, необязательной документированной истории редактирования, заданий, выполняемых прямо внутри платформы, независимых человеческих проверок и случайных контрольных выборок при необходимости — без навязчивого наблюдения за добровольцами.

Изначально созданные людьми данные (Human-Authored Data) должны быть отделены от текстов, в которых человек проверил машинный результат (Human-Verified AI Outputs), а те и другие — от синтетических данных (Synthetic Data).

Все три вида могут быть полезны, но это не одно и то же, и различия между ними не должны исчезать в общем хранилище.

Я предлагаю вести для каждого набора данных журнал происхождения, где будет указано, как он собран, на каких условиях лицензирован, как проверялся и насколько можно доверять его качеству.

Это согласуется с идеей, которую Тимнит Гебру (Timnit Gebru) и её коллеги изложили в известной работе *Datasheets for Datasets*: документировать наборы данных так же, как в промышленности документируют характеристики продукции и её компонентов.

Цель не в том, чтобы создать инструмент, хвастающийся способностью распознавать «запах искусственного интеллекта».

Цель — создать ресурс знаний, в котором нам известна история каждой части, понятно, чему она может научить и где проходят границы доверия к ней.

Запах — это впечатление.

А документированное происхождение — знание.

Вид материала	Происхождение	Что следует документировать?
Данные, созданные человеком (Human-Authored Data)	Человеческое авторство	Источник, процесс создания и права
Результаты ИИ, проверенные человеком (Human-Verified AI Outputs)	Машинный результат после человеческой проверки	Генератор, роль проверяющего и границы проверки
Синтетические данные (Synthetic Data)	Машинная генерация	Способ генерации, цель и ограничения использования

Концептуальное обобщение различий из этого раздела: принадлежность к одной из категорий не гарантирует достоверности или пригодности.

Когда разногласия между людьми — не ошибка в данных

Один из важнейших принципов «Дивана» должен состоять в том, чтобы не считать ошибкой, которую нужно стереть, любое расхождение между участниками.

Два человека могут увидеть одно изображение и описать его по-разному.

Два носителя из разных регионов могут не сойтись в значении выражения, потому что в их среде одно слово употребляется в разных смыслах.

Два исследователя могут прочесть неоднозначную фразу и прийти к двум обоснованным толкованиям.

Во многих традиционных системах разметки платформа стремится получить единственную окончательную метку, принимаемую за истинную (Ground Truth Label), и может опираться на решение большинства или профильного эксперта.

Это уместно в задачах с ясным объективным ответом: например, когда нужно прочесть цифру в неповреждённом документе или определить присутствие конкретного объекта на изображении.

Но для языка и культуры такой подход может подходить хуже.

Не всякое разногласие означает, что одна из сторон ошибается.

В исследовании *Investigating Reasons for Disagreement in Natural Language Inference*, опубликованном Нан-Цзян Цзян (Nan-Jiang Jiang) и Мари-Катрин де Марнефф (Marie-Catherine de Marneffe) в 2022 году, авторы изучили причины, по которым рецензенты по-разному толкуют отношения между текстами. Они установили, что некоторые расхождения объясняются реальной смысловой неоднозначностью, а не случайными ошибками разметки.

Для проекта, стремящегося представить саудовское общество, это имеет огромное значение.

Различие в употреблении слова между жителями разных регионов может быть именно той информацией, которая нужна машине.

Разное понимание выражения поколениями может указывать на языковое изменение, достойное документирования.

Поэтому система разметки должна уметь сохранять несколько ответов там, где в этом есть смысл, и связывать их с нужным географическим, возрастным или профессиональным контекстом, не превращая лишние персональные сведения в открытые записи.

Это не означает, что по любому вопросу все мнения одинаково верны.

Медицинские, научные и юридические вопросы требуют проверки специалистом, тогда как множественность может быть оправдана при описании языкового употребления или общественного обычая.

Иногда хорошие данные учат модель не одному ответу, а тому, почему люди отвечают по-разному.

Это один из самых больших возможных выигрышей от перехода со сбора текстов на сбор размеченных человеческих знаний.

Час добровольной работы может стать активом знаний

Теперь я подхожу к тому, что считаю решающим для успеха инициативы: как превратить участие в общественную привычку, а не во временный порыв, который закончится вместе с информационной кампанией?

Можно опереться на Национальную платформу добровольческой работы Королевства: в ней уже существуют механизмы организации добровольческих возможностей, учёта часов и выдачи сертификатов.

Отчёты Saudi Vision 2030 показывают значительный рост добровольчества: согласно отчёту за 2025 год, Национальная платформа зарегистрировала в общей сложности более 189 миллионов часов добровольной работы.

Представьте, что одной из официально признанных возможностей станет «участие в документировании саудовских диалектов», «проверка исторических арабских текстов», «разметка учебных документов» или «создание речевых данных для лучшего понимания арабского языка».

Доброволец заходит на платформу, выполняет задания своего уровня, а после проверки качества его работа засчитывается.

Положенные часы учитываются по опубликованным временным и профессиональным критериям, а не просто по числу кликов или слов.

В будущем, если ведомства официально одобряют такую интеграцию, благодаря институциональному взаимодействию эти часы и сертификаты можно будет отображать в государственных цифровых сервисах, в том числе в приложении Tawakkalna.

Нет необходимости создавать параллельную систему для всего, что уже существует.

Разумнее связать инициативу с национальной инфраструктурой добровольчества и образования и добавить к ней новую область — **цифровое волонтерство в сфере знаний (Digital Knowledge Volunteering)**.

Но в системе стимулов есть тонкость, которую необходимо понимать.

Если вознаграждать людей только за количество слов, мы можем получить много слов малой ценности.

Если вознаграждать за количество меток, некоторые начнут торопиться и отвечать без должной проверки.

Это напоминает закон Гудхарта (Goodhart's Law) в экономике и теории измерений: когда показатель становится целью вознаграждения, он может утратить способность отражать качество, для измерения которого служил.

Поэтому вознаграждение должно зависеть от качества вклада, уровня его сложности и результатов проверки.

Доброволец может получать очки опыта (Experience Points — XP), продвигаться по уровням знаний, открывать более специализированные задания, получать значки навыков (Skill Badges), официальные сертификаты и возможности стажировки в университетских лабораториях.

Университеты, школы и регионы могут соревноваться не по числу необработанных материалов, а по качеству, разнообразию и исследовательской значимости созданных ими ресурсов.

Стимулами могут стать исследовательские премии, стажировки, гранты на студенческие проекты и признание выдающихся участников.

Для работы, которая требует высокой квалификации или регулярных профессиональных усилий, должны существовать оплачиваемые формы — по примеру некоторых корейских проектов, — а не постоянная зависимость от добровольного труда.

Добровольчество не должно становиться предлогом для получения бесплатной работы специалистов.

Инициатива, которая уважает своих участников, лучше других способна надолго сохранить качество и их доверие.

Школа, которая становится лабораторией знаний

Я представляю школы и университеты сердцем этой инициативы.

Но это не означает, что учеников нужно заставлять производить данные или автоматически превращать школьные задания в обучающий материал.

Нужны добровольные и продуманные образовательные траектории, в которых участие помогает понимать язык, исследовательскую работу и искусственный интеллект.

В средней школе ученик может научиться описывать изображение, различать факт и толкование, понимать, почему машины не всегда одинаково читают тексты.

В старших классах он сможет участвовать в проекте по документированию местных выражений или научной лексики и узнавать, как составляют вопросы для оценки и проверяют результаты.

Студент университета может перейти к более сложным заданиям: проверять специализированные данные, изучать согласованность разметчиков (Inter-Annotator Agreement), обнаруживать смещение в наборах данных (Dataset Bias) и анализировать ошибки моделей.

Такая работа может стать выпускным проектом, научной статьёй или основой стартапа.

Учителя и преподаватели вузов будут играть ключевую роль в разработке и оценке заданий, а не просто осуществлять административный надзор.

Участие детей требует особых гарантий: согласия законного представителя, когда оно необходимо; минимизации персональных данных; предпочтения безопасных образовательных заданий; запрета на автоматическое использование голоса, изображения или личных текстов ребёнка в открытом обучении.

Саудовский Закон о защите персональных данных (Personal Data Protection Law — PDPL) и правила, регулирующие данные детей, подчёркивают важность законной цели, надлежащего согласия там, где оно служит правовым основанием обработки, минимизации и защиты данных.

Если инициатива основана на участии людей, уважение их права не участвовать — часть её законности.

Ученик не должен терять оценку или образовательную возможность из-за того, что решил не предоставлять свои данные.

Мы хотим, чтобы ребёнок узнал, как создаётся знание, а не сам стал сырьём, которое собирают, не объясняя ему, куда оно попадёт.

Кому принадлежит то, чем делятся люди?

На мой взгляд, это самый трудный и самый важный вопрос всей инициативы.

Если миллион человек предоставит тексты, голоса, изображения и объяснения, а компания использует их для разработки коммерческой модели, что получит общество?

Станут ли данные исключительной собственностью одной организации? Сможет ли независимый исследователь получить к ним доступ? Сможет ли участник отозвать персональные данные, обработку которых он больше не желает продолжать, в пределах надлежащих прав и правовых ограничений?

Допустимо ли превращать редкий культурный материал в товар, не проявляя уважения к сообществу, которое его предоставило?

Успокаивающей фразы «национальные данные» здесь недостаточно.

Национальный статус не отменяет прав людей и сообществ на их данные.

Я считаю, что инициативе нужна модель управления, близкая к концепции общественного доверительного управления данными (Data Trust): ответственная организация управляет ресурсами в рамках чётких полномочий и под правовым, научным и общественным надзором.

Каждый набор данных должен иметь паспорт, где указаны цель его использования, область лицензирования (Data Licensing), права участников и уровень доступа.

Некоторые наборы можно публиковать как открытые данные (Open Data), особенно если в них нет персональных сведений и соблюдены права на материалы.

Другие можно предоставлять исследователям на определённых лицензионных условиях и с ограничениями.

Третьи должны оставаться в защищённых исследовательских средах (Secure Research Environments), где нельзя копировать необработанные данные или узнавать личности их владельцев.

Полезным ориентиром могут служить принципы CARE для управления данными коренных народов (CARE Principles for Indigenous Data Governance), разработанные международными сетями. Они подчёркивают коллективную выгоду, право сообществ

контролировать данные, ответственность и этику.

Я не предлагаю буквально переносить эти принципы в саудовский контекст, а лишь обратиться к их общей философии: культурные знания нельзя изымать из сообщества так, чтобы получаемая от них польза больше к нему не возвращалась.

Женщина может предоставить запись, в которой объясняет изготовление традиционного изделия, но это не обязательно означает, что она разрешает компании использовать её голос для создания модели, имитирующей её голосовую идентичность.

Человек может разрешить записать местную историю, но не раскрывать имена членов его семьи или подробности их жизни.

Поэтому платформа должна различать согласие на сбор материала, его использование в исследованиях, использование для обучения моделей и коммерческое предоставление.

Согласие — не короткое слово, которое ставят внизу экрана.

Это этические отношения, от которых зависит, доверится ли общество инициативе и продолжит ли питать её своими знаниями.

Данные, собранные благодаря доверию людей, могут быть ценнее более крупного массива, владельцы которого не понимали, что с ним произойдёт.

Почему этот проект может быть важнее одной модели?

Представьте, что инициатива сможет создать за несколько лет.

Обширный ресурс саудовских голосов и диалектов, полезный для технологий распознавания речи (Speech Recognition).

Наборы подлинных и документированных арабских текстов для языкового моделирования (Language Modeling).

Размеченные изображения и документы для понимания документов (Document Understanding) и компьютерного зрения (Computer Vision).

Терминологические базы по медицине, инженерному делу, образованию и промышленности.

Скрытые оценочные наборы (Blind Evaluation Sets), с помощью которых университеты смогут проверять, насколько модели понимают общество.

Это разные ресурсы, и их нельзя без разбора складывать в один сосуд.

Одни данные пригодны для предварительного обучения (Pretraining), другие — для тонкой настройки (Fine-Tuning), третьи — только для оценки (Evaluation).

Оценочные данные необходимо тщательно изолировать, чтобы они не просочились в обучение: иначе модель уподобится ученику, который заранее узнал экзаменационные вопросы.

Но если управлять этими ресурсами как общей научной инфраструктурой, на их основе сможет возникнуть широкая экосистема исследований, компаний и продуктов.

Ею могли бы пользоваться существующие национальные модели, исследовательские лаборатории, компании глубинных технологий, университеты и стартапы.

Некоторые наборы могут стать востребованным за пределами Королевства ресурсом для обучения или тестирования многоязычных систем и разработки специализированных технологий.

Тогда достижение будет состоять не просто во владении самым большим количеством арабских токенов (Arabic Tokens).

Оно будет состоять во владении данными, происхождение, разнообразие и пределы которых нам известны, как известно и то, что именно они добавляют к возможностям модели.

Больше — не всегда лучше, а самое редкое — не всегда самое малочисленное.

Тысяча часов документированной естественной речи из среды, не представленной в мировых данных, может иметь огромную ценность, тогда как десять миллионов похожих предложений могут оказаться лишь дополнительной нагрузкой при обучении.

Новый капитал по имени общество

Я думаю, одним из важнейших экономических сдвигов в искусственном интеллекте станет переход от внимания лишь к объёму данных к их подлинности, происхождению, правам использования и познавательной ценности.

Чем легче и дешевле моделям генерировать тексты, тем труднее понять, представляют ли новые данные человеческий опыт, которого модели прежде не знали, или только перерабатывают вероятности самой модели.

Поэтому относительная ценность подлинных человеческих данных может вырасти, особенно в менее представленных областях, языках и видах опыта.

Не потому, что человек всегда создаёт более точные данные, чем машина, а потому, что он живёт в мире, который машина пытается представить.

Врач знает, как пациенты в действительности описывают симптомы. Учительница знает повторяющиеся ошибки учеников. Инженер знает рабочие термины, возникшие внутри его предприятия. Фермер знает слова земли, которых нет в техническом руководстве.

Этот опыт распределён между людьми, и ни одна центральная лаборатория не способна собрать его в одиночку.

Здесь мы возвращаемся к идее австрийского экономиста Фридриха Хайека (Friedrich Hayek) о рассеянном знании (Dispersed Knowledge), изложенной в его знаменитой статье *The Use of Knowledge in Society* 1945 года.

Хайек обращал внимание на то, что значительная часть практических знаний рассеяна среди членов общества и связана с местными обстоятельствами, которые ни один центр не способен охватить полностью.

Мы можем не соглашаться со всеми его экономическими выводами, но это наблюдение кажется чрезвычайно уместным в эпоху искусственного интеллекта.

Ни один центр обработки данных в одиночку не знает того, что знает целое общество.

Если инициатива сможет создать способы участия, проверки и сохранения, она откроет канал, который превратит эти рассеянные знания в ресурс для учёных и инженеров.

Это новый экономический актив.

Не просто огромное собрание файлов, а **инфраструктура непрерывного производства знаний, источник которой — общество и часть пользы от которой возвращается обществу.**

Начало, ради которого не нужно ждать будущего

Я не думаю, что в первый день инициатива должна просить миллион вкладов.

Лучше начать с конкретных проектов, воздействие которых исследователи смогут действительно измерить.

Первые направления можно выбрать среди диалектов и естественной речи, арабских документов, культурных изображений и профессиональной терминологии, привлекая университеты, школы, некоммерческие организации и исследовательские лаборатории.

Затем данные делят на обучающие и тестовые наборы и сравнивают результаты моделей до и после добавления человеческого вклада.

Лучше ли распознаётся конкретный диалект? Стало ли меньше ошибок при извлечении таблиц? Начала ли модель лучше понимать профессиональную лексику? Снизил ли разнообразие участников смещение?

Если проявится ясный положительный эффект, инициативу можно расширять на новые регионы и области.

Если результаты не улучшатся, сначала следует понять причину и лишь потом собирать больше данных.

Цель не в том, чтобы собрать тексты, объёмом которых мы сможем гордиться.

Цель — собирать знания, которые приносят пользу.

На несколько лет проекту можно поставить амбициозные цели: например, привлечь миллион добровольных участников, создать размеченные наборы человеческих данных в разных областях и разработать саудовские стандарты оценки, признанные исследовательским сообществом.

Но успех следует измерять качеством, разнообразием и воздействием данных, а также соблюдением прав их владельцев, а не одним числом участников.

Возможно, прекраснейшим показателем успеха станет момент, когда через пять лет саудовский университет обнаружит, что набор данных, созданный при участии школьников, учителей и добровольцев из разных регионов, превратился в эталон, которым пользуются международные лаборатории.

Тогда общество внесёт вклад в производство знания, вышедшего за пределы места, где оно возникло.

Когда люди дарят то, чего машина не способна изобрести

Я возвращаюсь к Финляндии.

Инициатива не просила обычного гражданина стать специалистом по информатике или понять математические уравнения, на которых основаны модели распознавания речи.

Она просила о более простом: говорить.

За этой просьбой стояло глубокое понимание: некоторые важнейшие ресурсы будущего уже находятся внутри общества, но для них ещё не создан путь, превращающий их в знание, доступное для обучения.

Я считаю, что у Саудовской Аравии есть возможность воплотить эту идею в более широком масштабе.

У нас развитое школьное и университетское образование, растущий некоммерческий сектор, национальные платформы добровольческой работы, исследовательские учреждения, технологические компании и общество с богатым географическим, социальным, профессиональным и языковым разнообразием.

Вопрос не в том, сумеем ли мы убедить граждан произвести миллионы предложений.

Вопрос в том, сумеем ли мы создать проект, где каждый малый вклад станет частью большого научного труда, а его автор сохранит право знать, куда отправился его материал и как его использовали.

Представьте ребёнка, который описывает изображение предмета, виденного им только в доме деда.

Затем дед объясняет название предмета и его прежнее назначение.

Учительница проверяет текст, лингвист исправляет некоторые слова, а специалист связывает его с документированным историческим материалом.

Несколько минут взаимодействия четырёх человек превратились в знание, которого в таком виде может не быть ни на одном сайте в интернете.

Это знание может стать примером в наборе данных, частью стандарта оценки, стать цифрового словаря или материалом для разработки модели, понимающей историю и культуру вещей.

Именно такой образ, на мой взгляд, достоин имени «Саудовский диван человеческих знаний».

Не огромное хранилище, куда сваливают всё, чем делятся люди.

А организация, которая умеет их слушать, через них обучает машины, а взамен даёт им навыки, признание и справедливую долю пользы от знания, созданию которого они помогли.

Тогда мать, объяснившая местное выражение, не останется лишь крошечным вкладом в базу данных.

А ученик, исправивший слово в рукописи, — добровольцем, проведшим пять минут у экрана.

Каждый из них по-своему поучаствует в создании чего-то большего.

Будущее, в котором машины смогут изучать нашу культуру, может зависеть от слов, ещё не сказанных никем перед микрофоном, от документов, ещё не прочитанных компьютером, и от знания, которое хранит женщина или мужчина, пока не подозревающие, что однажды оно станет ценным научным ресурсом.

Я хочу не просить общество говорить больше, чтобы заполнить память машины.

Я хочу, чтобы мы создали систему, которая слышит то, что общество действительно знает.

Народы передают искусственному интеллекту свой язык не простым переводом чужого знания; они отдают ему частицу себя, когда их сыновья и дочери становятся соавторами знания, на котором он будет учиться.

И, возможно, величайший саудовский проект в области искусственного интеллекта начнётся не в комнате, полной серверов, а в обычном доме, когда бабушка скажет внучке: «Иди сюда, я научу тебя слову, которого они не знают».

И внучка научится.

А следом научится машина.

Человек же останется первым учителем.

Источники и первичные материалы

1. [Donate Speech \(Lahjoita puhetta\)](#) — The Language Bank of Finland (Kielipankki).
2. [Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks](#) — Language Resources and Evaluation / Springer Nature.
3. [Over 10,000 contributors onboarded on Bhashini Samudaye](#) — Press Information Bureau, Government of India / Ministry of Electronics & IT.
4. [AI Data](#) — National Information Society Agency, Republic of Korea.
5. [AI-Hub data catalog](#) — AI-Hub / National Information Society Agency, Republic of Korea.
6. [Data Dam project begins, being key to Digital New Deal](#) — Ministry of Science and ICT, Republic of Korea.

7. [Data labeler established as a new future-oriented job](#) — National Information Society Agency, Republic of Korea.
8. [Introducing Aquarium: An Open Data Platform for Southeast Asian Languages](#) — AI Singapore / SEA-LION.
9. [We're expanding our presence in Singapore to advance AI in the Asia-Pacific region](#) — Google DeepMind.
10. [TI-Hüpe jõuab sügisest kutseharidusse ja põhikooliõpetajateni](#) — Estonian Ministry of Education and Research.
11. [AI Leap](#) — e-Estonia / Estonian Government.
12. [By the People: LC Crowdsourcing](#) — Library of Congress.
13. [Achievements \(Points\)](#) — Google Crowdsourcing Help.
14. [How does Google use my answers?](#) — Google Crowdsourcing Help.
15. [Common Voice](#) — Mozilla Common Voice.
16. [GPT detectors are biased against non-native English writers](#) — Patterns / Cell Press.
17. [Testing of detection tools for AI-generated text](#) — International Journal for Educational Integrity / Springer Nature.
18. [Datasheets for Datasets](#) — arXiv (preprint).
19. [Datasheets for Datasets](#) — Communications of the ACM (published article).
20. [Investigating Reasons for Disagreement in Natural Language Inference](#) — Transactions of the Association for Computational Linguistics / MIT Press.
21. [Saudi Vision 2030 Annual Report 2025](#) — Saudi Vision 2030.
22. [Ministry platforms — Volunteer Platform](#) — Saudi Ministry of Human Resources and Social Development.
23. [Personal Data Protection Law](#) — Saudi Data & AI Authority.
24. [Implementing Regulations of the Personal Data Protection Law](#) — Saudi Data & AI Authority / National Data Management Office.
25. [Children and Incompetents Data Protection Policy](#) — Saudi Data & AI Authority / National Data Management Office.
26. [The CARE Principles for Indigenous Data Governance](#) — Global Indigenous Data Alliance.
27. [The Use of Knowledge in Society](#) — American Economic Review / American Economic Association.
28. [Conference Volumes — Papers in Monetary Economics](#) — Reserve Bank of Australia.