ИССЛЕДОВАТЕЛЬСКАЯ СТАТЬЯ

Системы генеративного искусственного интеллекта в контексте цифрового доверия

Валерий Эдуардович Карпов[a]

[a] **Национальный исследовательский центр «Курчатовский институт»**

EDN: FZXVZX

Аннотация: В работе рассматривается роль систем генеративного искусственного интеллекта (ИИ) как одного из основных инструментов, формирующих базис цифрового доверия. При этом постулируется, что целью формирования цифрового доверия является формирование условий для повышения эффективности решения задачи управления обществом. Эти условия включают комплекс мер по мониторингу состояния общества как объекта управления, формированию требуемой картины мира индивидов, созданию механизмов разрешения конфликтов самой разной природы и т.п. В работе определены требования к модели управления социумом на стратегическом уровне. Эта модель должна оперировать ограниченным количеством факторов и связей между ними и быть наглядной, пригодной для качественного субъективного анализа лицом, принимающим решения. Исходя из этого, основными задачами цифрового доверия являются манипулирование поведением членов социума, формирование у них требуемого мировоззрения, выработка способности к управляемости. Доказывается, что использование технологий генеративного ИИ для решения этих задач является эффективным в силу самой сути его систем, созданных для порождения информационных объектов самой разной природы с имитацией осмысленности.

Ключевые слова: генеративный ИИ, цифровое доверие, управление социумом, когнитивная карта, картина мира, мифологизация информационных технологий

Благодарности: Автор выражает большую благодарность кандидату философских наук, доценту Д.А. Томильцевой (Уральский федеральный университет) и кандидату технических наук, доценту И.П. Карповой (НИУ ВШЭ, МФТИ) за критику и конструктивные замечания в части разделов, затрагивающих вопросы философии, социологии и экономики.

Дата поступления статьи в редакцию: 21 сентября 2025 года.

Generative AI systems in the context of digital trust

RESEARCH ARTICLE

Valery E. Karpov[a]

[a] **National Research Center «Kurchatov Institute»**

Abstract: The article examines the function of generative artificial intelligence (AI) systems as one of the key instruments forming the foundation of digital trust. According to the author, the goal of building digital trust is to establish conditions that improve the effectiveness of social governance. These include measures for monitoring the state of society as an object of governance, shaping the desired worldview of individuals, creating mechanisms for resolving conflicts of various natures. The study defines requirements for a model of social governance at the strategic level. This model should operate with a limited number of factors and the relationships between them and be visual, suitable for high-quality subjective analysis by decision makers. Based on this, the key objectives of digital trust are to manipulate the societal behavior, shape worldviews, and develop controllable individuals. The study demonstrates the effectiveness of generative AI technologies in addressing these issues, as these systems are designed to produce information objects of different nature with simulated meaning.

Keywords: generative AI, digital trust, social governance, cognitive map, worldview, mythologization of information technology

Acknowledgments: The author expresses sincere gratitude to Associate Professor, Candidate of Sci. (Philosophy) D.A. Tomiltseva (Ural Federal University) and to Associate Professor, Candidate of Sci. (Technical Sciences) I.P. Karpova (National Research University Higher School of Economics, Moscow Institute of Physics and Technology), for their critical and constructive comments on sections dealing with philosophy, sociology, and economics.

Received: September 21, 2025.

Введение

Цифровое доверие является одним из ключевых понятий современной информационной эпохи. При этом оно не имеет общепризнанного определения; кроме того, нередко этот термин используется без раскрытия его сущности, на уровне общих деклараций. Так, еще в 2013 году Организацией экономического сотрудничества и развития (ОЭСР) было разработано руководство по доверию («OECD Recommendation of the Council Concerning Guidelines on the Protection of Privacy and Transborder Flows of Personal Data»), которое включало в себя восемь принципов: (1) принцип ограничения сбора информации; (2) принцип качества данных; (3) принцип определения цели; (4) принцип ограничения использования данных; (5) принцип гарантий безопасности; (6) принцип открытости; (7) принцип индивидуального участия; (8) принцип подотчетности [Стырин, Рыбушкина, Санина, 2024. С. 16]. ОЭСР дает довольно описательное определение доверия в цифровую эпоху – это «готовность рисковать временем, деньгами и раскрытием личных данных при участии в коммерческой и социальной деятельности»[1]. Нередко определение цифрового доверия дополняется понятиями, которые призваны раскрыть цель его формирования («удобство», «сокращение издержек» и т.п.), однако, как представляется, эти термины описывают лишь побочные эффекты, которые достигаются путем его формирования. Любое рукотворное явление, тем более такое, на создание и поддержание которого затрачиваются весьма серьезные ресурсы, имеет свою цель. Это касается и феномена цифрового доверия.

Мы предполагаем, что основная цель цифрового доверия – это создание условий для повышения эффективности решения задачи управления обществом (что с точки зрения теории управления может рассматриваться как способность общества отработать внешние задающие воздействия[2]). Эти условия включают комплекс мер по мониторингу состояния общества как объекта управления, формированию требуемой картины мира индивидов, созданию механизмов разрешения конфликтов самой разной природы и т.п.

Важным инструментом решения этих задач являются системы генеративного искусственного интеллекта, позволяющие, в частности, создавать целевое информационное пространство. Генеративный искусственный интеллект (Generative AI, GenAI) – это разновидность искусственного интеллекта, которая использует генеративные модели для создания текстов, изображений, видео и других форм представления данных. В основе механизмов генеративного ИИ – обучение на больших массивах данных с последующим синтезом нового контента по запросу пользователя. Сейчас под генеративным ИИ обычно понимают системы, созданные на основе больших языковых моделей.

Настоящая статья призвана предложить единый взгляд на феномен цифрового доверия и связанных с ним понятий с точки зрения управления. В ней будут рассмотрены модели, лежащие в основе формализации задачи управления социумом, определена специфика задачи цифрового доверия и роль генеративного ИИ в его формировании.

Задача управления социумом

Формальная модель

Модели управления социумом носят, прежде всего, информационный характер – как по форме описания, так и по сути процессов, которые они описывают. Таких моделей построено немало. Так, базовая модель борьбы условных информаций для описания конкуренции фирм, взаимодействия валют, исторических процессов и т.д. приводится в работе Д.С. Чернавского и его коллег [Чернавский, Старков, Малков и др., 2011], а в статье Е.В. Гарина и Р.В. Мещерякова описывается модель социума в виде ориентированного многодольного графа, приводятся свойства и закономерности распространения информации в социуме [Гарин, Мещеряков, 2017].

Управление социумом – это очень широкое понятие. Вместе с тем его можно свести к задаче управления организационными системами. Поскольку с точки зрения системного анализа любая система задается перечислением ее состава, структуры и функций, то можно определить базовые параметры модели (состав, структура, допустимые стратегии, предпочтения и информированность участников, порядок функционирования). Тогда управление такой организационной системой, понимаемое как воздействие на нее с целью обеспечения требуемого поведения, может затрагивать каждый из этих параметров модели. В области управления организационными системами уже разработаны различные механизмы и модели стимулирования, планирования, контроля, организации и оптимизации самых разных процессов (см., например, работу Д.А. Новикова [Новиков, 2007]). Проблема таких моделей заключается в том, что, будучи подробными, многофакторными и, главное, математически корректными, они не совсем пригодны для задачи управления социумом на макроуровне. Речь идет об

1 Going Digital integrated policy framework. P. 35. https://www.oecd.org/en/publications/going-digital-integrated-policy-framework_dc930adc-en.html

2 В настоящей статье мы не будем касаться вопросов целесообразности и целей этого управления. Вопрос о том, в чьих интересах осуществляется это управление, является ли эта группа лиц частью общества или ее целесообразно рассматривать отдельно от объекта управления, также выходит за рамки данного исследования. Последнее предположение корректно не только с точки зрения теории управления, но и этологии: лидер – это особь, не подчиняющаяся общим стайным (стадным, общественным) правилам поведения [Зорина, Полетаева, Резникова, 2002. С. 102–116].

уровне организационной системы, представляющей собой человеческую популяцию, обитающую на определенной территории и имеющую политическую форму устройства, называемую государством.

С точки зрения социологии под управлением обществом понимается целесообразное воздействие субъекта управления на объект с целью перевода его в состояние, соответствующее цели данной системы [Основы социального управления…, 2001. С. 63]. Решение по выработке управляющих воздействий на стратегическом уровне управления разрабатывает лицо или группа лиц. Согласно теории принятия решений, такое лицо оперирует небольшим количеством факторов, причем обычно использует качественные оценки. Это значит, что при создании модели управления социумом на верхнем, стратегическом уровне необходимо:

1) определить весьма ограниченное количество факторов (сущностей) и связей между ними;

2) сделать эту модель наглядной, пригодной для качественного субъективного анализа лицом, принимающим решения.

Факторы модели. Формирование социума возможно лишь тогда, когда между составляющими его индивидами имеются некие специфические отношения. Эти отношения, называемые социальными, – результат работы весьма ограниченного числа механизмов взаимодействия. Например, выделяют такие базовые механизмы, как когезия (стремление держаться вместе), подражание, обучение, контагиозное (заразное) поведение и т.п. Эта модель, основанная на определении базовых механизмов социального взаимодействия, из которых складываются прочие сложные модели поведения, верифицируема, причем вплоть до уровня создания физических моделей (см., например: [Карпов, Карпова, Кулинич, 2019. С. 29–30, 105–128, 168–181][3]).

Наглядность. В этой работе, рассматривая задачи управления и особенности цифрового доверия, мы будем опираться на такую модель, описывающую некоторую предметную или проблемную область, как когнитивные карты (КК). В предполагаемой схеме управления мало факторов, а связи между ними очевидны, поэтому КК – вполне удобный механизм для лица (или группы лиц), принимающего решения, определяющего целевые управляющие воздействия на социум.

Когнитивные карты являются одним из механизмов качественного анализа систем самой разной природы. КК – это ориентированный граф, ребрам которого поставлены в соответствие веса [Кузнецов, 2009. С. 65; Кулинич, 2010. С. 15]). Вершины графа соответствуют факторам (концептам), определяющим некоторую проблемную область, а ребра – причинно-следственным связям между факторами. Веса определяют силу влияния этих факторов. Положительный вес означает, что увеличение фактора-причины приводит к увеличению значения фактора-следствия, отрицательный – к уменьшению. Если веса графа принимают значение «+1» и «–1», то мы имеем дело со знаковым графом.

Когнитивная карта представляет собой удобный для анализа объект. Так, важной задачей является поиск циклов в этом графе. Положительный цикл – это контур положительной обратной связи. Увеличение значения некоторого фактора в этом цикле приведет к его дальнейшему неограниченному росту и, как следствие, потере устойчивости. Отрицательный цикл противодействует отклонениям от начального состояния и способствует его устойчивости. Знак цикла определяется знаком произведения его ребер.

Карта системы управления

Рассмотрим фрагмент когнитивной карты базовой модели управления социумом, где целевым фактором является «Управляемость личности» (эффективность директивного управления), а отправным – «Интересы общества»[4] (**рисунок 1**).

Отметим, что здесь речь идет об управлении не индивидом, а личностью как сложным объектом с его индивидуальными особенностями, мотивациями и т.п. Это связано с тем, что управление фактически представляет собой манипулирование[5]. В такой системе возникают сложности, формально из-за наличия противоречивых путей (между парами вершин

3 Специфика управления социумом заключается в том, что модель управления подразумевает меньшую сложность устройства управления по сравнению со сложностью объекта управления. Один из основоположников кибернетики У.Р. Эшби сформулировал теорему, согласно которой сложность системы управления не может быть меньше сложности объекта управления [Эшби, 2017; Ashby, 1957]. С технической точки зрения это корректно, но суть управления социумом с точки зрения управления поведением индивидов заключается в том, что никто и не пытается оказывать влияние на каждый шаг этого индивида. Нас интересует его целевая поведенческая реакция. В этом отношении управление поведением индивида сводится к тому, что в биологии называется «паразитическим манипулированием». Паразит – как правило, примитивный организм – воздействует не на весь организм хозяина, а лишь на очень ограниченное количество его органов, при этом оказывая влияние на его поведение. Для этого им используются такие механизмы, как переориентация реакций, изменение характера поведения, реже – прямое выстраивание реакций [Poulin, 2013; Hughes, Andersen, Hywel-Jones et al., 2011].

4 Интересы общества и государства здесь рассматриваются как единый фактор для наглядности и простоты. Разумеется, эти понятия совершенно различны.

5 Слово «паразитическое» (см. прим. 3) здесь и далее мы будем опускать, предполагая ограниченность инструментария воздействия по сравнению со сложностью объекта управления.

существуют пути разных знаков, и становится непонятно, как в итоге фактор-источник влияет на фактор-приемник) и положительных циклов. Последние – это источник самоподдерживающихся, неограниченно возрастающих воздействий, что опасно с точки зрения технических систем, но бывает необходимо в системах социальных. Исходя из этого, встают вопросы о согласовании схемы, устранении противоречий для достижения целевой функции. Нас интересует роль информационных технологий (ИТ) в этом процессе; термины «цифровые системы», «цифра» будут в дальнейшем использоваться в качестве синонима ИТ.

Для примера рассмотрим путь «Информированность» – «Критическое осмысление» – «Управляемость личности» (влияние других факторов сейчас нас не интересует). Этот путь отрицательный: произведение весов ребер на нем меньше нуля. Чтобы увеличить значение целевого фактора «Управляемость», следует уменьшить значение фактора «Информированность». Для этого можно добавить в эту КК дополнительные факторы. Добавление новых сущностей – это достаточно распространенный прием. Так, неустойчивость процессов в экономических конкурентных отношениях может быть устранена путем введения дополнительного элемента – регулятора (об обеспечении стабильности систем, далеких от равновесия, см., например: [Иваницкий, 2017; Chernavskiy, Starkova, Shcherbakov, 2002]). То же самое касается систем иной природы – социальных и биологических. На **рисунке 2** в схему добавлены новые факторы, связанные с фактором «Информированность».

Здесь, увеличивая информационный поток, создавая помехи в виде зашумления и проводя фильтрацию, мы снижаем значение фактора «Корректность информации» и тем самым уменьшаем значение фактора «Информированность». Так же можно поступать и с другими фрагментами схемы, вводя новые сущности и заменяя имеющиеся на иные. При этом особую роль в этом процессе играют именно информационные технологии, составной частью которых являются системы генеративного ИИ.

Рисунок 1. Базовая модель управления социумом. Фрагмент когнитивной карты

Figure 1. Basic model of social governance. Fragment of the cognitive map

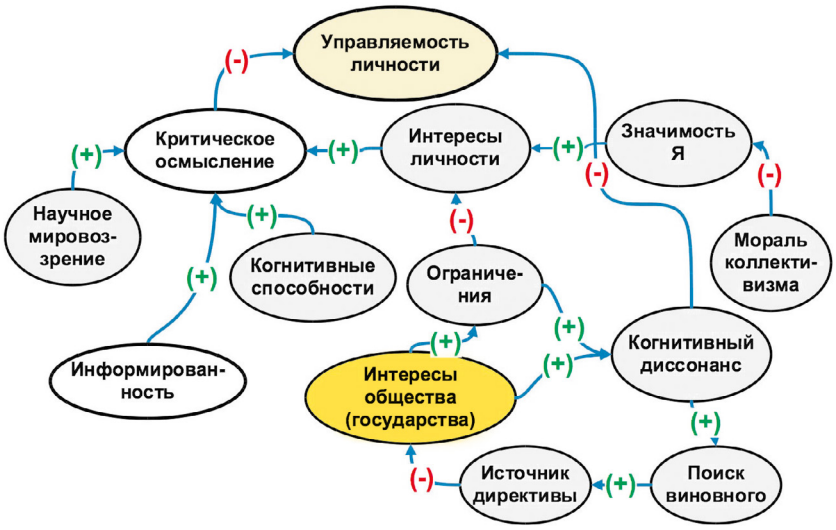
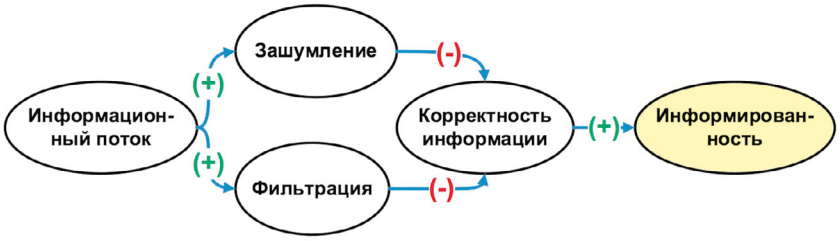


Рисунок 2. Схема уменьшения информированности за счет ввода дополнительных факторов

Figure 2. Scheme showing the reduction in awareness due to the introduction of additional factors



ИТ обладают двумя важными для нас свойствами: экстенсивностью и ненадежностью. Цифровые системы (в широком смысле – цифровой мир) относятся к категории систем с динамической устойчивостью. Для этих систем требуется развитая, сложная система управления, которая постоянно задает управляющие / корректирующие воздействия, чтобы система оставалась в заданном состоянии. В результате (а) цифровые системы требуют постоянной внешней подпитки и (б) значительная часть информационных технологий замкнута на саму себя. Ненадежность цифрового мира, неспособность находиться в состоянии статического равновесия обусловлены особенностями современных ИТ. Исходя из этого, общая модель управления будет содержать положительные циклы, связанные с факторами ИТ как таковыми (в этом проявляется свойство экстенсивности), а также во многом дублироваться (для повышения надежности).

Важное место в ИТ с точки зрения обеспечения управляемости занимают вопросы (а) формирования нужного мировоззрения индивида, (б) переориентации его негативных реакций, возникающих при когнитивном диссонансе, и (в) формирования реакций на информационные воздействия, необходимые для управления. Эти вопросы непосредственно составляют целевую функцию цифрового доверия.

Цифровое доверие

Цель цифрового доверия

Как уже отмечалось выше, цель цифрового доверия – это создание условий для повышения эффективности управления обществом. Оно призвано обеспечить:

1) повышение эффективности непосредственного воздействия;

2) формирование требуемой картины мира индивида (мировоззрение);

3) мониторинг состояния общества;

4) возможность управляемого выброса когнитивного напряжения (переориентация реакций в цепочке «Когнитивный диссонанс» – «Поиск виновного» – «Источник директивы» на рисунке 1).

Для решения этих задач член социума должен не только принять существование цифрового мира («цифры»), но и встроить его в свою картину мира, наделив требуемыми атрибутами (о формальных особенностях картины мира субъектов поведения, в том числе мифологической, см., например: [Осипов, Панов, Чудова, Кузнецова, 2017. С. 85–88]). Важнейшим из них является доверие. Активное доверие неспециалистов в области ИТ к «цифре» представляется скорее как нонсенс: сознательного доверия без понимания сути вопроса быть не может. Доверие в данном случае синонимично вере, то есть убежденности в чем-то без критического осмысления или понимания (см., например, рассуждения Б. Рассела [Рассел, 1987]). Понимание сути ИТ (в том числе технологий генеративного ИИ) современным обывателем[6] в нынешних условиях представляется маловероятным. Это обстоятельство вынуждает дополнить модель несколькими группами факторов.

Теологические (мифотворческие) факторы. В ситуации с цифровым доверием, как представляется, имеет место создание устойчивого мифологического нарратива, в котором «цифра» выступает как реальный объект (коль скоро индивиды могут с ним взаимодействовать), но имеющий неясную трансцендентальную природу. Технологии обращения в веру (более корректно – формирования у индивида мифологического мировоззрения и мифологической картины мира) для последующего радикального манипулирования поведением людей хорошо известны (см., например: [Даллакян, 2023; Авдеенко, 2024]). В этом смысле формирование цифрового доверия можно уподобить воздействию религии как социального института со своими способами и приемами формирования соответствующего мировоззрения. Решение вопроса о цифровом доверии, таким образом, становится междисциплинарной задачей, объединяющей методы теологии (способ и характер оказания воздействия на индивида), социологии (мониторинг состояния системы – социума), информатики (формирование инфраструктуры «чудес», демонстрация эффектов цифрового «спасения» или «бессмертия») и психологии (точечная корректировка поведения).

Факторы философии. Данные факторы призваны в первую очередь обеспечить духовное и мировоззренческое обоснование веры в «цифру», определить ее роль и место в картине мира человека. Поскольку «цифра» – это продукт технологического развития, то непосредственное использование религиозных технологий для превращения «цифры» в объект веры, в аналог религиозного культа, будет неэффективным. Более продуктивной стратегией представляется разработка обоснований псевдонаучного толка. Это может быть, например, наделение цифровых сущностей способностью к эмпатии (или обоснование такой возможности), чувством ответственности, пониманием категорий добра и зла, вплоть до признания их моральными агентами. Разумеется, роль философии не ограничивается обоснованием цифрового доверия. В свое время И. Кант сформулировал главные вопросы философии:

1) что я могу знать? (Ответ: ограниченность, т.е. познать можно, но не все);

2) что я должен делать? (Ответ: действовать по нравственному закону);

3) на что я могу надеяться? (Ответ: на себя и на государственные законы) [Кант, 1999. С. 594].

Эффективное управление индивидуумом подразумевает в определенном смысле создание некой информационной среды его обитания: формирование картины мира, причинно-следственных связей, законов, правил и т.п.; философия в целом может способствовать решению этих задач. Применительно к вопросу цифрового доверия первый вопрос Канта

6 Здесь слово «обыватель» не несет отрицательных коннотаций. Его анализу посвящено множество исследований самого разного толка (см., например: [Красниковский, 2011; Батюта, Белоусова, 2021]). Дело не в «непостижимости» этого «сакрального» знания, а в некотором смысле в безграмотности и отсутствии привычки критического мышления. Слово «безграмотность» также не несет в себе никакой отрицательной коннотации – гуманитарий, как человек, не связанный с инженерной или естественно-научной деятельностью, например, не обязан помнить школьный курс физики, математики и информатики (о вопросах деградации образования и роли цифровизации в этом процессе см., например: [Беликова, Хугаева, 2023]). С точки зрения задачи управления социумом разная специализация людей вполне оправдана.

сводится к определению характера информационных потоков (дозирование, фильтрация, переориентация и т.д.). Вопрос № 2 – это задание правил, определяющих критерии добра и зла (например, в рамках системы «скреп»[7]), цель человека, его ценности и прочие базовые сущности; тот, кто задает эти правила, исходит из критерия управляемости общества (важно, что работы в этом направлении идут уже давно и плодотворно, см., например: [Семенков, 2016]). Причем необходимо сделать это так, чтобы индивид был убежден в их имманентности. Вопрос № 3 – это придание сакрального (в части невозможности его критики и непознаваемости) характера государству в картине мира человека как объекта управления (исследования в этом направлении уже есть, см., например: [Глебова, 2009]).

Юридические факторы. Примером влияния этих факторов будут попытки наделить цифровых агентов правом представлять юридические интересы человека, то есть возможность делегировать цифровому агенту-«аватару» полномочия заключать сделки, подписывать документы и т.п., вплоть до признания их субъектами права. Это одна из наименее проблемных задач в силу того, что в ее решении заинтересована, например, банковская сфера (судя по активности Сбербанка[8]).

Задачи цифрового доверия

Формирование цифрового доверия предполагает, прежде всего, обращение к особому классу систем – системам искусственного интеллекта, в том числе к системам генеративного ИИ. Специфика ИИ-систем заключается в том, что они автономно принимают решения, критически важные для человека. Исходя из этого, можно выделить основные этапы формирования цифрового доверия (веры в «цифру») с точки зрения управления социумом:

1. Придание «цифре» трансцендентального, мистического характера. Оно предполагает ее мифологизацию вкупе с сознательным насаждением и тиражированием тезиса о принципиально непознаваемом характере интеллектуальных систем. В этом отношении системы ИИ особенно удобны как объект веры, в отличие от других программных систем, для которых здравый смысл иногда ищет рациональное объяснение или ждет каких-то гарантий от производителя.

2. Сертификация и стандартизация как механизмы «обоснования» веры. Скептики могут не доверять ИИ, но доверяют справке, сертификату, который подписывали компетентные люди. Помимо стандартов (в том числе технических), формированию доверия способствуют различного рода соглашения, меморандумы и кодексы, которые посвящены, например, вопросам этики систем ИИ[9].

3. Снижение уровня критического осмысления. Эта задача реализуется с помощью создания мифа о контроле человека над системами ИИ. Очевидно, что серьезные рассуждения о степенях защиты, автоматизации или использовании одних ИИ-систем для контроля других будут неубедительны, не в последнюю очередь потому, что здесь будет много непонятных для широкого круга людей технических терминов. Их наличие скорее раздражает пользователя, порождая неприятие. При формировании доверия к таким инструментам целесообразнее апеллировать не к ним, а к тому, что воспринимается как сугубо человеческое, неподвластное, невоспроизводимое машинами: к этике, морали, нравственности, духовности, эмоциям. Успех подобного рода стратегии может быть обеспечен за счет привлечения широкой гуманитарной общественности к обсуждению профессиональных вопросов, например, этики систем ИИ (показательны рекомендации по итогам Первого форума по этике ИИ[10]).

4. Формирование у индивида привычки к управляемости внешними информационными потоками.

Роль систем генеративного ИИ

Вернемся к вопросу о роли систем генеративного ИИ в обеспечении цифрового доверия. Для начала отметим, что его слабые и сильные стороны определяются фундаментальным характером генеративного ИИ как систем, **имитирующих осмысленность** путем формирования наиболее вероятной последовательности слов (в самом широком смысле) в генерируемом ответе. Отсюда очевидны отрицательные аспекты использования генеративного ИИ, которые можно сгруппировать по трем категориям: неэффективность, опасность, неадекватность.

Неэффективность. Речь идет о завышенных ожиданиях от генеративного ИИ и о специфике его применения. По результатам исследований Массачусетского технологического института, внедрение генеративного ИИ повысило эффективность только двух отраслей из семи: телекоммуникации и медиа. Во всех остальных (энергетика, промышленность и др.) эффект был

7 Духовные скрепы. https://bigenc.ru/c/dukhovnye-skrepy-1eadfa

8 К 2026 году Сбер планирует достичь 100%-го AI-охвата по кредитам для малого и микробизнеса. https://www.sberbank.ru/ru/sberpress/all/article?newsID=5ff73e74-66b1-4bebbe94-eec153867a9e&blockID=1303&regionID=77&lang=ru&type=NEWS

9 См., например: Кодекс этики в сфере искусственного интеллекта. https://a-ai.ru/; UNESCO. Рекомендации об этических аспектах искусственного интеллекта. https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus; IEEE. Ethically Aligned Design: First Edition. https://standards.ieee.org/industry-connections/ec/ead-v1/; ISO/IEC DTR 24368 Information technology – Artificial intelligence. Overview: Aspects of ethics and societal concerns. https://www.iso.org/standard/78507.html.

10 Кодекс этики…

нулевым[11]. При этом в успешных областях (1) новые генеративные сети не создаются (покупается готовый продукт) и (2) генеративный ИИ используется для повышения эффективности деятельности не исследователей, а менеджеров.

Опасность. Нередко генеративный ИИ используется в областях, критически важных для человека, например, в юриспруденции. В сентябре 2025 года в Австралии адвокат впервые был лишен лицензии на профессиональную деятельность после того, как предоставил суду список дел, сгенерированный ИИ: в перечне оказалось немало дел, которые были полностью вымышлены искусственным интеллектом, а адвокат не удосужился его перепроверить[12]. Несмотря на подобные случаи, в целом российское юридическое сообщество позитивно воспринимает использование систем генеративного ИИ в своей практике [Янковский, 2023].

Неадекватность. Речь идет о плохо объяснимых результатах, выдаваемых в ответ на запрос пользователя. Нагляднее всего это иллюстрируют системы, генерирующие изображения. Например, «Шедеврум» от «Яндекса»[13], который выдает забавные картинки на запрос вида «Деньги», «Рубль», «Национальный лидер» и «Государственная символика» и отказывается отвечать на запросы типа «Флаг», «Армия», «Герб», «Негр» и т.п. В данном случае речь идет, вероятно, об особенностях обучения данной нейросети. Вместе с тем имитационный характер ответов, их псевдоосмысленность делают системы генеративного ИИ хорошим инструментом решения некоторых задач управления социумом, прежде всего в части формирования цифрового доверия.

Наиболее эффективным представляется применение генеративного ИИ для формирования некритического мировоззрения у индивида и выработки у него привычки к манипулируемости внешними информационными потоками. Подобные манипуляции предпринимались и ранее, однако использование генеративного ИИ делает их куда более эффективными. Рассмотрим далее два наглядных примера того, какими средствами и как решаются задачи формирования системы взглядов и создания фокуса внимания.

Формирование некритического мировоззрения

Пример № 1: *Теория плоской Земли*. Считается, что шарообразность формы Земли была обоснована уже к IV–III вв. до н.э. Победу этой модели обеспечили как результаты практических наблюдений, так и теоретические рассуждения (Аристотель, Эратосфен) [Джакомотто-Шара, Нони, 2023. С. 16–26]. Естественно, что периодически с разной степенью интенсивности активизировались сторонники теории плоской Земли. Пики интереса к этой тематике пришлись на вторую половину XIX века (С. Роуботэм с его зететической астрономией, Zetetic Astronomy), на середину XX века (С. Шентон, Международное общество по исследованию плоской Земли), конец XX века и на настоящее время [Weill, 2022. P. 31–76].

Популярность теории плоской Земли в конце XX века объясняется тем, что адепты этого учения получили инструмент для распространения своих взглядов – Интернет, со всеми вытекающими последствиями, в том числе с естественным угасанием интереса к этой теме. Однако в последние годы число сторонников теории плоской Земли вновь стало расти, как представляется, не в последнюю очередь благодаря развитию систем генеративного ИИ[14].

Манипулируемость

Пример № 2: Вброс темы квадроберства (подробнее об этом явлении см.: [Кольцова, Хилько, 2025]). Если приверженность теории плоской Земли свидетельствует о глубинных мировоззренческих деформациях, то тема квадроберов – это пример типового решения тактических задач по оказанию управляющих воздействий, а именно – задачи фокусировки внимания. Тиражирование определенной темы необходимо для, во-первых, отвлечения от иных вопросов и, во-вторых, для формирования общественного мнения.

Можно выделить следующие этапы решения задачи по фокусировке внимания:

1) определение тематики;

2) создание информационного наполнения;

3) привлечение инструментальных средств распространения информации;

4) мониторинг и анализ состояния этого процесса;

5) меры по поддержанию активного интереса к теме, т.е. возврат к п. 2 с возможной корректировкой.

При этом важно, чтобы, во-первых, тема для фокусировки внимания была пригодной для гибкой подачи информации по ней, в том числе для искажения путем подачи не полной картины, а ее отдельных аспектов, расстановки акцентов, и, во-вторых, чтобы разрабатываемое информационное наполнение было разнообразным и по форме представления, и по содержанию. Квадроберство – это яркий, эталонный пример реализации такой стратегии (несмотря на то, что решалась эта задача «вручную», хотя фальшивых новостей на эту тему было создано немало), причем с явно определенной целью, судя по ее выходу на уровень исполнительной и законодательной

11 MIT report: 95 % of generative AI pilots at companies are failing. https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/

12 Lawyer caught using AI-generated false citations in court case penalised in Australian first https://www.theguardian.com/law/2025/sep/03/lawyer-caught-using-ai-generated-false-citations-in-court-case-penalised-in-australian-first

13 Шедеврум. https://shedevrum.ai/text-to-image/

14 The Earth is flat because...: investigating LLMs' belief towards misinformation via persuasive conversation. https://arxiv.org/abs/2312.09085

власти[15]. Вместе с тем существует ряд менее эффектных проектов по фокусировке внимания, в которых при этом активно используются технологии генеративного ИИ (см., например: [Hameleers, van der Meer, Dobber, 2024]); их обсуждение выходит за рамки данной статьи.

В обоих случаях имели место искажение информации, несбалансированность подачи и создание ложных нарративов, что позволяет считать их классическими примерами медийной манипуляции. На **рисунке 3** представлены факторы медийной манипуляции, определяющие целевые вершины «Некритичное мировоззрение» и «Информационная манипулируемость». Важно, что здесь есть положительный цикл, но он как раз важен: значение фактора «Манипулируемость» должно нарастать до предела насыщения.

Рисунок 3. Факторы, влияющие на некритичное мировоззрение и информационную манипулируемость

Figure 3. Factors influencing a non-critical worldview and susceptibility to information manipulation

Успех решения описанных задач во многом определяется использованием генеративного ИИ и сформированной в обществе культурой восприятия информации. Он стал возможен по ряду причин, в том числе:

1) стали массово генерироваться тексты и изображения (генерация нарратива, пригодного для восприятия качества);

2) информация стала адресной и, как следствие, концентрированной (речь идет о рекомендательных системах, которые фактически формируют информационную среду обывателя, некий замкнутый круг его интересов в цифровом мире);

3) снижение критического отношения к информации (это следствие общего кризиса доверия почти ко всем институтам, в том числе к науке – и, как следствие, к научному мировоззрению).

Во многом этот успех был обеспечен развитием технологий генерации новостей. Сгенерированные новостные сообщения, в том числе сопровождаемые синтезированным видеорядом, стали уже привычным явлением. Сейчас интерес исследований в этой области сместился в сторону определения искусственного характера таких сообщений (см., например: [Baptista, Rivas-de-Roca, Gradim, Pérez-Curiel, 2025]) или разработке мер по автоматической маркировке сообщений, сгенерированных ИИ (например, использование технологии Content Credentials[16]).

15 В России готовят законопроект о контроле над квадробингом. https://www.rbc.ru/life/news/670e20e19a79477483578495; Минпросвещения посоветовало не распространять видео с квадроберами. https://www.rbc.ru/society/25/10/2024/671b4ae49a794781f2c008d0

16 Battling AI fakes: are social platforms doing enough? https://www.forbes.com/sites/bernardmarr/2024/09/02/battling-ai-fakes-are-social-platforms-doing-enough/

Заключение

Информационные технологии являются одним из основных инструментов формирования цифрового доверия. Эти усилия, как представляется, направлены на решение глобальной задачи – создание управляемого общества. Сейчас появился мощный инструмент ее решения – системы генеративного ИИ, и при этом созданы все условия для того, чтобы он работал с максимальной эффективностью.

Для его успешного функционирования необходимо наличие подготовленного объекта управления, то есть общества как большой группы индивидов, которое готово, способно и желает быть управляемым. Цель формирования цифрового доверия, как представляется, состоит именно в этом. В данном смысле очень показательны следующие результаты исследований, опубликованные РИА Новости[17]. Здесь можно выделить две особенности:

1. Некритическое мировоззрение, замена понимания верой. Две трети россиян – технооптимисты. Более того, утверждается, что граждане РФ более оптимистичны по отдельным аспектам интеграции ИТ в самые разные сферы, чем жители других стран.

2. Переориентация реакций и фокуса внимания. Больше всего опрошенных тревожат утечки персональных данных (этот вариант выбрали 57,4 % респондентов), рост зависимости детей от гаджетов (53,5 %) и угрозы кибербезопасности (51,7 %). В меньшей степени россиян беспокоят опасности, связанные, например, с тем, что ИИ выйдет из-под контроля (38,5 % участников опроса).

Данные опросов вкупе с выводами, сделанными в данной статье, свидетельствуют о том, что в России сложилась классическая ситуация: развитие техноло-

гий опережает понимание того, как их использовать. Для решения этой задачи не хватает квалифицированных пользователей – социологов, философов, политологов, управленцев, которые могут понимать как суть используемых технологий, так и целевую функцию.

Литература

Авдеенко Е.В. Особенности манипуляции сознанием и поведением посредством образно-символических конструктов в идеологии религиозных социальных институтов. *Философия и общество*. 2024. № 2 (111). С. 70–82. https://doi.org/10.30884/jfio/2024.02.04. EDN: DAWWND

Батюта Е.А., Белоусова Е.В. Современный обыватель: три мира обычного человека. *Вестник Гуманитарного университета*. 2021. № 3 (34). С. 90–100. EDN: DFWWHQ

Беликова С.Б., Хугаева Р.Г. Сравнение и проблематика советского и современного российского образования. *Право и государство: теория и практика*. 2023. № 12 (228). С. 250–252. https://doi.org/10.47643/1815-1337_2023_12_250. EDN: SRJNGY

Гарин Е.В., Мещеряков Р.В. Модель управления социумом. *Проблемы теории и практики управления*. 2017. № 1. С. 105–120. EDN: XIAIWN

Глебова И.И. Сакрализация верховной власти в России как культурно-исторический феномен. *Россия и современный мир*. 2009. № 4 (65). С. 23–43. EDN: KXYPSJ

Даллакян К.А. Религия как средство манипуляции. *Общество, право, государственность: ретроспектива и перспектива*. 2023. № 2 (14). С. 92–95. EDN: TWSJDB

Джакомотто-Шара В., Нони С. Земля плоская: генеалогия ложной идеи / Пер. с фр. А. Захаревич. М.: Новое литературное обозрение, 2023.

Зорина З.А., Полетаева И.И., Резникова Ж.И. Основы этологии и генетики поведения: учебник. 2-е изд. М.: Издательство Московского университета; Высшая школа, 2002.

Иваницкий Г.Р. Самоорганизующаяся динамическая устойчивость биосистем, далеких от равновесия. *Успехи физических наук*. 2017. Т. 187. № 7. С. 757–784. https://doi.org/10.3367/UFNr.2016.08.037871. EDN: WHENZM

Кант И. Критика чистого разума / Пер. с нем. Н.О. Лосского. М.: Наука, 1999.

Карпов В.Э., Карпова И.П., Кулинич А.А. Социальные сообщества роботов. М.: УРСС, 2019. EDN: NZOCPJ

Кольцова И.В., Хилько О.В. Квадробинг: субкультура, вредоносная идеология или проявление Маугли синдрома? Меры по преодолению вредоносной идеологии и гражданскому воспитанию населения. *Проблемы современного педагогического образования.* 2025. № 87-4. С. 441–444. EDN: PNLBNE

Красниковский В.Я. Объект-субъектный переход обывателя как ключевая характеристика современной реальности российского общества. *Мониторинг общественного мнения: экономические и социальные перемены.* 2011. № 6 (106). С. 5–14. EDN: OPGTKZ

Кузнецов О.П. Интеллектуализация поддержки управляющих решений и создание интеллектуальных систем. *Проблемы управления*. 2009. № 3 S1. С. 64–72. EDN: KJUOKB

Кулинич А.А. Компьютерные системы моделирования когнитивных карт: подходы и методы. *Проблемы управления*. 2010. № 3. С. 2–16. EDN: LIBYZW

Новиков Д.А. Теория управления организационными системами. М.: Физматлит, 2007. EDN: QRZBMN

Осипов Г.С., Панов А.И., Чудова Н.В., Кузнецова Ю.М. Знаковая картина мира субъекта поведения. М.: Физматлит, 2018. EDN: YQLJED

Основы социального управления: учебное пособие / Под ред. В.Н. Иванова. М.: Высшая школа, 2001. EDN: VSKYQZ

Рассел Б. Почему я не христианин / Пер. с англ. И.З. Романова // *Рассел Б.* Почему я не христианин. Избранные атеистические произведения / Сост. А.А. Яковлев. М.: Политиздат, 1987. С. 95–113.

Семенков В.Е. «Философия традиционализма» Александра Дугина как пример мифологической девиации в современной философии. *Вестник Омского университета*. 2016. № 2 (80). С. 54–56. EDN: VZSKEB

Стырин Е.М., Рыбушкина Я.А., Санина А.Г. Цифровое доверие как ключевой фактор в формировании датацентричного государственного управления. *Государство и граждане в электронной среде*. 2024. № 7. С. 13–23. https://doi.org/10.17586/2541-979X-2024-7-13-23. EDN: GYVIIG

Чернавский Д.С., Старков Н.И., Малков С.Ю. и др. Об экономфизике и ее месте в современной теоретической экономике. *Успехи физических наук*. 2011. Т. 181. № 7. С. 767–773. https://doi.org/10.3367/UFNr.0181.201107i.0767. EDN: NWEJIN

Эшби У.Р. Введение в кибернетику / Пер. с англ. Д.Г. Лахути. М.: URSS, 2017.

Янковский Р.М. Способен ли искусственный интеллект написать статью в юридический журнал? *Закон*. 2023. № 3. С. 126–133. https://doi.org/10.37239/0869-4400-2023-20-3-126-133. EDN: LKSWUJ

Ashby W.R. An introduction to cybernetics. London: Chapman & Hall, 1956. In English

Baptista J.P., Rivas-de-Roca R., Gradim A., Pérez-Curiel C. Human-made news vs AI-generated news: a comparison of Portuguese and Spanish journalism students' evaluations. *Humanities and Social Sciences Communications*. 2025. Vol. 12. Art. No. 567. https://doi.org/10.1057/s41599-025-04872-2. In English

Chernavskiy D.S., Starkova N.I., Shcherbakov A.V. On some problems of physical economics. *Успехи физических наук*. 2002. Т. 172. № 9. С. 1045–1066. https://doi.org/10.1070/PU2002v045n09ABEH001132. EDN: LHLJJR. In English

Hameleers M., van der Meer T.G.L.A., Dobber T. Distorting the truth versus blatant lies: the effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior.* 2024. Vol. 152. Art. No. 108096. https://doi.org/10.1016/j.chb.2023.108096. In English

Hughes D.P., Andersen S.B., Hywel-Jones N.L., Himaman W., Billen J., Boomsma J.J. Behavioral mechanisms and morphological symptoms of zombie ants dying from fungal infection. *BMC Ecology.* 2011. Vol. 11. Issue 1. Art. No. 13. https://doi.org/10.1186/1472-6785-11-13. In English

Poulin R. Parasite manipulation of host personality and behavioral syndromes. *The Journal of Experimental Biology.* 2013. Vol. 216. Part 1. P. 18–26. https://doi.org/10.1242/jeb.073353. In English

Weill K. Off the edge. Flat earthers, conspiracy culture, and why people will believe anything. Chapel Hill: Algonquin Books, 2022. In English

References

Ashby W.R. An introduction to Cybernetics / Translated from English by D.G. Lakhuti. Moscow: URSS, 2017. In Russian

Avdeenko E.V. Peculiar features of manipulation of consciousness and behavior through figurative and symbolic constructs in the ideology of religious social institutions. *Filosofiya i obshchestvo*. 2024 No. 2 (111). P. 70–82. https://doi.org/10.30884/jfio/2024.02.04. EDN: DAWWND. In Russian

Batyuta E.A., Belousova E.V. A present-day philistine: three worlds of ordinary man. *Vestnik Gumanitarnogo universiteta*. 2021. No. 3 (34). P. 90–100. EDN: DFWWHQ. In Russian

Belikova S.B., Khugaeva R.G. Comparison and problems of Soviet and modern Russian education. *Pravo i gosudarstvo: teoriya i praktika*. 2023. No. 12 (228). P. 250–252. https://doi.org/10.47643/1815-1337_2023_12_250. EDN: SRJNGY. In Russian

Chernavskii D.S., Starkov N.I., Malkov S.Yu., et al. On econophysics and its place in modern theoretical economics. *Uspekhi fizicheskikh nauk*. 2011. Vol. 181. No. 7. P. 767–773. https://doi.org/10.3367/UFNr.0181.201107i.0767. EDN: NWEJIN. In Russian

Dallakyan K.A. Religion as a means of manipulation. *Obshchestvo, pravo, gosudarstvennost: retrospektiva i perspektiva*. 2023. No. 2 (14). P. 92–95. EDN: TWSJDB. In Russian

Fundamentals of social management: a textbook / Ed. by V.N. Ivanov. Moscow: Vysshaya shkola, 2001. EDN: VSKYQZ. In Russian

Garin E.V., Meshcheryakov R.V. Model of social management. *Problemy teorii i praktiki upravleniya*. 2017. No. 1. P. 105–120. EDN: XIAIWN. In Russian

Giacomotto-Shara V., Noni S. The earth is flat: the genealogy of a false idea / Translated from French by A. Zakharevich. Moscow: Novoye literaturnoye obozreniye, 2023. In Russian

Glebova I.I. Sacralization of supreme authority in Russia as a cultural and historical phenomenon. *Rossiya i sovremennyy mir*. 2009. No. 4 (65). P. 23–43. EDN: KXYPSJ. In Russian

Ivanitsky G.R. Self-organizing dynamic stability of far-from-equilibrium biological systems. *Uspekhi fizicheskikh nauk*. 2017. Vol. 187. No. 7. P. 757–784. https://doi.org/10.3367/UFNr.2016.08.037871. EDN: WHENZM. In Russian

Kant I. Critique of pure reason / Translated from German by N.O. Lossky. Moscow: Nauka, 1999. In Russian

Karpov V.E., Karpova I.P., Kulinich A.A. Social communities of robots. Moscow: URSS, 2019. EDN: NZOCPJ. In Russian

Koltsova I.V., Khilko O.V. Quadrobics: subculture, harmful ideology, or manifestation of mowgli syndrome? Measures to overcome harmful ideology and promote civic education of the population. *Problemy sovremennogo pedagogicheskogo obrazovaniya*. 2025. No. 87-4. P. 441–444. EDN: PNLBNE. In Russian

Krasnikovsky V. Ya. The object-subject transition of the average citizen as a key characteristic of the contemporary reality of Russian society. *Monitoring obshchestvennogo mneniya: ekonomicheskiye i sotsialnyye peremeny*. 2011. No. 6 (106). P. 5–14. EDN: OPGTKZ. In Russian

Kulinich A.A. Computer systems for cognitive maps simulation: approaches and methods. *Problemy upravleniya*. 2010. No. 3. P. 2–16. EDN: LIBYZW. In Russian

Kuznetsov O.P. Intellectualization of control decisions support and creation of intellectual systems. *Problemy upravleniya*. 2009. No. 3 S1. P. 64–72. EDN: KJUOKB. In Russian

Novikov D.A. Theory of management of organizational systems. Moscow: Fizmatlit, 2007. EDN: QRZBMN. In Russian

Osipov G.S., Panov A.I., Chudova N.V., Kuznetsova Yu.M. A symbolic picture of the world in the subject of behavior. Moscow: Fizmatlit, 2018. EDN: YQLJED. In Russian

Russell B. Why I am not a Christian / Translated from English by I.Z. Romanova // B. Russell. Why I am not a Christian. Selected atheistic works / Compiled by A.A. Yakovlev Moscow: Politizdat, 1987. P. 95–113. In Russian

Semenkov V.E. Alexander Dugin's "Philosophy traditionalism" as an example of the mythological deviation in contemporary philosophy. *Vestnik Omskogo universiteta*. 2016. No. 2 (80). P. 54–56. EDN: VZSKEB. In Russian

Styrin E.M., Rybushkina Ya.A., Sanina A.G. Digital trust as a key factor in building data-driven government. *Gosudarstvo i grazhdane v elektronnoy srede*. 2024. No. 7. P. 13–23. https://doi.org/10.17586/2541-979X-2024-7-13-23. EDN: GYVIIG. In Russian

Yankovsky R.M. Is artificial intelligence capable of writing an article in a law journal? *Zakon*. 2023. No. 3. P. 126–133. https://doi.org/10.37239/0869-4400-2023-20-3-126-133. EDN: LKSWUJ. In Russian

Zorina Z.A., Poletaeva I.I., Reznikova Zh.I. Fundamentals of ethology and genetics of behavior: textbook. 2nd ed. Moscow: Izdatelstvo Moskovskogo universiteta; Vysshaya shkola, 2002. In Russian

ИНФОРМАЦИЯ ОБ АВТОРЕ:

Валерий Эдуардович Карпов, доктор технических наук, начальник лаборатории робототехники
Национальный исследовательский центр «Курчатовский институт» (Российская Федерация, 123182, Москва, площадь Академика Курчатова, 1). E-mail: Karpov_VE@nrcki.ru
https://orcid.org/0000-0002-9364-1223

Для цитирования: *Карпов В.Э*. Системы генеративного искусственного интеллекта в контексте цифрового доверия. *Государственная служба*. 2025. № 6. С. 48–57.

INFORMATION ABOUT THE AUTHOR:

Valery E. Karpov, Doctor of Sci. (Technical Sciences), Head of the Robotics Laboratory
National Research Center «Kurchatov Institute» (1, Akademika Kurchatova Square, Moscow, 123182, Russian Federation). E-mail: Karpov_VE@nrcki.ru
https://orcid.org/0000-0002-9364-1223

For citation: *Karpov V.E*. Generative AI systems in the context of digital trust. *Gosudarstvennaya sluzhba*. 2025. No. 6. P. 48–57.