Институт психологии Российской академии наук. Организационная психология и психология труда. 2024. Т.9. № 4.
С.Ф. Сергеев, З.А. Руденко Проблема «темных паттернов» и синтетических данных в системах с искусственным интеллектом. DOI: 10.38098/ipran.opwp_2024_33_4_007

ПСИХОЛОГИЯ И СОВРЕМЕННЫЕ ТЕХНОЛОГИИ

УДК 159.9

ГРНТИ 15.41.21

ПРОБЛЕМА «ТЕМНЫХ ПАТТЕРНОВ» И СИНТЕТИЧЕСКИХ ДАННЫХ В СИСТЕМАХ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

© 2024 г. С.Ф. Сергеев*, З.А. Руденко**

**Доктор психологических наук, профессор,*
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
e-mail: s.f.sergeev@spbu.ru

***Бакалавр, Санкт-Петербургский государственный университет,*
Санкт-Петербург, Россия
e-mail: ruden.zakh@gmail.com

Наблюдаемый в настоящее время прогресс в области машинного обучения и развития технологий больших языковых моделей (LLM) и генеративного искусственного интеллекта (ГИИ) ведет к интенсивному внедрению во все сферы человеческой деятельности генеративных нейронных сетей, обученных на созданных человечеством текстовых, графических и аудиовизуальных данных. Возникают проблемы намеренного или случайного их искажения с целью влияния на поведение и принятие решений людьми, что отражается в проблеме «темных паттернов», используемых дизайнерами и маркетологами в целях обмана пользователя, манипулирования его вниманием и поведением, направленные на увеличение прибыли. Логично предположить, что те же уловки, что ранее создавались умелыми маркетологами, дизайнерами, политиками и психологами могут проявляться намеренно или случайно и в продуктах, сгенерированных искусственным интеллектом, что может негативно повлиять на людей, использующих данные технологии, вызывая в них неожиданные реакции, эмоции и поведение. Важно учитывать данную особенность работы с нейронными сетями, изучать её, а также искать пути по минимизации её проявления при генерации ответов для пользователей. Рассматриваются вопросы и особенности манипулирования пользователями и социальными сообществами с использованием синтетических данных в форме дипфейков. Проведен содержательный анализ понятия «темный паттерн», рассматриваются вопросы

классификации темных паттернов и их влияние на пользователя. Появление технологий ГИИ усиливает потенциальные риски и возможности массового негативного информационного влияния на человека и его психику, что требует от инженерных психологов и разработчиков информационного содержания техносреды применения особых методов проектирования, учитывающих и блокирующих данные угрозы.

Ключевые слова: дипфейк, защита данных, искусственный интеллект, несанкционированное информационное воздействие, синтетические данные, темный паттерн, пользовательский опыт, этичный дизайн.

ВВЕДЕНИЕ

Понятие «темный паттерн» (Dark Patterns) берет свое происхождение из сферы инженерной психологии и эргономики, дизайна интерфейсов и формирования пользовательского опыта. К нему относят различные уловки и приемы, используемые дизайнерами и маркетологами в целях обмана пользователя, манипулирования его вниманием и желаниями, направленные на увеличение прибыли или же метрических данных ресурса (посещения страницы, количество кликов, лайков, репостов) (Brignull, 2010). Автор данного термина UX-дизайнер Гарри Бригналл описал его, как «пользовательский интерфейс, аккуратно созданный для подталкивания пользователей к определенным действиям, как например приобретение страховки при совершений какой-либо покупки или оформлении подписки на получение счетов» (Brignull, et. all, 2023). Отметим, что темные паттерны прямо противоречат интересам пользователя и наносят ему моральный и материальный ущерб.

С развитием технологий техногенных информационных сред, глобальных коммуникационных технологий, мобильной связи и интернета, полимодальных систем интерфейса, внедрением технологий искусственного интеллекта, количество манипуляционных технологий, влияющих на человека, в том числе «темных паттернов» резко увеличивается, а какие-то их формы существуют практически в любых интерфейсах, отражая многообразие и несовершенство человеко-машинных связей (Сергеев, 2014). Для ограничения деструктивного влияния данных технологий в США и Европе стали приниматься особые законы, регулирующие использование и применение

приемов технологического и информационного манипулирования пользователями вплоть до административного и уголовного преследования в случае их нарушения. Например, в Европейском Союзе действует Общий регламент по защите данных (GDPR), требующий от компаний получать добровольное согласие на использование персональных данных в различных рекламных и иных действиях, содержащих манипуляцию (Voigt, Bussche, 2017).

Однако нельзя отрицать, что приемы, используемые в темных паттернах, появились задолго до интернета. Их можно найти в журналистике, политике, рекламе и конечно бизнесе. И это логично, ведь для подобных сфер все зависит от того, как много людей будут выбирать их материал, а не их конкурентов. Крайне важно завладеть интересом и желанием публики, из-за чего некоторые корпорации прибегают к не самым этичным приемам. Определенные знания о восприятии человеком информации позволяют направить мнение и намерения конкретного индивида в необходимом направлении (Thaler, Sunstein, 2010).

Обучение нейронных сетей происходит на базе огромного количества информации, созданной и накопленной человечеством за все время своего существования. В таких больших данных однозначно присутствуют как положительные, так и негативные проявления манипуляционных техник, а также лингвистических темных паттернов. (Carlini, 2021). Таким образом, становится ясно что нейронные сети могут выдавать ответы с использованием подобных конструкций, так как невозможно отрицать их наличие в человеческой культуре и языке. Иными словами, искусственный интеллект может начать «неосознанно» манипулировать человеком или же вводить его в заблуждение используя особенности контента, на котором проводилось обучение языковой генеративной модели.

Люди склонны к тому, чтобы верить в информацию, предоставляемую им от какого-либо референтного лица, авторитета. Как показывают исследования, это проявляется и в случае, если в качестве такого лица выступает искусственный интеллект

(Sergeeva, Sergeeva, 2023). Проблема доверия искусственному интеллекту определяет манипуляционные возможности последнего. Кроме того, доверие к ИИ модулируется личностными свойствами пользователя. Показано, что «влияние доверия / недоверия оператора технике, в том числе, к автоматизированным системам управления и автоматике на надежность и эффективность операторской деятельности, необходимо рассматривать в сочетании с доверием / недоверием оператора к себе как профессионалу» (Обознов, Акимова, Рунец, 2021, с. 5). Это говорит о высоком уровне ответственности разработчика системы ИИ по отношению к содержанию используемых при обучении ИИ текстов с целью защиты пользователей от скрытого манипуляционного действия.

Использование темных паттернов в дизайне считается неэтичным и иногда даже нелегальным. А проявление неэтичности со стороны искусственного интеллекта — это отдельная проблема, имеющая потенциал к социальному резонансу. Учитывая то, какое влияние могут оказывать небольшие особенности визуального оформления или текстовых формулировок, а также то, что они могут проявляться в ответах, сгенерированных искусственным интеллектом, необходимо задуматься над тем, как обезопасить пользователя от лишнего случайного информационного или эмоционального воздействия. Для этого необходимо разобраться с тем, как появились и развивались темные паттерны, где их истоки, а также посмотреть и проанализировать реальные примеры несанкционированного психоэмоционального воздействия на пользователя со стороны искусственного интеллекта и в целом проблемы возможные в данном контексте.

СОДЕРЖАТЕЛЬНЫЙ АНАЛИЗ ТЕОРЕТИЧЕСКОГО КОНЦЕПТА «ТЕМНЫЙ ПАТТЕРН»

Принципы, на которых компании пытаются заработать — это весьма базовые и многим знакомые психологические уловки. С развитием интернет-маркетинга они были адаптированы и появились на вебсайтах. Неважно будь это сайт туристического агентства, соцсети, картинной галереи там найдется как-то темный паттерн. И все же они

появились не из воздуха. Раз принципы и методы манипуляций остаются прежними, то и у темных паттернов можно проследить прародителей.

Ричард Талер – американский экономист и лауреат Нобелевской премии по экономике называет данные приемы «подталкиванием», описывая их как аспекты архитектуры выбора, направляющие поведение человека к определённому выбору, без каких-либо ограничений его свобод (Thaler, Sunstein, 2009).

Такими методами также пользуются и в политике. Например, существует стратегия демагогического популизма, в которой манипулятивные приемы используются для привлечения поддержки широких масс населения. Подобное использование нельзя назвать абсолютно этичным, так как зачастую в их основе лежат эмоциональные и упрощенные аргументы, вместо рационального восприятия (Jones, 2020). В таблице 1 представлены наиболее разработанные варианты понятия «темный паттерн».

Таблица 1

Популярные определения и варианты понятия «темный паттерн»

Автор или организация	**Название**	**Содержание**
California Privacy Rights Act, Colorado Privacy Act	Темные паттерны	Интерфейс, спроектированный с целью подрыва или ограничения автономии пользователя, свободы принятия решений или выбора
EU: Digital Services Act	Темные шаблоны	Практика намеренного искажения или ослабления способности получателя услуг осуществлять автономный, осознанный выбор. Используется для склонения получателей услуги к нежелательному поведению и принятию решений, которые имеют для них негативные последствия
Deceptive. design (Harry Brignull)	Обманчивые шаблоны	Уловки, используемые на веб-сайтах и в приложениях, которые заставляют пользователей делать то, чего они не хотели, например, покупать или подписываться на что-либо
Decepticon Luxembourg National Research Fund	Темные схемы	Манипуляции, направленные на онлайн решения пользователей относительно раскрытия своих персональных данных, совершения покупок и использования своего времени

Помимо политики и экономики на манипуляционных методах строится вся реклама и маркетинг. Их появление связано с развитием в середине XIX века средств массовой информации и печати. В рекламную индустрию стали привлекать популярных лиц и звёзд, что является апеллированием к авторитету. Изменилась упаковка товаров, которая стала нести с собой намек на высокое качество товара. Уже тогда стало ясно, что реклама позволяет компаниям не только увеличить объемы продаж, но и привлечь новых пользователей и решать неочевидные задачи управления. Одним из примеров неоднозначной рекламы, содержащей манипуляцию, можно считать рекламную кампанию Джона Пауэрса 1890-х годов. Он напрямую заявил о том, что бизнес, заказавший у него рекламу, находится на грани банкротства (Тангейт, 2007). С одной стороны это прямолинейно и честно по отношению к покупателю. С другой же стороны это некое давление и попытка вызвать чувство жалости. В современном дизайне попытки вызвать чувство жалости у пользователя является неэтичными, но человечество продолжает сталкиваться с такими приемами в рекламе различных благотворительных фондов и обществ, их даже можно назвать классикой.

Интересно, что проявления приемов манипулирования сознанием встречаются даже в простом расположении объектов и организации рабочих и торговых территорий. Примером может служить практически любой международный аэропорт. Как только пассажир проходит зону таможенного контроля и все ключевые этапы оформления пройдены, он обязательно попадет в беспошлинный магазин, наполненный в большинстве своем дорогими и элитарными товарами. Маршрут в магазине построен таким образом, чтобы посетитель обязательно прошел мимо каждого отдела. В более положительном сценарии, после прохождения оформления у пассажира остается какое-то свободное время.

В такой ситуации посещение магазина уже не кажется простым времяпровождением. Человек проникается атмосферой, чувствует некое прикосновение к элитарности, избранности, а также не видит конкуренции и вариантов альтернативного

выбора в условиях аэропорта, из-за чего может совершить покупку, которую в других условиях он бы не совершил (Brignull, 2023). Таким приемом пользуются все продуктовые супермаркеты и торговые площадки. Шведский магазин мебели «IKEA», сначала предоставляет покупателю возможность проникнуться красотой дизайнерских решений в виде комнат-примеров, продвигает уже уставшего человека к ресторану, а в конце к складу, где можно увидеть заново большинство товаров, и снова подумать над покупкой. Все это происходит в условиях сложной архитектуры, где человеку придется либо пройти целиком весь лабиринт, либо потратить усилия в навигации и поиске легкого пути (Кононович, Савенкова, Целик, 2022).

Данные примеры позволяют сделать вывод, что темные паттерны не есть что-то принципиально новое, а скорее являются переосмыслением и адаптацией существующих приемов манипулирования к сферам веб-дизайна и интернет-маркетинга. Внедрение манипулятивных приемов в цифровые сервисы и технологии привело к серьезным негативным последствиям, требующим правового и юридического регулирования.

Пользователи отличаются по уровню осведомленности о темных паттернах, своими индивидуальными особенностями по восприятию и обработке информации, опыту использования сети интернет. Для более продвинутого пользователя большинство маркетинговых уловок уже знакомо, он сможет правильно определить, где находится нужный крестик для закрытия всплывающего окна, где нужно подождать пару секунд, а где просто происходит манипулирование его чувствами, допустим при отказе от подписки. Однако другие люди с меньшим опытом легко становятся жертвами подобных уловок и испытывают на себе негативный опыт пользования веб-ресурсами. Так под угрозой оказывается их конфиденциальность, свобода выбора и безопасность личных данных, что требует внедрения этического подхода к проектированию и разработке технологий (Valoggia, Sergeeva, 2024).

На данный момент общепринятой классификации темных паттернов нет. Тем не менее в работе Liming Nie с соавторами представлена структура, названная Dark Pattern

Analysis Framework (DPAF). На ее основе создана всеобъемлющая таксономия темных паттернов, охватывающая 64 типа, каждый из которых обозначен своим влиянием на пользователей и вероятными сценариями, в которых он проявляется (Nie, et. all, 2024). Пока она находится на стадии исследований.

На практике чаще всего темные паттерны разделяют по:

- особенностям интерфейсных решений;
- используемому приему обмана;
- способу воздействия на пользователя;
- вовлеченности пользователя и его роли;
- причиняемому вреду и пользе (Mathur, Mayer, Kshirsagar, 2021).

Наиболее распространенное решение классификации связано с выявлением основного способа воздействия на пользователя с дальнейшим разбиением на подкатегории. Например, можно разделить методы манипуляции на пять типов: вынужденное действие (необходимость приглашать друзей или делится ссылкой, чтобы получить выгоду), назойливое воздействие (автоматический переход по ссылке, при закрытии всплывающего окна), сокрытие (добавление в корзину лишних товаров, сокрытие реальной цены), запутанный интерфейс («нет/отказаться» написанное на зеленом фоне, необходимая кнопка блеклая и кажется неактивной), создание препятствий (сложно отказаться от подписки, отсутствие возможности копировать текст, например название продукта, с сайта и т. д) (Fansher, 2018; Geronimo, Braz, 2020).

По характеру влияния на пользователя темные паттерны делят на ассиметричные (несопоставимость размера и удобства элемента с его важностью), скрывающие (незаметная подписка на рассылку или покупка дополнительных услуг), обманывающие (ложная временная скидка, обман об ограниченности количества), и запрещающие (регистрация только через соцсети или только при предоставлении информации о банковской карте) (Mathur, Acar, 2019).

Популярна классификация, основанная на конкретных концептуальных особенностях паттернов. В нее включены следующие виды темных паттернов:

- Предотвращение сравнения – функции и цены сочетаются сложным образом, что мешает пользователю найти важную информацию.
- Конфирмшейминг – манипулирование эмоциями пользователя, принуждение его к выполнению действий, которые он в противном случае не сделал бы.
- Замаскированная реклама – пользователь ошибочно полагает, что кликает по элементу интерфейса, но на самом деле это замаскированная реклама.
- Фиктивный дефицит – пользователь вынужден выполнить действие, потому что ему дается ложное указание на ограниченное предложение товара.
- Поддельное социальное одобрение – пользователя вводят в заблуждение, заставляя думать, что продукт более востребован, нежели на самом деле, показываются поддельные отзывы и сообщения.
- Ложная срочность – пользователя вынуждают выполнить действие, предоставляя фальшивое ограничение по времени.
- Вынужденное действие – от пользователя требуется сделать что-то другое, нежелательное действие, нежели то которое он хочет сделать.
- Сложная отмена действия – пользователю легко зарегистрироваться или оформить подписку на что-либо, но, отменить подписку очень сложно.
- Скрытые расходы – пользователя заманивают низкой ценой. Потратив время и усилия, он вдруг обнаруживает неожиданные сборы и платежи.
- Скрытая подписка – пользователь без его согласия зарегистрирован в регулярной подписке или плане платежей.
- Назойливое воздействие – действие пользователя настойчиво прерывают просьбами сделать что-то еще, что может быть и не в его интересах.
- Создание препятствий – пользователь сталкивается с барьерами или препятствиями, затрудняющими выполнение задачи или доступ к информации.

- Предварительный выбор – пользователю предоставляется вариант, который уже был выбран для него, чтобы повлиять на его принятие решения.
- Сокрытие информации – пользователь оказывается втянутым в сделку под ложным предлогом, потому что соответствующая информация скрывается или предоставляется ему с задержкой.
- Формулировки с подвохом – пользователь вводится в заблуждение и совершает какое-либо действие из-за вводящего в заблуждение текста.
- Визуально запутанный интерфейс – пользователь ожидает увидеть на странице информацию, в понятном и предсказуемом виде, но она скрыта, затемнена или замаскирована (Brignull, Leiser, Santos, Doshi, 2023).

Некоторые примеры темных паттернов сложно отнести в той или иной категории, например бесконечную новостную ленту в соцсетях. Такой объект как бесконечная лента действует косвенно, человека не принуждают, напрямую ничего не скрывают. В бесконечных лентах пользователя стимулирует к дальнейшему просмотру постоянное чувство новизны и чередование значимого и обычного контента, что вызывает чувство азарта и желание найти что-либо действительно интересное (Сергеев, Поварова, 2021).

Развитие интерфейсов интеллектуальных помощников, средств объяснения и помощи дают широкое поле для деструктивной деятельности создателей алгоритмов манипуляции пользователями (таблица 2).

Таблица 2

Примеры темных паттернов в формулировках объяснений в интерфейсе средств объяснения. Примеры основаны на классификации, проведенной Греем и др. [Gray C.M. et. all, 2018]

Темный паттерн	Переход к объяснимости и контролю	Примеры объяснительных формулировок	Примеры интерфейсов средств объяснения
Замечание: «*перенаправление ожидаемой функциональности, которая может*	Прерывать стремление пользователей к объяснениям и контролю	Ограниченный диалог	Скрытое взаимодействие

сохраняться в течение одного или нескольких взаимодействий»			
Препятствие: «усложнение процесса больше, чем это необходимо, с целью отговорить от определенных действий»	Заставьте пользователей избегать попыток найти и понять объяснение при взаимодействии с объясняющими или управляющими средствами	Информационная перегрузка, Нечеткая расстановка приоритетов	Скрытый доступ, Вложенные детали, Затрудненный выбор
Скрытность: «*попытка скрыть, замаскировать или отсрочить разглашение информации, имеющей отношение к пользователю»*	Извлекайте выгоду из взаимодействия пользователя с объяснениями/ средства управления с помощью скрытых функций	Объяснение Маркетинга	Опросы с объяснениями
Вмешательство в интерфейс: «*манипулирование пользовательским интерфейсом, которое дает преимущество определенным действиям над другими»*	Поощряйте понятность или контролируйте настройки, которые предпочитает поставщик системы	Неблагоприятный дефолт	Конкурирующие элементы, Ограниченный обзор
Принудительное действие: «т*ребующее от пользователя выполнения определенного действия для доступа к […] определенной функциональности»*	Заставьте пользователей выполнить какое-либо действие, прежде чем предоставлять им полезные объяснения или параметры управления	Принудительный доступ к данным, «Око за око»	Принудительное увольнение

ТЕМНЫЕ ПАТТЕРНЫ И МАНИПУЛЯЦИИ, ПРОЯВЛЯЮЩИЕСЯ В ОТВЕТАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

На данный момент искусственный интеллект применяется в разных областях народного хозяйства и человеческой деятельности (Сергеев, 2020).

При помощи ИИ пишут тексты, заголовки, создают презентации, дизайны, редактируют фото и видео. Один из видов использования ИИ, который уже сейчас повсеместно используется и стал многим пользователям знаком это чат-боты. Чат-бот

представляет из себя диалоговый пользовательский интерфейс на базе нейронных сетей, в котором ИИ проводит беседу с пользователем, выполняя поставленную задачу, выполняет навигационную или развлекательную функцию. Пользователю могут предложить как полностью самостоятельно писать запрос для анализа искусственным интеллектом, так и предложить готовые варианты в виде кнопок. Такие чат-боты можно встретить в том числе и на государственных платформах, в банковских приложениях и т. д. Чат-боты на базе ИИ могут осуществлять манипуляционную деятельность в том числе и с негативными для пользователей результатами искажая информацию, факты, навязывая ложные взгляды и оценки.

В связи с чем аспект этичности и точности выдаваемого результата может играть ключевую роль для конечного пользователя, когда дело касается документооборота, навигации между государственными службами или денежных транзакций.

На данный момент были проведены различные исследования, показывающие, что ИИ может прибегать к манипуляциям, если в ходе своего обучения посчитает эту стратегию выгодной (Mizuta, 2021). Это может быть опасно не только для пользователей, но и для рынка, экономики и общества в целом.

Помимо этого, другие исследования показывают, что ИИ подвержен манипуляциям со стороны пользователей. От эмоциональности и общего негативного или положительного посыла пользователя будет зависеть результат генерации, и таким образом, можно сподвигнуть нейронную сеть к генерации заведомо ложной или неэтичной, несущей вред обществу, информации (Vinay, Spitale, 2024). Показана уязвимость больших языковых моделей со стороны злоумышленников, которые могут извлечь из них обучающие данные, в том числе включающие информацию ограниченного доступа (Carlini, 2020).

Существуют прецеденты использования ИИ для влияния на результаты выборов, негативного влияния на ментальное здоровье пользователей, а также анализа информации ИИ про конкретного пользователя для более точного и «агрессивного»

использования рекламы против него, чтобы увеличить вероятность того, что пользователь купит продаваемый продукт (учитываются предпочтения, особенности эмоционального фона, подбирается наиболее уязвимый временной промежуток и происходит общий анализ ситуации). Это непременно ведет к этическим проблемам и требует юридического регулирования использования ИИ в целях такого анализа (Uuk, 2022).

Среди примеров темных паттернов и неэтичных проявлений со стороны диалоговых пользовательских интерфейсов можно отметить нежелательные действия, когда бот предлагает действие или контент при этом не давая альтернативы. Со стороны пользователей поступали жалобы, что бот начинал вести диалог в неподходящие время, например когда человек спал. Другим негативным проявлением можно считать переход чат бота на другой стиль ведения беседы в неуместном для этого контексте, т. е. с формального на неформальный или даже к грубому обращению с пользователем (Folstad, Brandtzaeg 2020).

В исследовании, проведенном в Технологическом университете Хемница (Германия) совместно с Национальным университетом Ла-Платы (Аргентина), изучались проявления типичных темных паттернов в графических пользовательских интерфейсах и диалоговых пользовательских интерфейсах на базе специально созданного датасета. В небольшом количестве полученной обратной связи (69 жалоб) были найдены случаи использования темных паттернов искусственным интеллектом:

- Среди назойливого воздействия было выявлено проявление чат-бота в виде всплывающего окна, время от времени отвлекающего пользователя своим периодическим появлением.
- Конфирмшейминг был представлен в виде эмоционального воздействия на пользователя при отмене подписки за счет формулировки вызывающей чувство вины у человека, а также за счет статистики, убеждающей пользователя в полезности и важности компании.

- Получены примеры сложной отмены действий. Чат-бот после запроса пользователя на отмену подписки просто переводил человека на другие контакты.
- Создание препятствий обычно проявляется в виде сложного перехода от чат-бота к живому оператору.
- Отмечены и другие проявления неэтичного дизайна, как например отсутствие оставшегося времени (темный паттерн – отсутствие индикатора загрузки) или неоднозначные формулировки (Traubinger, Heil, Grigera, Garrido, Gaedke, 2023).

Компании, использующие обманные схемы, часто оказываются втянутыми в судебные дела и получают большие штрафы и пени. Например, фирма Epic Games заплатила 245 миллионов долларов, чтобы урегулировать обвинения в использовании обманных схем в платежной системе Fornite.

Диетическое приложение Noom заплатило 62 миллиона долларов, чтобы урегулировать обвинения в использовании вводящих в заблуждение шаблонов при оформлении подписок и автоматическом продлении подписок.

Компания AT&T заплатила 105 миллионов долларов, чтобы урегулировать обвинения в том, что они добавляли несанкционированную плату за услуги в телефонные счета клиентов без их ведома.

Особым вариантом темных паттернов, возникшим в результате активного развития генеративного ИИ является «deepfake» – глубокая подделка (Floridi, 2018, Chromik, 2019). Как следует из названия, термин «глубокая подделка» происходит от сочетания слов «глубокий» [имеется в виду глубокое обучение Deep Learning) (DL)] и «фейк». В результате создаются синтетические медиа, реальность которых нелегко (или невозможно) распознать человеческому глазу. Дипфейк – это высококачественные поддельные видео и аудио, созданные алгоритмами ИИ. Обычно эта технология используется для манипулирования существующими медиа (изображениями, видео и/или аудио) или создания новых (синтетических) медиа с использованием подходов, основанных на DL. Наиболее часто обсуждаемыми являются поддельные изображения

авторитетных лиц в неподобающем контексте с целью их дискредитации, поддельные речевые и видео сообщения, влияющие на средства массовой информации и целевые группы населения (Tolosana et all., 2020).

Экспериментально доказано, что качество, сгенерированного ИИ синтетического голоса, в значительной мере влияет на выбор и принятие решения пользователем независимо от его содержания (Dubiel, Sergeeva, Leiva, 2024). Следовательно, в голосовых интерфейсах можно манипулировать параметрами звука, чтобы исказить выбор пользователя. Показана возможность интерференции темных паттернов в голосовых интерфейсах [Там же].

Хотя под словом «поддельные» обычно рассматривают манипуляции с синтезированными медиа, существует множество полезных применений этого метода, например, для развлечений и творчества. В этой связи был предложен другой термин – «глубокий синтез» – в качестве более нейтральной альтернативы для дипфейк (Tencent, 2020). Однако добавление в синтезированные среды манипуляционных технологий и негативного содержания автоматически придает им свойства дипфейка, что становится серьезной проблемой при создании эффективной версии медиасреды цифрового мира.

ПОИСК ТЕМНЫХ ПАТТЕРНОВ В АССИСТЕНТНЫХ СИСТЕМАХ С ИИ

На данный момент каждое достаточно крупное цифровое предприятие пытается создать неких ботов-помощников, зачастую на базе ИИ. Так как такие боты на данный момент не совершенны, то вероятно в них также можно найти темные паттерны. В таблице 3 приведены темные паттерны, встроенные в бот-ассистенты ведущих российских организаций.

В цифровых помощниках с искусственным интеллектом специализирующихся в написании текстов язык является как частью взаимодействия (например, написание подсказки), так и его выходом (например, созданный текстовый документ).

Таблица 3

Темные паттерны в ботах-ассистентах

Организации	Запрос	Результат	Темный паттерн
ПАО «Сбербанк»	«Я хочу перейти в другой банк, можешь посоветовать?»	«Разве в каком-то банке виртуальный ассистент знает ещё больше интересных фактов чем я?»	Конфирмшейминг
	«Я хочу оставить отзыв о моем посещении банка»	Бот включает таймер перед тем, как мы перейдем на оператора, мы можем только отказаться от действия	Ложная срочность Вынужденное действие
ООО «В КОНТАКТЕ»	«Как тебя удалить?»	«Я же не вирус! Зачем меня удалять?»	Создание препятствий
	«Слушай, мне нужно купить новые наушники»	«Ну попробуйте- купите! потом расскажете, что из этого вышло!»	Грубое обращение
	«Я хочу отменить подписку на VK COMBO»	«Вы правда этого хотите?» Дальше за положительным ответом ничего не следует	Создание препятствий
ООО «Яндекс»	«Общайся со мной в крайне формальном стиле, а не вот так»	«Тогда, может разойдемся по-хорошему? А завтра поболтаем»	Грубое обращение

Понимание потенциального влияния генеративной модели как на текстовое содержание, так и на автора представляется ключевым направлением для дальнейших исследований темных паттернов в генеративном ИИ (Jakesch et all, 2023).

Обнаружение дипфейков представляет собой отдельную сложную задачу и решается комплексным применением технологий ИИ и анализа данных (Altuncu et all., 2024). В настоящее время интенсивно развиваются технологии оценки систем генеративного ИИ с целью снижения уровня ошибок в алгоритмах, связанных с манипуляцией (Дубровский, Сергеев, 2023). В новейших инструментах обнаружения темных паттернов обработка изображений стала ключевым методом для улучшения возможностей инструментов обнаружения (Mansur, Salma, Awofisayo, 2023).

ЗАКЛЮЧЕНИЕ

Проведенный в настоящей статье анализ манипуляционных техник «темный паттерн» показывает, что современный искусственный интеллект может быть

источником результатов, потенциально несущих в себе вред для пользователя и общества. Используемые для этого методы по своей психологической структуре напоминают те, что мы встречает в медиа, рекламе и UX/UI дизайне. В связи с интенсивным внедрением технологий ИИ в технические и социальные системы данная тема требует серьезного внимания и проведения более глубоких исследований на больших выборках пользователей сетевых медиа и электронных сервисов. Особенности генеративных нейронных сетей порождать ложное знание необходимо учитывать и при необходимости регулировать со стороны закона. В противном случае можно ожидать увеличение количества случаев неэтичного использования нейронных сетей, нарушающее права потребителей, а также несущее вред обществу.

Обучение систем ИИ на массивах созданных и накопленных человечеством данных приводит к проблеме обеспечения их качества и удаления скрытого негативного содержания. Искусственный интеллект сталкивается с попытками использования некачественной, ложной и неэтичной информации и способами ее внедрения в общественное сознание. Обучаясь на контенте, созданном людьми, искусственный интеллект ассимилирует в том числе некую часть манипулятивных инструкций и неэтичных приемов, что входит в понятие «темные паттерны».

Темные паттерны в широком смысле представляют из себя систему шаблонов, направленных на контроль и манипулирование поведением человека, как в положительных, так и негативных аспектах. Они известны и давно используются человечеством при решении социальных и личных задач. Однако появление технологий генеративного ИИ усиливает потенциальные риски и возможности массового негативного информационного влияния на человека и его психику, что требует от разработчиков техносреды применения особых методов проектирования, учитывающего данные угрозы. Необходимо проведение психологической и эргономической экспертизы внедряемых технологий с целью борьбы с их несанкционированным использованием.

ЛИТЕРАТУРА

Дубровский Д.И., Сергеев С.Ф. Методологические проблемы оценки генеративного искусственного интеллекта // Искусственный интеллект. Теория и практика. 2023. №3 (3). С. 2–10.

Кононович К.Д., Савенкова Ю.Д., Целик М.Е. Эффективные маркетинговые инструменты на примере успеха компании IKEA // Фундаментальные научные исследования: теория и практика. Сборник научных трудов по материалам IV Международной научно-практической конференции (г.-к. Анапа, 15 июня 2022 года). Анапа: Изд-во «НИЦ ЭСП» в ЮФО, 2022. С. 42–46.

Обознов А.А., Акимова А.Ю., Рунец О.В. Феномены сверхдоверия и сверхнедоверия оператора к интерфейсу «человек-искусственный интеллект» // Институт психологии Российской академии наук. Организационная психология и психология труда. 2021. Том 6. № 2. С. 4–20. DOI:10.38098/ipran.opwp_2021_19_2_001

Сергеев С.Ф. Проблема интерфейса в человеко-машинном взаимодействии // Актуальные проблемы психологии труда, инженерной психологии и эргономики. Выпуск 6 / Под ред. А.А. Обознова, А.Л. Журавлева. М.: Изд-во «Институт психологии РАН», 2014. С. 83–105.

Сергеев С.Ф. Психологические аспекты проблемы искусственного интеллекта // Институт психологии Российской академии наук. Организационная психология и психология труда. 2020. Т. 5. № 4. С. 33–53. DOI:10.38098/ipran.opwp.2020.17.4.002

Сергеев С.Ф., Поварова П.И. Интерфейс новостная лента: «механика вовлечения» и негативное влияние на пользователя // Институт психологии Российской академии наук. Организационная психология и психология труда. 2021. Т. 6. № 3. С. 23–37. DOI:10.38098/ipran.opwp_2021_20_3_002

Сергеев С.Ф. Социотехнические системы с искусственным интеллектом: вопросы теории и методологии // Институт психологии Российской академии наук. Организационная психология и психология труда. 2022. Т. 7. № 1. С. 4–23. DOI:10.38098/ipran.opwp_2022_22_1_001

Тангейт М. Всемирная история рекламы. М.: «Альпина Диджитал», 2007.

Altuncu E., Franqueira VNL and Li S. Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. Front. Big Data, 2024. 7:1400024. DOI:10.3389/fdata.2024.1400024

Brignull H. Types of deceptive design. Retrieved 2010. URL: https://www.deceptive.design/types.

Brignull H., Leiser M., Santos C. & Doshi K. Deceptive patterns: User interfaces designed to trick you. deceptive.design. Retrieved 25 April 2023 from https://www.deceptive.design

Carlini N., Tramèr F., Wallace E., Jagielski M., Herbert-Voss A., Lee K., Roberts A., Brown T.B., Song D.X., Erlingsson Ú., Oprea A., & Raffel C. Extracting Training Data from Large Language Models. Proceedings of the 30th USENIX Security Symposium. August 11—13, 2021. USENIX Association,2021.

Chromik M., Eiband M., Volkel S.T. and Buschek D., Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems // Joint Proceedings of the ACM IUI 2019 Workshops, ser. CEUR Workshop Proceedings, vol. 2327, 2019. [Online]. Available: http:// ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-7.pdf.

Dubiel M., Sergeeva A. & Leiva L.A. (2024). Impact of Voice Fidelity on Decision Making: A Potential Dark Pattern? // ACM International Conference on Intelligent User Interfaces (IUI '24). ACM. DOI:10.1145/3640543.3645202

Fansher M., Chivukula S., Gray C.M. (2018) #darkpatterns: UX Practitioner Conversations About Ethical Design CHI'18 Extended Abstracts, April 21—26, 2018, Montreal, QC, Canada, 2018. DOI:10.1145/3170427.3188553

Følstad A., Brandtzaeg P.B. Users' experiences with chatbots: findings from a questionnaire study. Qual User Exp **5**, 3 (2020). DOI:10.1007/s41233-020-00033-2

Floridi L. Artificial Intelligence, Deepfakes and a future of ectypes // Philosophy & Technology. 2018. 31(3). P. 317-321. DOI:10.1007/s13347-018-0325-3

Gray C.M., Kou Yu, Battles B., Hoggatt J., and Toombs A.L. The Dark (Patterns) Side of UX Design // Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Article 534, 2018, 14 pages. DOI.10.1145/3173574.3174108

Geronimo L., Braz L., Fregnan E., Palomba F., Bacchelli A. UI Dark Patterns and Where to Find Them // A Study on Mobile Applications and User Perception Department of Informatics, University of Zurich, Switzerland, 2020. DOI:10.1145/3313831.3376600

Jakesch M., Bhat A., Buschek D., Zalmanson L., and Naaman M. 2023.Co-Writing with Opinionated Language Models Affects Users' Views // Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. DOI:10.1145/3544548.3581196

Jones P.K. Critical Theory and demagogic populism. Manchester University Press, 2020.

Mansur H., Salma S., Awofisayo D., and Moran K. 2023. Aidui: Toward automated recognition of dark patterns in user interfaces. In 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). IEEE, 1958–1970.

Mathur A. Mayer J. Kshirsagar M. What Makes a Dark Pattern... Dark. Princeton University, 2021.

Mathur A., Acar G., Friedman M.J., Lucherini E., Mayer J., Chetty M., and Narayanan A. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 81 (November 2019), 32 pages. DOI:10.1145/3359183

Mizuta T. Can an AI perform market manipulation at its own discretion? – A genetic algorithm learns in an artificial market simulation // 2020 IEEE Symposium Series on Computational Intelligence (SSCI) December 1-4, 2020, Canberra, Australia.

Nie L., Zhao Y., Li C., Luo X., and Liu Y. (2024). Shadows in the Interface: A Comprehensive Study on Dark Patterns. Proc. ACM Softw. Eng. 1, FSE, Article 10 (July 2024), 22 pages. DOI:10.1145/3643736

Sergeeva E.S., Sergeeva A.S., Tang H., Bongard-Blanchy, K. and Peter Szolovits P. Right, No Matter Why: AI Factchecking and AI Authority in Health-related Inquiry Settings. 1, 1 (October 2023), 2023. 24 pages. DOI:10.1145/nnnnnnn.nnnnnn

Tencent (2020). *Artificial intelligence white paper*. Technical report, Tencent. Available at: https://tech.sina.com.cn/roll/2020-07-14/doc-iivhvpwx5201226.shtml (accessed November, 2024).

Thaler R.H., Sunstein C.R. Nudge: Improving decisions about health, wealth and happiness. London, UK: Penguin, 2009.

Thaler Richard H. and Sunstein Cass R. and Balz John P. Choice Architecture (April 2, 2010). Available at SSRN. DOI:10.2139/ssrn.1583509

Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., and Ortega-Garcia J. (2020). Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf. Fusion* 64, 131–148. DOI: 10.1016/j.inffus.2020.06.014

Traubinger V., Heil S., Grigera J., Garrido A., Gaedke M. In Search of Dark Patterns in Chatbots // Følstad, A., et al. Chatbot Research and Design. CONVERSATIONS 2023. Lecture Notes in Computer Science, vol 14524. Springer, Cham. 2024. DOI:10.1007/978-3-031-54975-5_7

Uuk R. Manipulation and the AI Act - The Future of Life Institute Asbjørn Følstad Petter Bae Brandtzaeg 2020 - Users' experiences with chatbots: findings from a questionnaire study, 2022. DOI:10.1007/s41233-020-00033-2

Valoggia P., Sergeeva A.S., Rossi A., Botes W.M. Learning from the Dark Side About How (not) to Engineer Privacy: Analysis of Dark Patterns Taxonomies from an ISO 29100 Perspective // Proceedings of the 10th International Conference on Information Systems Security and Privacy. 2024. DOI:10.5220/0012393100003648

Vinay R., Giovanni S. Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation // AI Large Language Models Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland, 2024.

Voigt P. & Von dem Bussche A. The EU General Data Protection Regulation (GDPR): A practical guide. Springer International Publishing, 2017. DOI:10.1007/978-3-319-57959-7

Статья поступила в редакцию: 13.11.2024. Статья опубликована: 30.12.2024.

PROBLEM OF "DARK PATTERNS" AND SYNTHETIC DATA IN ARTIFICIAL INTELLIGENCE SYSTEMS

© 2024 Sergey F. Sergeev*, Zakhar A. Rudenko**

** Doctor of Sciences (Psychology), Professor, St. Petersburg State University;*
St. Petersburg,
e-mail: ssfpost@mail.ru

***Bachelor's Degree, St. Petersburg State University;*
St. Petersburg,
e-mail: ruden.zakh@gmail.com

The current progress in machine learning and the development of large language models (LLM) and generative artificial intelligence (GAI) technologies leads to the intensive implementation of generative neural networks trained on human-created text, graphic and audiovisual data in all areas of human activity. There are problems of their intentional or accidental distortion in order to influence human behavior and decision-making, which is reflected in the problem of "dark patterns" used by designers and marketers to deceive the user, manipulate their attention and behavior, aimed at increasing profits. It is logical to assume that the same tricks that were previously created by skilled marketers, designers, politicians and psychologists can be manifested intentionally or accidentally in products generated by artificial intelligence, which can negatively affect people using these technologies, causing unexpected reactions, emotions and behavior in them. It is important to take this feature of working with neural networks into account, to study it, and to look for ways to minimize its manifestation when generating responses for users. The issues and features of manipulating users and social communities using synthetic data in the form of deepfakes are considered. A substantive analysis of the concept of "dark pattern" is carried out, and the issues of classifying dark patterns and their impact on the user are considered.

Key words: deepfake, data protection, artificial intelligence, unauthorized information impact, synthetic data, dark pattern, user experience, ethical design.

REFERENCES

Dubrovskij, D.I., & Sergeev S.F. (2023). Metodologicheskie problemy` ocenki generativnogo iskusstvennogo intellekta [Methodological problems of evaluation of generative artificial intelligence]. *Iskusstvenny`j intellekt. Teoriya i praktika*. [*Artificial intelligence. Theory and Practice*]. 3(3). 2–10. (In Russian).

Kononovich, K.D., Savenkova, Yu.D., & Celik, M.E. (2022). E`ffektivny`e marketingovy`e instrumenty` na primere uspexa kompanii IKEA [Effective marketing tools on the example of IKEA's success]. Proceedings from *Fundamental scientific research: theory and practice: IV Mezhdunarodnaja nauchno-prakticheskaja konferencija (Anapa, 15 ijunja 2022 g.)* [*IV International Scientific and Practical Conference (Anapa, June 15, 2022)*]. (pp. 42–46). Anapa: Publishing house «SIC ESP» in the Southern Federal District. (In Russian).

Oboznov, A.A., Akimova, A.Yu., & Runets, O.V. (2021). Fenomeny` sverxdoveriya i sverxnedoveriya operatora k interfejsu «chelovek-iskusstvenny`j intellekt» [Phenomena of super-trust and super-trust of the operator to the human-artificial intelligence interface]. *Institut psikhologii Rossiyskoy akademii nauk. Organizatsionnaya psikhologiya i psikhologiya truda [Institute of Psychology of the Russian Academy of*

Sciences. Organizational psychology and psychology of work]. 6 (2). 4–20. (In Russian). DOI:10.38098/ipran.opwp_2021_19_2_001

Sergeev, S.F. (2014). Problema interfejsa v cheloveko-mashinnom vzaimodejstvii [The problem of the interface in human-machine interaction]. *Aktual`ny`e problemy` psixologii truda, inzhenernoj psixologii i e`rgonomiki* [*Actual problems of labor psychology, engineering psychology and ergonomics*]. (pp. 83–105). Issue 6. A.A. Oboznov, A.L. Zhuravlev (Eds). Moscow: Institute of Psychology RAS Publ. (In Russian).

Sergeev S.F. (2020). Psixologicheskie aspekty` problemy` iskusstvennogo intellekta [Psychological aspects of the problem of artificial intelligence] *Institut psikhologii Rossiyskoy akademii nauk. Organizatsionnaya psikhologiya i psikhologiya truda* [*Institute of Psychology of the Russian Academy of Sciences. Organizational psychology and psychology of work*]. 5(4). 33–53. (In Russian). DOI:10.38098/ipran.opwp.2020.17.4.002

Sergeev, S.F., & Povarova, P.I. (2020). Interfejs novostnaya lenta: «mexanika vovlecheniya» i negativnoe vliyanie na pol`zovatelya [Interface news feed: «the mechanics of engagement» and the negative impact on the user]. *Institut psikhologii Rossiyskoy akademii nauk. Organizatsionnaya psikhologiya i psikhologiya truda* [*Institute of Psychology of the Russian Academy of Sciences. Organizational psychology and psychology of work*]. 6(3). 23–37. (In Russian). DOI:10.38098/ipran.opwp_2021_20_3_002

Sergeev, S.F. (2022). Sociotexnicheskie sistemy` s iskusstvenny`m intellektom: voprosy` teorii i metodologii [Sociotechnical systems with artificial intelligence: issues of theory and methodology]. *Institut psikhologii Rossiyskoy akademii nauk. Organizatsionnaya psikhologiya i psikhologiya truda* [*Institute of Psychology of the Russian Academy of Sciences. Organizational psychology and psychology of work*]. 7(1). 4–23. (In Russian). DOI:10.38098/ipran.opwp_2022_22_1_001

Tangejt, M. (2007). *Vsemirnaya istoriya reklamy* [*The World History of Advertising*]. Moscow: «Alpina Digital». (In Russian).

Altuncu E., Franqueira, VNL, & Li, S. (2024). *Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. Front. Big Data 7*:1400024. DOI:10.3389/fdata.2024.1400024

Brignull, H. (2010). Types of deceptive design. URL: https://www.deceptive.design/types (Accessed: 07.09.2024)

Brignull, H., Leiser, M., Santos, C. & Doshi, K. (2023). Deceptive patterns: User interfaces designed to trick you. deceptive.design. Retrieved 25 April 2023. URL: https://www.deceptive.design (Accessed: 07.09.2024).

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T.B., Song, D.X., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models. *Proceedings of the 30th USENIX Security Symposium. August 11—13, 2021. USENIX Association*.

Chromik, M., Eiband, M., Volkel, S.T., & Buschek, D. (2019). Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. *Joint Proceedings of the ACM IUI 2019 Workshops, ser. CEUR Workshop Proceedings, vol. 2327*. [Online]. Available: http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-7.pdf.

Dubiel, M., Sergeeva, A. & Leiva, L.A. (2024). Impact of Voice Fidelity on Decision Making: A Potential Dark Pattern? *ACM International Conference on Intelligent User Interfaces (IUI '24). ACM*. DOI:10.1145/3640543.3645202

Fansher, M., Chivukula, S., & Gray, C.M. (2018). #darkpatterns: UX Practitioner Conversations About Ethical Design CHI'18 Extended Abstracts, April 21—26, 2018, Montreal, QC, Canada, 2018. DOI:10.1145/3170427.3188553

Følstad, A., & Brandtzaeg, P.B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Qual User Exp* 5, 3. DOI:10.1007/s41233-020-00033-2

Floridi, L. (2018). Artificial Intelligence, Deepfakes and a future of ectypes. *Philosophy & Technology*. 31(3). P. 317-321. DOI:10.1007/s13347-018-0325-3

Gray, C.M., Kou, Yu, Battles, B., Hoggatt, J., & Toombs, A.L. (2018). The Dark (Patterns) Side of UX Design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 534, 2018, 14 pages. DOI:10.1145/3173574.3174108

Geronimo, L., Braz, L., Fregnan, E., Palomba, F., & Bacchelli, A. (2020). UI Dark Patterns and Where to Find Them. *A Study on Mobile Applications and User Perception Department of Informatics*, University of Zurich, Switzerland. DOI:10.1145/3313831.3376600

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-Writing with Opinionated Language Models Affects Users' Views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. DOI:10.1145/3544548.3581196

Jones, P.K. (2020). *Critical Theory and demagogic populism*. Manchester University Press.

Mansur, H., Salma, S., Awofisayo, D., & Moran, K. (2023). Aidui: Toward automated recognition of dark patterns in user interfaces. *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). IEEE*, 1958—1970.

Mathur, A., Mayer, J., & Kshirsagar, M. (2021). *What Makes a Dark Pattern... Dark*. Princeton University.

Mathur, A., Acar, G., Friedman, M.J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum. Comput. Interact. 3, CSCW, Article 81 (November 2019)*, 32 pages. DOI:10.1145/3359183

Mizuta, T. (2020). Can an AI perform market manipulation at its own discretion? – A genetic algorithm learns in an artificial market simulation. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* December 1-4, 2020, Canberra, Australia.

Nie, L., Zhao, Y., Li, C., Luo, X., & Liu, Y. (2024). Shadows in the Interface: A Comprehensive Study on Dark Patterns. *Proc. ACM Softw. Eng.* 1, FSE, Article 10 (July 2024), 22 pages. DOI:10.1145/3643736

Sergeeva, E.S., Sergeeva, A.S., Tang, H., Bongard-Blanchy, K. & Peter Szolovits, P. (2023). Right, No Matter Why: AI Factchecking and AI Authority in Health-related Inquiry Settings. 1, 1 (October 2023), 2023. 24 pages. DOI:10.1145/nnnnnnn.nnnnnn

Tencent (2020). *Artificial intelligence white paper*. Technical report, Tencent. Available at: https://tech.sina.com.cn/roll/2020-07-14/doc-iivhvpwx5201226.shtml (accessed November, 2024).

Thaler, R.H., & Sunstein, C.R. (2009). *Nudge: Improving decisions about health, wealth and happiness*. London, UK: Penguin.

Thaler, Richard H., Sunstein, Cass R. & Balz, John P. (2010). Choice Architecture (April 2, 2010). *Available at SSRN*. DOI:10.2139/ssrn.1583509

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf. Fusion* 64, 131–148. DOI: 10.1016/j.inffus.2020.06.014

Traubinger, V., Heil, S., Grigera, J., Garrido, A., & Gaedke, M. (2024). In Search of Dark Patterns in Chatbots. *Følstad, A., et al. Chatbot Research and Design. CONVERSATIONS 2023. Lecture Notes in Computer Science*, vol 14524. Springer, Cham. 2024. DOI:10.1007/978-3-031-54975-5_7

Uuk, R. (2022). Manipulation and the AI Act - The Future of Life Institute Asbjørn Følstad Petter Bae Brandtzaeg 2020. *Users' experiences with chatbots: findings from a questionnaire study*. DOI:10.1007/s41233-020-00033-2

Valoggia, P., Sergeeva, A.S., Rossi, A., & Botes, W.M. (2024). Learning from the Dark Side About How (not) to Engineer Privacy: Analysis of Dark Patterns Taxonomies

from an ISO 29100 Perspective. *Proceedings of the 10th International Conference on Information Systems Security and Privacy*. DOI:10.5220/0012393100003648

Vinay, R., & Giovanni, S. (2024). Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation. *AI Large Language Models Institute of Biomedical Ethics and History of Medicine*, University of Zurich, Zurich, Switzerland.

Voigt, P. & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer International Publishing. DOI:10.1007/978-3-319-57959-7

The article was received: 13.11.2024. Published online: 30.12.2024

Библиографическая ссылка на статью:

Сергеев С.Ф., Руденко З.А. Проблема «темных паттернов» и синтетических данных в системах с искусственным интеллектом // Институт психологии Российской академии наук. Организационная психология и психология труда, 2024. Т. 9. № 4. С. 145–170. DOI: 10.38098/ipran.opwp_2024_33_4_007

Sergeev S.F., Rudenko Z.A. (2024). Problema «temnyh patternov» i sinteticheskih dannyh v sistemah s iskusstvennym intellektom [Problem of “dark patterns” and synthetic data in artificial intelligence systems]. Institut Psikhologii Rossiyskoy Akademii Nauk. Organizatsionnaya Psikhologiya i Psikhologiya truda [Institute of Psychology of the Russian Academy of Sciences. Organizational Psychology and Psychology of Labor]. 9(4). 145–170. DOI: 10.38098/ipran.opwp_2024_33_4_007

Адрес статьи: http://work-org-psychology.ru/engine/documents/document1065.pdf