Киселев Е.Ю. (науч. рук. Межина М.В.)

Когнитивная безопасность в веб-среде: гибридный метод обнаружения угроз и интерфейсных манипуляций (Dark Patterns) на стороне клиента

УДК тезиса: 004.056

Аннотация: рассматривается проблема когнитивной уязвимости пользователей перед интерфейсными манипуляциями. Предложена концепция сервиса «Цифровой Иммунитет» - браузерного расширения с архитектурой Client-Side, использующего гибридную модель детектирования. Внедрение решения повышает цифровую грамотность и снижает риски мошенничества в веб-среде

Человек, независимо от степени образованности, рода деятельности, способностей и иных черт имеет фундаментальную проблему в виде ограниченности когнитивных ресурсов, при которой мозг, стараясь минимизировать умственные затраты – использует эвристики для принятия решений, выбирая «путь наименьшего сопротивления».

Если в физической реальности существуют устоявшиеся паттерны распознавания угроз (резко проезжающая машина в нескольких сантиметрах от человека, устоявшаяся фраза-клише, услышанная от мошенника и т.п.) – по которым можно явно понять опасность и включить полную бдительность в конкретной обстановке. То вот в веб-пространстве пользователь часто сталкивается с обилием технических угроз и интерфейсных манипуляций «темных паттернов», менее подверженные обнаружению.

Многообразия типов угроз и манипуляций в сочетании со сниженным уровнем бдительности пользователя (например, находясь в состоянии утомления или пассивного потребления контента) создает условия – при которых опасность в веб-среде может оставаться незамеченной. Это в свою очередь приводит к когнитивным ошибкам пользователя и, как следствие, к принятию иррациональных решений, которые в частности могут иметь экономический убыток.

Учитывая, что по России, согласно статистики DataReportal, на 2025 год ежедневно пользуются Интернетом более 133 млн. человек, включая в том числе и поколения, что с раннего времени находятся в виртуальной среде, например «Альфа». Дети и взрослые не в полной мере понимают всю ту степень угрозы, которая может исходить из Интернета – пока она лично с ними не случится. Ситуация и осложняется тем, что классические антивирусы (Avast, Kaspersky, Dr.Web) могут отлично замечать вирусы – но вот интерфейсные манипуляции и технические угрозы, в большинстве своем – нет.

Даже если человек понимает опасность веб-среды, как отмечалось выше – он не может иметь постоянную бдительность. А значит всегда будет определенный шанс, когда он потеряет фокус внимания и не заметит, например, техническую угрозу в виде подмены домена «rnicrosoft» вместо «microsoft», что приведет к негативным последствиям.

Решение столь комплексной проблемы формирует целую свободную нишу, на данный момент времени, в виде сервисов для повышения когнитивной безопасности пользователей в веб-среде.

Мы представляем игрока этой ниши в первую очередь в виде браузерного расширения, поскольку его технология позволяет с меньшими издержками получать содержимое веб-страницы и проверять на наличие угроз.

Помимо этого сервис-расширение не должно собирать данные на сервера – поскольку иначе это будет трактоваться вторжением в личное пространства с юридическими последствиями для разработчика сервиса. Учитывая это, разрабатываемый продукт должен быть на основании «Client-Side» подхода. Это означает, что с одной стороны, на стороне сервера должна формироваться определенная

база данных алгоритмов и знаниях об угрозах и темных паттернах манипуляций, а затем подгружаться в ядро клиентской части расширения – чтобы анализ выполнялся на стороне пользователя, никуда не передаваясь от него.

Помимо ясной архитектуры, продукт должен предупреждать пользователя – но при этом не нарушать политики магазина расширений, например согласно политике Google Chrome Web Store, расширениям запрещено радикально изменять внешний вид или функциональность веб-страниц без предварительного разрешения пользователя.

Следовательно – продукт должен мягко предупреждать пользователя или также иметь под собой настройку для включения жесткого предупреждения (с полным перечнем того – как именно это настройка изменяет веб-страницу для пользователя при своей активации).

Наконец, расширение должно анализировать веб-страницу на угрозы через алгоритмы сопоставления элемента веб-страницы с тем – что мы ожидаем. Поскольку темные паттерны все еще мало изучены – можно использовать трактовки, данные Гарри Бригнуллом, где он разделял эти манипуляции на различные группы, в зависимости от их сути. Например «Sneaking» - темные паттерны, связанные с маскировкой информации – пример «тайпсквоттинг» или «отсутствие политики конфиденциальности» на странице с формой контактов.

В тоже время, работая с веб-средой, важно разделять алгоритмы детекции на две группы: детерминированные и обученные экспертным путем.

Детерминированные – это те алгоритмы, где при обучении мы гарантированно знаем что с чем сопоставлять. Угрозы, где применяются эти алгоритмы – имеют либо один элемент, либо строго четкий перечень конструкций с которыми нужно сравнивать. Как правило – дальнейшей доработки, корректировки здесь не требуется. Например, сюда можно отнести паттерн «отсутствие политики конфиденциальности», где можно проверить наличие слов «политика конфиденциальности», «оферта», и парочку других.

Другое дело с алгоритмами, обученными экспертным путем. Здесь паттерн манипуляций имеет большое количество различных вариаций и, чтобы увеличить точность срабатывания сервиса – важно привлечь к ним экспертов в виде людей или организаций. Сюда можно отнести угрозы в виде «домена из черного списка ЦБ РФ» (признак мошеннической деятельности организаций) или «апелляция к авторитету без источника» с regex-запросами, которые должны быть разнообразными.

Наконец – расширение не должно по одному признаку делать диагноз по всему сайту. Расширение обязано анализировать совокупно все угрозы, ведь в большинстве приложений темные паттерны используются, но бить тревогу по каждой из них (жестко уведомляя) – это может отторгнуть клиента и не принести должной отдачи.

В качестве практической реализации предложенных архитектурных и алгоритмических принципов нами представлена концепция сервиса «Цифровой Иммунитет». Данный прототип демонстрирует возможность автоматизированного выявления 7 видов манипуляций без использования ресурсоемких ML-моделей на стороне клиента.

Предлагаемый подход к когнитивной безопасности, сочетающий мягкое информирование и жесткое прерывание сценария при высоких рисках, может стать эффективным инструментом снижения ущерба от мошенничества и повышения цифровой грамотности пользователей, не нарушая при этом пользовательский опыт.

Список использованных источников:

1. Types of deceptive pattern // Deceptive Patterns URL: https://www.deceptive.design/types (дата обращения: 16.02.2026).

2. Program Policies // Chrome Web Store URL: https://developer.chrome.com/docs/webstore/program-policies/policies (дата обращения: 16.02.2026).