AI Dark Pattern Detection System For Fair Web And App UX

A.Alagarh[1], V.Sanjay[2], E.Sundara Vignesh[3], S.Sriram[4]
[1]Assist prof, Dept of CSE
[2, 3, 4] Dept of CSE
[1, 2, 3, 4] K.L.N.College of Engineering,Sivagangai, India

Abstract- *The increasing prevalence of deceptive user interface designs, commonly referred to as dark patterns, poses significant challenges to user autonomy and transparency in digital environments. These patterns manipulate users into making unintended decisions, such as accepting unnecessary permissions or engaging with misleading offers. Traditional detection approaches are often manual, reactive, or limited in scalability, making them ineffective for real-time user protection.*

To address these limitations, this project proposes an automated dark pattern detection system implemented as a Chrome browser extension using Manifest V3 architecture. The system leverages a rule-based detection mechanism grounded in the FoSIP framework to identify multiple categories of manipulative design, including social engineering, forced actions, interface interference, fake discounts, and persistent elements. By integrating real-time DOM analysis with the MutationObserver API, the extension dynamically detects and highlights suspicious elements without requiring user interaction.

Furthermore, the system incorporates a fairness scoring model that quantifies the ethical quality of web pages based on detected patterns and their severity levels. The modular architecture, built using lightweight web technologies such as JavaScript, HTML, and CSS, ensures scalability and extensibility, enabling future integration with AI-based contextual analysis. The proposed solution provides a proactive, user-centric approach to enhancing transparency in web interactions, promoting ethical design practices, and empowering users to make informed decisions while browsing.

I. INTRODUCTION

With the rapid expansion of digital platforms and online services, user interface design has evolved to play a critical role in shaping user behavior. While many design strategies aim to enhance usability and engagement, a growing number of websites employ deceptive techniques known as *dark patterns* to manipulate users into unintended actions, such as accepting cookies, making purchases, or subscribing to services. These practices undermine user autonomy, reduce transparency, and raise significant ethical concerns in modern web ecosystems.

Despite increasing awareness, the detection of dark patterns remains largely manual, inconsistent, and dependent on user vigilance. Existing solutions are either limited in scope, lack real-time capabilities, or require explicit user interaction, making them ineffective for seamless integration into everyday browsing experiences. Furthermore, modern web applications frequently utilize dynamic content rendering through frameworks such as React and Vue, making traditional static detection approaches inadequate.

This project introduces an automated, browser-based system implemented as a Chrome extension that provides real-time detection of dark patterns without requiring any user intervention. Built using Manifest V3 architecture and lightweight web technologies, the system leverages rule-based detection aligned with the FoSIP framework to identify and classify multiple categories of manipulative design. By incorporating dynamic DOM monitoring through the MutationObserver API, the solution ensures adaptability across modern, interactive websites.

Additionally, the system introduces a fairness scoring mechanism that quantifies the ethical quality of web pages, offering users an intuitive understanding of potential manipulation. The proposed solution is scalable, efficient, and user-centric, providing a practical approach to promoting transparency, improving digital literacy, and encouraging ethical interface design practices across the web.

II. LITERATURE SURVEY

Recent studies highlight the increasing concern surrounding deceptive user interface practices, commonly referred to as dark patterns, and their impact on user decision-making and digital trust. Research in the field of Human-Computer Interaction (HCI) demonstrates that such patterns exploit cognitive biases, leading users to unintentionally consent to actions such as data sharing, subscriptions, or

purchases. Studies indicate that a significant percentage of popular websites incorporate at least one form of dark pattern, emphasizing the scale of the issue.

Several approaches have been proposed for detecting dark patterns, including manual audits, crowdsourced reporting systems, and rule-based browser extensions. While manual and survey-based methods provide valuable insights, they are not scalable for real-time detection. Existing browser-based tools offer partial automation but often rely on static rules and fail to adapt to dynamic web environments, particularly in modern single-page applications (SPAs).

Recent advancements explore the use of machine learning and natural language processing techniques to identify deceptive patterns by analyzing textual and visual cues.

Although these approaches improve detection accuracy, they require large annotated datasets and significant computational resources, limiting their practicality for lightweight, real-time deployment in browser environments.

Frameworks such as the FoSIP classification model provide a structured approach to categorizing dark patterns based on factors such as social influence, forced actions, and persistence. However, many existing systems do not fully integrate such frameworks into automated detection pipelines. Additionally, most solutions lack continuous monitoring mechanisms to handle dynamically loaded content such as popups, overlays, and asynchronous UI updates.

These studies collectively highlight the need for a system that combines structured classification with real- time, automated detection in dynamic web environments. The proposed project addresses these gaps by integrating a FoSIP-based rule engine with continuous DOM monitoring, enabling scalable, efficient, and user-friendly detection of dark patterns directly within the browsing experience.

III. PROPOSED METHODOLOGY

The proposed system adopts a modular, browser-based architecture to enable real-time detection of dark patterns across web interfaces. The extension is developed using Chrome Extension Manifest V3, ensuring efficient background processing and enhanced security. The frontend component, implemented through a popup interface using HTML, CSS, and JavaScript, provides users with a clear summary of detected patterns along with a computed fairness score.

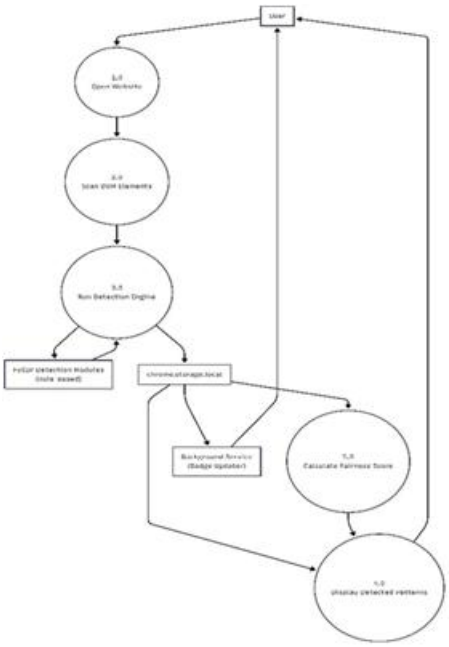
The core detection mechanism is implemented within the content script, which is automatically injected into every webpage visited by the user. This module performs real-time analysis of the Document Object Model (DOM) to identify manipulative design elements. A rule-based detection engine, structured according to the FoSIP framework, classifies patterns into categories such as social engineering, forced action, interface interference, fake discounts, and persistence. Each category is detected using predefined heuristics, including keyword matching, structural analysis, and attribute inspection.

To ensure compatibility with modern dynamic websites, the system integrates the MutationObserver API, which continuously monitors changes in the DOM. This enables the detection of patterns that appear after page load, such as popups, overlays, and asynchronously rendered components commonly found in single-page applications.

The background service worker operates independently to manage auxiliary tasks such as updating the extension badge with the number of detected patterns and coordinating communication between components. Detected results are stored locally using the Chrome Storage API, allowing efficient retrieval and persistence across sessions.

Additionally, the system incorporates a scoring mechanism that evaluates the ethical quality of a webpage based on the number and severity of detected patterns. This fairness score provides users with an intuitive, quantitative measure of potential manipulation.

The overall architecture is lightweight, scalable, and extensible, allowing future integration with advanced techniques such as machine learning or natural language processing for improved detection accuracy and contextual understanding.



IV. SYSTEM PROCESS FLOW

The system follows a sequential yet dynamic workflow to ensure real- time detection and visualization of dark patterns during web browsing. The process is designed to operate automatically in the background, requiring no explicit user interaction.

1. Website Access and Script Injection

The process begins when the user opens a webpage through the browser. At this stage, the Chrome extension automatically injects the content script into the webpage using

Manifest V3 specifications. This enables direct interaction with the Document Object Model (DOM) of the loaded page.

2. DOM Scanning

Once injected, the content script initiates a comprehensive scan of the webpage's DOM structure. It analyzes elements such as text nodes, buttons, popups, banners, and form components to identify potential indicators of manipulative design patterns**.**

3. Execution of Detection Engine

Following DOM analysis, the detection engine is triggered. This engine consists of multiple rule- based modules aligned with the FoSIP framework. Each module operates independently to detect specific categories of dark patterns, including:

- Social engineering cues (e.g., urgency messages)
- Forced actions (e.g., mandatory cookie acceptance)
- Interface interference (e.g., misleading UI text)
- Fake discounts (e.g., manipulated pricing)
- Persistence mechanisms (e.g., autoplay content)

The detection process applies heuristic rules such as keyword matching, structural analysis, and attribute evaluation.

4. Data Storage and Background Processing

Detected patterns are stored locally using the Chrome Storage API (chrome.storage.local). This ensures persistence of results and efficient data retrieval. Simultaneously, the background service worker processes this data to update the extension badge, displaying the number of detected patterns in real time.

5. Fairness Score Calculation

The system then computes a fairness score based on the number and severity of detected patterns. Each pattern category contributes differently to the score, enabling a quantitative assessment of the webpage's ethical design quality

6. Popup Interface and User Interaction

When the user clicks on the extension icon, the popup interface is displayed. This interface retrieves stored results and presents:

- A list of detected dark patterns
- The computed fairness score
- Additional controls such as rescan, clear, or blocking options

V. RESULTS AND DISCUSSION

The developed Chrome extension was evaluated across a variety of real-world websites, including e- commerce platforms, news portals, and service-based applications, to assess its effectiveness in detecting dark patterns. The system successfully identified multiple categories of manipulative design, including urgency-based messages, forced cookie consent mechanisms, misleading interface elements.

User-level testing indicated that the extension operates seamlessly in the background and requires minimal interaction, making it suitable for everyday use. The popup

interface was found to be intuitive, effectively summarizing detected patterns and providing actionable insights.

However, the reliance on rule- based detection introduces certain limitations, including the possibility of false positives in ambiguous cases and reduced effectiveness against highly sophisticated or context- dependent patterns. Despite these limitations, the system demonstrates strong practical utility and highlights the effectiveness of integrating real-time detection, structured classification, and user-centric visualization within a lightweight browser-based solution.

Overall, the results validate the proposed approach as an efficient and scalable method for enhancing transparency in web interactions and promoting ethical user interface design practices.

VI. CONCLUSION

This project presents a browser-based system for the automated detection of dark patterns in web interfaces, implemented as a Chrome extension using Manifest V3 architecture. By leveraging lightweight web technologies and a rule-based detection engine grounded in the FoSIP framework, the system provides real-time identification and classification of manipulative design practices. Unlike traditional approaches that rely on manual inspection or user intervention, the proposed solution operates seamlessly in the background, ensuring continuous monitoring and user awareness.

REFRENCES

[1] D. Di Geronimo, A. Braz, and P. Barcellos,“Manipulative Interface Designs in Mobile Applications: A Systematic Review,”IEEE Transactions on Human-Machine Systems, vol. 52, no. 3, pp. 421–433, 2022.

[2] S. Zimmeck, C. M. Gray, J. L. Schaub, and B. N. Schaub,“Automated Analysis of Privacy Policy Text and Consent Interfaces,”IEEE Security & Privacy Magazine, vol. 18, no. 4, pp. 42–50, 2020.

[3] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs,“The Dark (Patterns) Side of UX Design,”Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2018.

[4] A. Mathur, G. Acar, M. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan,“Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites,”Proceedings of the ACM on Human-Computer Interaction (CSCW), 2019..

[5] S.Mellivora, R. Wang, J. Reidenberg, and H. Nissenbaum,“Automated Privacy Compliance Analysis of Mobile App Interfaces,”IEEE Transactions on Software Engineering, vol. 48, no. 6, pp. 2105–2120, 2022.