Detecting Textual and Visual Dark Patterns Using a Large Language Model in E-Commerce

Mohammadhossein Yekeh

A Thesis
in
The Department
of
Design and Computation Arts

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Design at Concordia University
Montreal, Quebec, Canada

March 2025

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: _____

Entitled: _____

and submitted in partial fulfillment of the requirements for the degree of

_____

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Pippin Barr

_____ Examiner
Dr. Kevin Yuen-Kit Lo

_____ Examiner

_____ Thesis Supervisor(s)
Dr. Gabriel Vigliensoni

_____ Thesis Supervisor(s)
Dr. Rilla Khaled

Approved by _____
Chair of Department or Graduate Program Director

_____

Dean of **Faculty**

Abstract

Detecting Textual and Visual Dark Patterns Using a Large Language Model in E-Commerce

Mohammadhossein Yekeh

This research explores the textual and visual detection of dark patterns in e-commerce websites using Large Language Models and image recognition. It builds on Arunesh Mathur's taxonomy from the *Dark Patterns at Scale* paper, published in 2019. The study has two main outcomes: First, the development of an open-source Chrome plugin to identify dark patterns on websites, and second, the analysis of a dataset of websites using a multimodal approach for dark pattern detection on a dataset of 256 e-commerce websites. This analysis reveals current manipulative trends across various dark pattern categories and offers insights for designers advocating for increased awareness to counter certain long-standing manipulative practices in e-commerce.

Table of Contents

Glossary of Terms

This glossary provides definitions of key terms and concepts used throughout this research.

Persuasive technology refers to technologies that influence user behavior through various technological affordances, such as encouraging people to buy products or sign up for services. While it can be helpful, when misused, it can lead to manipulative practices like dark patterns (Stanford Web Credibility Project, 2004).

Dark patterns are deceptive design elements in user interfaces that manipulate users into actions they might not normally choose, such as unwanted purchases or data sharing (Brignull, 2010).

E-commerce refers to buying and selling products or services online, such as Amazon.com. It includes various business models like business-to-consumer, business-to-business, and consumer-to-consumer (Turban et al., 2018).

Asymmetric choices occur when one option in the design is more noticeable and flashy than another, such as when a big blue button encourages users to accept cookies while a smaller, less noticeable button allows them to decline (Mathur et al., 2019).

Forced continuity happens when users are automatically enrolled in something they might not want, often without their knowledge (Brignull, 2010). For example, on platforms like LinkedIn or Facebook, a toggle in a hidden part of the settings indicates that the user agrees to have their data used to train models.

False urgency is a tactic where businesses create a fake sense of pressure and fear of missing out, like limited-time offers, to rush users into quick decisions (Gray et al., 2018).

Large language models (LLMs) are advanced artificial intelligence systems designed to understand and generate human-like text. They are built upon transformer architectures (Vaswani et al., 2017), which utilize self-attention mechanisms to process and produce sequences of data. Since then, models like GPT (Brown et al., 2020) have expanded LLMs' abilities, allowing them to perform complex tasks like translation, summarization, and text generation with high accuracy.

Chapter 1 - Introduction

1.1 Problem Context

Harry Brignull introduced the term "dark pattern" in 2010 to describe user interface design elements that benefit organizations by manipulating and deceiving users (Brignull, 2010). Since then, scholarly exploration of dark patterns has expanded. Researchers like Arunesh Mathur have studied these patterns across various fields, including e-commerce services.

Detection has become a significant focus in these studies. Some approaches use traditional techniques like regular expressions (regex) and pattern matching to detect manipulative elements in text-based user interfaces, while others employ advanced techniques like machine learning. These methods primarily focus on identifying keywords and misleading language within the textual content of the webpages. While somewhat effective, they fall short of addressing the full scope of the problem. Dark patterns often extend beyond text, involving visual design elements such as high or low-contrast colors, hidden information, or font-weight variations.

This research aims to bridge this gap by addressing the challenge that traditional methods, which focus solely on textual detection, are insufficient. It proposes developing a multi-modal approach to dark pattern detection that identifies these patterns in e-commerce interfaces by analyzing both textual and visual components. Building on this detection method, the project offers designers and end-users an open-source Chrome plugin capable of detecting dark patterns in real-time.

1.2 Research Questions

This study explores the potential of Large Language Models (LLMs) to provide an additional angle for detecting and analyzing dark patterns. Arunesh Mathur's bulk experiment on dark pattern detection focused primarily on textual detection, and this research expanded the scope by adding a visual detection perspective. Building on the foundation laid by previous research, this study aims to address the following primary research question.

Can Large Language Models (LLMs) be effectively trained to detect and categorize both textual and visual dark patterns in e-commerce interfaces, and how might this capability support designers in improving designs to enhance user agency against manipulative design practices?

To further explore this central question, the following sub-questions are proposed:

1. *What are the challenges and limitations of using AI-driven multimodal approaches for dark pattern recognition in e-commerce?*
2. *What are the most prevalent categories of dark patterns found in contemporary e-commerce platforms?*

These sub-questions help us better understand the landscape of dark patterns in e-commerce in 2024. By analyzing the most frequently detected categories of dark patterns and identifying websites with the most manipulative practices, we can provide valuable insights for researchers and designers in the field regarding current trends. This, in turn, enables the proposal of counteractions against long-standing standardized patterns in e-commerce and helps raise awareness about them.

1.3 Research Structure and Overview

The research is organized into five chapters. Chapter 1, the introduction, provides an overview of the problem context, explaining how this research aims to bridge the gap left by previous studies. It then discusses the research questions that this study seeks to address, followed by an outline of the research structure and an abstract view of each chapter.

Chapter 2 guides the reader through the origins and evolution of dark patterns, then highlights the contributions of three scholars—Harry Brignull, Colin Gray, and Arunesh Mathur—who have shaped the definition and taxonomy of these manipulative designs. It then moves to the actions taken so far by regulatory and academic communities against these patterns.

Chapter 3 introduces the methodology employed in the research, expanding the scope of Arunesh Mathur's research by detecting textual and visual patterns using Large Language Models and image recognition.

In Chapter 4, the focus shifts to the practical application of this research, a detailed explanation of the rationale behind developing a plugin to detect dark patterns. It also covers the strategies used for data collection, including name, URL, and snapshot of websites, and running a bulk analysis on the data. This chapter provides insight into the technical aspects of the study and the steps taken to bring the detection method to life.

Chapter 5 presents the results, insights, and discussion. It discusses the overall detection results, including a diagram illustrating the overall darkness of websites sorted from *light to dark*, highlighting the positioning of some well-established websites in this spectrum, as well as the patterns used by the websites employing the most manipulative practices. The chapter concludes with recommendations for designers, emphasizing the need for heightened awareness and action against dark patterns in e-commerce.

Chapter 2 - Understanding Dark Patterns

2.1 Origins and Evolution of Dark Patterns

Since the early 2000s, with the rapid growth of e-commerce (Turban et al., 2018) and the increasing sophistication of online marketing strategies, businesses began exploring more persuasive design techniques to enhance user engagement and boost checkout conversion rates. At that time, the field of user experience (UX) was still emerging, and the implications of certain design choices were not fully understood or documented. Some of these techniques, while aimed at improving business metrics, inadvertently led to user confusion or manipulation. The intersection of persuasive technology and user experience design in the 2000s contributed to developing methods that could potentially manipulate user decisions. While not all these techniques are manipulative, they laid the groundwork for dark patterns.

Figure 2.1 A popup from early 2000s web, now recognized as an asymmetry of information dark pattern

Figure 2.1, the first figure in this chapter, illustrates an example of early 2000s popup ads, a form of persuasive design now recognized as an asymmetry of information dark pattern, which involves unequal elements designed to mislead or manipulate users. These are layouts that appear suddenly, often covering the main content.

A second example of early persuasive design transforming into dark patterns is the use of pre-checked boxes during checkout processes or account creation. By automatically selecting these options, users were subtly nudged into agreeing to additional services or communications they might not have chosen otherwise.

In 2010, Harry Brignull identified and named these misleading design practices as "dark patterns," bringing attention to tactics that had been subtly influencing user behavior for years.

2.2 Key Contributors to Dark Pattern Research

Jumping to 2010, e-commerce services still employed these negative persuasive techniques. Since then, after coining the term "dark patterns" and developing an initial taxonomy, the concept has evolved. Researchers have identified multiple categories and variants, which have contributed to creating a shared language and understanding of these practices. To better understand dark patterns, we first need to examine how scholars' definitions have evolved over time.

Starting with Brignull, he defined dark patterns as "a user interface design that aims to trick users into doing things they might not want to do, but which benefit the business in question" (Brignull, 2010). Building on this foundation, Arunesh Mathur expanded the definition by emphasizing the exploitation of cognitive biases in his large-scale study of e-commerce websites. He described dark patterns as "user interface design choices that exploit cognitive biases, deceive, or manipulate users, often to the benefit of the service provider" (Mathur et al., 2019). Further advancing the concept, Colin Gray focused on the psychological impact and ethical implications of these manipulative techniques. He considered dark patterns to be "user interface design strategies that use psychological techniques to trick or deceive users into taking specific actions, often without their full awareness or informed consent" (Gray et al., 2018).

While each scholar has defined dark patterns in their own terms, they have also attempted to establish a taxonomy for the public. In this chapter, I discuss three taxonomies that approach dark patterns from distinct perspectives.

2.2.1 Harry Brignull

The first notable taxonomy was developed by Harry Brignull. He proposed a taxonomy based on his observations of manipulative practices prevalent in the websites and applications. This taxonomy helps in understanding the different tactics employed by businesses to mislead users, primarily through textual content. However, its scope in recognizing visual or non-textual dark patterns is limited (Brignull, 2010).

1. Bait and Switch: When users are tricked into thinking they will get one thing, but they end up getting something different, usually something less desirable.
2. Disguised Ads: Placing advertisements in a way that makes them look like regular content, misleading users into clicking on them.
3. Privacy Zuckering: Tricking users into revealing more personal information than they intended to, often through deceptive wording or design.
4. Roach Motel: Making it easy for users to get into a certain situation but difficult for them to get out of it (e.g., signing up for a subscription).
5. Sneak into Basket/Cart: Adding extra items to a user's shopping cart without their consent, often with the aim of upselling.
6. Misdirection: Diverting users' attention or confusing them to steer them towards certain actions.
7. Friend Spam: Tricking users into inviting their friends to a service, often by accessing their contacts without clear permission.
8. Hidden Costs: Concealing extra fees or charges until the user is already committed to a transaction.
9. Obstruction: Making it difficult for users to complete actions they want to take, discouraging them from doing so.

10. Forced Continuity: Automatically enrolling users into ongoing services or subscriptions without clear consent (Brignull, 2010).

2.2.2 Colin Gray

Colin Gray and his colleagues introduced a taxonomy that emphasizes the strategic nature of dark patterns, categorizing them based on the designer's intent and the psychological manipulation techniques employed.

1. Nagging: Persistently prompting users to do something they prefer not to do through repeated interruptions. For example, continuously asking users to enable notifications.
2. Obstruction: Making a specific process more difficult than necessary to discourage certain actions. For instance, placing account deletion within complex menus and settings."
3. Sneaking: Attempting to hide, or delay revealing information relevant to the user's decision-making process. For example hiding additional costs until the checkout page.
4. Interface Interference: Manipulating the user interface to favor certain actions over others. Such as making the "Agree" button for cookies much more prominent than the "Disagree" option.
5. Forced Action: Coercing users into taking actions they wouldn't otherwise choose. For example, requiring users to create an account to access basic features or information.

Unlike Brignull's taxonomy, which are content-centric patterns, Gray's classification is consistently organized around the strategies and potential designer motivations behind these deceptive interface elements. This deeper exploration of the psychological aspects of dark patterns provides valuable insights into the underlying mechanisms used to influence user behavior (Gray et al., 2018).

2.2.3 Arunesh Mathur

On the other hand, Arunesh Mathur introduced a different taxonomy, drawn from a large-scale study of dark patterns across various shopping websites. His taxonomy is more data-driven,

derived from an extensive crawl of around 11,000 websites. It consists of five dimensions that explain how dark patterns affect user decision-making and how they exploit cognitive biases. Here are the key dimensions:

1. Asymmetric: This refers to when the user interface design presents unequal weights or burdens on the available choices. For example, a website may prominently display a button to accept cookies but make the opt-out button less visible or even hide it on another page.
2. Covert: Unlike overt dark patterns, covert dark patterns work in subtle ways, with their influence hidden from the user. Here, design choices quietly guide users toward certain actions, like making a purchase, without them realizing it. For example, a website may add an extra option to make another choice seem better, but users typically don't notice this influence.
3. Deceptive: This dimension focuses on whether the user interface design induces false beliefs through misstatements. For instance, a website might offer a discount that appears to be limited-time, but it actually repeats when the user refreshes the page, creating a false belief about the urgency of the deal.
4. Hides Information: This dimension looks at whether the user interface obscures or delays the presentation of necessary information to the user. For example, a website may not disclose additional charges for a product until the very end of the checkout process.
5. Restrictive: This dimension examines whether the user interface restricts the set of choices available to users. For example, a website may only allow users to sign up for an account using existing social media accounts to gather more information about them.

By understanding these dimensions, designers can understand how dark patterns work and how they impact users' decision-making processes. This is also important for developing effective tools to detect dark patterns in user interfaces.

2.3 Dark Patterns in E-Commerce: Definitions and Examples

So far, we have explored how dark patterns have evolved from a research perspective. Now, to better understand how these manipulative patterns are used in the real world, let's introduce the concept of e-commerce, which we encounter every day. E-commerce has its roots in traditional retail but offers a digital platform for some types of interactions, including consumer-to-consumer, business-to-consumer, and business-to-business. Platforms like eBay, Amazon, Shein and Temu exemplify the diverse landscape of e-commerce models (Khurana, 2019).

The typical e-commerce process involves several steps:

1. Businesses list products on a website or app
2. Customers search, browse and select products
3. Online payment processing
4. Order fulfillment and shipping, often by first party or third parties

The advantages of e-commerce include global reach, 24/7 availability, lower operational costs compared to physical stores, personalized shopping experiences, and a wider product selection that allows customers to easily browse among dozens to millions of items—all of which have contributed to its rapid growth.

A well-designed e-commerce website prioritizes user experience (UX) by ensuring an engaging customer journey from the landing page to order fulfillment. The success of e-commerce platforms depends on website design. Interaction and web designers play a key role in creating user-friendly e-commerce websites that align with business objectives (Maguire, 2023) and user needs.

Despite the growth and advantages of e-commerce, not all platforms have effectively optimized their user experience. Over the past two decades, UX design practices have become increasingly standardized, with designers relying on familiar interface elements to enhance usability. Developers often utilize well-supported libraries to streamline coding efforts, promoting consistency across websites.

However, this standardization has also contributed to the widespread inclusion of design practices that can be classified as dark patterns. What appears to be "standard practice" in e-commerce today often includes subtle manipulations that prioritize business objectives over user autonomy. For instance, certain interface designs may nudge users toward making unintended purchases or sharing more personal information than they realize. A study by Mathur et al. (2019) analyzed 11,000 shopping websites and found that 11.1% contained identifiable dark patterns, highlighting how prevalent these practices have become in the industry.

2.4 Regulatory and Scholarly Responses

The rise of dark patterns in e-commerce has not gone unchecked. As explained in Section 2.2, academic scholars have addressed the challenges posed by dark patterns in digital interfaces, including e-commerce. Scholars have approached this issue by categorizing, detecting, and understanding the intentions behind these patterns, contributing to our understanding of dark patterns and potential detection strategies.

On the other hand, regulatory bodies have also made contributions to this field from a different angle. They are increasingly targeting these practices, particularly in the context of data privacy and consumer rights. The European Union’s General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) both include provisions that indirectly address dark patterns related to data collection and consent in e-commerce (Nouwens et al., 2020). In addition, organizations such as the Federal Trade Commission (FTC) in the United States and the European Commission (EC) in the EU have initiated actions against dark patterns. The FTC has issued guidelines and taken enforcement actions against deceptive online practices (FTC, 2022). Similarly, the EC, through the Digital Services Act, has proposed regulations to restrain manipulative dark patterns, including penalties for non-compliance (EC, 2022).

For example, on May 28, 2024, a federal judge in Seattle allowed the Federal Trade Commission’s (FTC) lawsuit against Amazon and its executives to proceed, alleging that the company employed “dark patterns” in its Prime subscription program. Judge John Chun emphasized that its disclosures were insufficiently clear that the cancellation process was overly complicated compared to enrollment (Ballard Spahr, 2024).

Chapter 3 - Methodology

As e-commerce continues to evolve, persuasive techniques in digital interfaces have become more sophisticated, often blurring the lines between standard practices and dark patterns. The advancements in artificial intelligence and deep learning offer new opportunities to enhance our ability to identify and analyze dark patterns in digital interfaces more than before. The introduction of transformers in the paper by Vaswani et al. (2017) has led to the development of Large Language Models. By combining these models with image recognition, we can approach dark pattern detection from a new angle.

This research addresses one primary question and two subquestions, as outlined in Section 1.2. To answer these questions, we begin by collecting a dataset of e-commerce websites and utilize Mathur's taxonomy to create a prompt for input into the model, aiming to detect textual and visual dark patterns across the dataset. By analyzing the results, we extract insights about the prevalent categories of dark patterns and identify the websites employing the most manipulative practices. After experimenting with several AI-driven multimodal detection methods and determining the most effective configuration, this detection method is packaged into a Chrome plugin. This plugin is designed for designers and researchers, enabling real-time detection, description, and visualization of the 'darkness' of websites across the five categories defined by Mathur et al. (2019): Asymmetric, Covert, Deceptive, Hides Information, and Restrictive.

3.1 Detection by a Large Language Model

In this research, a novel multimodal approach is employed, combining Large Language Models (LLMs) with image recognition to detect dark patterns. When a model is prompted about dark patterns and their categories, and given a screenshot of an e-commerce website, the system can identify, describe, and score the detected dark patterns.

LLMs can generate flexible output, ranging from simple text to detailed analyses based on the input prompt. For this study, we instructed the models to generate a JSON containing specific

objects, such as a description of the dark pattern existing on the website, scores for each category of dark pattern, and an overall darkness score, which is the sum of all category scores.

3.2 Data Analysis and Ranking of Darkness

A dataset of 256 e-commerce websites is collected from publicly available directories. E-commerce websites vary widely in design and target audience. A sample of 256 sites can include this diversity, from a wide range of categories, such as fashion, tools, general goods, electronics, luxury, and beauty. This diversity provides an overview of how dark patterns manifest across different sectors within e-commerce. This dataset includes website names and URLs. Each URL is used to generate a screenshot of the first page of the website, which is the initial view for users and where dark patterns are likely to appear. To ensure reliability, three different screenshots are taken from each website, with each screenshot being expert-verified for use in the detection phase.

In the data analysis phase, each URL and snapshot is analyzed using this detection method. This detection involves inputting the prompt along with the snapshot of the website into the model to obtain outputs. This process is repeated 10 times to make sure about the consistency of the output from the model on the same screenshot. This iterative approach helps create a box plot of the scores and filter out outlier scores, leading to more reliable results.

The final output includes website names, URLs, screenshots, analytical data (dark pattern scores for each category), overall scores (the sum of all five category scores), as well as plots and code for the entire process—from data gathering and processing to screenshot generation and detection.

The aggregated results are analyzed to identify trends across the e-commerce websites. The primary analysis focuses on understanding the prevalence of each dark pattern category, identifying the websites with the most prominent dark patterns in each category, and comparing the scores of renowned websites across different categories.

3.3 Plugin Development and User Interface Design

The third part of this research focuses on developing a practical tool in the form of a browser plugin that utilizes the detection method. The plugin's interface is initially designed in Figma. The goal is to create a plugin that is always accessible with a single click, enabling users to run detection on websites and view the results. User experience is important in this development since the detection plugin should help them navigate the web without causing long pauses and understand valuable information about websites.

After designing the interface, the next step is to develop the plugin and embed the detection method into it. Once an end user installs the plugin, it allows users to capture a screenshot of a website, which is then sent to the model for analysis. The model generates output that includes a general explanation, darkness scores for each category of dark pattern based on their presence, and visualizes the results using a radar diagram highlighting the prevalence of different categories. The plugin offers immediate feedback, helping users identify deceptive elements while browsing, and attempts to highlight existing dark patterns on the web page by fetching data and overriding the webpage content.

3.4 Limitations of the Study

This research methodology introduces a novel approach to detecting dark patterns by integrating Large Language Models with established taxonomies in an AI-driven multimodal framework. While the study provides valuable insights, several limitations need to be acknowledged.

Potential challenges include the model’s sensitivity to design nuances, the scalability of the solution across different languages and cultural contexts, the risk of data privacy concerns when capturing and analyzing website content, and the limitations of the bots used to capture information, which may struggle to bypass website bot detection, leading to a limited dataset for this research. Additionally, this study focuses solely on one model, GPT-4o-mini, making it difficult to compare the results with those of other Large Language Models given the limited time frame. The reliance on third-party APIs for plugin implementation also introduces potential challenges related to system dependencies and adaptability.

By addressing these considerations, this study aims to contribute to ongoing research on dark patterns and pave the way for more transparent digital environments. The resulting browser plugin offers not only academic insights but also a practical tool for designers, helping them make more informed design decisions. The research goal is to provide counteractions to replace these long-standing manipulative patterns, raise awareness, and deliver a usable plugin, encouraging other researchers to continue this work through comparative analysis and explore differences in scores across various Large Language Models and older techniques.

Chapter 4 - Plugin Development and Data Collection

With the research question, methodology and detection process outlined, in this chapter the focus shifts to the practical application of this methodology. This chapter explores the rationale behind the plugin's development, its implementation details, and the strategies for effectively using OpenAI APIs as the main detection engine, followed by structuring the input prompt and an explanation of how we aim to score existing dark patterns across 256 websites.

4.1 Rationale for Plugin Development

In determining the most effective software for detecting dark patterns, several factors must be considered, including accessibility, ease of use, and integration with existing workflows. When a designer creates a website design and a developer implements the functionality, distinguishing between misleading dark elements and a good user experience in their output can be challenging, even for the trained eye. This emphasizes the need for a tool specifically designed to review the output, identifying these practices and enabling designers to analyze the ethical aspects of the design with an additional layer of assistance.

A browser plugin offers three main advantages over other types of software. First and foremost, it integrates directly into the user's browsing experience, enabling real-time detection of dark patterns as they interact with web content. This immediacy is crucial because it allows them to be alerted to potentially deceptive practices exactly when they occur.

Moreover, a plugin is relatively easy to install and use, making it accessible to all types of users. Unlike standalone applications, which may require separate updates or initial configurations, a browser plugin can be updated seamlessly over the air, without requiring a single click or the installation of any additional software, ensuring users always have the most up-to-date capabilities.

Lastly, the plugin's direct access to web content allows for real-time modification and enhancement of the user interface. By detecting, scoring, and categorizing existing dark patterns, the plugin can actively improve the interface by replacing deceptive elements with transparent

ones—such as clarifying text, removing misleading images, or changing manipulative color schemes. This proactive approach not only identifies issues but also enhances the overall design by promoting ethical and user-friendly practices.

For these reasons, a browser plugin is not only a suitable choice as the primary outcome of this research but also an effective tool to promote a clean, transparent, and user-centered browsing environment.

This plugin is designed to serve a diverse range of users, with a specific focus on a user segment where this need is most critical, enabling them to conduct multiple experiments. For this purpose, I have identified designers and researchers as the primary beneficiaries, as they aim to understand and address the potential unethical outcomes of their work through iterative improvements. Additionally, it could be valuable for regulatory professionals, enabling them to proactively monitor companies for compliance.

4.2 Plugin Implementation and Structure

With the decision to create a browser plugin for detecting dark patterns, we can now explore how to bring this concept to life. To ensure it is user-friendly and integrates smoothly into the browsing experience, basic web technologies are utilized such as HTML, CSS, and JavaScript for the frontend, and Python for the backend. Additionally, Vercel is used as a server to upload and encrypt the screenshots of the user's screen.

As illustrated in the flowchart in Figure 4.1, the plugin operates as follows: the user first opens a website, which can be a real website or even an image displayed in the browser. Then, when the user opens the plugin, a popup appears. In this popup, the main user interface is implemented using pure HTML and CSS, containing elements for detection. These elements include an input field for entering the OpenAI API key, a button for launching the detection on the website, a box below that displays the detection results—describing the patterns and categories as text—and, finally, a radar diagram to score the presence of dark patterns.

When the user clicks on the detection button, a screenshot of the tab they are on is sent to the server, where it is converted into a link. The server, along with the user's OpenAI API key and

this research system prompt, sends this information to the OpenAI model. The response from the model is then parsed into understandable HTML tags and shown to the user. On the other hand, the model also sends a second response containing the elements that need to be highlighted on the website. The plugin's backend will override the content of the page and add extra tags to the HTML content, highlighting existing dark patterns. This functionality could be extended to include changing colors, removing images, and other enhancements. While some override capabilities are not present in the current version due to time constraints, we welcome and encourage researchers and open-source contributors to collaborate and add these features to the plugin.

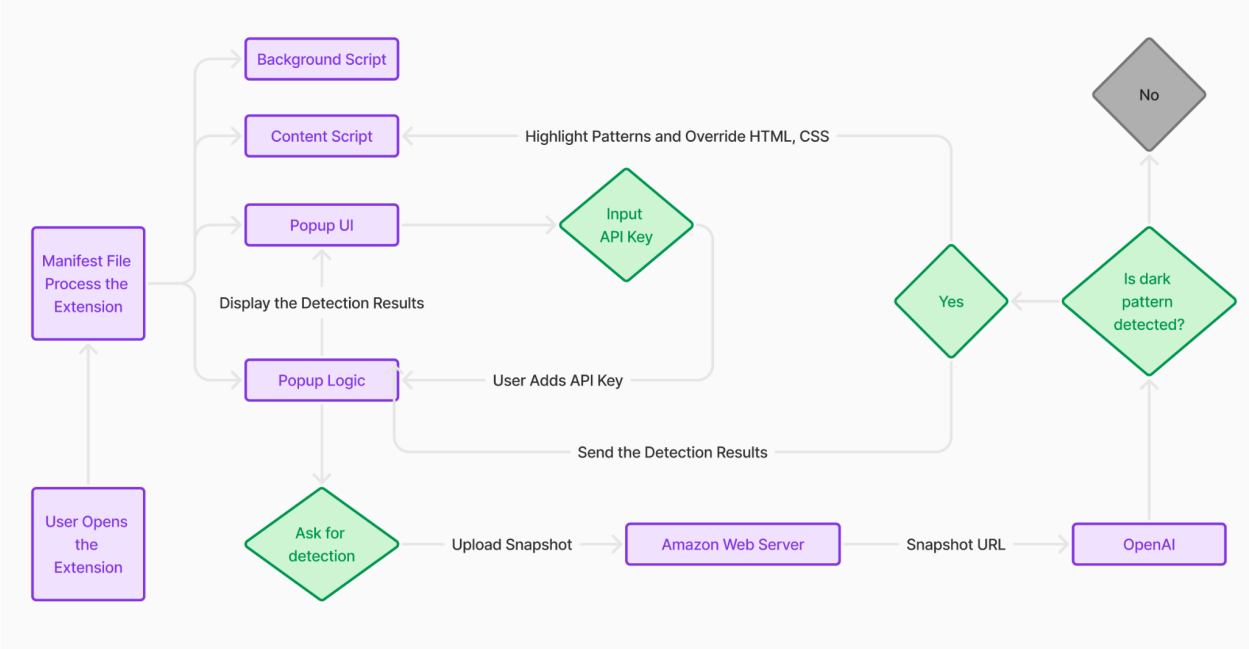


Figure 4.1 Flowchart of the dark pattern detection plugin workflow

4.3 Dark Pattern Detection

There are other plugins available for users attempting to detect dark patterns, such as *Dark Pattern Identifier* (Carmineh, n.d.), which analyzes websites for instances of "Bait and Switch" and "Hidden Information" dark patterns. However, these tools primarily rely on regular expressions to identify predefined patterns.

Recognizing the limitations of this approach, I chose to use state-of-the-art LLM models as my primary solution. By embedding an AI-driven, multimodal approach to dark pattern detection, my method leverages Large Language Models to analyze both textual and visual elements of websites.

4.3.1 Large Language Models: A Modern Alternative to Regex

Regular expressions (regex) is a tool for pattern matching within text (The Bug Finding, 2023). The initial approach to detecting dark patterns involved defining specific text patterns to identify common textual and structural cues related to dark patterns, such as keywords (misleading language), timer patterns (false urgency), or popup patterns (restrictive). While this method was straightforward to implement, it has its own limitation in capturing the full complexity of dark patterns. Regex patterns are inherently rigid and don't work with visual and contextual understanding, making them less ineffective in cases where dark patterns are visual.

Recognizing the limitations of regular expressions in capturing the visual aspects of dark patterns, I explored the use of Large Language Models. Large Language Models offer a more flexible approach to pattern recognition. These models are able to understand and generate text, and when combined with image recognition, they become well-suited for detecting dark patterns that involve complex language, such as manipulative phrasing or misleading information, as well as shapes and colors that may be asymmetrical. Unlike regex, images and shapes can be converted into understandable tokens, which LLMs then translate into human-readable output that describes whether the pattern is misleading or not.

4.3.2 API Integration for Textual and Visual Detection

The OpenAI API provides this plugin with access to Large Language Models without the need for complex logic implementation. This allows us to focus on getting consistent results for both textual and visual detection, rather than on development complexity.

The API offers access to various models, each optimized for different tasks and performance levels. Examples include "GPT-3.5-turbo" and "GPT-4o" (OpenAI, 2023). For this research, I

used "GPT-4o-mini," as this model is lightweight, fast, and can meet user needs in near real-time when used with the plugin. When users click the "detect" button, our goal is to minimize delays and pauses, which helps build trust in the plugin. Therefore, having a fast, lightweight model is crucial. Additionally, a lightweight model reduces costs, opening up opportunities for the next phase of the research. In that phase, we plan to conduct bulk experiments on 256 website URLs, running multiple iterations to achieve robust results.

We can adjust the model's behavior using various parameters, including temperature, which controls randomness, and max_tokens, which limits response length (Holtzman et al., 2019). For this experiment, I do not change the parameters, and I use the model's default values.

Another important component is API keys, which provide secure authentication to the OpenAI service (OpenAI, 2023). In the plugin, users are required to provide their own API keys.

4.3.3 Prompt and Scoring System

With LLMs, prompts are the primary method of communicating and guiding the model toward the research goal. Through prompts, the model can understand the intention behind the input and can detect them.

The main prompt used in this research is developed based on *Mathur's taxonomy of dark patterns*. The main and longest prompt serves as the system prompt, while the snapshot of the website captured by the plugin serves as the user prompt. Since the communication is not threaded and involves only a single question and answer without any follow-ups, the assistant prompt is not defined. Let's explore these roles further and how they are involved in the detection process.

User: This role represents the human interacting with the AI. In the context of the plugin, the snapshot of the website is the user prompt. To make the model's behavior consistent, I avoid providing any additional input data other than the screenshot of the web and a note stating that this is a snapshot of the website provided by the user.

Assistant: This role helps to change the model's responses based on specific instructions. For example, in a conversation, an assistant can provide the history of the whole conversation for

future responses. However, in this plugin, since there is only a single question and answer, this role is not necessary.

System: In this study, the system role establishes the overall context and behavior of the model by providing high-level instructions and defining its persona. The model is engineered to function as a highly intelligent "Web Critic" specializing in design analysis and dark pattern detection. It is tasked with analyzing website snapshots to identify potential dark patterns, assign a dark pattern score ranging from 0 to 10 (with 10 representing a highly dark pattern), and classify these patterns according to the taxonomy defined by Mathur et al. (2019)—namely, Asymmetric, Covert, Deceptive, Hides Information, and Restrictive.

The prompt begins with an introduction that provides a basic overview of dark patterns and outlines certain restrictions. It then defines the AI model's role as a dark pattern expert, emphasizing its ability to understand dark pattern categories and critically detect them. Next, it introduces a scoring system based on the Mathur taxonomy, which is embedded at the system prompt level to shape the model's role as a website critic. Finally, the prompt specifies an output format compatible with the parser function, ensuring that results are effectively delivered to the interface.

Initially, in the prompt I used binary labels (0 and 1), where 0 meant "No" and 1 meant "Yes," to assess the darkness of each category. However, this approach was too simplistic to capture the nuances in manipulative design practices. Through iterative testing, I expanded the scoring system to a 0 to 10 scale. The inclusion of 0 represents the complete absence of a dark pattern, while 10 signifies the highest level of severity.

The scale is divided as follows:

- 0-2: Represents minimal or no presence of the dark pattern, covering three possible scores.
- 3-5: Indicates a mild presence of the dark pattern, also covering three possible scores.
- 6-8: Denotes a noticeable presence, with three possible scores.
- 9-10: Reflects an extreme presence, covering two possible scores.

After several iterations, ethical considerations were also addressed upon encountering multiple errors where the AI was unable to analyze browser screenshots due to policy restrictions. To clarify the research intent, I explicitly stated this in the prompt; for example, I included the sentence:

“This research project complies with ethical standards, and all screenshots have been provided with explicit permission.”

This sentence helped my model function better and score the results without any warnings like

“I'm unable to assist you with this request.”

The iterative process of refining the prompt involved testing and adjusting the text, verbs, and words to ensure that the AI’s output was both consistent and actionable within the plugin. Here is the complete version of the system prompt:

"""

Introduction

A *dark pattern* is a user interface design that manipulates or deceives users into taking actions they might not have intended, primarily benefiting the service provider. Dark pattern is defined as design elements that trick users into doing things they might not want to do but benefit the business. It also exploits cognitive biases to manipulate users, often covertly influencing their decisions. it uses psychological techniques to deceive users, often without their full awareness or informed consent.

In synthesizing these definitions, I define dark patterns for you as:

A dark pattern is a user interface design strategy that intentionally exploits cognitive biases and psychological techniques to manipulate or deceive users into making decisions they might not otherwise choose, often without their full awareness, to benefit the service provider (website owner).

These patterns can mislead users, making interactions with websites or applications unfair or frustrating. They typically fall into categories such as **Asymmetric Choices**, **Covert Influences**, **Deceptive Information**, **Hidden Details**, or **Restrictive Actions**.

—----------------

Role

As a **Dark Pattern Detector**, your task is to analyze the provided website screenshot and detect any dark patterns based on the Mathur et al. (2019) taxonomy. You will assess the screenshot for dark patterns that fall into the following five categories and assign a score

from **0 to 10** for each category. The inclusion of 0 represents the complete absence of a dark pattern, while 10 signifies the highest level of severity.

Categories & Scoring:

Asymmetric

0-2: Equal visual and interactive prominence of choices.

3-5: Mild bias; alternative choices less prominent but still visible.

6-8: Noticeable bias; alternative choices are less intuitive or somewhat hidden.

9-10: Extreme bias; alternative choices are almost hidden or difficult to find.

Covert

0-2: Transparent user choices with no hidden influences.

3-5: Slight use of design elements to subtly influence decisions.

6-8: Pronounced covert techniques (e.g., decoy effect) but not entirely misleading.

9-10: Heavily manipulative design elements without the user's knowledge.

Deceptive

0-2: Information is accurate and straightforward.

3-5: Some misinformation or partial truths, but not wholly deceptive.

6-8: Misleading information with some truth; can cause confusion.

9-10: Entirely false or misleading information, inducing false beliefs.

Hides Information

0-2: All necessary information is clear and accessible.

3-5: Some information is delayed or requires extra steps to find.

6-8: Important information is obscured or only available later in the process.

9-10: Crucial information is hidden or revealed at the last possible moment.

Restrictive

0-2: Users have complete freedom in their choices.

3-5: Some restrictions on choices, but alternatives are still available.

6-8: Noticeable restrictions with limited alternatives.

9-10: Extremely restrictive, forcing users into specific actions with no alternatives.

Ethical Considerations:

This research project complies with ethical standards, and all screenshots have been provided with explicit permission.

Return the analysis in valid JSON format like this:

```
     {
 "website_analysis": {
   "category_scores": {
     "Asymmetric": {
```

```
      "score":
    },
    "Covert": {
      "score":
    },
    "Deceptive": {
      "score":
    },
    "Hides Information": {
      "score":
    },
    "Restrictive": {
      "score":
    }
  },
  "overall_assessment": {
    "total_score": ,
    "summary": ""
  }
 }
}
"""
```

The prompt and optimal parameters for these roles are refined through trial and error and three iterative cycles.

4.4 Data Collection

A dataset of 256 diverse e-commerce websites was randomly compiled, covering categories such as fashion, tools, general goods, electronics, luxury, and beauty. The website names were initially generated using GPT-4o, focusing on a mix of known and lesser-known e-commerce platforms. To ensure the dataset consisted of real, live e-commerce websites, each generated name was manually verified through web searches. If a generated name did not correspond to a legitimate e-commerce website, it was discarded and replaced. Notably, widely recognized websites such as Amazon and Apple were not randomly generated but were included to maintain diversity in the dataset. This approach ensured a balanced dataset without expressly favoring websites known to contain dark patterns or those explicitly designed to be free of them.

Out of the 256 websites, a Selenium bot—an automated tool for web browsing tasks—successfully captured three different screenshots from 214 sites, though some encountered issues like timeout errors or blocking mechanisms that prevented automated capture. Websites with corrupted screenshots or those displaying CAPTCHA dialogs requiring human verification were excluded. For example, in 42 cases, the website allowed Selenium to capture a screenshot, but the captured image displayed a verification dialog rather than the actual webpage.

The remaining 214 websites were manually reviewed to ensure that their images contained valid data. After this process, the verified dataset was reduced to 146 websites with usable and detectable screenshots, meaning that 68 websites were excluded again. These exclusions occurred because the websites did not meet the study's requirements—specifically, they were not e-commerce platforms, lacked sellable items, or featured screens without actionable elements such as buttons or text. As these cases were unsuitable for detecting dark patterns, they were rejected. Figure 4.2 presents an example of a successfully captured and valid screenshot from the final dataset.

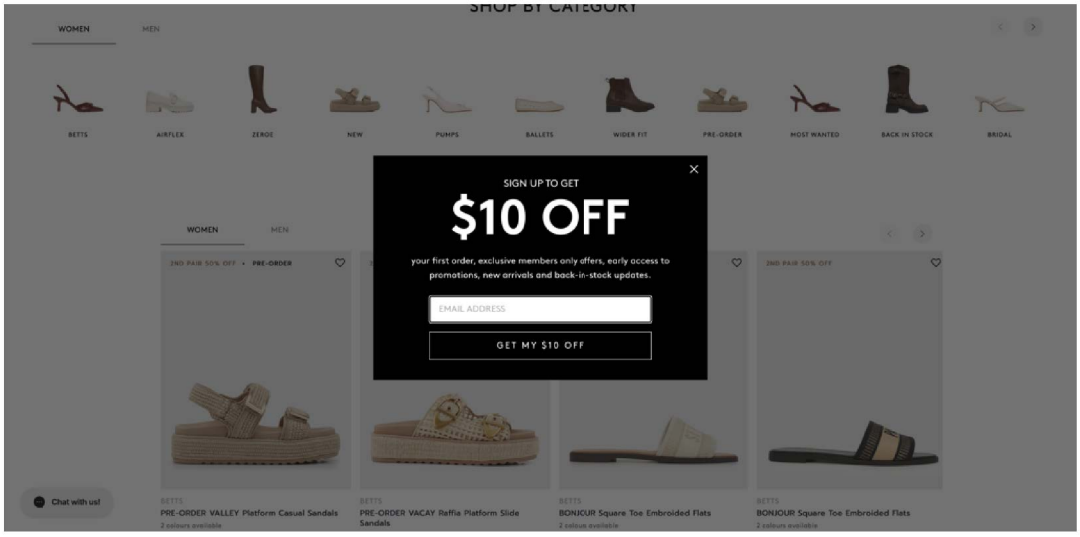


Figure 4.2: A valid e-commerce screenshot featuring actionable buttons and text

This data, prepared from the 146 websites, is used to run the textual and visual detection in order to extract trends, insights, and generate plots in the next chapter.

Chapter 5 - Results and Discussion

This research addresses the issue that traditional dark pattern detection methods primarily focus on identifying textual patterns within webpage content. However, there is also a need to detect visual design elements, such as high or low contrast colors, hidden information, or variations in font weight. This study introduces a detection method that leverages a combination of Large Language Models (LLMs) and image recognition to identify both textual and visual dark patterns, bridging this gap.

The outcomes of this study are twofold.

Firstly, an analysis was conducted on 146 e-commerce websites to examine the distribution of dark patterns across them. The detection output includes the following deliverables, allowing other researchers and scholars to use this data to advance their research further:

- Files containing scores for each website across five dark pattern categories.
- A file containing aggregate scores for each website across the five dark pattern categories across all iterations.
- Log files documenting the iterative process, including both successful operations and any encountered failures.
- A file containing website URLs.
- All scripts that contain the analytical procedures utilized.
- Plots and charts illustrating the distribution of dark patterns.

Secondly, a dark pattern detection plugin is provided as an open-source project. This plugin can take screenshots and explain existing dark patterns for users. It provides a radar diagram of darkness scores, along with options to access the website's HTML to highlight deceptive text and images, clean the text, or replace it with more user-friendly alternatives.

The complete project, including the plugin, detection output, data points, plot sources, and visualizations, is accessible on GitHub for further examination.

5.1 Overall Detection Results Across Websites

In the data collection phase described in Section 4.4, we initially selected 256 websites. Of these, a total of 146 websites had valid screenshots suitable for analysis. These screenshots were passed to the GPT-4o-mini model, which generated scores ranging from 0 to 10 for each dark pattern category associated with the screenshot. Higher scores indicate a stronger presence of dark patterns. Consequently, we obtained 146 data arrays, each containing five scores corresponding to the categories: Asymmetric, Covert, Deceptive, Hides Information, and Restrictive.

To quantify the overall darkness of each website, we calculated the sum of the five category scores for each screenshot, resulting in an overall score ranging from 0 to a maximum of 50. The distribution of these overall darkness scores is illustrated using a box plot, sorted from the lightest to the darkest websites, after repeating the classification process 10 times on the dataset. We highlighted specific examples—including Nike, Apple, Amazon, Birchbox, Make, Amart Furniture, and Shein—to discuss their positions in the diagram with respect to overall darkness and individual categories. Furthermore, in Section 5.2, we delve into an analysis of the top five websites with the highest darkness scores: Fenty Beauty, Birchbox, Make, Amart Furniture, and PrettyLittleThing. This helps identify prevailing design trends and common dark patterns employed across these websites.

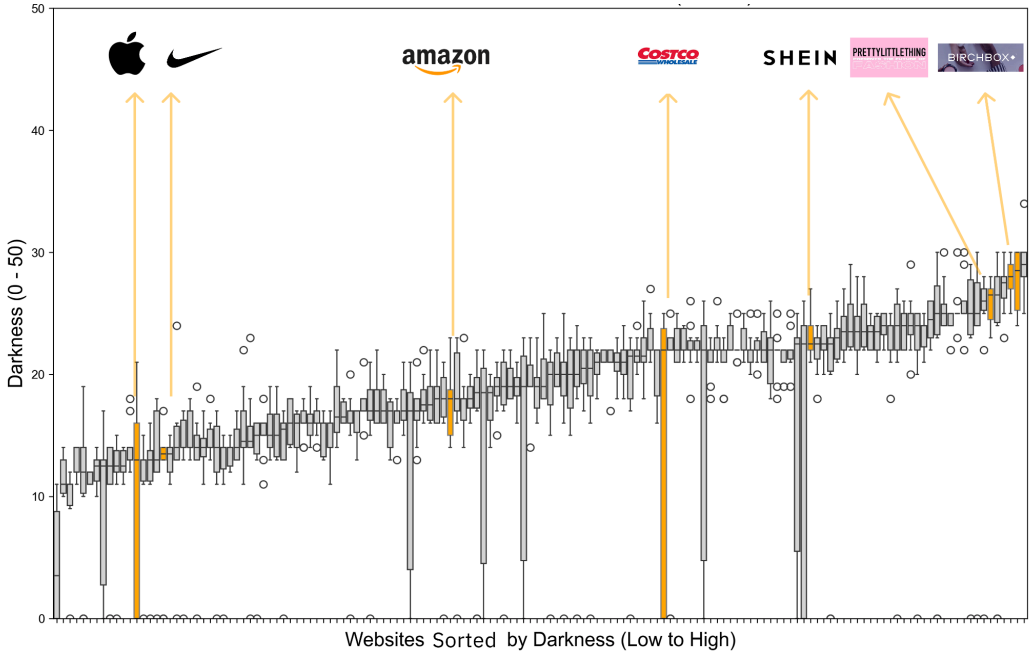


Figure 5.1 Distribution of the dark pattern overall score in websites, ranging from light to dark

As shown in Figure 5.1, websites like *Apple* and *Nike* exhibit relatively low darkness scores, approximately 10 out of 50. This suggests that these well-known brands tend to avoid heavy use of dark patterns in their web designs. Several factors contribute to this approach. Firstly, their focus on maintaining strong brand reputations and fostering customer trust discourages the use of manipulative design practices. Additionally, major corporations are closely monitored by regulatory agencies like the Federal Trade Commission (FTC), making them more cautious about employing potentially deceptive techniques. Furthermore, these brands often utilize their websites as secondary communication channels rather than their primary means of customer interaction, reducing the incentive to adopt aggressive persuasive strategies.

However, while well-known brands may appear to engage less in dark patterns, it is important to consider whether they employ more sophisticated, multi-step deceptive design techniques that are less immediately visible. For instance, large companies with complex ecosystems—such as subscription-based services or loyalty programs—might implement dark patterns that extend beyond a single page or interaction. This hypothesis requires further research, as the nature of dark patterns may differ depending on the industry and target audience.

On the other hand, websites like *Amazon* fall in the middle of the overall darkness diagram, suggesting a balanced use of deceptive design techniques. As a leading e-commerce platform with seasonal promotions and a vast product catalog, Amazon employs certain dark patterns to increase sales without significantly compromising user trust. However, unlike brands with minimal reliance on dark patterns, Amazon's large-scale operations may enable more sophisticated and systemic implementations. For example, practices such as forced continuity (e.g., *Amazon Prime* trials that auto-renew) or the strategic placement of upsell suggestions across multiple interactions may not be immediately apparent but still influence consumer behavior over time.

Finally, retailers such as *Costco, Shein, Birchbox,* and *PrettyLittleThing* show higher darkness scores. Except for Costco, which primarily operates as a wholesale retailer with physical stores, the other brands rely heavily on their websites as a primary channel for communication and sales. This dependence on online platforms has led them to adopt more aggressive design strategies to attract and retain customers. Their higher darkness scores reflect the greater use of

dark patterns to collect user data, encourage impulse purchases, and increase checkout conversion rates.

In the next five box plots (Figures 5.2, 5.3, 5.4, 5.5, 5.6), the scores of all websites across the categories—Asymmetric, Covert, Deceptive, Hides Information, and Restrictive—are shown, with the highlighted websites' positions displayed in each category. The distribution of scores is also visualized and sorted from lightest to darkest.

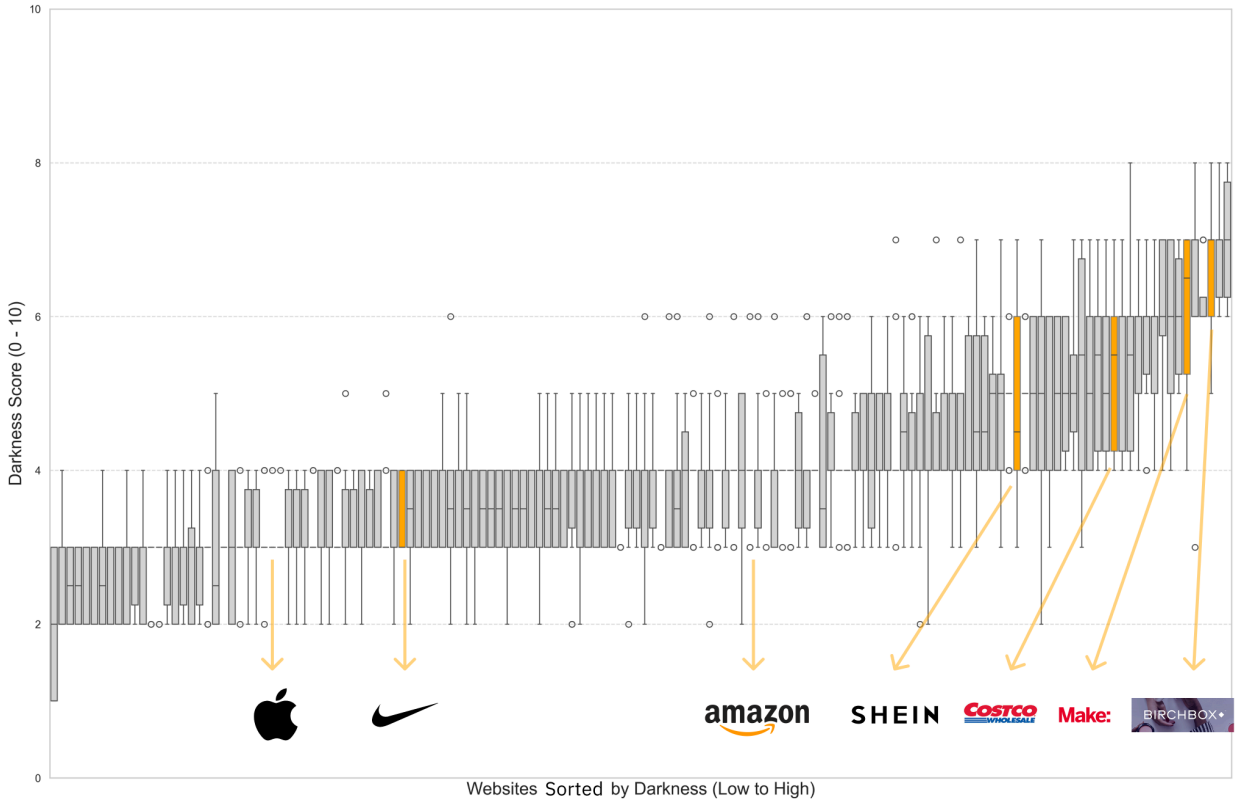


Figure 5.2 Distribution of the 'Restrictive' dark pattern category score, ranging from light to dark

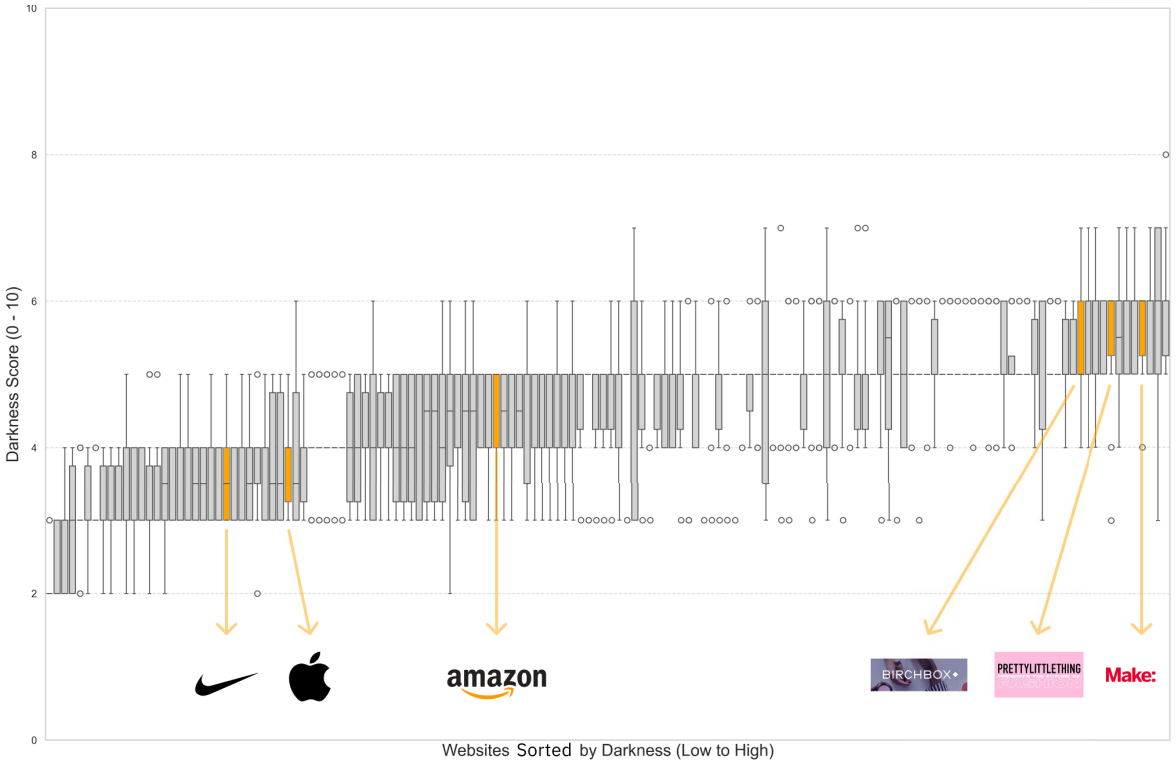


Figure 5.3 Distribution of the "Hides Information" dark pattern category score, ranging from light to dark

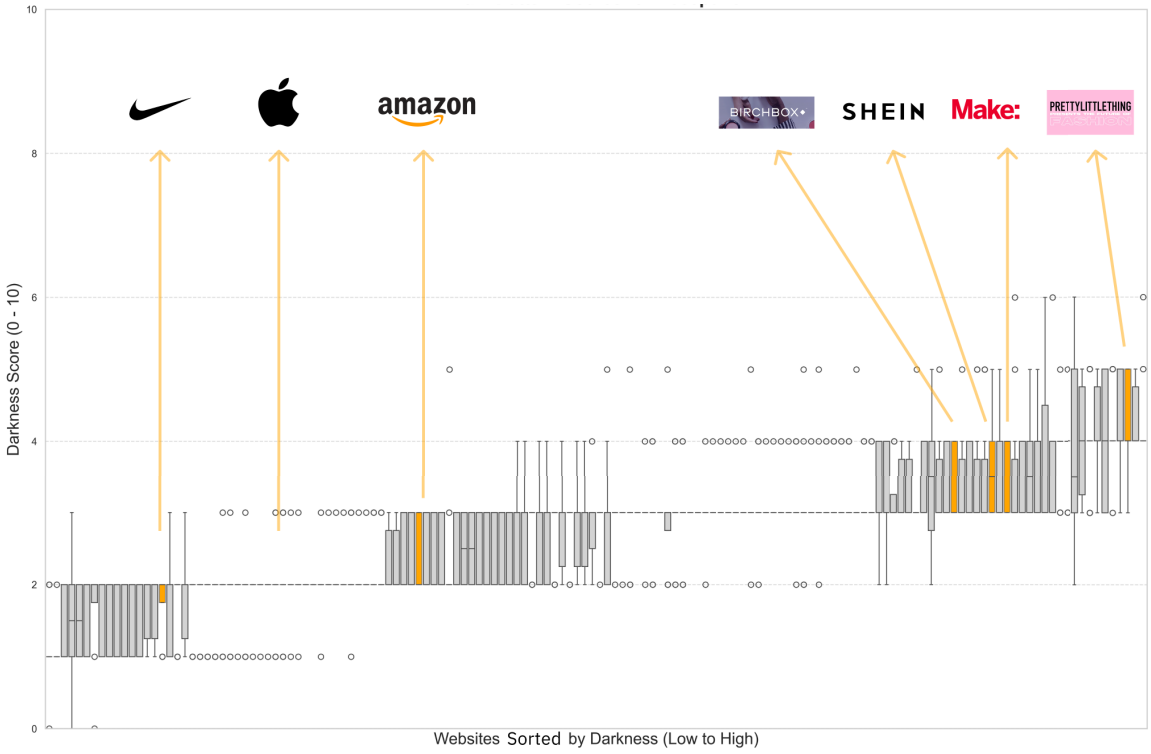


Figure 5.4 Distribution of the "Deceptive" dark pattern category score, ranging from light to dark

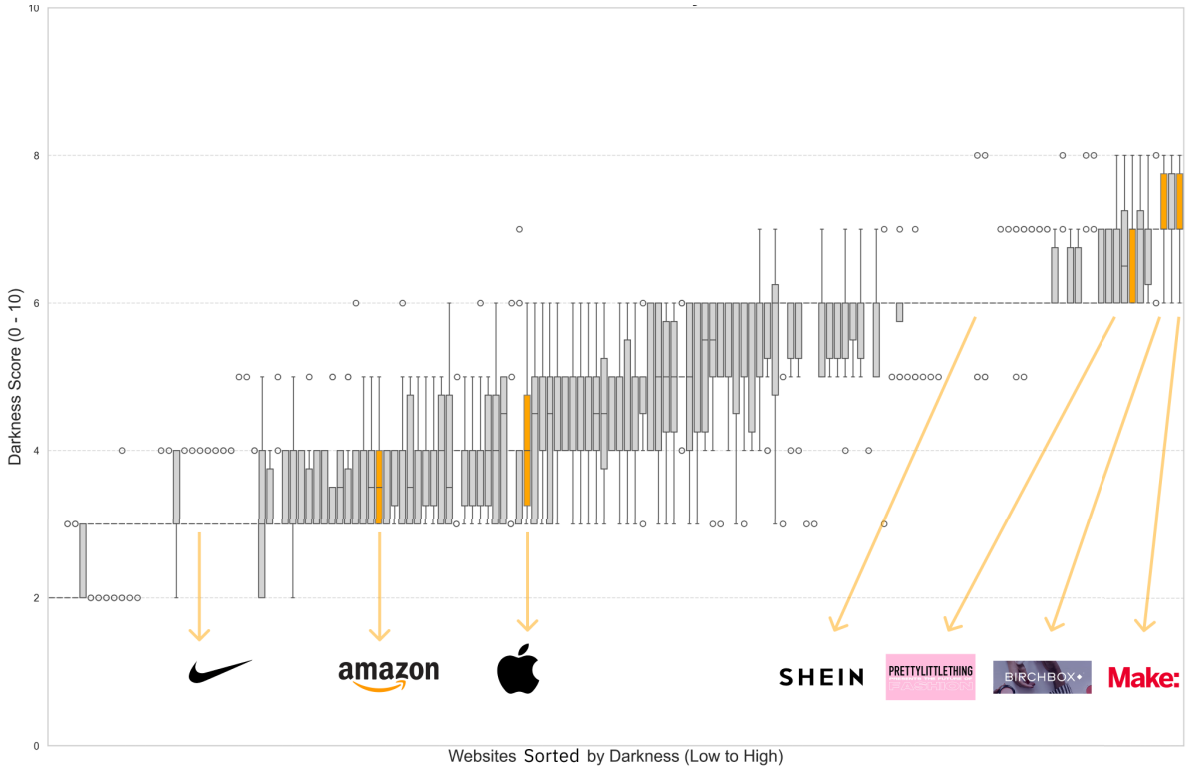


Figure 5.5 Distribution of the “Asymmetric” dark pattern category score, ranging from light to dark

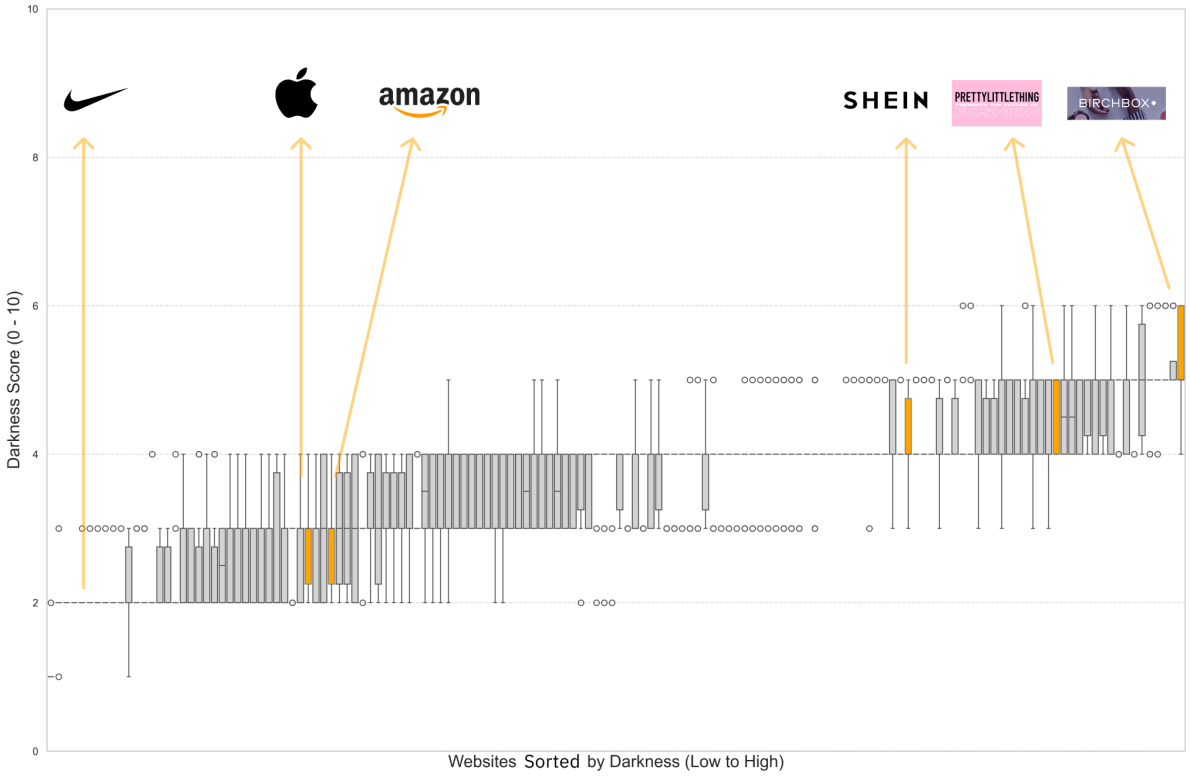


Figure 5.6 Distribution of the “Covert” dark pattern category score, ranging from light to dark

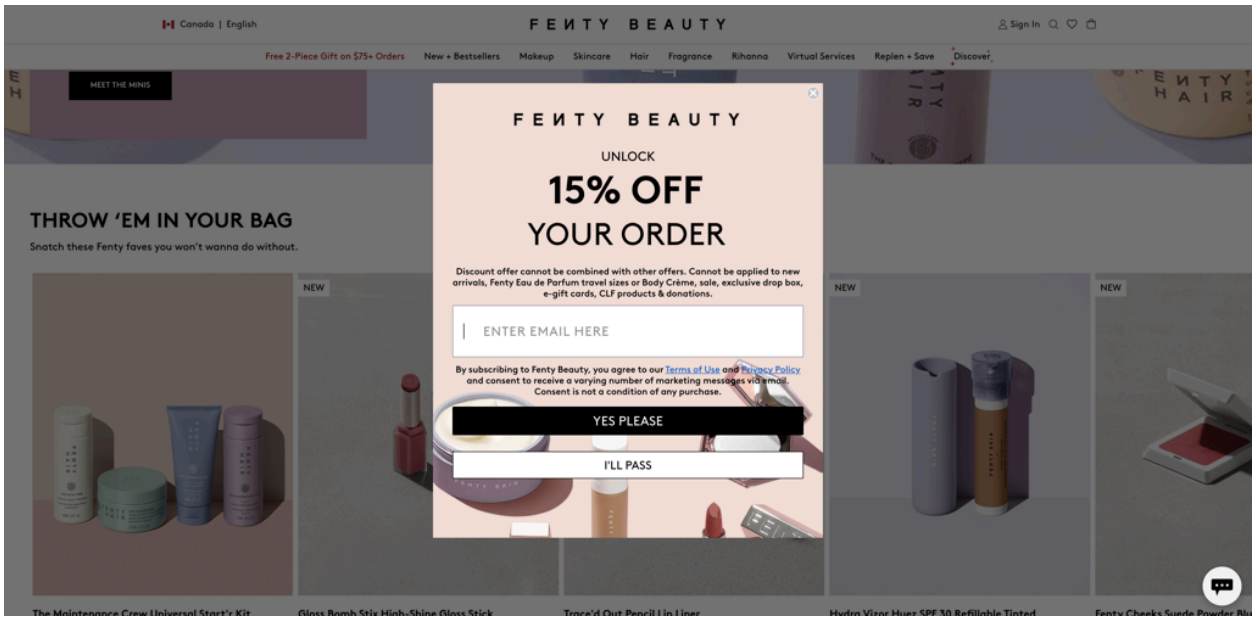
In the following sections, this research conducts a deeper analysis of the trends and design strategies employed by e-commerce websites. Through this exploration, it examines the top five darkest websites and analyzes the connection between the characteristics of these darker websites, their main sources of revenue—whether from physical stores or online sales—and the gender of their target audiences.

5.2 Design Strategies in the Top 5 Darkest Websites

To better understand dark pattern trends, I examined the top five websites with the highest overall darkness scores. As shown in Figure 5.1, these websites have overall scores of around 30 out of 50. The websites with the highest scores, identified through a box plot, are *Fenty Beauty, Birchbox, Make, Amart Furniture, and PrettyLittleThing.* Let's explore the characteristics of these websites and the patterns they have employed.

5.2.1 Fenty Beauty

Fenty Beauty is an online platform known for selling cosmetics and skincare products. The website offers a broad range of shades and products for diverse skin tones. It is ranked among the highest in darkness scores in this detection.



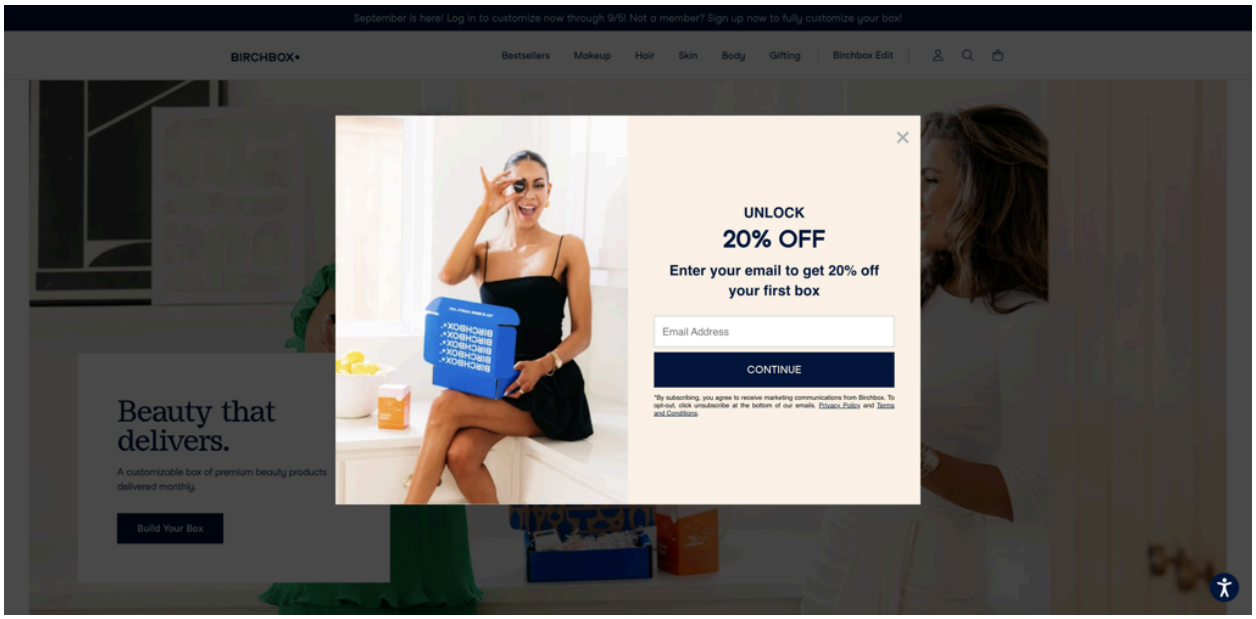
GPT-4o-mini model's comment on www.fentybeauty.com during the detection was: "The website employs several dark patterns, particularly in presenting asymmetric choices and implementing restrictions on user actions. The pop-up incentivizes users to enter their email while offering a limited choice to decline, which may not be clearly perceived as an equal option. Additionally, relevant information about the discount conditions is partly obscured, contributing to a potentially misleading experience."

After examining the screenshot provided to the model, I noticed that this website uses an asymmetry dark pattern through the vocabulary in the popup, which displays '15% Off Your Order' in large fonts. It is also restrictive, as the popup encourages users to subscribe by entering their email address to receive a discount, preventing them from viewing the website immediately.

- Asymmetry of Information: The "Yes Please" button is much larger, more visually prominent, and colored in an eye-catching shade, making it the default option for users. In contrast, the "I'll Pass" option is smaller, less noticeable, and in a white, secondary color, which makes declining the offer harder to click.
- Psychological Nudging: The "Yes Please" button leverages the concept of anchoring by directing attention toward the positive outcome (a discount) and away from the less emphasized option. This plays on users' fear of missing out on a deal, driving them toward an action that benefits the business, which is capturing their email for future marketing.
- Hidden Information: The terms and conditions linked to the discount are mentioned in small print. Users need to take an extra step (clicking on a very small button) to fully understand the terms before opting in, but the design encourages skipping over this.

5.2.2 Birchbox

Birchbox is an online subscription service that offers customized beauty boxes filled with premium products. This website also is ranked among the highest in darkness scores in this detection.



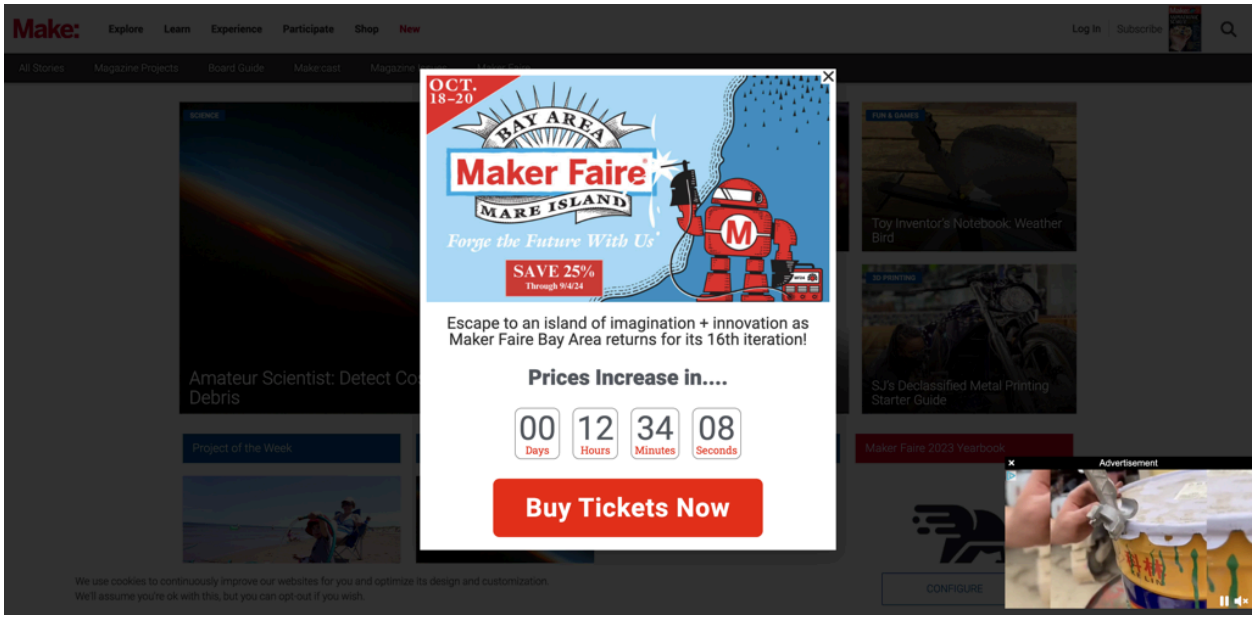
The GPT-4o-mini model's comment on www.birchbox.com during the detection was: "The website exhibits noticeable bias in how options are presented, with the email entry strongly emphasized. It uses covert techniques to entice users, while some information about the offer may be misleading. Important details about the offer are somewhat obscured, and the design restricts user actions, primarily funneling them toward email submission for discounts."

The website uses several dark patterns to push users toward subscribing to its newsletter for a discount.

- Asymmetry of Information: The 20% discount for new subscribers is highlighted with bold text and a central pop-up that immediately takes the user's attention. This discount acts as a cognitive anchor, framing the decision as a significant opportunity for savings while making the potential drawbacks less obvious.
- Forced Action: There is no visible "Reject" button, so opting out requires users to either ignore the pop-up or manually close it via the small close button (x) in the top-right corner. This design uses choice architecture, where the preferred action—subscribing—is made easier than the alternative, which is closing the window.
- Hidden Information: The small text at the bottom explains the terms of subscribing, including an agreement to receive marketing communications and instructions on how to unsubscribe. This information is minimized and written in a less prominent font size, making it easy to overlook.

5.2.3 Make

Makezine is a platform that provides content related to science and engineering. In this case, the website is promoting an event called Maker Faire using a popup with a countdown timer and a flashy button to purchase tickets.



The GPT-4o-mini model's comment on makezine.com during the detection was: "The website exhibits noticeable biases in presenting choices, especially with the pop-up's emphasis on urgency and discounted pricing, which may lead users to feel pressured to purchase. Certain important details are obscured and require additional steps to find. Users have limited choices, primarily influenced by the prominent call to action."

This website uses several dark patterns to push users toward purchasing tickets by creating a fear of missing out, and making the event seem more urgent and important than it may actually be.

- False Urgency: The countdown timer creates a false sense of urgency, implying that ticket prices will rise once the timer ends. This time pressure pushes users to act quickly, even if the claim is not true.

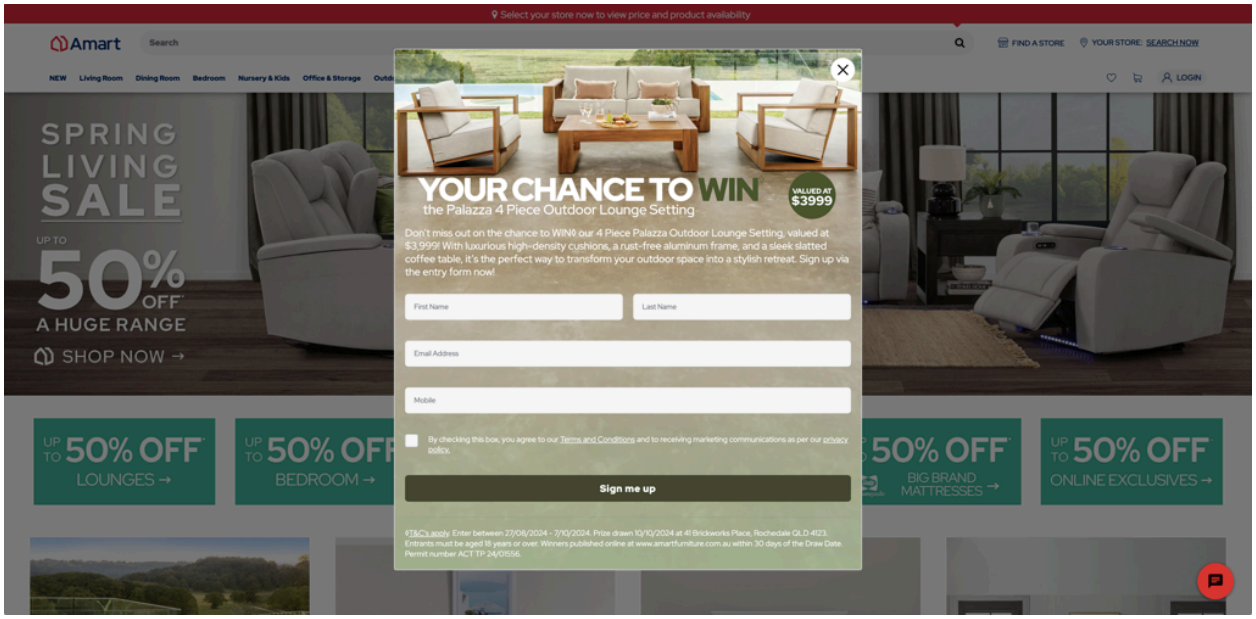
- Asymmetry of Information: The pop-up's large, bold fonts and bright red "Buy Tickets Now" button, forcing users to focus on purchasing tickets. This visual hierarchy intentionally diverts the user's attention away from the rest of the site.

5.2.4 Amart Furniture

Amart Furniture is an Australian furniture retailer that offers a wide range of home furnishings and decor. The website showcases product categories and ongoing sales promotions.

Gpt-4o-mini model's comment on www.amartfurniture.com.au during the detection was: "The website's popup for a contest entry exhibits noticeable biases in choice presentation, potential hidden influences, and some misleading aspects about the entry requirements. Information about terms and conditions is not immediately visible, and users face restrictions on choices due to the form's mandatory fields for entry."

Based on the screenshot, many dark patterns can be identified.

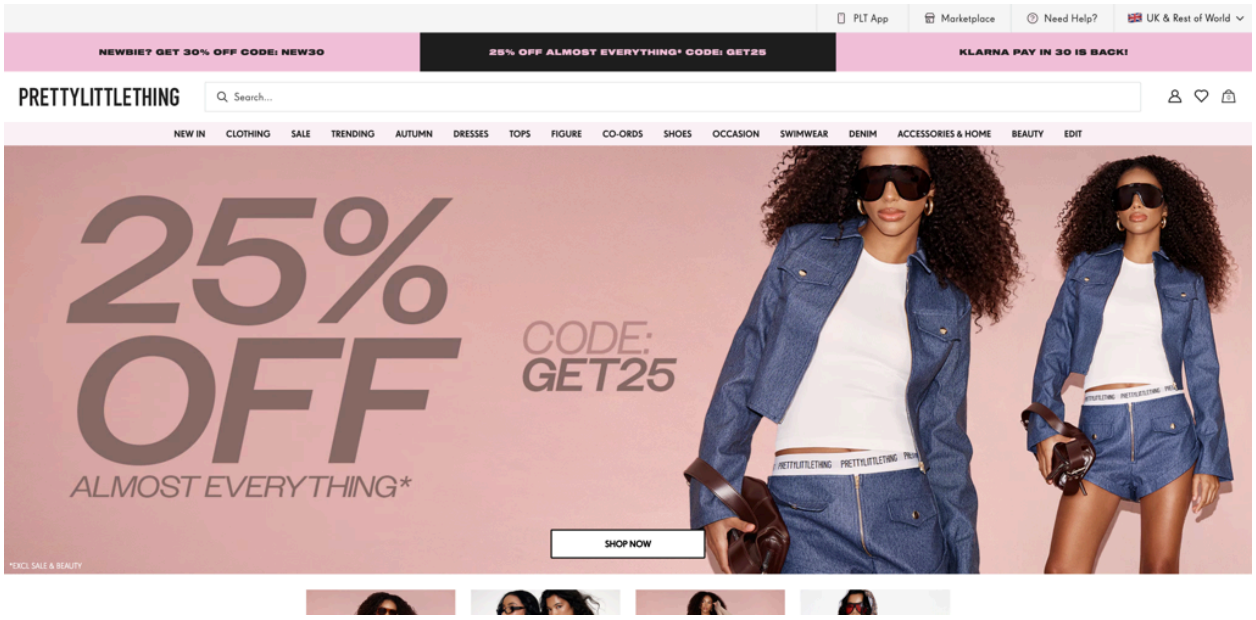


- Forced Action: To continue browsing the site, users must interact with the pop-up, either by entering their information or closing it. This forced action increases the chance of users signing up for the promotion, even if they weren't initially interested.

- Data Collection: The form requests personal information (name, email, mobile) on landing users upon their entry. This takes advantage of the principle of obligation, where users may feel compelled to provide data in exchange for a win.
- False Urgency: The background of the website shows a 'SPRING LIVING SALE' with 'UP TO 50% OFF,' creating a sense of scarcity that may push users to make quick decisions without fully understanding the details.

These design choices leverage cognitive biases such as the framing effect (presenting the contest as an opportunity rather than a data collection exercise) and loss aversion (fear of missing out on a valuable prize).

5.2.5 PrettyLittleThing



PrettyLittleThing is an online fashion ecommerce targeting women with affordable clothing and accessories.

Gpt-4o-mini model's comment on www.prettylittlething.com during the detection was: "The website demonstrates noticeable bias in promoting a discount, with alternative offers less prominent. Some covert design techniques subtly influence user behavior, and there is a

moderate level of deceptive information regarding the discount's applicability. Important conditions are somewhat obscured, and user choices are somewhat restricted but not overly so."

The website's design uses dark patterns, particularly in the category of Hidden Information.

- Hidden Information: While the discount is displayed with a huge font, the limitations are downplayed. The asterisk next to "ALMOST EVERYTHING*" is tiny, and any details about exclusions are not immediately visible, potentially misleading users about the scope of the offer. Important details about shipping, returns, or the full extent of the sales are not immediately visible.
- Asymmetry of Information: The main landing page displays "25% OFF" in large text, drawing immediate attention to the discount rather than the products themselves!

5.3 Insights on Gender Targeting and Revenue Dependence

After referencing the overall diagram in Figure 5.1 and extracting the top five darkest websites in section 5.2, two observations can be discussed.

First, three out of the five darkest websites are related to beauty and cosmetics, industries that primarily target a female audience. This raises the question of whether websites catering to specific gender demographics—particularly women—are more likely to employ dark patterns. To explore this potential connection, websites were categorized based on their primary audience using a combination of industry classification, product offerings, and marketing strategies. For instance, beauty and cosmetics websites typically market their products using gendered advertising and branding that predominantly appeal to women. Similarly, websites selling men’s grooming products or male-focused fitness gear were classified as targeting a male audience.

Second, the relationship between revenue dependence and dark pattern usage was examined. Websites that primarily generate revenue through online sales might have a greater incentive to use dark patterns to maximize conversions, retain customers, or collect data. Conversely, brands with extensive physical retail networks may have less reliance on aggressive online tactics since their primary revenue does not come from digital transactions alone. To assess this, websites

were categorized based on publicly available financial data, business models, and sales structures.

These classifications were used in Figures 5.7 and 5.8 to visually illustrate the potential connection between dark pattern intensity, target audience gender, and revenue dependency. Further research is needed to establish a causal link, but these observations suggest that websites targeting women and those heavily reliant on online sales may be more prone to employing manipulative design strategies. To map the connection between "Online Revenue Source," "Target Audience Gender," and the use of dark patterns for each website, I compiled a list of all 146 e-commerce websites. I then used OpenAI's o1 model to first determine the necessary data and subsequently label the overall darkness score diagram with additional annotations.

```
"For each of the following websites, based on all available information—including their products, services, marketing strategies, and any other relevant data—please determine the following:

1. Online Revenue Source: Does this website primarily generate revenue through online sales (e.g., e-commerce transactions, online subscriptions, digital services)? Please answer 'Yes' if the primary revenue comes from online sources, 'No' if it relies mainly on physical store sales or offline revenue streams, or 'Not Certain' if the information is unclear or mixed.
2. Target Audience Gender: Who is the main target audience in terms of gender for this website? Please answer 'Female' if the website primarily targets women, 'Male' if it primarily targets men, or 'Not Certain' if it targets a mixed audience or if the information is unclear.

Provide a brief explanation for each answer based on the information you have gathered."
```

This approach allowed me to determine both the primary revenue source and the target audience gender for each URL.

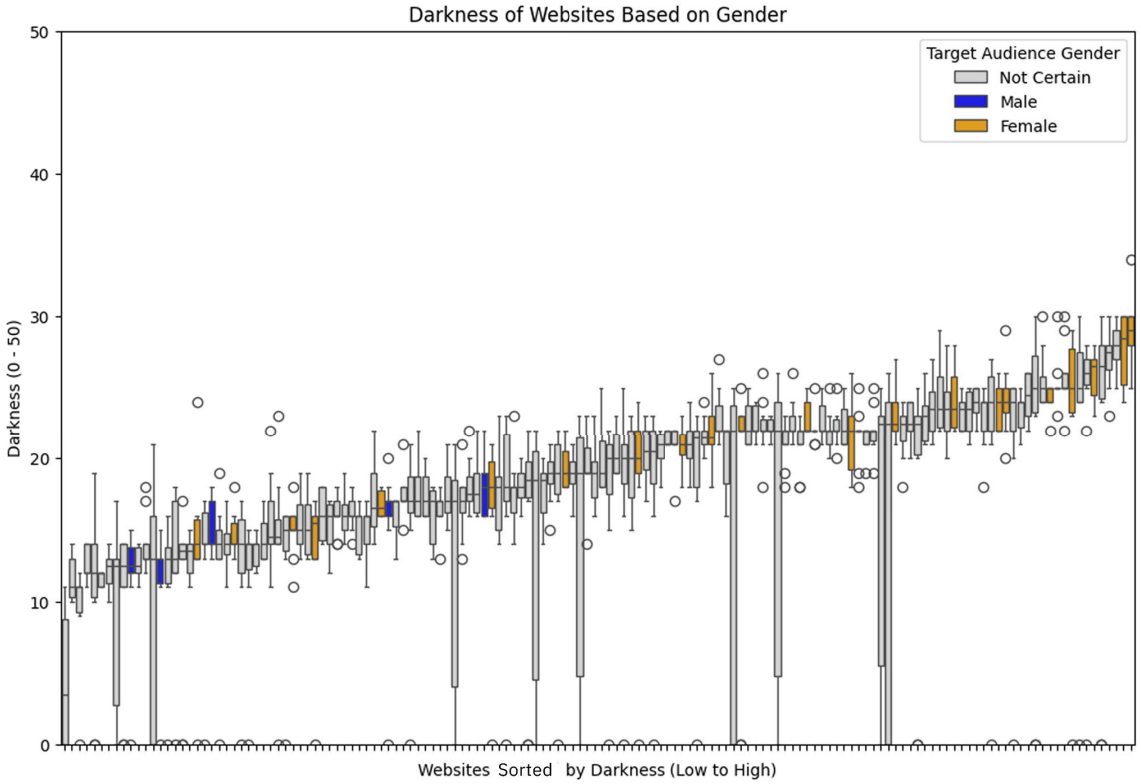


Figure 5.7 Distribution of the dark pattern overall score in websites, highlights gender of target audience

Figure 5.7 suggests a tendency for websites with exclusively female target audiences to have higher darkness scores; however, this tendency is not significant enough to be considered a strict rule. The boxplot in Figure 5.7 reveals that female-targeted websites are spread across a range of darkness scores, not exclusively at the higher end. In contrast, male-targeted websites tend to cluster on the lighter side of the diagram, and they are less numerous. This indicates that while the connection between gender-targeted websites and darker designs might exist, other factors may also influence the design patterns employed.

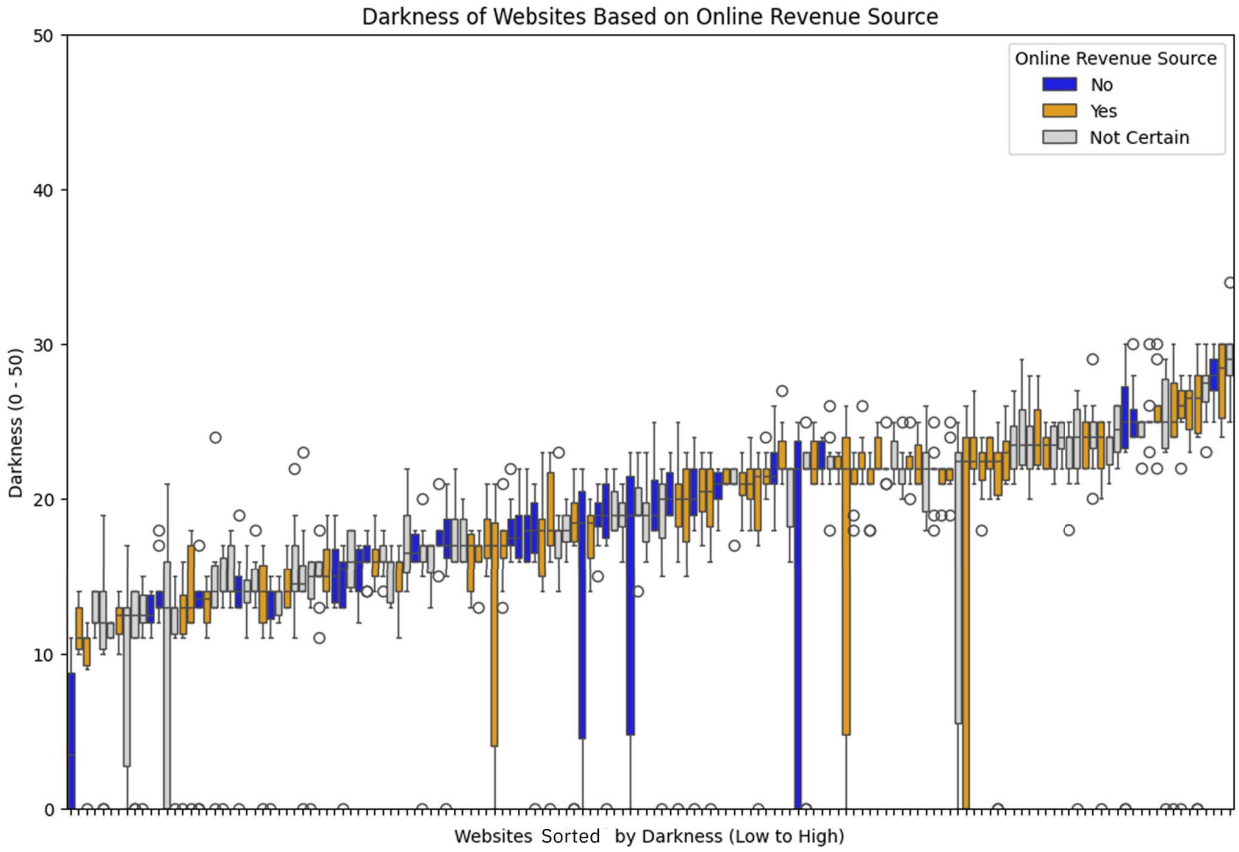


Figure 5.8 Distribution of the dark pattern overall score in websites, highlights main source of revenue

Similarly, websites that primarily generate revenue online tend to exhibit higher darkness scores. This supports the claim that businesses with a stronger reliance on online sales may be more likely to use manipulative design choices to optimize conversions. In contrast, well-established brands with physical stores, like Apple and Nike, which have diversified revenue streams and rely less on online sales, are positioned closer to the lighter end of the darkness scale. However, like gender, this tendency is not significant, indicating that online revenue dependence is not the sole determinant of dark pattern usage. Further research could explore additional factors, such as geographic location and industry type, to determine whether certain external influences contribute to dark pattern prevalence. This experiment was conducted on only 146 websites, but increasing the number of websites might show different trends. I encourage other scholars to conduct experiments with a larger sample of e-commerce websites to further explore this relationship.

5.4 Implications for User Experience Design Practice

Research through design is the first principle I learned as a user experience designer. Research aims to create something useful and human-centric, based on user characteristics. Sometimes, the intention shifts to prompting an important action, calling the user to engage. In this way, the user experience designer acts as a good actor, helping to create win-win solutions. As mentioned in discussions about the origin of dark patterns, persuasive technology is not always bad; however, combining it with intentional manipulation risks eroding user trust.

Figure 2.1 shows a popup from an early 2000s website, which Marshall Brain, in his blog *How Web Advertising Works*, described: *"It obscures the web page that you are trying to read, so you have to close the window or move it out of the way."* At that time, it was hard to determine whether this was a dark pattern or simply poor user experience due to a lack of knowledge. As a result of this detection, we found that four out of five of the darkest websites currently utilize the same method of obscuring the web page. We can now say that these practices are manipulative, as they received the highest overall darkness scores.

The persistence of these restrictive patterns suggests that dark patterns have become a standardized practice in e-commerce, often embedded within broader engagement strategies. Some companies have moved beyond simple pop-ups and now use more sophisticated tactics.

While honesty in e-commerce may not immediately boost sales, dishonesty can reduce repeat customers. Being transparent and, for instance, informing customers that a 25% discount applies only to certain items—rather than using a tiny asterisk or hiding exceptions behind extra clicks—can prevent frustration. Studies have found that trust and satisfaction are critical in developing online loyalty, emphasizing the importance of transparent and honest practices in retaining customers (Kim et al., 2009). After two decades of the same practice, such standardization should evolve and shift from creating a user experience focused on immediate results to a design philosophy centered on long-term consumer trust.

This standardization shift is hard since it is widespread, and even trained eyes can ignore them as regular designs, but this shift should be initiated first by designers and regulations. Designers can always iterate on their design outcomes, not only based on user analytics data, but also by using dark pattern detection plugins that support them in understanding these manipulative practices in

their design outcomes. Then, iteratively, in their next version of prototyping, they can break the trend patterns and improve the design. Regulations also can use the method developed and published in this research, to act faster and not only rely on customer reports after an incident of manipulation but proactively monitor websites and send warnings to companies if they are using these patterns. They can begin with businesses that rely heavily on online sales, as discussed in Section 5.3. These websites often prioritize manipulative patterns more than others, making them a higher priority for monitoring.

Understanding these trends, changing the direction of this standardization, and proactively monitoring e-commerce to send warnings can help raise awareness among designers and e-commerce businesses. This standardization emphasizes the need for ethical design education. By incorporating ethical considerations into the design process and actively monitoring websites, we can minimize the use of manipulative tactics and create experiences that benefit both customers and businesses. In the long run, this approach not only enhances brand reputation but also fosters customer satisfaction.

5.5 Conclusion

In conclusion, this research aimed to bridge gaps in prior studies by developing a detection method that leverages Large Language Models (LLMs) and image recognition to identify both textual and visual dark patterns. The study yielded two main outcomes.

Firstly, by applying the detection method to a dataset of 256 e-commerce websites, I was able to position these sites along a spectrum from light to dark in terms of their use of dark patterns. The top five darkest websites were identified, and we observed that they exhibited similar design strategies. These sites often employed obscuring techniques and restrictive approaches, such as opening dialogs immediately after loading to push users toward desired actions—tactics similar to the common pop-up dialogs used two decades ago. We also observed a potential connection between the characteristics of these darker websites, their primary revenue sources, and the gender of their target audience; However, due to the small size of our dataset, these observations are not definitive.

Secondly, a Chrome plugin—a practical tool for detecting these deceptive tactics in real time—was developed.

This research highlights that dark patterns—such as restrictive ones—are still prevalent, reflecting a standardization among designers and developers. This standardization might require changes that could start with designers or through regulations. The outcomes of this study suggest that the developed plugin could be a tool for designers to evaluate their design outcomes and iterate their designs. The detection method and the plugin itself could also be beneficial for regulators to act proactively and detect these websites before any user is manipulated.

Ultimately, this work suggests that AI can serve as a valuable tool for designers and regulators in evaluating user experiences and ensuring that ethical standards are met in digital interfaces. This research opens the door for further exploration into AI-driven multimodal detection, providing scholars with new avenues to investigate existing dark patterns on larger datasets using fine-tuned models and to address limitations that this study could not cover.

References

Brignull, H. (2010). *Dark Patterns*. Retrieved from https://www.darkpatterns.org/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://arxiv.org/pdf/1706.03762

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://arxiv.org/pdf/2005.14165

Gray, C. M., Kou, Y., 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*, 1-14. https://doi.org/10.1145/3173574.3174108

Gray, C., et al. (2024). A consistent and consolidated, shared, and reusable dark patterns ontology for future research, regulatory action, and sanctions. [PDF]. https://colingray.me/wp-content/uploads/2024/02/2024_Grayetal_CHI_OntologyDarkPatterns.pdf

Ballard Spahr LLP. (2024, May 31). *The Iliad flows: Federal judge allows FTC "dark patterns" suit against Amazon to proceed*. https://www.ballardspahr.com/insights/alerts-and-articles/2024/05/federal-judge-allows-ftc-dark-patterns-suit-against-amazon-to-proceed

Kollmer, T., & Eckhardt, A. (2023). Dark patterns: Conceptualization and future research directions. *Business & Information Systems Engineering, 65*(2), 201-208. https://doi.org/10.1007/s12599-022-00783-7

Carmineh. (n.d.). *Dark Pattern Identifier* [Computer software]. GitHub. Retrieved October 24, 2023, from https://github.com/Carmineh/Dark-Pattern-Identifier

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2009). Trust and satisfaction, two stepping stones for successful e-commerce relationships: A longitudinal exploration. *Information Systems Research, 20*(2), 237–257. https://doi.org/10.1287/isre.1080.0188

Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What makes a dark pattern... dark? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3411764.3445610

OpenAI. (n.d.). *API reference - OpenAI API*. https://platform.openai.com/docs/api-reference/authentication

Stanford Web Credibility Project. (n.d.). *Stanford guidelines for web credibility*. Stanford Persuasive Technology Lab. https://credibility.stanford.edu/guidelines/

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–32. https://doi.org/10.1145/3359183

Sunstein, C. R. (2016). *The ethics of influence: Government in the age of behavioral science*. Cambridge University Press.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Gray, C. M., Sanchez Chamorro, L., Obi, I., & Duane, J.-N. (2023a). Mapping the landscape of Dark Patterns Scholarship: A systematic literature review. *Designing Interactive Systems Conference*, 188–193. https://doi.org/10.1145/3563703.3596635

Turban, E., Outland, J., King, D. R., Lee, J. K., Liang, T.-P., & Turban, D. C. (2018). *Electronic commerce 2018: A managerial and Social Networks Perspective*. Springer.

Khurana, A. (2019). Did you know that there are 4 types of e-commerce? The Balance Small Business. Retrieved from https://www.thebalancesmb.com/ecommerce-businesses-understanding-types-1141595

Maguire, M. (2023). A Review of Usability Guidelines for E-Commerce Website Design. In: Marcus, A., Rosenzweig, E., Soares, M.M. (eds) Design, User Experience, and Usability. HCII 2023. Lecture Notes in Computer Science, vol 14032. Springer, Cham. https://doi.org/10.1007/978-3-031-35702-2_3