International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Detecting Deception: An AI-Driven Approach to Identify Dark Patterns

Sakshi Taaresh Khanna*[1], Anukool Johri[2], Vashu Tangri[3]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: Dark patterns, which are deceptive design approaches used by websites and applications to affect user behavior, raise important ethical and user experience issues. This work presents an artificial intelligence (AI) system designed to identify many forms of dark patterns, such as coerced activities, social manipulation, diversion, obstruction, limited availability, deceitful tactics, and time pressure. Using machine learning algorithms and natural language processing techniques, our system examines website and application interfaces to detect aspects that indicate the use of dark patterns. The authors showcase the effectiveness of their approach in precisely identifying and classifying dark patterns through thorough experimentation and validation. This enables regulatory compliance and promotes a digital world that is more transparent and ethical.

Keywords: *Artificial intelligence, Dark patterns, Deceptive design, User experience, Machine learning, Natural language processing*

1. Introduction

Dark patterns are deceptive practices employed in websites and apps to manipulate people into engaging in unexpected actions, such as making purchases or subscribing to services. These strategies exploit the principles of human psychology and interface design to discreetly guide individuals toward choices they may not have consciously chosen [1].

Dark patterns manifest in several forms, including misleading visual cues, hidden information, or language that bewilders users into unintended behaviours. Examples of this phenomenon include intentional utilization of vague phrasing, deliberately placed checkboxes, or design components that guide users towards unintentional selections.

These deceitful methods subvert clear and moral design standards by exploiting cognitive biases and manipulating users' subconscious inclinations. Consequently, they undermine the independence and capacity of individuals to make well-informed choices when engaging with digital interfaces. In order to cultivate user trust, promote ethical design principles, and establish a good and respectful online experience for all users, it is imperative to recognize and address these deceptive and manipulative design techniques known as dark patterns.

Dark patterns are a growing concern in the E-commerce sector of the internet, as new stores emerge on a daily basis. Appmysite.com reports that there are approximately 12 to 24 million e-Commerce websites [2]. Given the large number of online stores, it is not unexpected that some engage in misleading practices to increase their earnings and outperform their competitors in sales.

Dark patterns possess the ability to affect user behaviours and can result in inadvertent and unsustainable consumption patterns. This phenomenon occurs when individuals, under the influence of deceptive design strategies, engage in excessive shopping, collecting more products than necessary or obtaining items that have no practical utility to them. These manipulative tactics exploit the vulnerabilities of consumers and subtly guide them towards making choices that exceed their actual requirements. Dark patterns employ false visual cues, persuasive language, or hidden information to manipulate individuals into surpassing their initial goals, resulting in the unneeded acquisition of products or services [3].

The inadvertent and uncontrollable usage of this has consequences not only for individual users but also adds to broader societal and environmental issues. Overconsumption, driven by deceptive tactics, can lead to the exhaustion of resources, deterioration of the environment, and an increase in waste generation. Hence, it is essential to raise awareness about the detrimental impact of deceptive tactics on consumer behaviour and advocate for ethical design methods that prioritize user well-being and sustainability.

Dark patterns are deceptive design patterns in user interfaces that exploit user behaviour for the profit of the website or application owner, while disregarding user autonomy. These patterns have gained more attention in recent years. These patterns leverage psychological principles to compel people to take behaviours they may not otherwise choose voluntarily. Typical instances comprise of

[1] *Department of Computer Science, Ram Lal Anand College, University of Delhi-110021, INDIA*
Corresponding Author Email: sakshitaareshkhanna@gmail.com
ORCID ID: 0009-0006-2936-8708
[2] *Department of Computer Science, Ram Lal Anand College, University of Delhi-110021, INDIA. Email: Anukool4028@rla.du.ac.in*
ORCID ID: 0009-0002-9028-8011
[3] *Department of Computer Science, Ram Lal Anand College, University of Delhi-110021, INDIA. Email: Vashu4146@rla.du.ac.in*
ORCID ID: 0009-0004-0398-0195

deceptive cues, undisclosed expenses, and manipulative wording designed to guide people towards unexpected results. With the increasing prevalence of dark patterns on digital platforms, there is a rising demand for efficient methods to identify and counteract them in order to safeguard user interests and maintain ethical design practices.

1.1. Background and Motivation

User interfaces are extremely important in the digital era because they determine how people interact with and use different types of online platforms. Concern over dark patterns, which are misleading design techniques, has been on the rise in tandem with the popularity of digital interactions. Websites and applications use dark patterns, which are manipulative strategies in interface design, to influence user behavior in ways that might not be in their best interests. In order to trick consumers into giving personal information or making rash judgments, these manipulative strategies take advantage of cognitive biases and psychological concepts.

Since dark patterns weaken user autonomy, trust, and transparency in digital interactions, they offer substantial ethical and user experience concerns. Dark patterns can take many forms, including deceptive prompts, hidden costs, coercive language, and social proofs. Their common purpose is to maximize engagement, conversions, or data collecting for the profit of the applications or websites owner. Users may experience negative emotions like deceit, frustration, or misinformation, which can damage their faith in the digital ecosystem and cause problems like financial loss, privacy breaches, or reduced user pleasure.

This research is driven by the pressing necessity to tackle the widespread problem of dark patterns and devise efficient techniques to identify and minimize their influence on users. By utilizing advancements in artificial intelligence (AI) and machine learning, it is possible to automate the identification of dark patterns in digital interfaces. This would give users more transparency, control, and protection against misleading design tactics. This research project seeks to make significant contributions to the progress of ethical design principles, advocate for user-centered digital experiences, and cultivate a digital environment that is more transparent and trustworthy for all users.

1.2. Overview of Dark Patterns and Their Impact

Websites and applications often utilize misleading design strategies called dark patterns to influence user behavior and interactions in ways that might not be in their best interests. At the cost of user agency and openness, these manipulative strategies take advantage of cognitive biases and psychological principles to sway user decisions in favour of the website or app owner.

Dark patterns occur in many varieties, each with its own sneaky tricks and desired effects. Users may unwittingly sign up for subscriptions they didn't want or make purchases they didn't mean to make when they're subjected to forced activities. By appealing to users' sense of popularity or urgency, social proofs take use of peer pressure and social influence. Misdirection is the practice of drawing users' attention away from what they need to do, which might cause them to make mistakes or become confused.

Obstruction dark patterns hide unsubscribe buttons or make cancellation processes too complicated on purpose to prevent users from accomplishing activities or getting needed information. Showing limited supply or time-restricted promotions are examples of scarcity strategies that use people's perceptions of scarcity and urgency to induce impulsive purchases. As an example, pre-checked boxes or hidden costs are examples of sneaking dark patterns, which are sneaky or misleading activities that hide real intentions or consequences. Countdown timers or deceptive notifications are common examples of urgency tactics used to generate a feeling of urgency or FOMO in order to induce rapid action.

Dark patterns can have a substantial impact on users, resulting in reduced trust and happiness, as well as potential money losses and privacy violations. Deceptive design approaches can cause users to feel fooled, frustrated, or misled, which can erode trust in the digital ecosystem. This erosion of trust can possibly result in bad outcomes, such as unintentional purchases, subscription sign-ups, or the revelation of personal data. Moreover, dark patterns can exacerbate the absence of transparency and accountability in digital interactions, maintaining a recurring pattern of unethical design practices and limiting the user's ability to exercise agency and control.

Understanding the widespread occurrence and significant consequences of dark patterns highlights the necessity of creating efficient methods for identifying and reducing their impact on user experiences. By increasing knowledge of manipulative design strategies and utilizing advancements in artificial intelligence and machine learning, we may give users more visibility, authority, and defense against misleading interface design tactics. The objective of this research project is to make a valuable contribution to the progress of ethical design methods and promote a digital environment that is more open and reliable for all users.

1.3. Importance of Detecting and Mitigating Dark Patterns

The detection and mitigation of dark patterns in digital interfaces are paramount for several reasons, encompassing ethical, user-centric, and regulatory considerations. Understanding the importance of these efforts is crucial for promoting transparency, trust, and user autonomy in the

digital landscape. Fig 1 given below shows the key points highlighting the significance of detecting and mitigating dark patterns:

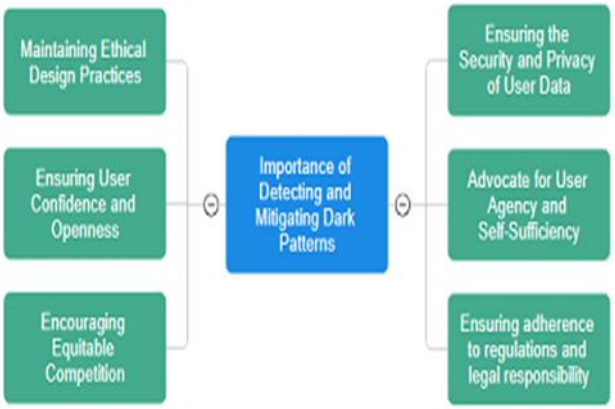

Fig 1. Importance of detecting dark patterns

In order to maintain ethical design practices, protect user trust, and preserve privacy, it is essential to identify and mitigate dark patterns in digital interfaces. Dark patterns utilize cognitive fallacies to influence user behavior, frequently leading to unintended consequences that prioritize engagement over user welfare. By implementing these strategies, it is possible to establish a culture of transparency and trust by ensuring that digital interfaces are designed ethically. This not only improves user experiences but also establishes long-term relationships with users, as they feel respected and are not deceived by manipulative interface elements [18].

Moreover, it is imperative to mitigate negative patterns in order to protect user privacy and foster autonomy. Unauthorized data acquisition is a component of certain deceptive practices, which compromise user control over personal information. Designers reinforce users' autonomy by enabling them to make informed decisions based on their genuine preferences by eliminating these strategies. Furthermore, it is imperative to ensure compliance in order to prevent legal penalties and maintain ethical accountability in light of the growing regulatory scrutiny of deceptive design practices. The digital ecosystem can be made more transparent, trustworthy, and user-centric by prioritizing the detection and mitigation of dark patterns, which will be advantageous to

both users and businesses [19].

1.4. Objectives and Contributions

1.4.1 Objectives

- The objective is to present a novel AI-based method for identifying deceptive design trends in digital interfaces.
- The objective is to provide a framework that utilizes machine learning algorithms and natural language processing techniques to detect different forms of deceitful design strategies.
- The objective is to examine the ethical ramifications of automated dark pattern recognition and deliberate methods to enhance transparency and empower users in digital interactions.

1.4.2. Contributions

- The research study presents an innovative AI-driven method for spotting dark patterns in digital interfaces. This methodology provides a scalable and efficient solution for detecting misleading design strategies.
- The study introduces a framework that combines machine learning algorithms and natural language processing approaches to assess digital interfaces and categorize deceptive components.
- This research contributes to the ongoing discussion on digital ethics and user-centered design practices.

2. Literature Review

Prior studies on dark patterns have predominantly concentrated on discerning distinct forms of deceitful design methods and their consequences for user experience and consumer rights. Although many researchers have suggested using manual inspection methods or heuristic evaluations to identify dark patterns, these approaches frequently lack scalability and may fail to discover tiny instances of manipulation. In recent times, the utilization of machine learning and artificial intelligence methods has demonstrated potential in automating the identification of dark patterns, providing enhanced efficiency and precision in detecting deceitful design components.

2.1. Taxonomy of Dark Patterns

Dark patterns are deceptive design strategies used by websites and apps to control user behavior; these strategies frequently lead to unintentional behaviors or outcomes that serve the platform owner's interests at the expense of the user's [6]. These design strategies force users to make decisions or do behaviors they might not otherwise choose voluntarily by taking advantage of cognitive biases and psychological concepts [7].

Digital interface misleading design strategies can be classified and categorized with the aid of a taxonomy of dark patterns. fig 2 shows the categories of the taxonomy of dark patterns:

1. Misdirection: This category includes design features that are deliberately used to mislead or divert users from their intended behaviors, such as misleading visual cues or

deceptive navigation paths [8].

2. Social Proof: Dark patterns in this category exploit social proof principles to manipulate user behavior, using deceptive tactics such false social endorsements or falsified user testimonials to influence purchasing decisions [9].

Fig 2. Taxonomy of dark patterns

3. Forced behaviors: These dark patterns manipulate people into performing behaviors they may not desire, typically by using deceptive prompts or misleading opt-out procedures [10].

4. Friend Spam: Digital design's dark pattern of friend spam involves a service tricking users into sending undesired or false communications to their connections without their agreement or knowledge. This strategy uses social media to promote content, sign-ups, and website and app traffic [20].

5. Obstruction refers to the employment of dark patterns that deliberately create hurdles or impediments to prevent users from successfully completing desired tasks. Examples of obstruction are purposely complex or confusing checkout processes [11].

6. Scarcity: Dark patterns employ tactics to artificially create a sense of scarcity in order to provoke urgency or FOMO (fear of missing out) among users. These tactics may include showing limited stock or countdown timers to compel users to make hasty decisions [12].

7. Sneaking: Sneaking dark patterns utilize covert or unreported strategies to control user behavior, such as automatically selected checkboxes or undisclosed charges added to transactions without the user's explicit knowledge [13].

8. Urgency: Urgency dark patterns manipulate time-sensitive signals to induce a feeling of urgency or terror in users. This is achieved by presenting deceptive countdown timers or deadlines that compel users to take quick action [7].

9.Privacy Zuckering: Privacy Zuckering. Tricked into disclosing more personal information than intended [20].

10. Roach Motel: In this one can readily enter a situation, but it is difficult to exit it (e.g., a premium subscription) [20].

11. Bait and Switch: is a dark pattern in which people are deceived into clicking a button or following a link, only to have something bad happen. Users may click on a link expecting a product page but instead getting an advertisement or sign-up form [20].

12. Confirm Shaming: is a deceptive practice that shames or guilts consumers into accepting something. The decline option makes consumers feel awful or irresponsible for choosing it, pressuring them to comply. This strategy uses social and emotional influences to influence user behavior toward service goals [20][21].

13. Disguised Ads: "Disguised ads" are adverts that look like website content or navigation to trick users into clicking on them. This dark design exploits users' confidence and familiarity with webpage elements to make content and adverts hard to differentiate. Ads may disguise themselves as news articles or download buttons to deceive visitors into interacting with them. This deceives users and damages the website's reputation [21].

14. Hidden Cost: Hidden Costs: The user is attracted by the low advertised price. When they arrive at the register, they are met with unforeseen fees and charges, despite the time and effort they have invested [21].

2.2. Previous Approaches to Dark Pattern Detection

Detecting dark patterns in digital interfaces has been the subject of research in various fields, including human-computer interaction, user experience design, and computer science. Previous approaches to dark pattern detection have employed a variety of methods, ranging from manual inspection to automated algorithms. Understanding these approaches can provide valuable insights into the challenges and opportunities in detecting deceptive design tactics. The fig 3 shows the approaches to dark patterns:

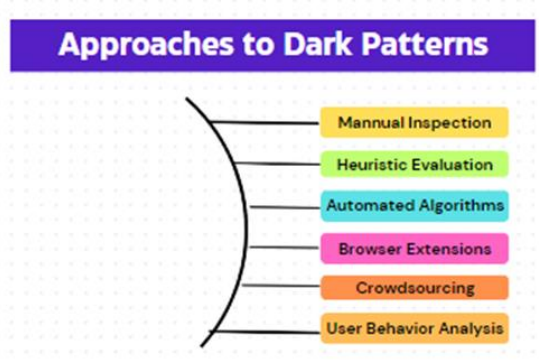

Fig 3. Different approaches to dark patterns

1. Manual Inspection: Researchers conduct manual analyses of interfaces to identify deceptive design elements. This procedure is time-consuming and impractical for large-scale evaluations, despite the fact that it provides detailed insights [13].

2. Heuristic Evaluation: Researchers evaluate interfaces for dark patterns by employing predefined usability heuristics. This approach offers valuable feedback; however, it may be lacking in specificity and consistency [14].

3. Automated Algorithms: Machine learning techniques and algorithms are employed to analyze interfaces for deceptive patterns, thereby efficiently processing large datasets and identifying subtle manipulations [8].

4. Extensions for Browsers: Browser extensions provide users with real-time notifications regarding deceptive design elements, thereby enabling them to make well-informed decisions. Nevertheless, they may not be accessible to all users and necessitate frequent updates [10].

5. Crowdsourcing: Researchers utilize crowdsourcing platforms to recruit users to identify and categorize dark patterns, which offers a variety of viewpoints but may encounter data quality and reliability concerns [12].

2.3. Limitations of Existing Detection Methods

Although prior methods for detecting dark patterns have achieved substantial advancements, they still have certain limitations. Comprehending these constraints is crucial in directing future research endeavors and enhancing the efficiency of detection techniques. The following are several prevalent constraints linked to current detection techniques:

1. Subjectivity and Interpretation Bias: Manual inspection and heuristic evaluation approaches require human evaluators to spot and interpret misleading design tactics. Subjectivity can cause detection inconsistencies and dark pattern identification difficulties [6].

2. Scalability and Efficiency: Manual inspection and heuristic evaluation are too time-consuming and resource-intensive for massive dataset analysis or digital interface evaluations. Scalable and efficient automated algorithms may struggle to catch nuanced or context-dependent deceit [14].

3. Lack of Transparency and Explainability: Machine learning-based automated detection systems lack transparency and explainability in their decision-making process. Researchers and practitioners may struggle to understand and trust these strategies without explicit explanations of detection [8].

4. Limited Coverage and Generalization: Detection algorithms created for certain datasets under controlled situations may not apply to real-world events. Dark patterns can take many shapes across digital interfaces, making it difficult to develop detection systems that can catch every deceptive design [16].

5. Accessibility and uptake: Browser extensions and crowdsourcing platforms may struggle with accessibility and uptake for dark pattern identification. Browser extensions require users to install and maintain software, and crowdsourcing platforms may struggle to recruit and keep diverse contributors [10].

6. Ethical and Legal Considerations: Dark pattern detection systems must balance privacy, consent, and data protection. Automation algorithms that evaluate user activities may accidentally gather sensitive data or violate user privacy, posing data security and regulatory compliance risks [15].

2.4. Role of Artificial Intelligence in Dark Pattern Detection

Artificial intelligence (AI) is critically important in the identification and mitigation of deceptive design strategies in digital interfaces, as it provides sophisticated computational techniques for detecting dark patterns. To analyze digital interfaces and identify patterns indicative of deceptive design elements, AI-driven approaches employ machine learning algorithms, natural language processing techniques, and computer vision technologies. The fig 4 shows the function of AI in the detection of dark patterns:

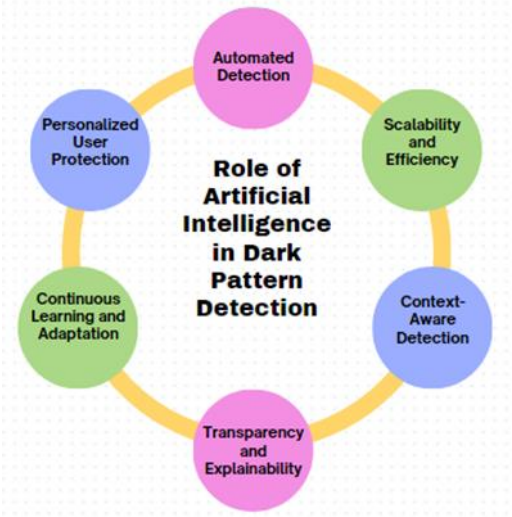

Fig 4. Role of AI in dark pattern detection

AI systems facilitate the automated identification of dark patterns by evaluating extensive datasets and detecting subtle indicators of misleading design strategies. Machine learning models can undergo training using labeled datasets to identify patterns linked to many categories of dark patterns, including misdirection, forced acts, and shortages. AI-powered methods provide the potential to scale and improve efficiency in identifying dark patterns. This enables the analysis of extensive datasets and the real-time monitoring of digital interfaces. Automated algorithms have

the ability to analyze large quantities of user interactions and detect possible occurrences of deceptive design strategies with great precision and efficiency. This allows for prompt interventions to safeguard users from manipulation. Artificial intelligence algorithms can be educated to comprehend the specific circumstances in which dark patterns appear, thus facilitating more precise identification and categorization of deceitful design strategies. Natural language processing techniques can be employed to examine textual content and discover instances of manipulative language or forceful prompts. Similarly, computer vision technology can be utilized to scan visual aspects and find deceptive visual cues or concealed information. In spite of the intricate nature of AI algorithms, endeavors are underway to guarantee transparency and comprehensibility in the identification of dark patterns. Explainable AI (XAI) approaches offer consumers the capacity to gain insights into the decision-making process of AI models. This allows users to comprehend how detections are generated and evaluate the dependability of the results. AI-driven approaches promote confidence and accountability in detecting dark patterns by improving transparency and explainability. Artificial intelligence algorithms have the ability to consistently acquire knowledge and adjust to changing misleading design patterns. This allows for the proactive identification and reduction of emerging dark patterns. By employing repeated training and feedback mechanisms, machine learning models can enhance their performance gradually and adjust to variations in digital interfaces and user behaviors. This enables them to effectively identify deceptive design approaches with resilience and efficiency.

3. Proposed Methodology

Our proposed AI system for the detection of dark patterns incorporates a diverse array of natural language processing techniques and machine learning algorithms. The system will extract features that are indicative of common dark pattern implementations by analyzing website and application interfaces. These characteristics may encompass the presence of coercive user prompts, deceptive visual signals, or misleading language. The system will subsequently categorize the detected elements into specific categories of dark patterns, including forced actions, social substantiation, misdirection, obstruction, scarcity, sneaking, and urgency. The fig 5 shows the complete flow chart of the proposed approach:

The AI-driven detection system proposed in this research paper aims to identify and mitigate dark patterns in digital interfaces using advanced computational techniques and machine learning algorithms. The system operates by analyzing user interactions with digital interfaces and identifying patterns indicative of deceptive design tactics. The overview of the AI-driven detection system can be summarized as follows:

1. Data Collection: The detection system collects data from various sources, including digital interfaces, websites, and applications, to analyze user interactions and identify potential instances of dark patterns. Data may include user behavior logs, website content, visual elements, and textual information. During this stage, the

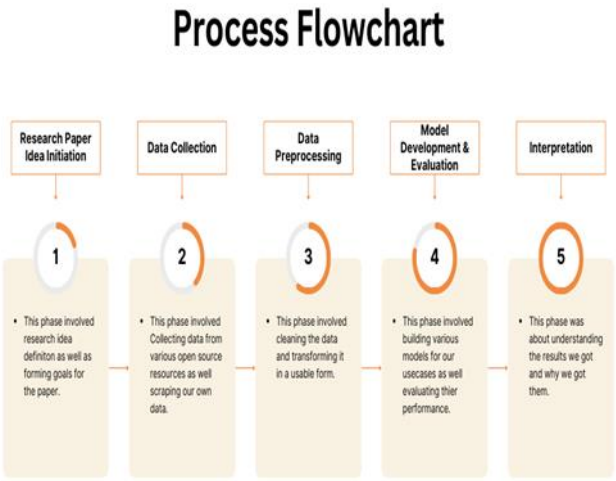

Fig 5. Entire research paper work flow

authors focus was on gathering relevant information from open resources available on the internet, which is known as data acquisition. The primary goal was to speed up the preprocessing phase by utilizing existing datasets. The authors were successful in finding a comprehensive open-source repository of data specifically related to dark patterns in e-commerce [4][5], which greatly enhanced the diversity of the dataset. To ensure the efficacy of our approach, we employ a multi-faceted strategy for data collection. Utilizing web scraping techniques to gather data from various online platforms known to employ dark patterns. This involves accessing and extracting information from websites, including e-commerce platforms, social media sites, and mobile applications.

2. Preprocessing: The collected data undergoes preprocessing to clean and prepare it for analysis. Preprocessing steps may include noise removal, data normalization, feature extraction, and data transformation to ensure compatibility with the detection algorithms.

3. Feature Extraction: The detection system extracts relevant features from the preprocessed data to capture patterns indicative of deceptive design tactics. Features may include user interactions, textual content, visual elements, and contextual information extracted from digital interfaces.

4. Machine Learning Models: The detection system employs machine learning models to analyze the extracted features and identify patterns associated with various types of dark patterns. In this stage of the process, the authors utilized the data collected from their research to create a

machine-learning model that can identify the existence of dark patterns on websites and measure their degree. To carry out this phase, the authors have selected different classifiers and performed Grid Search to fine-tune the hyperparameters for the classifiers. Then the classifier with the best results was selected based on evaluation metrics.

5. Detection Algorithms: The machine learning models are used to develop detection algorithms that can automatically identify instances of dark patterns in digital interfaces. These algorithms analyze user interactions and identify patterns indicative of deceptive design tactics, such as misdirection, forced actions, and scarcity.

6. Evaluation and Validation: The performance of the detection system is evaluated and validated using benchmark datasets and real-world scenarios. Metrics such as precision, recall, accuracy, and F1-score are used to assess the effectiveness of the detection algorithms in identifying dark patterns.

3.1. Building Machine Learning Model Classifying Dark Patterns into deceptive or non-deceptive

Non-deceptive dark patterns refer to design techniques that might lead users to unintended actions, but without any malicious intent, Deceptive dark patterns, on the other hand, involve intentionally misleading or psychologically manipulating users to achieve a specific outcome that benefits the designer or organization at the cost the user's interest.

In this phase the authors leveraged the same dataset employed in classifying patterns as either dark or non-dark; however, the preprocessing methodology underwent modification.

This model facilitates the creation of a robust framework for discerning between deceptive and non-deceptive patterns, enhancing user experience and trust, additionally, by elucidating the characteristics indicative of dark patterns, organizations gained insights into unethical design practices, enabling them to mitigate potential reputational risks.

3.2. Evaluation Metrics and Performance Analysis

In this section, the authors delve into the evaluation metrics utilized to assess the performance of the proposed methodology. The evaluation process is essential to gauge the effectiveness and reliability of the model in achieving its intended objectives. The metrics employed encompass a comprehensive analysis, considering various aspects of classification accuracy and predictive capability.

Accuracy: Measures the model's overall correctness by comparing properly predicted instances to total instances. A high accuracy score implies all-class prediction proficiency.

Precision: Determines positive prediction accuracy by comparing true positives to true and false positives. When minimizing false positives is important, it helps.

Recall: Sensitivity measures the model's capacity to catch all positive cases by determining the ratio of true positives to true positives and false negatives. When identifying all positives, recall is crucial.

F1 Score: The harmonic mean of precision and recall balances model performance. It helps when positive and negative instances are imbalanced, including false positives and negatives.

ROC AUC Score: The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) score provide insights into the model's ability to discriminate between classes across various thresholds. The ROC AUC score quantifies the model's performance across different levels of sensitivity and specificity, offering a comprehensive evaluation of its predictive capability.

4. Results and Discussion

Initial findings from the research indicate that the AI system has a favorable performance in accurately identifying different forms of dark patterns. The system's capacity to detect nuanced instances of manipulation and classify them into several pattern categories demonstrates its potential to assist in regulatory compliance and promote transparent design practices.

4.1. Preprocessing of dark/non-dark classification

Feature Engineering: Created a new binary target variable "dark pattern?" indicating the presence of dark patterns. Merged and concatenated the datasets to create the final dataframe.

Data Cleaning: Removed rows from df2 where the "Deceptive?" column equals 1. Renamed the "Deceptive?" column in df2 to "dark patterns?".

Handling Missing Values and Duplicates: Counted and handled any duplicate rows in the data frame.

Encoding: Encoded the "dark patterns?" column in the final data frame using Label Encoder.

Text Data Preprocessing: Generated word clouds for both dark and non-dark patterns.

Exploratory Data Analysis: Created a pie chart to visualize the distribution of dark and non-dark patterns. Calculated and printed the class distribution and class imbalance ratio.

The database consisted of 2 columns after removing unnecessary columns from the datasets. After preprocessing 1708 data points were left and then the authors merged the open-source dataset with the collected data. Out of these, 196 (11.5%) were dark patterns and 1512 (88.5%) were non-dark patterns indicating a class-imbalance ratio of 0.13. To

combat this, the authors first transform the training data into numerical vector data and apply SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic data for addressing the problem of class imbalance.

Fig 6. (a) Word cloud for dark

Fig 6. (b) Word cloud for non-dark patterns

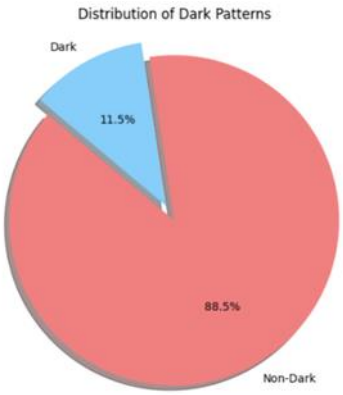

Fig 7. Distribution of dataset

4.2. Classification Results for dark and non-dark pattern classification

The table 1 presents a comparison of different classifiers' performance in differentiating between dark and non-dark patterns. It evaluates their accuracy, precision, recall, F1 score, ROC AUC, and sensitivity. The Bernoulli Naive Bayes classifier exhibits balanced metrics with a value of 0.95, however, it has a lower ROC AUC of 0.78. Gradient Boosting exhibits a modest improvement overall (0.96) with a higher ROC AUC (0.83). Logistic Regression has strong performance with a ROC AUC of 0.90, achieving an accuracy of 0.97. The SVM and AdaBoost algorithms are the highest performing models, both obtaining a 0.98 accuracy, precision, recall, and F1 score. The SVM algorithm has a perfect sensitivity of 1.00, while AdaBoost maintains a high ROC AUC of 0.91. Random Forest demonstrates robust performance, with a ROC AUC of 0.91, which is the highest of the models. The K-Nearest Neighbors (KNN) algorithm exhibits a high level of precision, with a value of 0.95. However, it falls short in terms of recall (0.89) and overall accuracy (0.89), suggesting that it fails to identify a significant number of positive examples when compared to other classifiers. In general, Support Vector Machines (SVM) and AdaBoost provide the optimal equilibrium and superior performance in identifying dark patterns.

Table 1. Results for Dark and Non-Dark Pattern Classification

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 score*	*ROC AUC*	*Sensitivity*
Bernoulli Naive Bayes	0.95	0.95	0.95	0.95	0.78	0.99
Gradient Boosting	0.96	0.96	0.96	0.96	0.83	0.99
Logistic Regression	0.97	0.97	0.97	0.97	0.90	0.99
SVM	0.98	0.98	0.98	0.98	0.89	1.00
Random Forest	0.97	0.97	0.97	0.97	0.91	0.99
AdaBoost	0.98	0.98	0.98	0.98	0.91	0.99
KNN	0.89	0.95	0.89	0.90	0.94	0.88

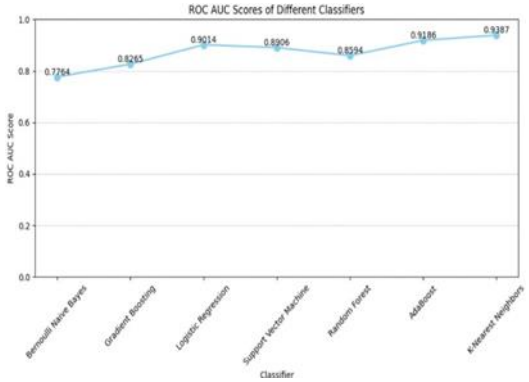

Fig 8. ROC AUC Scores of Different Classifiers

4.3. Preprocessing for deceptive/non-deceptive classification

Data Cleaning: Handled missing values, duplicates, and irrelevant features. Encoded the target variable ('Deceptive?') into uniform integers.

Data Concatenation: Merged multiple datasets to create the

final dataset.

Text Data Preprocessing: Cleaned and normalized the textual data. Removed special characters, punctuation, and stop words. Lowercased the text and applied stemming/lemmatization.

Class Distribution Analysis: Provided an overview of the distribution of deceptive and non-deceptive patterns. Used visualizations (e.g., pie chart) to illustrate class distribution.

Class Imbalance Analysis: Analyzed the imbalance ratio between deceptive and non-deceptive patterns. Discussed the implications of class imbalance on model training and evaluation

The final database consisted of 2 columns after removing unnecessary columns from the datasets, one for the strings and the other indicating if they are a deceptive dark pattern. In total 1913 data points are left after preprocessing and then these are merged with the open-source dataset. Out of these, we had 358 (18.7%) deceptive dark patterns and 1555 (81.3%) non-deceptive dark patterns indicating a class-imbalance ratio of 0.23. To combat this, the authors first transform the training data into numerical vector data and apply SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic data for addressing class imbalance.

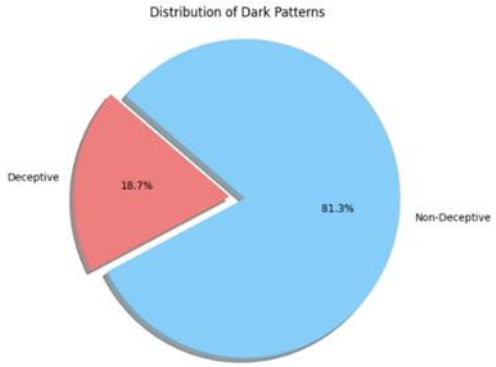

Fig 9. Distribution of dataset for deceptive and non-deceptive patterns

Fig 10. (a) Word Cloud for Deceptive

Fig 10. (b) Non-Deceptive Patterns

4.4. Results for deceptive/non-deceptive classification

The table 2 provides a concise overview of the performance metrics of different classifiers utilized for discriminating between deceptive and non-deceptive patterns. The Bernoulli Naive Bayes model demonstrates an accuracy of 0.78, along with a commendable precision of 0.82, indicating its proficiency in reducing false positives. However, it exhibits a relatively lower sensitivity of 0.65. Gradient Boosting and Logistic Regression exhibit comparable performance, achieving accuracies of 0.77 and 0.78 respectively. However, they demonstrate lower ROC AUC values, suggesting a moderate capability in distinguishing between classes. The Support Vector Machine (SVM) algorithm demonstrates superior performance with the highest accuracy (0.79) and well-balanced metrics. However, it has a comparatively lower sensitivity of 0.47. The Random Forest model consistently achieves an accuracy of 0.78. However, it has the lowest sensitivity of 0.34 and ROC AUC of 0.61, suggesting that it faces challenges in correctly identifying positive events. AdaBoost achieves comparable performance to Logistic Regression and SVM, with an accuracy of 0.78 and balanced metrics. The K-nearest neighbors (KNN) algorithm exhibits the lowest accuracy rate of 0.61. However, it demonstrates excellent precision (0.80) and sensitivity (0.79), suggesting its effectiveness in minimizing false positive results, albeit with reduced overall efficiency. Overall, Support Vector Machines (SVM) and Bernoulli Naive Bayes provide a favorable equilibrium in terms of performance for identifying deceptive patterns.

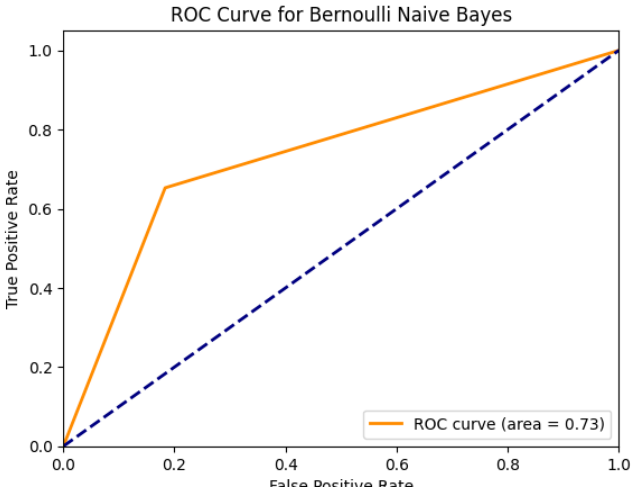

Fig 11. AUC-ROC curve for Bernoulli Naive Bayes

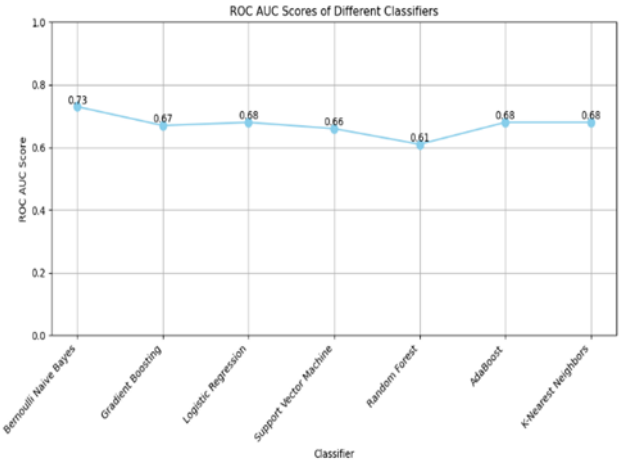

Fig. 12 ROC AUC Scores of Different Classifiers

Table 2. Results for Deceptive and Non-Deceptive Pattern Classification

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 score*	*ROC AUC*	*Sensitivity*
Bernoulli Naive Bayes	0.78	0.82	0.78	0.80	0.73	0.65
Gradient Boosting	0.77	0.79	0.77	0.78	0.67	0.50
Logistic Regression	0.78	0.80	0.78	0.79	0.68	0.51
SVM	0.79	0.79	0.79	0.79	0.66	0.47
Random Forest	0.78	0.77	0.78	0.77	0.61	0.34
AdaBoost	0.78	0.79	0.78	0.79	0.68	0.51
KNN	0.61	0.80	0.61	0.66	0.68	0.79

dark patterns

The preprocessing steps undertaken to prepare the dataset obtained from the open-source repository provided by Mathur et al. for subsequent analysis and modeling tasks. The preprocessing phase plays a critical role in ensuring data quality, consistency, and suitability for machine learning algorithms.

Data Cleaning: Null values were removed, and duplicate rows were eliminated to ensure data integrity.

Feature Selection: Only the "Pattern String" column, containing textual descriptions of patterns, was retained as the input feature.

Label Encoding: Categorical labels in the selected classification column were encoded into numerical representations using label encoding.

Class Distribution Analysis: An analysis of class distribution provided insights into the frequency of each class within the dataset.

Class Imbalance Assessment: The imbalance ratio between classes was computed to assess potential class imbalance challenges

4.6 Classification Results for categories of dark patterns

The table 3 displays the performance of a variety of classifiers in the classification of dark and non-dark patterns. The Multinomial Naive Bayes algorithm achieves a high accuracy (0.95) with a balanced precision, recall, and F1 score (0.95), as well as an outstanding ROC AUC (0.997). However, its sensitivity was slightly lower (0.93). The highest accuracy (0.97) and balanced metrics (all 0.97) are exhibited by Logistic Regression, which also has the highest ROC AUC (0.997) and sensitivity (0.95). Both Random Forest and SVM exhibit robust performance and high ROC AUC (0.997 and 0.996, respectively), with accuracies of 0.96 and equivalent balanced metrics. Gradient Boosting exhibits a slightly lower performance, with an accuracy of 0.93 and all metrics at 0.93. Additionally, it has a slightly lower ROC AUC (0.991) and sensitivity (0.89). KNN exhibits comparable performance to Multinomial Naive Bayes, with an accuracy of 0.95 and balanced metrics. However, it has a lower ROC AUC (0.985) and sensitivity (0.92). The Decision Tree exhibits excellent performance, with an accuracy of 0.94 and balanced metrics. However, its ROC AUC (0.960) and sensitivity (0.91) are lesser. The classification task is not effective for AdaBoost, as it exhibits the lowest accuracy (0.58), precision (0.54), recall (0.58), F1 score (0.52), ROC AUC (0.75), and sensitivity (0.47). In general, the most effective methods for classifying dark patterns are Logistic Regression, SVM, and Random Forest.

Table 3. Results for dark/non-dark classification

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 score*	*ROC AUC*	*Sensitivity*
Multinomial Naive Bayes	0.95	0.94	0..95	0.95	0.997	0.93
Logistic Regression	0.97	0.97	0.97	0.97	0.997	0.95
SVM	0.96	0.96	0.96	0.96	0.997	0.94
Random forest	0.96	0.95	0.96	0.95	0.996	0.94
Gradient Boosting	0.93	0.93	0.93	0.93	0.991	0.89
KNN	0.95	0.95	0.95	0.95	0.985	0.92
Decision Tree	0.94	0.94	0.94	0.94	0.960	0.91

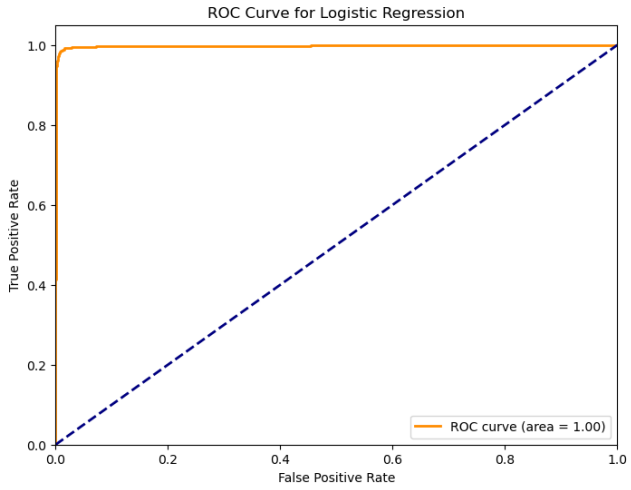

Fig 13. ROC curve for logistic regression

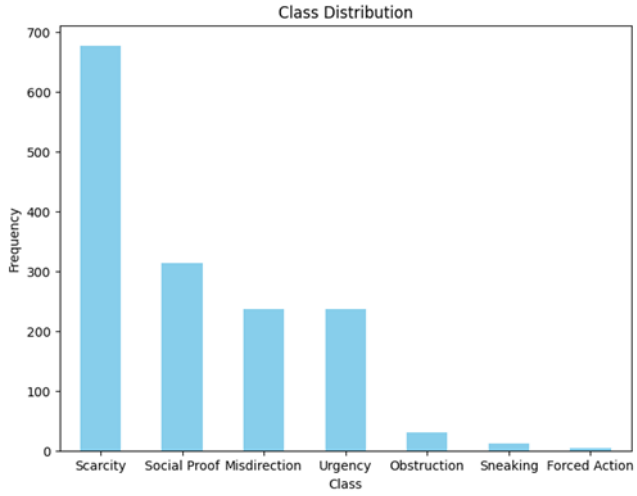

Fig 14. Distribution of various types of dark patterns in the data

5. Conclusion

This study assessed the efficacy of different classifiers in recognizing obscure and misleading patterns in digital interfaces. The classifiers were evaluated on various performance metrics, such as Accuracy, Precision, Recall, F1 Score, ROC AUC, and Sensitivity, for different classification tasks.

Differentiation between Dark and Non-Dark Patterns:

Support Vector Machine (SVM) and Adaptive Boosting (AdaBoost) appeared as the two best performing models in discerning between dark and non-dark patterns; their accuracy was excellent, alongside precision, recall, and F1 score. These are highly reliable algorithms for spotting dark patterns and ensuring they do not significantly affect user experience.

Differentiation between Deceptive and Non-Deceptive Patterns:

SVM and Bernoulli Naive Bayes have been found to strike a good balance between their performance capabilities when it comes to distinguishing deceptive from non-deceptive patterns. These two algorithms with high accuracy and metrics equilibrium can offer effective ways of detecting what makes up deceitful digital aspects. Gradient Boosting, Logistic Regression, and AdaBoost similarly present similar results although they slightly differ in sensitivity from one another as well as ROC AUC values. All three provide sensible choices in spotting deceitful design elements; even though not all are equally effective, there is certainly an option available for every situation.

Classification of Dark Patterns:

Logistic Regression, SVM, and Random Forest were identified as the most effective models in classifying dark and non-dark patterns. These models exhibited good accuracy, with Logistic Regression achieving a score of 0.97 and SVM and Random Forest achieving scores of 0.96. Furthermore, these models demonstrated balanced metrics across all categories. These models also demonstrated exceptional ROC AUC scores (0.997 for Logistic Regression and SVM, 0.996 for Random Forest), which indicates their strong ability to differentiate across classes. Multinomial Naive Bayes and KNN demonstrated strong performance, albeit slightly inferior to the leading classifiers. On the other hand, AdaBoost exhibited markedly inferior performance, rendering it inappropriate for this particular assignment.

Overall implications:

The findings underscore the importance of utilizing machine learning techniques for detecting and mitigating dark patterns in digital interfaces. By leveraging sophisticated algorithms such as SVM, Logistic Regression, and Random Forest, designers and developers can enhance user trust, transparency, and satisfaction while minimizing the prevalence of deceptive design practices. Additionally, the study highlights the need for ongoing research and refinement of classification models to address evolving challenges in digital ethics and user-centered design.

References

[1] "Deceptive Patterns - Home." Accessed: Jan. 23, 2024. [Online]. Available: https://www.deceptive.design/

[2] "The ultimate list of 70+ eCommerce facts and statistics for 2024 - AppMySite." Accessed: Jan. 23, 2024. [Online]. Available: https://www.appmysite.com/blog/ultimate-ecommerce-facts-and-statistics/

[3] W. C. Koh and Y. Z. Seah, "Unintended consumption: The effects of four e-commerce dark patterns," *Clean. Responsible Consum.*, vol. 11, no. 3, p. 100145, Dec. 2023, DOI: 10.1016/j.clrc.2023.100145

[4] "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites." Accessed: Jan. 23, 2024. [Online]. Available: https://webtransparency.cs.princeton.edu/dark-patterns/

[5] "dark-patterns/data/final-dark-patterns/dark-patterns.csv at master · aruneshmathur/dark-patterns." Accessed: Jan. 23, 2024. [Online]. Available:https://github.com/aruneshmathur/dark-patterns/blob/master/data/final-dark-patterns/dark-patterns.csv

[6] Brignull, H. (2013). Dark Patterns: Inside the Interfaces Designed to Trick You. Retrieved from https://darkpatterns.org/

[7] A. Bhattacherjee, "Understanding consumers' aversion to deceptive online advertising: A model and its validation," Journal of the Association for Information Science and Technology, vol. 71, no.10, pp.1264-1278, 2020.

[8] Z. Zhang, S. Han, S and Y. Li, "Detecting dark patterns on the web using machine learning and human computation," In Proceedings of the 2020 Conference on Computer-Supported Cooperative Work and Social Computing, pp. 1-11, 2020.

[9] P. Garaizar, J. F. Bonnefon and E. R. Igou, "A roadmap for the study of dark side phenomena in information systems," Computers in Human Behavior, vol. 86, pp. 387-396, 2018.

[10] A. Hakkak, T. Latham and L. Hines, "Designing and evaluating a dark patterns detection browser extension," In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-13, 2018.

[11] J. Bergstrom and A. Blomberg, "Designing to evade dark patterns: Investigating how designers perceive and cope with unethical persuasion attempts," In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-13, 2020.

[12] Q. V. Liao, Y. Yuan, and S. Wang, "Dark patterns at scale: Findings from a crawl of 11K shopping websites," In Proceedings of the 2018 World Wide Web Conference, pp. 1053-1062, 2018.

[13] C. Hansen and F. Motti-Stefanidi, "Understanding and mitigating the impact of dark patterns in user interaction design," In Proceedings of the 2021 Conference on Human Factors in Computing Systems, pp. 1-15, 2021.

[14] E. Luger and T. Rodden, "Exploring deceptive interfaces that manipulate task completion times," In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3609-3618, 2015.

[15] A. Mathur, A. Vance and M. Neff, "Evaluating dark patterns in games: A first empirical study," In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-11, 2019.

[16] P. Garaizar, J. F. Bonnefon and E. R. Igou, "Crowdsourcing the detection of dark patterns in user interfaces," ACM Transactions on Computer-Human Interaction (TOCHI), vol. 27, no. 5, pp. 1-27, 2020.

[17] A. Nasr, "Dark patterns: The story of deceptive design," In Proceedings of the 25th International Conference on Pattern Recognition, pp. 1-7, 2019.

[18] S. Mills and R. Whittle, "Detecting Dark Patterns Using Generative AI: Some Preliminary Results," Oct. 2023 Available at SSRN:https://ssrn.com/abstract=4614907 or DOI: 10.2139/ssrn.4614907

[19] S. R. Kodandaram, M. Sunkara, S. Jayarathna, and V. Ashok, "Detecting Deceptive Dark-Pattern Web Advertisements for Blind Screen-Reader Users," J. Imaging., vol. 9, no. 11, 239, 2023. DOI: 10.3390/jimaging9110239.

[20] Quigley-Simpson. Understanding Dark Patterns, and How They Impact Your Brand's Consumer Experience, 2021.

[21] Axelerant. Design Ethics: Navigating Dark Patterns and Building Trust. Retrieved from https://www.axelerant.com/blog/design-ethics-navigating-dark-patterns-and-building-trust, Jan. 2024.