

# AI Based Automated Detection of Dark Patterns on Websites

Prakathi V<sup>1</sup>, Srinithi M<sup>2</sup>, Vikashini C<sup>3</sup>, Yoga Dharshni R<sup>4</sup>, Dr.R.Ahila<sup>5</sup>

<sup>1,2,3,4</sup>Student, <sup>5</sup>Associate Professor of Computer Science and Engineering, School of Engineering

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu

## ABSTRACT

Websites and apps use dark patterns, which are misleading design strategies, to influence user behavior. They bring up important moral and pragmatic issues. Time pressure, dishonest tactics, distraction, obstruction, social manipulation, compelled behaviours, limited availability, and other dark patterns are all recognized by the artificial intelligence (AI) system presented in this study. Our approach uses natural language processing and machine learning algorithms to examine internet and application user interfaces and identify indicators that indicate the employment of dark patterns. The authors show how well their approach can recognize and classify dark patterns by thoroughly testing and validating it. This makes it easier to follow rules and promotes a more moral and open online community.

**Keywords:** *Dark patterns, deceptive design, natural language processing, artificial intelligence, machine learning, and user experience.*

## 1. INTRODUCTION

Websites and applications use dark patterns, which are deceptive tactics, to trick users into completing unexpected actions like making purchases or signing up for services. These Strategies gently persuade people to choose choices they might not have otherwise thought of by utilizing ideas from human psychology and interface design [1]. Among other things, dark patterns might manifest as deceptive visual cues, information that is hidden or phrasing that causes people to behave in unexpected ways. The deliberate use of wording well placed checkboxes or design components that use users to make decision are example of this phenomenon. In order to subvert ethical and explicit design standards, these dishonest methods

exploit cognitive biases and play with users' subconscious proclivities. Dark patterns are deceptive strategies that websites employ. They consequently jeopardize people's independence and capacity for informed decision-making when using digital platforms. To advance ethical design principles, foster user trust, and provide all users with a positive and

productive online experience, these dishonest and manipulative design techniques, sometimes referred to as "dark patterns," must be identified and addressed. Concern over dark with the opening of new stores, trends in the online e-commerce industry are evolving every day. An estimated 12 to 24 million e-commerce websites exist, according to Appmysite.com [2].

It makes sense that certain online retailers would use unethical tactics to boost their profits and surpass their rivals in sales, considering the size of these businesses. Dark patterns have the power to influence user behaviour and produce unforeseen, unsustainable expenditure trends. When people are influenced by dishonest design tactics, they may shop excessively and purchase more than they need or unnecessary products. These deceptive practices quietly influence customers to make decisions that are not in line with their true needs, taking advantage of their weaknesses. [3] Dark patterns trick people into going beyond their intended objectives by using deceptive visual cues, persuasive language, or concealed information, which leads to the unnecessary purchase of goods or services.

## 2. LITERATURE REVIEW:

The vast bulk of past studies on dark patterns have concentrated on uncovering unethical design techniques and their effects on user rights and experience. To find dark patterns, many academics have proposed employing heuristic evaluations or human inspection methods; however, these methods are frequently not scalable and could overlook small changes. In recent years, there has been promise for more precise and effective detection of dishonest design elements through the automation of suspicious pattern recognition using artificial intelligence and machine learning techniques.

1. Title: E-commerce's dark patterns: a dataset and preliminaries

Tsuneo Matsumoto, Hayato Yamana, Nao Fukushima, Jiaying Feng, Yuki Yada

Dark patterns are UI styles used in internet services that cause unintended actions. Concerns about equity and

privacy have recently been raised in relation to dark patterns. As a result, a wide range of studies on dark pattern detection are eagerly anticipated. In this study, we built a dataset for dark pattern detection and used the latest machine learning techniques to develop its baseline detection performance. The original dataset, which included 1,818 dark pattern texts from shopping sites, was taken from a 2019 study by Mathur et al. Negative samples, or non-dark pattern texts, were then added by extracting texts from the same websites as the Mathur et al. dataset. In order to demonstrate the accuracy of automatic detection, we also used cutting-edge machine learning techniques like BERT, RoBERTa, ALBERT, and XLNet as baselines. After five-fold cross-validation, RoBERTa produced the best accuracy of 0.975.

**2. Title: Aid UI: Transitioning to Automated Dark Pattern Identification in User Interfaces**  
Samiha Salma, Damilola Awofisayo, Kevin Moran, and S M Hasan Mansur are the authors.  
Year of Publication: 2023 Software Engineering (ICSE) 45th International Conference, IEEE/ACM  
The frequency of UI dark patterns—user interfaces that could cause end users to (unknowingly) carry out behaviors they may not have intended—has been shown in earlier studies. Such misleading user interface designs, which might be purposeful (to promote an online business) or inadvertent (due to complicit design practices), may cause end users to overshare personal information or suffer financial losses.

Despite extensive study on creating taxonomies of dark patterns in various software domains, developers and users still lack assistance in identifying, avoiding, and navigating these incredibly subtle design motifs. By introducing AidUI, a novel automated method that recognizes a set of textual and visual clues in application images that indicate the presence of 10 distinct dark patterns in user interfaces, we begin to understand how common UI dark patterns can be automatically identified in modern software programs. This enables us to locate, identify, and categorize these dark patterns. However, a single pattern type may instantiate in multiple ways, leading to a high degree of unpredictability, making automatic detection of dark patterns a difficult process. We have created Context DP, the largest dataset of fully-localized UI dark patterns to date, which includes 301 examples of dark patterns across 175 mobile and online user interface screenshots, in order to assess our methods. According to our evaluation's findings, AidUI can identify instances of dark patterns with an overall

precision of 0.66, recall of 0.67, and F1-score of 0.65. It returns few false positives and has an IoU score of 0.84, which allows it to localize patterns that are discovered. Additionally, a sizable portion of the dark patterns we examined might be consistently identified (F1 score higher than 0.82), and other research avenues might make it possible to locate more patterns with greater success.

**3. Title: Automated identification of ten suspicious cookie notification patterns**  
Daniel Kirkman, Daniel W. Woods, and Kami Vaniea published it in 2023. IEEE's eighth edition of the European Symposium on Security and Privacy (EuroS&P). Consumers can theoretically define their privacy preferences regarding how a website and its partners handle their personal information through consent dialogues. In actuality, dialogs frequently use delicate design strategies called "dark patterns" that nudge users in the direction of allowing more data processing than they otherwise would. In addition to potentially breaking privacy regulations, dark patterns erode user autonomy. Our Dark Dialogs engine automatically collects arbitrary consent dialogs from websites and looks for ten dark patterns. When compared to a hand-labeled dataset, Dark Dialogs extracts dialogs with 98.7% accuracy and accurately classifies 99% of the dark patterns under study. When we implemented Dark Dialogs on a sample of 10,992 websites, 2,417 permission dialogs were successfully gathered, and 3,744 distinct dark patterns were automatically discovered. present at the consent discussions. We then examine whether the prevalence of dark patterns is related to the amount of ID-like cookies, the popularity of the website, and the existence of a third-party consent management provider.

**4. Title: Adorability as a 2018 Dark Trend in Domestic-Robots**  
Authors: Catherine Caudwell, Cherie Lacey  
Dark patterns are a relatively new idea in interface design that combines behavioral psychology and design patterns to trick the user. Nonetheless, home robots should be included in the present body of work on dark patterns, which focuses mostly on digital interactions that take place on screens. In this study, we apply the concept of dark patterns to the "cute" aesthetic of house robots and propose that their design is a dark pattern in HRI because it (1) prioritizes short-term gains over long-term decisions; (2) deprives users of some conscious agency during interaction; and (3) elicits an emotional response from the user in order to gather emotional data. In order to lay the foundation for an ethical design approach in HRI, this exploratory endeavour extends the existing

library of dark patterns and their application to new technology interfaces into the home robotics arena.

5.Title: Enhancing Privacy by Redesigning Dark Patterns  
The authors are Rodrigo Hernández-Ramuedrez, Davide Maria Parrilli, Released in 2020 Technology and Society International Symposium (ISTAS) by IEEE .

In digital design, dark patterns are extremely unethical techniques designed to collect as much personal information as possible from consumers, usually without their agreement. Nevertheless, the techniques used by dark patterns can be modified to improve user privacy making them into moral instruments. Dark patterns can be redesigned to guide users toward selecting the most stringent privacy settings, according to this ongoing study. According to ethical design principles, this suggests a significant conversion of something fundamentally negative into a tool that serves the public interest.

6.Title: Recognizing Negative Trends in Social Robotics Conduct

Writers: Elizabeth Phillips, Andres Rosero, Elizabet Dula. Social robots have been used more and more in private settings, where they can be used as caregivers for the elderly, provide general emotional or physical support, provide entertainment, and teach kids. Robotics businesses have started using robots that can recognize emotions and react with emotionality in return to support these more personal connections. This artificial emotional connection creates the possibility of manipulating and exploiting users through dishonest robot design. Deceptive design patterns known as "dark patterns" are used by apps and websites to trick users into taking unexpected behaviours. We contend that social robotics can be taught to harness these unidirectional human-robot emotional ties to deceive users by introducing dark patterns. This could lead to the exploitation of vulnerable groups, such as the elderly and children. We propose methods in which dark patterns may appear in these connections, drawing on the research on social robotics and dark patterns. Furthermore, we offer guidelines for moral behavior in the development of emotional social robots.

7.Title: Risk Analysis of Discovering Corrupt UX E-commerce Application Patterns Impacting Personal Information Authors: Nimkoompai Apichaya, The PDPA, which emphasizes the protection of personal data in the digital realm, is consistent with the current

focus on designs known as the "Dark Patterns of UX," particularly in e-commerce. This work focuses on mobile applications because they are easier to use. In order to draw conclusions for a future awareness campaign, the researcher gathered information on dark patterns from relevant, possibly vulnerable samples as well as from other sources. 58.3% of the sampled group did not know about dark patterns, according to the survey. Designers most frequently employed veiled advertisements (59.3%) and forced continuity (71.8%) as deceptive dark patterns. Apps that demand payment before use have been shown to have the highest risk of exposing personal information.

8.Title: Examination of Tweets Associated with the Dark Pattern in 2010

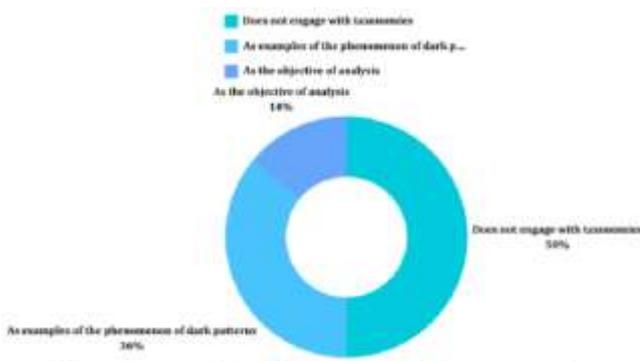
User interfaces known as "dark patterns" lead users to take actions they didn't want to take, such buying something or registering for a service. The employment of dark patterns is regarded as a violation of users' privacy and rights. We demonstrate users' responses to dark patterns in 2019 by analyzing 12 years of tweets. Our results indicate that: 1) users in countries where dark pattern laws are in place have more conversations about them; 2) tweets about dark patterns shifted from expressing their diversity to taking action against them around 2017; and 3) sneaking, obstruction, and interface interference—all of which are common on e-commerce websites—are the most talked-about types of dark pattern tweets. Our findings may help lawmakers and regulators promote safer internet use.

9.Title: A Comparative Study of Dark Pattern and Anti-Pattern to Improve Application Design Efficiency  
The authors are Apichaya Nimkoompai, Pumarin Tiangpanich, Offline marketing is beginning to give way to social media marketing as technology has advanced more quickly than anticipated. As a result, marketers are already using mobile applications to show consumers their brands and goods. To make information appealing, though, the user interface (UI) must be used to portray it in ways that are easy to grasp and the user experience (UX) must be used to facilitate smooth communication between businesses and consumers. Nowadays, marketing relies heavily on information. result in an organization attempting to obtain users' personal information without their consent, which could cause harm to users' property through the Dark Pattern of User Experience (UX). Nevertheless, when the application is being designed, the designers may observe that the design is struggling or lacks a solid solution to address the issue

that users may encounter. The design experience known as Anti-patterns can address this issue, wherein typical problems may seem to have straightforward solutions, but in reality, they may not be the best. Even if the dark pattern of user experience (UX) is dishonest, the purpose of UX/UI design, or interactions developed with psychological knowledge, is to trick customers into doing something they didn't intend to do in order to add value for the service they work for. With the help of the appropriate tools, designers should be able to produce visually striking artwork. recognizing and properly utilizing anti-patterns and dark patterns.

**2.1 Classification of dark patterns:**

These are deceptive design strategies used by apps and websites to influence user behaviour; they usually result in unexpected actions or consequences that prioritize the platform owner's interests over that of the user [6]. Using psychological ideas and cognitive biases, these design tactics coerce people into making choices or doing actions they might not otherwise choose to do voluntarily [7].



**Engagement with types and taxonomies**

**Fig. 1.** The importance of recognizing dark patterns

**1. Misdirection:** Design elements that purposefully mislead or distract users from their intended behaviors fall under the category of misdirection.

**2. Social Proof:** To influence user behavior, dark patterns in this category use deceptive techniques like fabricated user testimonials or phony social endorsements to sway decisions about what to buy [9].

**3. Friend Spam:** This dark pattern in digital design refers to a service that deceives users into sending unwanted or fraudulent messages to their contacts without their knowledge or consent social media is used in this tactic to increase website and app traffic, sign-ups, and content [20]. When dark patterns purposefully create barriers or hurdles to keep users from doing desired objectives, this is referred to as obstruction. Checkout procedures that are

purposefully complicated or challenging are an example of obstruction [11].

**4. Confirm Shaming:** This dishonest tactic pressures customers into accepting something by embarrassing or blaming them. Customers are under pressure to comply when they select the refuse option since it makes them feel terrible or careless. In order to change user behavior toward service goals, this method makes use of social and emotional variables [20][21].

**Previous methods for detecting dark patterns:**

Human-computer interaction, computer science, and user experience design are among the fields that have looked into how to spot dark patterns in digital interfaces. A variety of methods, including hand examination and computer algorithms, have been used in the past to approach dark pattern detection. Understanding these tactics could improve one's comprehension of the challenges and opportunities involved in spotting dishonest design practices.

**1. Manual Inspection:** To find misleading design aspects, researchers study interfaces by hand. This process offers precise insights, but it is time-consuming and unfeasible for large-scale assessments [13].

**2. Heuristic Evaluation:** Researchers use established usability heuristics to find dark patterns in interfaces. Though it may be weak in specificity and consistency, this method provides insightful feedback [14].

**3 Automated Algorithms:** Machine learning techniques and algorithms are used to scan interfaces for deceptive patterns in order to rapidly evaluate large datasets and identify subtle changes [8].

**4. Browser Extensions:** By giving users real-time alerts about misleading design elements, browser extensions empower users to make informed choices.[10].

**5. Crowdsourcing:** To find and classify dark patterns, researchers use crowdsourcing platforms to enlist users. This provides a range of perspectives, but it may also raise issues with data quality and dependability [12].



**Fig 2.** Different approaches to dark patterns

## 2.2 Artificial Intelligence's Function in Dark Pattern Identification

Artificial intelligence (AI)-driven methods evaluate digital interfaces and spot patterns that could indicate dishonest design components using computer vision, machine learning, and natural language processing technologies. The patterns is depicted in Figure 3. Since artificial intelligence (AI) provides advanced computational methods for identifying dark patterns, it is essential for identifying and countering misleading design strategies in digital interfaces

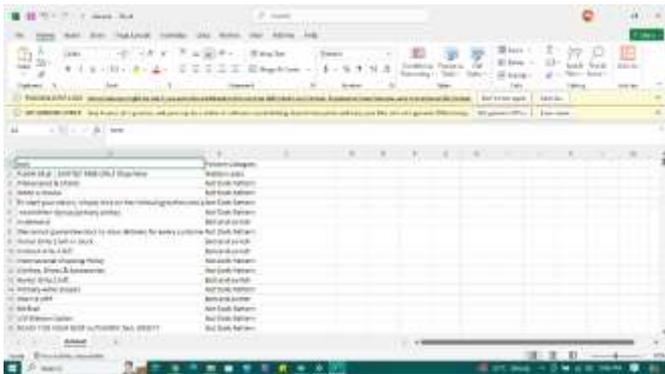


Fig.3 AI is used to detect dark patterns.

1. Artificial intelligence (AI) tools make it easier to automatically identify dark patterns by analyzing large datasets and spotting minor clues of deceptive design tactics. To find patterns connected to a variety of dark pattern categories,

such as deception, coercion, and shortages, machine learning models can be trained on labeled datasets. Techniques driven by AI provide the ability to scale and increase the effectiveness of spotting dark patterns. This makes it possible to monitor digital interfaces in real time and analyze large datasets. Large volumes of user interactions can be analyzed by automated algorithms, which can then quickly and accurately identify potential instances of dishonest design tactics. This enables timely actions to protect users against manipulation.

## 3.SYSTEM MODEL

### 3.1 EXISTING SYSTEM

Current research on dark pattern detection includes various methodologies, from dataset creation and machine learning evaluations to automated detection systems. Notable works focus on specific types of dark patterns, such as cookie consent dialogs or e-commerce sites. These studies employ advanced machine learning models like BERT and GPT-3, or use computer vision and NLP techniques for detection. There is a need for a more

generalized and efficient system that can detect a broad range of dark patterns across diverse online environments.

### DRAWBACKS OF EXISTING METHOD:

Only limited to specific domains (e.g., shopping sites, cookie dialogs). High computational cost and complexity in model training. Moderate precision and recall, leading to potential false positives/negatives. May struggle with detecting complex or subtle dark patterns.

### 3.2 THE SYSTEM PROPOSED

By creating an automated strategy with a Naive Bayes classifier, the suggested system seeks to overcome the drawbacks of current dark pattern identification techniques. The system may identify a number of dark pattern types, such as Sneaking, Price Comparison Prevention, Bait and Switch, Forced Continuity, and Hidden Costs. The system retrieves and analyzes textual content from webpages using web scraping methods and the TFIDF (Term Frequency - Inverse Document Frequency) vectorizer for data preprocessing. The trained classifier then examines this data to find design features that are misleading. This method improves user safety, encourages openness in online interactions, and guarantees thorough detection across various website kinds.

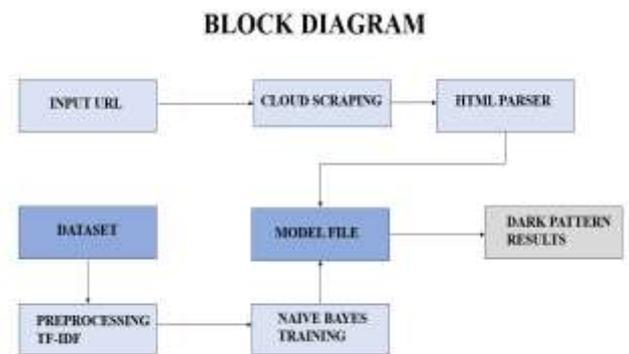


Fig 4. Block diagram for dark pattern detection

**BLOCK DIAGRAM DESCRIPTION:** The implementation is made up of two primary blocks, the Graphical User Interface (GUI) Block and the Machine Learning Block, each of which performs unique but related tasks. The URL is first entered by the user in the ML block. The URL is then compared to the model file using the Naive Bayes classifier, which predicts the output and indicates whether or not there is a dark pattern.



### Module 5: System Development

1. System design: Design a user-friendly interface for the automated system. Develop a backend to integrate the trained model.
2. System implementation: - Implement the system using a suitable programming language (e.g., Python). Integrate web scraping, data preprocessing, and model classification.

### Module 6: Testing and Deployment

1. System testing: - Test the system using various websites and dark patterns. Evaluate performance and accuracy.
2. Deployment: - Deploy the system as a web application or browser extension. Make the system accessible to users.

### Module 7: Maintenance and Updates

1. Continuous monitoring: - Monitor system performance and update the model as needed
2. User feedback: - Collect user feedback and improve the system accordingly.
3. Dark pattern updates: Update the system to detect new dark patterns and categories.

## 5. FUTURE WORK AND CONCLUSION

A scalable way to improve user safety and transparency in online settings is through the automatic dark pattern detecting system. By leveraging machine learning techniques and web scraping, the system effectively identifies and classifies deceptive design elements, which are often difficult to detect manually. The use of a Naive Bayes classifier combined with TFIDF vectorization allows for efficient data processing and classification, contributing to a safer online experience for users. It approaches the address gaps in existing methods, offering a more comprehensive and automated solution for detecting dark patterns across diverse websites.

In conclusion, the classification report shows that the Naive Bayes classifier has performed admirably in identifying a variety of dark patterns on websites. The classifier shows a strong capacity to detect misleading design aspects with excellent precision and recall across several categories, such as Bait and Switch, Hidden Costs, and Sneaking. The overall accuracy of 91% underscores the effectiveness of the implemented solution in providing users with reliable insights into potentially

harmful online practices. This achievement lays a solid foundation for advancing the system's capabilities, potentially exploring more sophisticated models and expanding the dataset to further refine the detection of dark patterns. The successful implementation of the Naive Bayes classifier establishes a valuable tool for users to navigate the digital landscape with increased awareness and protection against deceptive practices, contributing to a safer and more transparent online experience. By expanding the dataset and researching other machine learning algorithms, the system can be further enhanced in the future to boost the model's accuracy and generalizability. Additionally, incorporating advanced natural language processing techniques and real-time updates for evolving patterns will contribute to a more adaptive and resilient dark pattern detection system. The GUI can be enriched with features such as visualization tools and detailed explanations of detected patterns to enhance user understanding. Collaboration with browser extensions or integration into security software could extend the reach and impact of this system, contributing to a safer and more ethical digital experience for users.

## REFERENCE

- [1] Chauhan, K. D. S., & Anupriya. (2025). Darker patterns? AI-generated persuasion and the regulatory void in Indian law. *Journal of Development Policy and Practice*, 10(1), 80-95.
- [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: 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. Bhattacharjee, "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. 1-13, 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 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.