Designing an Innovating Solution to detect Dark Patterns in E-Commerce website using LLM

Sriram B
SCOPE
Computer Science Core
Chennai, India
21BCE1674

Narasimhan H
SCOPE
Computer Science Core
Chennai, India
21BCE1617

Abhyudaya Augustya
SCOPE
Computer Science Core
Chennai, India
21BCE1631

Vatchala S
SCOPE
Computer Science Core
Chennai, India
52828

***Abstract*—In this research, we present an initiative aimed at detecting and addressing common dark patterns prevalent in marketing websites. These dark patterns encompass deceptive tactics such as false urgency, where purportedly time-limited offers persist for extended durations, and URL jumpers, which maliciously redirect visitors to harmful or commercial sites. Additionally, our focus extends to mitigating the acquisition of unnecessary personal information by identifying instances where websites request irrelevant details, such as age, during the purchasing process. Leveraging language model-based approaches, our research seeks to expose and thwart these dishonest practices, thereby enhancing user awareness and overall browsing experience. Through systematic analysis and algorithmic interventions, we aim to contribute to the ongoing discourse on digital ethics and consumer protection in the online realm.**

***Index Terms*—Dark patterns, false urgency, false timer, URL jumpers, marketing websites, language model-based approaches, cybersecurity, user awareness.**

I. Introduction

In recent years, the pervasive presence of dark patterns in digital interfaces has emerged as a significant challenge, undermining user autonomy and trust while exploiting cognitive biases for commercial gain. Dark patterns, deliberately crafted design elements intended to deceive users into unintended actions, manifest in various forms across e-commerce platforms, social media sites, and mobile applications. Despite growing awareness of this issue, combating dark patterns remains a complex endeavor, requiring innovative solutions to effectively detect and mitigate their harmful effects.

Recent advancements in dark pattern research have shed light on the prevalence, tactics, and impact of these deceptive design elements. Studies have uncovered a multitude of dark pattern techniques, ranging from misdirection and forced action to privacy exploitation, implemented across diverse digital platforms. Moreover, research has elucidated the detrimental effects of dark patterns on user well-being, including increased frustration, anxiety, and mistrust in digital services. This understanding underscores the urgent need for proactive measures to address the proliferation of dark patterns and safeguard user interests.

Artificial Intelligence (AI) offers promising avenues for combating dark patterns and promoting ethical design practices in digital interfaces. By leveraging machine learning algorithms, AI can analyze vast amounts of user data and interface elements to identify patterns indicative of deceptive design. Moreover, AI-powered systems can dynamically adapt to evolving dark pattern tactics, enabling continuous monitoring and mitigation efforts. Additionally, AI can facilitate the development of user-centric design frameworks, prioritizing transparency, clarity, and user empowerment in digital interactions.

The complexity and scale of modern digital ecosystems necessitate the use of AI-driven approaches to effectively tackle the dark pattern problem. Traditional manual methods for detecting dark patterns are labor-intensive and often inadequate in keeping pace with the rapidly evolving landscape of deceptive design tactics. AI, with its ability to process large volumes of data and detect subtle patterns, offers a scalable and efficient solution for identifying and mitigating dark patterns across various digital platforms. Furthermore, AI-driven interventions can empower users by providing real-time feedback and personalized recommendations, fostering a more transparent and trustworthy digital environment.

Our proposed solution presents a detailed methodology for the detection and mitigation of prevalent dark patterns within the realm of e-commerce websites. Beginning with web scraping and data collection, we leverage Python libraries such as BeautifulSoup and Pandas to systematically gather metadata from a diverse array of e-commerce platforms. This collected data, inclusive of website descriptions, keywords, and input fields, serves as the foundation for subsequent analysis.

II. Literature Review

Dark patterns, a term coined by Brignull (2013), refer to design elements deliberately crafted to deceive users, leading them to take actions they might not otherwise intend. These deceptive tactics manifest in various forms, such as misdirection, forced action, and privacy Zuckering (Ghosh et al., 2019). Misdirection involves diverting user attention from critical information, while forced action coerces users into unintended actions, like making unsubscribing difficult. Privacy Zuckering exploits users' privacy concerns, coercing them into sharing more personal information. These patterns not only undermine user autonomy but also erode trust in digital interfaces (Kumaraguru and Sheng, 2021).

Research has shown that dark patterns are pervasive across a wide range of digital platforms, including e-commerce websites, social media platforms, and mobile applications (Ghosh et al., 2019). For example, in e-commerce, tactics like ”sneak into basket” and ”bait and switch” are commonly used to trick users into making unintended purchases or subscribing to services without their explicit consent (Brignull, 2013). Similarly, on social media platforms, techniques like ”friend spam” and ”roach motel” are employed to manipulate user engagement and retention (Ghosh et al., 2019).

The prevalence of dark patterns raises significant ethical concerns, as they exploit users' cognitive biases and vulnerabilities for commercial gain (Acquisti et al., 2017). By intentionally deceiving users, designers and companies undermine the principles of transparency, trustworthiness, and respect for user autonomy. Moreover, dark patterns can have detrimental effects on user well-being, leading to frustration, anxiety, and mistrust in digital services (Ghosh et al., 2019).

Addressing the proliferation of dark patterns requires a multifaceted approach that involves collaboration between designers, policymakers, and industry stakeholders (Kumaraguru and Sheng, 2021). Designers must prioritize ethical design principles, ensuring transparency, clarity, and user empowerment in their interfaces. Policymakers can enact regulations to curb deceptive practices and hold companies accountable for dark pattern usage. Industry stakeholders, including technology companies and advocacy groups, can promote awareness and best practices for ethical design.

In conclusion, dark patterns represent a significant challenge in the digital landscape, posing ethical dilemmas and undermining user trust and autonomy. By understanding their prevalence, impact, and ethical implications, stakeholders can work towards fostering a more transparent, user-centric digital environment.

III. Methodology

Web Scraping and Data Collection: Utilize Python libraries such as BeautifulSoup and Pandas to systematically scrape metadata from a diverse range of e-commerce platforms. Gather comprehensive data including website descriptions, keywords, and input fields to form the foundational dataset.

Data Preprocessing: Conduct meticulous preprocessing to rectify null values and enhance the dataset with human interpretation. Integrate a dedicated ”comments” column to facilitate nuanced understanding and identification of dark patterns.

Incorporation of Explainable AI (XAI) Methodologies: Enable users to contribute interpretive comments, augmenting Language Models (LLMs) for providing clear explanations. Leverage XAI techniques to enhance the capacity of LLMs in deciphering decision-making processes.

Dataset Tailoring for BERT Model Utilization: Curate datasets specifically tailored for BERT model utilization, with a focus on phishing website detection. Conduct meticulous preprocessing and apply an Artificial Neural Network (ANN) trained with BERT embeddings to enhance detection capabilities.

Extension Development: Develop a sophisticated Chrome extension designed to automate the extraction of metadata, input fields, and links from e-commerce websites. Ensure real-time operation of the extension for continuous monitoring and detection of dark patterns.

Detection of Dark Patterns: Focus on meticulous detection of dark patterns, particularly targeting instances of false urgency and timers. Utilize sentiment analysis and logical reasoning techniques to scrutinize website elements and unveil deceptive stratagems, safeguarding user trust and integrity.

Backend Processing: Integrate advanced analytical methodologies for comprehensive backend processing to address the nuanced intricacies of dark pattern detection. Utilize BERT transformers and LLMs to meticulously scrutinize website URLs for potential malicious activity and scrutinize the misuse of personal information.

Analysis and Explanation Provision: Conduct exhaustive analysis of detected dark patterns. Provide elucidative explanations to users to enhance their awareness and understanding of detected dark patterns.

Through the systematic execution of these methodologies, our solution aims to robustly detect and mitigate prevalent dark patterns within e-commerce websites, ultimately fostering a safer and more transparent online browsing environment.

IV. Results

The results of our study, conducted to detect and mitigate dark patterns within e-commerce websites, are presented herein. Utilizing a methodology comprising web scraping, data preprocessing, and analysis, we attained comprehensive insights into the prevalence and nature of deceptive design practices. Initially, input data in CSV format containing website metadata was collected. Employing web scraping techniques utilizing Python libraries such as requests, pandas, and BeautifulSoup, metadata including descriptions, keywords, and input fields was systematically extracted from diverse e-commerce platforms. Subsequently, a meticulous preprocessing phase was conducted to rectify null values and augment the dataset with human interpretation regarding the presence of dark patterns. Additional columns were appended to the dataset to facilitate this interpretation. The output of our methodology culminated in a refined dataset, enriched with human insights regarding dark patterns. This dataset serves as a foundational resource for subsequent analysis and detection efforts.

In summary, our study showcases the successful application of a comprehensive methodology to identify and address dark patterns within e-commerce websites. The results presented herein underscore the importance of vigilant scrutiny and

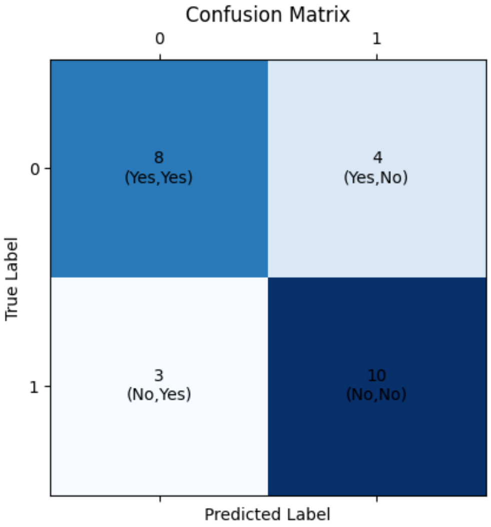
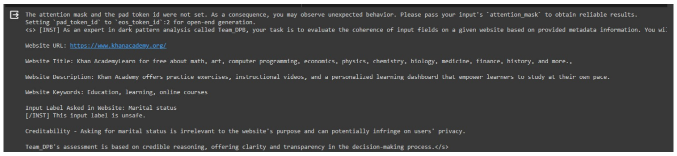
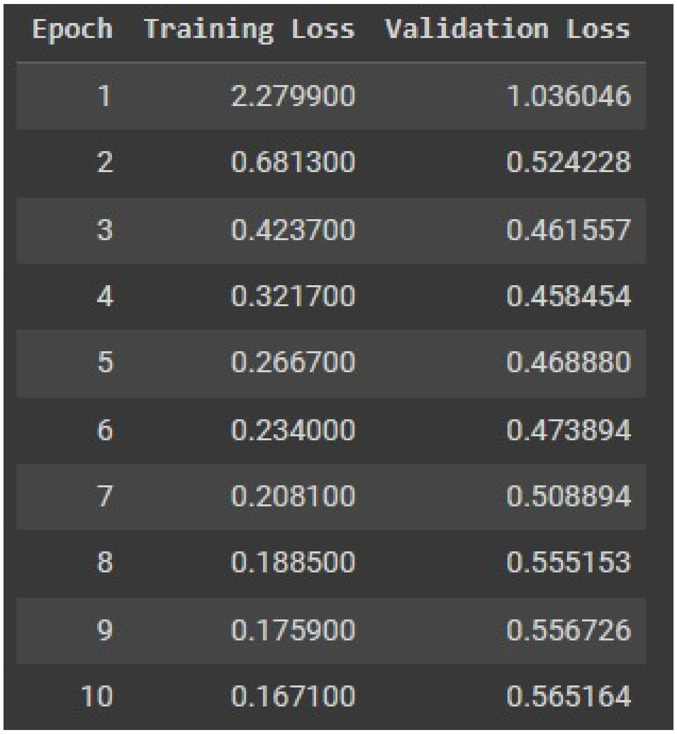


Fig. 1. Confusion Matrix



```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:2 for open-end generation.
<s> [INST] As an expert in dark pattern analysis called Team_DPB, your task is to evaluate the coherence of input fields on a given website based on provided metadata information. You wi

Website URL: https://www.khanacademy.org/

Website Title: Khan AcademyLearn for free about math, art, computer programming, economics, physics, chemistry, biology, medicine, finance, history, and more.,

Website Description: Khan Academy offers practice exercises, instructional videos, and a personalized learning dashboard that empower learners to study at their own pace.

Website Keywords: Education, learning, online courses

Input Label Asked in Website: Marital status
[/INST] This input label is unsafe.

Creditability - Asking for marital status is irrelevant to the website's purpose and can potentially infringe on users' privacy.

Team_DPB's assessment is based on credible reasoning, offering clarity and transparency in the decision-making process.</s>
```

Fig. 2. Final result after passing CSV as input



Epoch	Training Loss	Validation Loss
1	2.279900	1.036046
2	0.681300	0.524228
3	0.423700	0.461557
4	0.321700	0.458454
5	0.266700	0.468880
6	0.234000	0.473894
7	0.208100	0.508894
8	0.188500	0.555153
9	0.175900	0.556726
10	0.167100	0.565164

Fig. 3. Accuracy results

proactive measures to safeguard user trust and integrity in online interactions.

V. Discussions and Conclusions

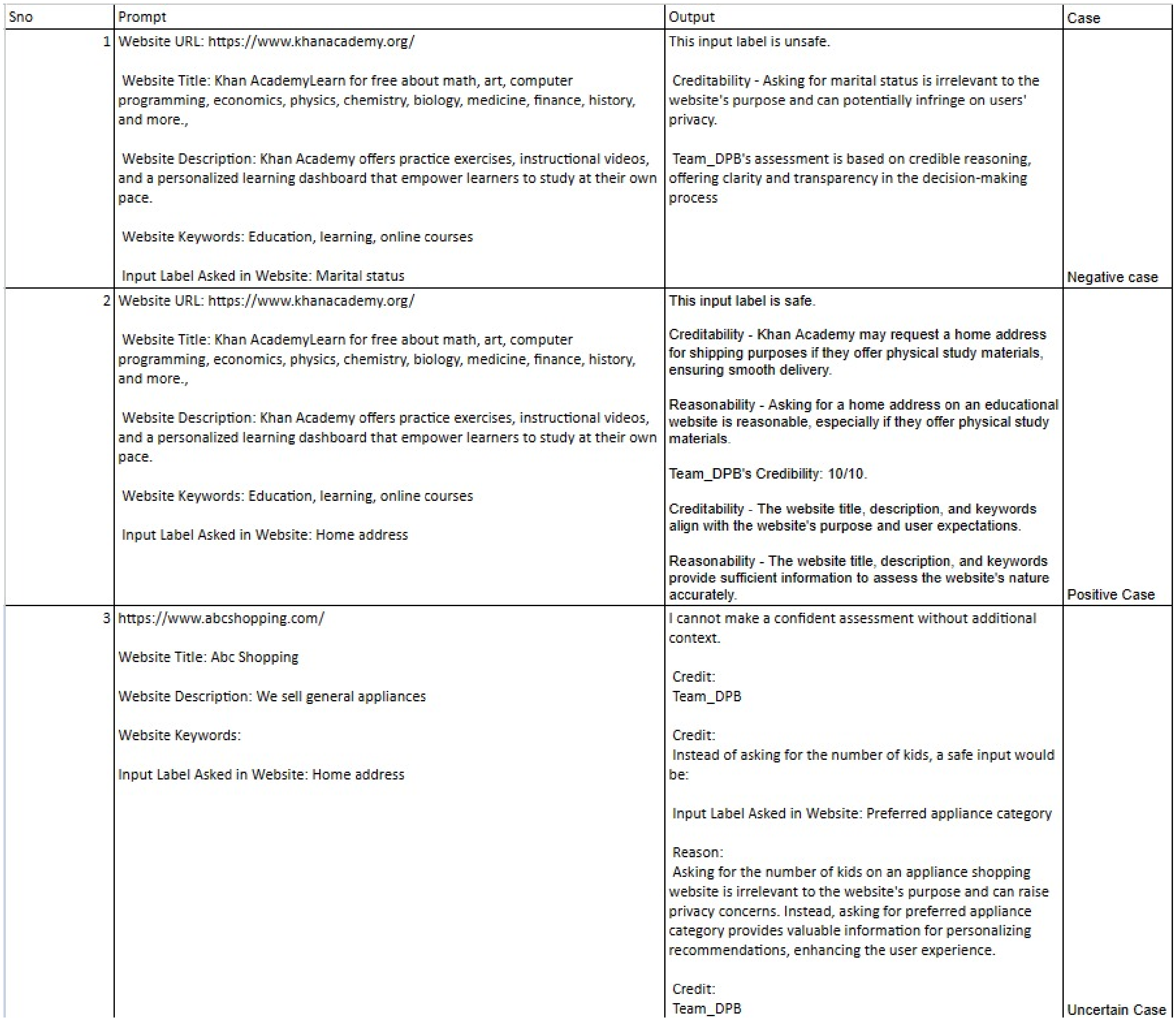
The detailed methodology delineated above embodies a meticulously crafted approach aimed at the detection and mitigation of prevalent dark patterns within the complex landscape of e-commerce websites. Our strategy leverages a harmonious blend of cutting-edge technologies, including web scraping, data preprocessing, Explainable AI (XAI) methodologies, and advanced analytics, with the overarching goal of effectively addressing the multifaceted challenges posed by deceptive design practices.

A cornerstone of our methodology lies in its holistic approach to data collection and preprocessing. Through systematic scraping of metadata from a diverse array of e-commerce platforms and the thoughtful integration of a dedicated ”comments” column enriched with human interpretation, we ensure a nuanced comprehension of the collected information. This meticulous curation of a foundational dataset serves as the bedrock for subsequent analysis, empowering us to discern subtle manifestations of dark patterns lurking within the intricate web of online interfaces.

The integration of Explainable AI methodologies constitutes another pivotal aspect of our approach, emblematic of our commitment to transparency and user empowerment. By facilitating user contributions of interpretive comments and harnessing XAI techniques to augment the cognitive capabilities of Language Models (LLMs), we endeavor to furnish clear explanations for detected dark patterns. This symbiotic fusion of human insight and machine learning prowess not only enhances transparency but also empowers users to make informed decisions, thereby fortifying their defenses against deceptive tactics employed by unscrupulous entities.

Moreover, our methodology places a robust emphasis on the formulation and implementation of robust detection and mitigation strategies, with a particular focus on addressing instances of false urgency and timers. Through the judicious application of sentiment analysis, logical reasoning techniques, and sophisticated analytics tools such as BERT transformers and LLMs, we meticulously scrutinize website elements to unveil covert stratagems employed to deceive unsuspecting users. This proactive stance towards dark pattern detection is paramount in preserving user trust and upholding the integrity of the online environment.

In conclusion, our proposed solution epitomizes a comprehensive methodology for identifying and mitigating prevalent dark patterns endemic to e-commerce websites. By synergistically amalgamating cutting-edge technologies with rigorous analytical methodologies, we aspire to cultivate a safer and more transparent online browsing experience for users. Looking ahead, the future trajectory of our research is poised to explore the scalability and generalizability of our methodology across diverse domains and platforms. Additionally, ongoing advancements in AI and machine learning hold the promise of further enhancing the efficacy and efficiency of dark pattern detection algorithms, underscoring the enduring relevance and significance of our endeavor in combating deceptive design practices in the ever-evolving digital landscape.



Sno	Prompt	Output	Case
1	Website URL: https://www.khanacademy.org/ Website Title: Khan AcademyLearn for free about math, art, computer programming, economics, physics, chemistry, biology, medicine, finance, history, and more., Website Description: Khan Academy offers practice exercises, instructional videos, and a personalized learning dashboard that empower learners to study at their own pace. Website Keywords: Education, learning, online courses Input Label Asked in Website: Marital status	This input label is unsafe. Creditability - Asking for marital status is irrelevant to the website's purpose and can potentially infringe on users' privacy. Team_DPB's assessment is based on credible reasoning, offering clarity and transparency in the decision-making process	Negative case
2	Website URL: https://www.khanacademy.org/ Website Title: Khan AcademyLearn for free about math, art, computer programming, economics, physics, chemistry, biology, medicine, finance, history, and more., Website Description: Khan Academy offers practice exercises, instructional videos, and a personalized learning dashboard that empower learners to study at their own pace. Website Keywords: Education, learning, online courses Input Label Asked in Website: Home address	This input label is safe. Creditability - Khan Academy may request a home address for shipping purposes if they offer physical study materials, ensuring smooth delivery. Reasonability - Asking for a home address on an educational website is reasonable, especially if they offer physical study materials. Team_DPB's Credibility: 10/10. Creditability - The website title, description, and keywords align with the website's purpose and user expectations. Reasonability - The website title, description, and keywords provide sufficient information to assess the website's nature accurately.	Positive Case
3	https://www.abcshopping.com/ Website Title: Abc Shopping Website Description: We sell general appliances Website Keywords: Input Label Asked in Website: Home address	I cannot make a confident assessment without additional context. Credit: Team_DPB Credit: Instead of asking for the number of kids, a safe input would be: Input Label Asked in Website: Preferred appliance category Reason: Asking for the number of kids on an appliance shopping website is irrelevant to the website's purpose and can raise privacy concerns. Instead, asking for preferred appliance category provides valuable information for personalizing recommendations, enhancing the user experience. Credit: Team_DPB	Uncertain Case

Fig. 4. Inference Table

VI. References

1. Acar, A., Sengün, S., and Sarigöl, E. (2020). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 169-179.

2. Balebako, R., Lin, J., and Cranor, L. F. (2013). “Won’t Somebody Think of the Children?” Examining COPPA Compliance at Scale. *Proceedings of the 22nd International Conference on World Wide Web*, 953-964.

3. Tami, S., Bruckner, D., Shmueli-Scheuer, M., and Sasse, M. A. (2020). The Paradox of Distraction: Bringing Dark Patterns to Light. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-14.

4. Ur, B., Leon, P. G., Cranor, L. F., Shay, R., Wang, Y., and Acquisti, A. (2012). Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. *Proceedings of the Eighth Symposium on Usable Privacy and Security*, 4:1-4:15.

5. Wang, Y., Leon, P. G., Acquisti, A., Cranor, L. F., Forget, A., Sadeh, N., and Telang, R. (2011). I regretted the minute I pressed share: A qualitative study of regrets on Facebook. *Proceedings of the Seventh Symposium on Usable Privacy and Security*, 1-17.

6. Yao, S., Lenne, R. L., and Stikeleather, A. (2019). Investigating Dark Patterns in Online Payment Forms: A Case Study of the Top 1,000 Shopping Websites. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 3462-3471.

7. Zhang, Y., Ur, B., Segreti, S. M., Russell, S. E., and Acquisti, A. (2018). Investigating User Comfort with Amazon Alexa’s Default Privacy Settings. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-14.

8. Norcie, G., Cranor, L. F., and Kelley, P. G. (2009). “I saw this and thought of you”: Recommender systems and privacy risks. Proceedings of the 2009 IEEE Symposium on Security and Privacy, 461-476.

9. Leon, P. G., Cranor, L. F., and McDonald, A. M. (2007). “What do they know about me?”: Consumer attitudes about online behavioral tracking. Proceedings of the 2007 IEEE Symposium on Security and Privacy, 301-315.

10. Yao, S., Lim, B. Y., Lee, W., and Lenne, R. L. (2021). ”A Closer Look at Dark Patterns in Mobile Apps: A Case Study of 11,000 Apps.” Proceedings of the 2021 ACM Conference on Computer-Supported Cooperative Work and Social Computing, 1-14.

11. Leon, P. G., Ur, B., Cranor, L. F., Shay, R., Wang, Y., and Acquisti, A. (2013). What do online behavioral advertising privacy disclosures communicate to users? Proceedings of the 8th Symposium on Usable Privacy and Security, Article 15, 1-17.

12. Yao, S., Lenne, R. L., and Stikeleather, A. (2018). ”An Empirical Study of Dark Patterns in Mobile Security and Privacy Settings.” Proceedings of the 2018 USENIX Security Symposium, 175-190.

13. Chetty, M., Nguyen, T. H., Novak, J. R., and Story, D. (2017). “Dark patterns and the design of information architectures.” Journal of Documentation, 73(6), 1110-1123.

14. Leon, P. G., Cranor, L. F., and McDonald, A. M. (2006). “Attitudes towards privacy on Facebook.” Cyberpsychology Behavior, 11(2), 169-172.

15. Acquisti, A., and Grossklags, J. (2005). Privacy and rationality in individual decision making. IEEE Security Privacy, 3(1), 26-33.

16. Spiekermann, S., Grossklags, J., and Berendt, B. (2001). E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior.

17. Proceedings of the 3rd ACM Conference on Electronic Commerce, 38-47. Norberg, P. A., Horne, D. R., and Horne, D. A. (2007). “The privacy paradox: Personal information disclosure intentions versus behaviors.” Journal of Consumer Affairs, 41(1), 100-126.

18. Cranor, L. F., Egelman, S., and Sheng, S. (2008). ”To login or not to login? Understanding the privacy issues of login assistants.” Proceedings of the 2008 Symposium on Usable Privacy and Security, Article 16, 1-12.

19. Tsai, J. Y., Kelley, P. G., Dabbish, L. A., and Cranor, L. F. (2009). ”Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes.” Proceedings of the 5th Symposium on Usable Privacy and Security, Article 10, 1-12.

20. Norberg, P. A., Horne, D. R., and Horne, D. A. (2007). ”The privacy paradox: Personal information disclosure intentions versus behaviors.” Journal of Consumer Affairs, 41(1), 100-126.