The variety of dark patterns on social media and their effect on user behavior: the development of a tool to detect and mitigate the effect of dark patterns

Jara Rodriguez, D22127275

1. Background, Context, Scope

On Brignull et al.'s (2023) website, https://www.deceptive.design (formerly darkpatterns.org), the term Dark Patterns was first introduced and defined as "tricks used in websites and apps that make you do things that you didn't mean to" and are present across the internet. In today's growing digital world, social media is rapidly becoming a central part of daily life. Dark patterns are prevalent across many social media platforms and can lead to a negative effect on user trust and engagement with the platform (Stavrakakis et al., 2021). The use of dark patterns in modern social media can also lead to impacts on user interactions and behavior, privacy, and well-being by exploiting cognitive biases (Karagoel & Nathan-Roberts, 2021).

This proposal situates itself within the context of human-computer interaction (HCI) in the computer science domain. Research into HCIs has shown that choices in the design of a social media platform can have a negative physiological and psychological impact on the platform's users. (Mildner & Savino, 2021; Trice & Potts, 2018). While dark patterns can be found on many websites, this proposal will focus on those found on social media platforms. Examples of these dark patterns are Privacy Zuckering: users being tricked into sharing more information than they wanted, Friend Spam: tricking or encouraging users to invite friends onto the platform, and Infinite Scroll: constantly loading content for a user to keep them engaged for longer periods of time; these can all currently be found on Facebook, Twitter, and TikTok (Cara, 2019; Waldman, 2020).

This proposal's intent is to outline a project in which to explore the variety of dark patterns on modern internet-based social media and their effects on user behavior and to create a tool to help users detect and avoid them.

2. Problem Description

Dark patterns are used by social media platforms to retain users, collect personal data, maximize engagement, drive traffic to the platform, and govern or control user behavior (Mildner, Savino, et al., 2023). In Cara's (2019) review dark patterns are categorized into different types to gain a better understanding and awareness of dark patterns and

highlights that "[dark] patterns are designed to benefit certain stakeholders, not the user".

Similarly, Bösch et al. (2016) specifies that some patterns are used to exploit a user's data and privacy under the disguise of user benefit or increased functionality (Nouwens et al., 2020). Even more worrisome, dark patterns can be strategically utilized to subtly manipulate the user's decision-making to the goals of the social media platform.

The project aims to address these problems by performing a study of dark patterns on current social media platforms, their effect on users, and creating an automated dark pattern detection tool for social media platforms. Initially, an empirical analysis of current social media platforms using quantitative research methods will be used to identify existing dark patterns. Next, surveys will be used to gauge the effects of these dark patterns on users. Finally, an automated detection tool will be developed using machine learning algorithms and user interface analysis. The tool will be designed to recognize and flag dark patterns in real-time, displaying alerts to users that they may be manipulated or exploited and give suggestions so users can protect themselves. If time permits, users will be asked to test the tool and an exit survey administered to determine the tool's effectiveness at dark pattern detection and effect reduction on users.

2.1. Approaches to Solving the Problem

The problem of dark patterns in social media platforms can be approached through a variety of methods, such as empirical analysis, cognitive studies, and the development of detection frameworks and tools.

Empirical analysis can lay the groundwork for understanding dark patterns. An approach used in the literature is the detailed examination of specific social media platforms to identify existing dark patterns and how they're used to manipulate user behavior. In Cara's (2019) paper, the various types of dark patterns prevalent in social media and other digital media sources are discussed and categorized using a systematic review. Through the review of over 30 original digital media sources, the dark patterns found were able to be categorized into different groups such as their popularity, how serious their consequences are, and their strategic purposes. Within each group the patterns were categorized into further sub-groups, for example, the strategic purpose group was further split into sales, data gathering, views, time spent on the product, and miscellaneous, helping to provide information on the dark patterns being used.

In Mildner, Freye, et al.'s (2023) work, dark patterns were similarly analyzed in how they were used on social media platforms. They stressed the importance of empirical research and evidence in identifying dark patterns and understanding how they affect user trust and engagement. Through examination of specific social media platforms and the dark

patterns found there, dark patterns can be documented, categorized, and analyzed for their effects on user behaviors.

An argument can be made for an interdisciplinary approach for tackling this problem. Mathur et al. (2021) make note that other disciplines such as psychology, economics, philosophy, and law have fought against "modifications to users' choice architecture" at a deeper level and understanding of the normative implications that appear from these modifications when compared to HCI. Most of these disciplines, bar psychology, fall outside of the scope of this proposal, however. Understanding what cognitive biases dark patterns take advantage of and how they do so can help counter their existence. The exploration of how cognitive biases like hyperbolic discounting, overchoice, and anchoring make users susceptible to dark patterns can be of great import in this endeavor (Waldman, 2020). These biases can impair rational decision-making, leading to less than favorable choices regarding privacy and disclosure by users. Approaching the design of a framework or tool through a cognitive and behavioral lens could target the root cause of an individual's susceptibility to dark patterns.

While not intrinsically a dark pattern, emotionally manipulative language (EML) can also have a manipulative impact on users' emotions, behaviors, and media consumption depending on the algorithms in use. The forms in which content is fed to a user can have a profound negative effect, as evidenced by recent events such as GamerGate, Russian troll farms, the Cambridge Analytica scandal, fake news surrounding the COVID-19 pandemic and vaccine, and the current conflicts in Ukraine and Palestine. Crowdsourcing detection of EML could be an effective way of identifying and countering the effects of manipulative language, content, and algorithms (Huffaker et al., 2020). As detection of EML would unfortunately fall outside of this proposal's scope, leveraging user feedback and crowdsourced data can be an effective approach to detect and mitigate the effect of dark patterns.

The creation of a framework to detect dark patterns is a practical way to approach the problem of dark patterns and possibly be hands-on if a tool that follows the framework is created. In Curley et al.'s (2021) paper, a "software tool to detect dark patterns on websites, social media platforms and mobile applications" was outlined. This framework mentions automated detection methods such as web crawling, web scraping, and natural language processing to detect dark patterns that can be adapted for use specifically on social media platforms. This framework also includes proposed features for dark patterns that can either be manually detected or cannot be detected at all, such as an ancillary window feature to highlight potential concerns for user review and a reporting feature for users to report suspected dark patterns, respectively.

Additionally, a large-scale analysis of dark patterns across thousands of shopping websites was conducted by Mathur et al. (2019). The researchers iteratively designed and built a Selenium-based web crawler built on top of OpenWPM to identify dark patterns and developed a taxonomy for classification purposes. This approach could be adapted for social media platforms and scaled for fast and easy detection on platforms at a large scale. However, this may be hampered by the recent restrictions and price gauging for API access seen on certain platforms (Bruns, 2019).

The use of technological intervention methods is another possible route to approach this problem. Roffarello and De Russis (2022) outline the creation of a browser extension that detects and intervenes in an attention-capturing dark pattern they termed social investment; this pattern uses likes, comments, shares, etc. to cause users to continue use of the platform to avoid loss of perceived progress, like gamification. The browser extension slowly removed instances of social investment until the pattern was completely removed. Adapting this for use in other dark patterns is a possibility but may not be received well as it could remove features that users may want to access.

2.2. Gaps in Research

While there has been significant research into identifying and categorizing dark patterns, some gaps remain. One gap that this project hopes to tackle in part is the lack of a comprehensive tool for real-time detection of dark patterns across diverse social media platforms. As previously discussed, Mathur et al. (2019) used automated methods to conduct large-scale identification of dark patterns for shopping websites, but no such scalable solutions exist for social media.

There is also a need for more longitudinal empirical study on the psychological effects dark patterns on social media platforms may have on users. Existing research currently focuses on the effects dark patterns have on immediate user interactions, and not on the possible long-term effects on user trust, behavior, psychology, or physiology (Bongard-Blanchy et al., 2021). This is in stark contrast to the abundance of research conducted on the psychological effects of social media consumption itself and unfortunately, it could be difficult to filter the two (Mildner & Savino, 2021).

Though thematic analyses like the one provided by Obi et al. (2022) are insightful in the evolution of dark pattern discourse on Twitter, there tends to be a lack of practical applications or tools to detect and protect against dark patterns.

3. Research Question

What dark patterns are in use on modern social media and what effects do they have on user behavior? Can an automated real-time detection tool be developed to detect dark patterns on social media and mitigate their effect on users?

4. Hypothesis

H_0 (Null Hypothesis):

1. Dark patterns on social media do not significantly impact user behavior.
2. An automated real-time detection tool cannot accurately and reliably detect dark patterns on social media or mitigate their effects on users.

H_A (Alternative Hypothesis):

1. Dark patterns on social media significantly impact user behavior.
2. An automated real-time detection tool can accurately and reliably detect dark patterns on social media and effectively mitigate their effects on users.

5. Design and Build

The first step will be to conduct a literature review to combine and analyze existing research on dark patterns in social media. The types of dark patterns that exist on social media, their characteristics, and on what platforms they appear will be identified and documented. Taxonomies and categorizations developed by Mathur et al. (2019), Cara (2019), (Schaffner et al., 2022), etc. will be used to properly distinguish and document dark patterns based on their purposes or effects. This can be extended to home in on patterns that are specific to social media.

Secondly, an empirical analysis of major social media platforms such as Facebook, Twitter, Instagram, and TikTok will be performed to document the extent and types of dark patterns in use. If possible, web scraping techniques such as the one used by Mathur et al. (2019) will be employed to collect data on each platform's user interface. This data will then be analyzed to determine which patterns are used and their possible effects on user behavior. Users will then be asked to perform surveys such as those used by Bongard-Blanchy et al. (2021) to gain insight into how users interact with the platforms and the patterns found there, their awareness of dark patterns, and how the patterns affect their behavior. Users could also be asked to perform the walkthrough method or other similar methods as they are observed using the platforms, and notes are taken on their interactions, behavior, and feedback.

Next, iterative development of machine learning model(s) and algorithms to detect dark patterns on social media platforms will begin. Techniques from the work of Roffarello and De Russis (2022) can be adapted to create a model that can detect more dark patterns and on more platforms. The model(s) will be trained and based on the data collected in the previous step. This data may have to be manually labeled and supervised machine learning model(s) will be used because of this labeled data. Taking inspiration from Donnelly et al.'s (2022) completed dark pattern detection tool, multiple models and algorithms may have to be used due to differences in the dark patterns to be detected. The algorithms to be tested and used are random forest, convolutional neural networks, logistic regression, and support vector machine. Refinements and tuning of the models will occur during this development step by evaluating their precision, recall, and F_1 scores.

The models will then need to be integrated into a user-friendly tool, with browser extensions being the first deployment and app-overlays as a possibility in future works. The tool will then provide automated real-time detection of dark patterns, alerting users of their presence when encountered on social media platforms. Drawing from Curley et al.'s (2021) framework, the tool will feature dark pattern highlighting, a window that can display features of potential concern, and reporting and educational functionalities.

Finally, if time allows, users will be asked to test the tool and share their feedback. They will also be asked to perform a survey to assess how much the tool mitigates the effects dark patterns on social media have on their behavior, if at all. The tool's effectiveness at detecting dark patterns accurately and its impact on user behavior will also be assessed.

6. Evaluation

The project will be measured through a mixed-methods approach. Quantitative evaluation will primarily answer the second research question and hypothesis. Precision, recall, and F_1 scores will be the primary metrics for evaluating the reliability and accuracy of the machine learning models in identifying dark patterns. Precision will measure the proportion of true positive detections among all detected instances, while recall will measure the proportion of true positive detections among all actual instances of dark patterns. The F_1 score will measure the balance between false positives and false negatives since it is the harmonic mean of precision and recall. User interactions will also be quantitively measured to determine the tool's effect on user behavior. This will be through tracking the reduction in user engagement with dark patterns through methods such as time spent on dark patterns and the frequency that a user triggers the detection of a dark pattern.

The first research question and hypothesis will be measured primarily through qualitative evaluation of user feedback gathered during the surveys. Users will be asked as

to how dark patterns on social media affect them. The information gathered here will help to also determine the tool's effectiveness and what the user experiences. Users will be asked to provide feedback on the clarity of alerts and their reduction in the impact dark patterns have on them.

References

Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., & Lenzini, G. (2021). "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!" - Dark Patterns from the End-User Perspective. *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 763-776. https://doi.org/10.1145/3461778.3462086

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies*, *2016*(4), 237–254. https://doi.org/10.1515/popets-2016-0038

Brignull, H., Leiser, M., Santos, C., & Doshi, K. (2023, April 25). *Deceptive Patterns - user interfaces designed to trick you*. deceptive.design. https://www.deceptive.design/

Bruns, A. (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118x.2019.1637447

Cara, C. (2019). DARK PATTERNS IN THE MEDIA: A SYSTEMATIC REVIEW. *Network Intelligence Studies*, *VII*(14), 105–113. https://doaj.org/article/e3d388b79eba4d66a92b12b0e0e2dc78

Curley, A., O'Sullivan, D., Gordon, D., Tierney, B., & Stavrakakis, I. (2021, July). The design of a framework for the detection of Web-Based dark patterns. *ICDS 2021: The 15th International Conference on Digital* Society. Retrieved from https://arrow.tudublin.ie/ascnetcon/3/

Donnelly, J., Downley, A., Liu, Y., Su, Y., Sun, Q., Zeng, L., Curley, A., Gordon, D., Kelly, P., O'Sullivan, D., & Becevel, A. (2022, June 19). "Be a Pattern for the World": The Development of a Dark Patterns Detection Tool to Prevent Online User Loss. *Proceedings of Ethicomp, 20th International Conference on the Ethical and Social issues in Information and Communication Technologies*. https://doi.org/10.21427/2Y2Q-6323

Huffaker, J. S., Kummerfeld, J. K., Lasecki, W. S., & Ackerman, M. S. (2020). Crowdsourced Detection of Emotionally Manipulative Language. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 248. https://doi.org/10.1145/3313831.3376375

Karagoel, I., & Nathan-Roberts, D. (2021). Dark patterns: social media, gaming, and E-Commerce. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 752–756. https://doi.org/10.1177/1071181321651317

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-computer Interaction*, *3*(CSCW), 81. https://doi.org/10.1145/3359183

Mathur, A., Mayer, J., & Kshirsagar, M. (2021). What Makes a Dark Pattern... Dark?: Design Attributes, Normative Considerations, and Measurement Methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 360. https://doi.org/10.1145/3411764.3445610

Mildner, T., Freye, M., Savino, G., Doyle, P. R., Cowan, B. R., & Malaka, R. (2023). Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. *Proceedings of the*

2023 ACM Designing Interactive Systems Conference, 2362-2374. https://doi.org/10.1145/3563657.3595964

Mildner, T., & Savino, G. (2021). How social are social media: The dark patterns in Facebook's interface. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2103.10725

Mildner, T., Savino, G., Doyle, P. R., Cowan, B. R., & Malaka, R. (2023). About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 192. https://doi.org/10.1145/3544548.3580695

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 194. https://doi.org/10.1145/3313831.3376321

Obi, I., Gray, C. M., Chivukula, S. S., Duane, J., Johns, J., Will, M., Li, Z., & Carlock, T. (2022). Let's Talk About Socio-Technical Angst: Tracing the History and Evolution of Dark Patterns on Twitter from 2010-2021. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2207.10563

Roffarello, A. M., & De Russis, L. (2022). Towards Understanding the Dark Patterns That Steal Our Attention. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 274. https://doi.org/10.1145/3491101.3519829

Schaffner, B., Lingareddy, N. A., & Chetty, M. (2022). Understanding account deletion and relevant dark patterns on social media. *Proceedings of the ACM on Human-computer Interaction*, *6*(CSCW2), 417. https://doi.org/10.1145/3555142

Stavrakakis, I., Curley, A., O'Sullivan, D., Gordon, D., & Tierney, B. (2021, December 30). A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically. *International Journal On Advances in Internet Technology, 14(1& 2),* 36-45. Retrieved from https://arrow.tudublin.ie/scschcomart/149/

Trice, M., & Potts, L. (2018). Building Dark Patterns into Platforms: How GamerGate Perturbed Twitter's User Experience. Present Tense, 6(3). Retrieved from https://www.presenttensejournal.org/volume-6/building-dark-patterns-into-platforms-how-gamergate-perturbed-twitters-user-experience/

Waldman, A. E. (2020). Cognitive biases, dark patterns, and the 'privacy paradox.' *Current Opinion in Psychology, 31,* 105–109. https://doi.org/10.1016/j.copsyc.2019.08.025