Neighborhood Watch: Crowdsourcing for Dark Pattern Detection

Nathan B.
nbhak@stanford.edu

Cameron L-H.
cameroni@stanford.edu

Daniel WR.
danwr@stanford.edu

INTRODUCTION

For an individual navigating the web, it has become an increasingly difficult task to maintain one's agency as a user. Manipulative UI design patterns, known as "dark patterns", deceive users into behaving against their own best interests, such that the actions we take online don't always have the consequences we intend. Companies use these tactics to achieve various objectives at the expense of users, such as directing privacy choices and hindering attempts to unsubscribe from their services [1]. As such, it is alarming that dark patterns have become increasingly prevalent across sites and platforms, while users themselves remain largely unaware of them [2].

In order to mitigate the threat of dark patterns and protect user agency, we need to employ methods to detect and point them out to users on the web. There exist detection algorithms that use machine learning to classify specific categories of dark patterns, but these methods are not easily generalizable to dark patterns as a whole [3, 4]. Others have been useful for the measurement of dark patterns across the Internet, but only serve to establish a lower bound [5]. Researchers have found that certain dark patterns cannot be detected through automated methods at all due to variation in their definition and implementation [6]. As such, none of these methods provide a suitable foundation to alert users of dark patterns in real time. At the same time, users continue to identify dark patterns as a source of frustration, even without knowing their formal definition [12]. We therefore suggest that the users themselves may hold the key to a more effective detection strategy.

In this paper, we propose the use of crowdsourcing to detect and flag dark patterns across the web, protecting users through collective identification. Crowdsourcing has proven to be an effective method of collecting information and detecting issues on a larger scale [7]. Given the wide array of interfaces that they experience holistically while browsing the web, we believe that their contributions can cover more ground than existing AI-driven mechanisms, and therefore serve as an effective foundation for a defense against dark patterns.

We introduce Neighborhood Watch, a Chrome extension that improves user awareness of dark patterns. Our system allows users to select elements on websites to tag them as a dark pattern, and view tags submitted by others when they visit the webpage. We show that our simple tagging system causes users to be more conscientious in their browsing habits, and less susceptible to manipulative interfaces. We address limitations of Neighborhood Watch, but conclude that crowdsourcing can serve as a versatile detection and protection mechanism against dark patterns.

RELATED WORKS

Background

From automatically added items to your online shopping cart to the difficulty of unsubscribing from a newsletter, manipulative online experiences are becoming increasingly prevalent across the web. Dark patterns are deliberate user interface (UI) designs that are made to manipulate a user's behaviour against their own interests [9]. In 2010, Cognitive Science PhD Harry Brignull coined this term for "deceptive user experiences in digital products" [1]. Basic examples of dark patterns include disproportionately sized buttons that bias the user towards making a specific choice, or using a small font size to conceal important information for users.

Though in isolation these design patterns may seem purely bothersome, they can have pernicious effects at scale. Dark patterns feature in our online interactions much more than one might expect. One study found that dark patterns were found on 95 percent of the free Android apps in the US Google Play Store [2].

Since Brignull's popularization of the term, there has been substantial work in defining and taxonomizing dark patterns [9, 5]. According to one of the most recent influential studies, dark patterns are "user interface design choices that benefit an online service by coercing, steering, or deceiving users into making unintended and potentially harmful decisions" [10].

Defining & Detecting Dark Patterns

Online manipulation strategies can take many different forms. Studies have explored and taxonomized variations of dark patterns that can be found on the web. Mathur et al. used AI-driven methods to conduct this investigation at scale, analyzing over fifty thousand product pages "to characterize and quantify the prevalence of dark patterns." They provide a comprehensive breakdown of dark pattern characteristics based on their collected dataset. Their clustering algorithm is designed to obtain a lower bound for the number of dark patterns at scale and is therefore ill-suited to more comprehensive detection on individual sites [5].

In fact, despite the prevalence of dark patterns on the web, effective detection systems for users are few and far between. Researchers have used machine learning mechanisms to target dark patterns in cookie banners with some success, but they lack the generalizability that is required to alert users effectively [4, 3]. However, literature suggests that users themselves

are able to effectively identify certain types of dark patterns [11] and that individuals are often aware of manipulative designs, regardless of their familiarity with the terminology [12]. Researchers also propose methods to increase ease of detection and strengthen resistance to them through design measures, economic incentives, and regulatory solutions.

Leveraging the Crowd to Find Light in the Dark

Based on prior research, we believe that crowdsourcing would be a powerful tool in dark pattern detection. Crowdsourcing is an invitation for users online to contribute meaningful information to under-studied areas [7]. Furthermore, given the vastness of the web, crowdsourcing provides a valuable "divide-and-conquer" tactic that enables groups to more rapidly detect and solve problems on a larger scale [13]. However, crowdsourcing mechanisms come with several challenges; they require incentives for users to contribute to them, and a method to determine the trustworthiness and accuracy of their contributions. As such, it's often useful to include a user evaluation system to improve the efficacy of crowdsourcing methods, such as the detection of well-intentioned agents and the ability for users to grant credibility to one another [14]. Studies have shown that it can be difficult to distinguish between benign users and malicious actors [15], but [16] demonstrated that non-experts perform similarly to experts and algorithms in a study that measured different skill-levels in crowdsourcing. Research has also been published on leveraging crowdsourcing for detection methods; in [17], researchers effectively mitigated skepticism regarding their computer vision models by using the crowd to detect sampling biases.

A USER-DRIVEN DARK PATTERN DETECTOR

To demonstrate the potential of crowdsourcing in this problem space, we introduce Neighborhood Watch, a Chrome extension that relies on users to identify and report dark patterns. When loading a webpage, the extension annotates the DOM to highlight the dark patterns tagged by previous visitors, then allows the user to contribute their own tags.

Design Considerations

From a practical standpoint, a browser extension serves as a viable way to offer functionality to users and influence their behavior as they navigate the web. Based on preliminary feedback on our low fidelity wireframes (Figure 1), we designed Neighborhood Watch with an emphasis on low friction, enabling users to report dark patterns with few intermediary steps.

Visually, the extension popup provides a short description of dark patterns and includes a link to additional examples (Figure 2). A box outlined in bright red offers a brief explanation as to how the extension works, matching the color that Neighborhood Watch uses to highlight dark patterns (Figure 3), so that users are able to recognize when an element has been tagged.

The Tagging System

Neighborhood Watch consists of several different components. The extension interfaces with a Firebase Realtime Database and is itself comprised of multiple scripts with their own

Figure 1. Variations of our low fidelity mockups on Figma.

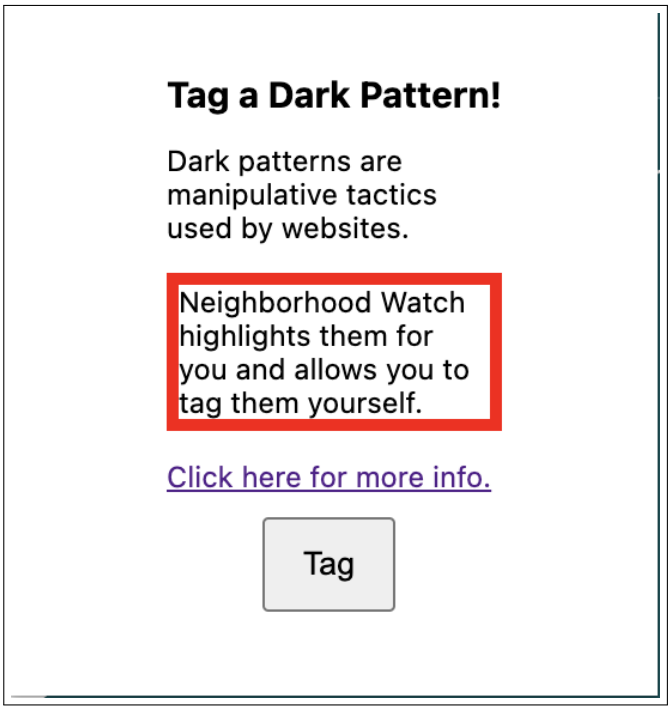

Figure 2. The extension popup of Neighborhood Watch.

unique capabilities. The popup script (`popup.js`) allows the user to interact with the extension popup, the content script (`content.js`) interacts with the webpage itself, and the background script (`firebase.js`) provides the functionality to store and retrieve information from the database.

In Figure 4 we present the architecture of Neighborhood Watch and an example of how parts of the system interact with one another. When Alice, a user of the extension, loads evil.com, `content.js` requests the crowdsourced dark pattern information from `firebase.js` ①. This causes `firebase.js` to query the database, requesting all tagged dark patterns for evil.com ② and upon receiving them ③, the script passes this data back to `content.js` ④. Using this information, `content.js` annotates the dark patterns by highlighting elements on the DOM and displays these changes to Alice ⑤.

If Alice wants to tag a dark pattern herself, she clicks a button on the popup ⑥. `popup.js` conveys this to `content.js` ⑦, which allows Alice to select on element on the page (Figure

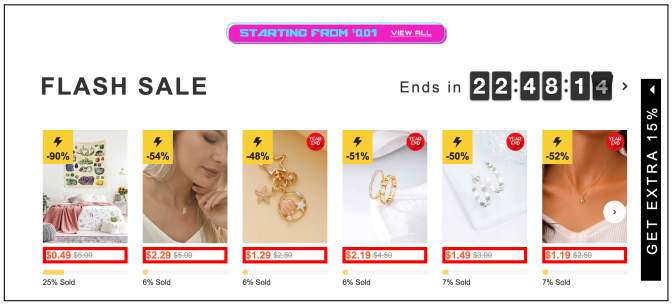

Figure 3. Dark patterns tagged by Neighborhood Watch.

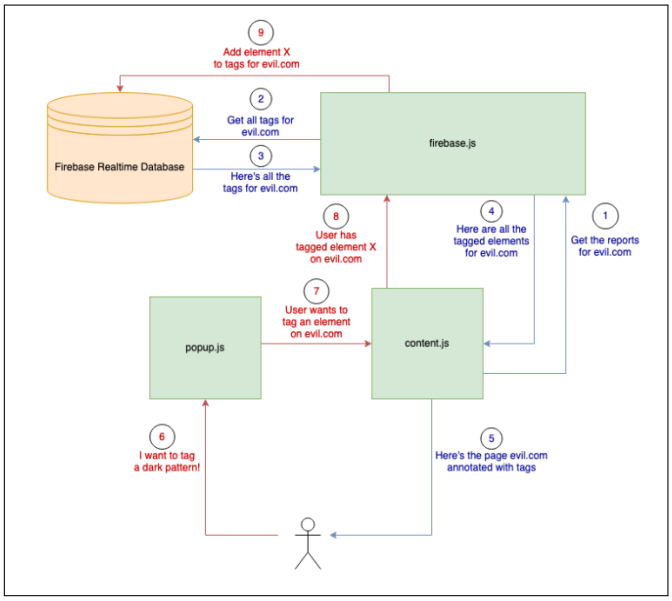

Figure 4. Neighborhood Watch architecture: tagging process highlighted in red and tag-viewing in blue.

5). When Alice has clicked on the dark pattern, `content.js` passes the relevant information to `firebase.js` ⑧, which in turn adds an entry to the database ⑨. The reported element will be stored under "evil.com", alongside the others shown on the page. Each is uniquely identified by its HTML tag (e.g. 'div') combined with its order among other elements on the page with the same tag.

EVALUATION

We evaluated Neighborhood Watch to test our claim that crowdsourcing can be used to improve user awareness of dark patterns and detect them more effectively than automated methods. The design of our evaluation examined two distinct experiences for users of our tool. The first of these experiences was the tagging of dark patterns. At the core of our thesis, we assumed that our tool would be able to harness the power of crowdsourcing, and as such it was necessary that users were able to intuitively understand and navigate the tagging functionality, and be willing to use it as they browse the Internet on a daily basis. As part of our study, we specifically addressed this question through the qualitative analysis of the participants' experiences through survey and observation. This study also provided insight into the capabilities of users, and therefore crowdsourcing, in dark pattern detection.

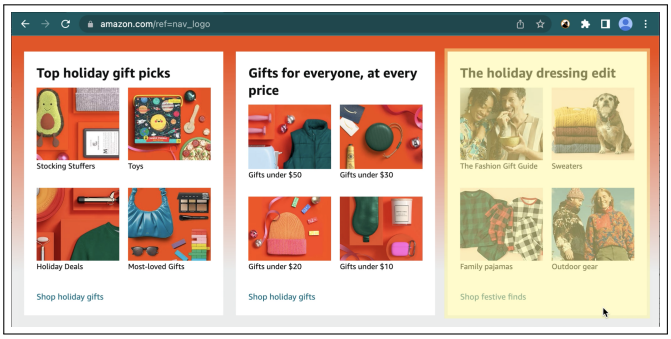

Figure 5. A user can report a dark pattern by selecting an element on a webpage.

The second of these experiences concerned the way in which viewers receive dark pattern information through the crowdsourced tags, as well as how it affected their online experience and decisions as they navigated the web. Through this study we gathered qualitative data regarding the usability of our tool, and developed an understanding of the extent to which crowdsourced information affects user awareness and browsing behavior. To do so, we also observed and surveyed the participants and employed a think-aloud protocol to gather their thoughts as they completed a task on a website. We performed this study on one group with our extension enabled, and another without, identifying the differences between them.

Method

To evaluate the effectiveness of Neighborhood Watch, we observed how susceptible users are to falling for dark patterns by testing two different experiences. Our first experiment was to test the experience of annotating. We directed participants to a flower shopping site with several dark patterns and asked them to complete several tasks, first checking the extension popup window, then proceeding to make a purchase. They were also instructed to tag any dark patterns they found throughout the process. Their experience was screen-recorded with a think-aloud protocol and following the exercise, users were interviewed to gather qualitative observations on their experience. We used this study to illustrate whether the use of our tool would encourage more active engagement in considering the deceptiveness of a user interface, and more generally whether users are less susceptible to manipulation by dark patterns when actively searching for them.

For our second experiment, we tested the experience of viewing these annotations on a site with dark patterns by observing two different groups of users completing the same task on the flower shopping site. The experimental group navigated the site with the extension enabled, allowing them to view the popup and the tagged dark patterns, while the control group performed the task without the extension. Neither group was briefed on the definition of a dark pattern in advance, and both groups were instructed to do a screen-recorded, think-aloud protocol. The purpose of this was to test whether our tool is intuitive and effective in preventing users from interacting with dark patterns on an interface.

After testing both our experimental and our control group in our second experience, we conducted post-study interviews to

gather more qualitative data. We asked users from both groups to describe their experience completing the task. We also asked users if they were able to define a dark pattern, which was used to help us gauge the influence of computer literacy and dark pattern awareness on our results. In our control group, we asked users if they could describe any of the dark patterns they thought that they may have experienced while completing the task. In our experimental group, we asked them how intuitive they found the tool to use, if they would make any adjustments to the tool to make it more accessible to users, and if they think that they would have been able to detect the dark patterns they found had they not been using the extension.

Participants

We attempted to recruit participants with technical and non-technical backgrounds (the intention being to seek participants without deep prior knowledge on dark patterns). In gathering feedback on our initial sketches and prototypes, our participants consisted of 3 college students, who were interviewed for feedback on our initial sketches and prototypes.

For our main study, we gathered 10 participants total; all participants were college students, four described themselves as having technical backgrounds, and two knew the definition of a dark pattern. Two participants were chosen to tag dark patterns, and the remaining 8 were divided equally among the control and experimental groups for the non-tagging experience. The division of the participants into the experimental and control groups was done randomly. Participants were gathered through flyers seeking volunteer participants via social media and text messaging.

FINDINGS

Experience 1: Tagging

In the first experience (where participants were asked to tag dark patterns), we learned several important insights about our tool and possible design implications. Firstly, we learned that prior knowledge of dark patterns did not dictate whether a participant successfully tagged multiple dark patterns. Interestingly, out of the two participants who tagged, the one who had no prior knowledge of dark patterns was much more diligent in seeking out dark patterns to tag on the site.

Secondly, upon reviewing the extension, when offered the opportunity to click for more information, the participant without prior knowledge studied the taxonomy of the dark patterns in detail and used it as a resource throughout their tagging task. Though this behaviour may not be entirely indicative of what a user may do in real life (given the participant was operating under the context of a study where they were explicitly told to tag dark patterns), it demonstrated the potential educational value of our tool to users.

Another insight we gained from the tagging experience was from the other participant who reported that they had prior knowledge of dark patterns. Upon asking them to explain what they were, the participant gave a thorough explanation. However, despite their demonstrated understanding and recognition of dark patterns throughout their navigation (shown in their verbalized thought process), they did not use the system to tag most dark patterns. Reflecting on this result, we gained valuable insight into the possible lack of clarity in the instructions of our extension, and the possible issue in incentive structure for users of our tool. Even if users may be knowledgeable about the existence of dark patterns, they may not take the initiative to tag them. Incentivization strategies would be critical to investigate in future design iterations of this tool.

Experience 2: Viewing

In the second experience, one observation we made was a disparity between the users' perception on the dark patterns featured on the site. In the control group, several participants noted that the user reviews were a great addition to the UI, and only critiqued the way in which they were displayed. In the experimental group, each participant showed suspicion when observing that every review had a 5-star rating. Moreover, several participants commented that the reviews were likely AI-generated. When looking at the sale prices displayed, a user in the control group made a positive remark when seeing several flowers on sale, while a user in the experimental group noticed a dark pattern in how the sale for the red poinsettia was displayed as "From $39.99" despite making no note of the price difference. From these findings, we have observed that participants using the tool overall demonstrated more skepticism and were more likely to view the site through a critical lens.

Another observation we made was that users using the extension were more likely to investigate malicious elements on the page that users in the control group tended to gloss over. For example, when tasked to select a bouquet size for checkout, every user in the control group did not question the lack of clarity between the different size options. Meanwhile, every user in the experimental group questioned what "As shown" meant as a size option. When tasked to return to the homepage, users were instructed to click the site logo which prompted a modal that attempted to mislead users to stay on the checkout page. The modal gave the option to stay on the checkout page as a green button, and to return home as a red button. In the experimental group, every user remarked on the deceptive modal. In the control group, the users faced some friction when attempting to return to the homepage, but overall only one user made note of how the colors were misleading them to stay on the checkout page.

Another result that we observed from the control and experimental groups was the noted disagreement on certain dark patterns. After being asked to add a bouquet to their cart, users on the site were led to another page suggesting more items to add to their purchase (Figure 6). While all users in our experimental group classified this as a dark pattern and made verbal remarks about it, such as "Obviously they're trying to get me to buy more things", some participants in our control group had the opposite reaction. One participant in particular remarked that the suggested items seemed "nice to add", so much so that they asked us for permission to add a greeting card to their cart (given that they were not instructed to do so). This result gave us insight into the noted ambiguous nature of dark patterns. While a feature in the design of a UI may

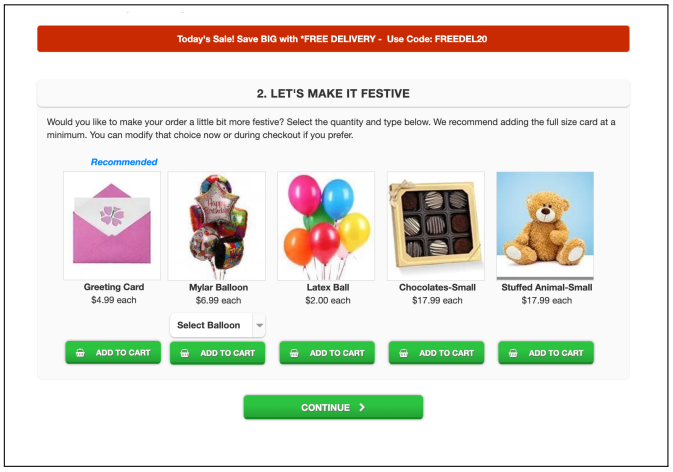

Figure 6. Page suggesting more items to add into a user's cart.

appear to be a dark pattern (in this case an additional page before checking out, suggesting more items to add to one's cart), for some it may be seen as a useful encouragement. Acknowledging this ambiguity in our alert of dark patterns may be a possible design choice to consider moving forward.

DISCUSSION

Dark Patterns and Implications For Users

Given much of the existing data set of dark patterns is collected through automated systems, our tool's unique user-generated dataset will help contribute to the growing library of documented dark patterns from previous work. With our current tool, we envision many different possibilities to leverage the system assuming its widespread adoption. One such vision we have is for the extension to be used equally as an educational tool as a flagging tool. Neighborhood Watch currently provides a brief definition of dark patterns, but is ultimately made more effective when users possess existing domain knowledge. The public is largely unaware of these patterns, and even for those that understand their definition, the variations in the ways they manifest may make them more difficult to recognize across different site designs. In future iterations, the extension could potentially present the user with educational advice on how to both recognize and respond to dark patterns as they navigate a webpage.

Based on the results of our study and the existing literature regarding interventions, it would be important for Neighborhood Watch to acknowledge the potential ambiguity of certain dark pattern tags, thereby distinguishing them from more egregious instances.

Iterating on Incentive Model & Design

Given the findings in our study, we learned critically that incentivizing users to tag is key to the success of our tool. In future iterations, redesigning and testing the extension based on various incentive models would help learn more about the most effective method to garner successful traction in our crowdsourcing model.

Future work

While Neighborhood Watch illustrates the power of crowdsourcing in dark pattern detection, our research has several limitations. Though we recognize that dark patterns occur much in the same manner on mobile views, we narrow the scope of our tool to web browsers. Furthermore, given the limited scope due to time constraints, there is room for improvement regarding the robustness of our system, as element annotation and selection doesn't always work correctly for certain site features. In addition, if Neighborhood Watch were to operate at scale, it would require defensive measures to combat bad actors that might want to subvert our tagging mechanisms. Our tool also assumes the user has substantial knowledge on the definition of a dark pattern and will voluntarily contribute to the system. As such, additional features to educate and incentivize users would need to be added before deploying the tool in the real world.

CONCLUSION

Researchers have explored the vast array of dark patterns online through detailed taxonomies and comparisons across modalities. In proposing mitigation strategies, existing literature has suggested implementing policies and educational measures. Detection mechanisms from prior work use automated systems and specifically machine learning models to alert users, but ultimately fall short in terms of generalizability and comprehensiveness. Instead, Neighborhood Watch leverages crowdsourcing, which has proven to be effective in social computing systems, particularly in the context of assigning users simple tasks that collectively address a large scale issue. We observe that our tool protected users through collective identification and education, and its success demonstrates the capabilities of crowdsourcing for dark pattern detection, a task that is but one step towards diminishing the impact of malicious design. In developing Neighborhood Watch, we hope to not only increase awareness of dark patterns, but also encourage the use of crowdsourcing in more social computing systems, contributing to a widespread restoration of user agency on the web. The code for Neighborhood Watch is released under the Apache 2.0 license at https://github.com/nbhak/neighborhood-watch.

REFERENCES

[1] Harry Brignull. 2019. Dark Patterns. https://www.darkpatterns.org/.

[2] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In CHI Conference on Human Factors in Computing Systems (CHI '20), April 25–30, 2020, Honolulu, HI, USA. ACM, New York, NY, USA 14 Pages. https://doi.org/10.1145/3313831.3376600

[3] Than Htut Soe, Cristiana Teixeira Santos, and Marija Slavkovik. 2022. Automated detection of dark patterns in Cookie banners: How to do it poorly and why it is hard to do it any other way. (April 2022). Retrieved October 13, 2022 from https://arxiv.org/abs/2204.11836

[4] Philip Hausner and Michael Gertz. 2021. Dark Patterns in the Interaction with Cookie Banners. Position Paper at the

Workshop "What Can CHI Do About Dark Patterns?" at the CHI Conference on Human Factors in Computing Systems, May 8-13, 2021, Yokohama, Japan.

[5] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 81 (November 2019), 32 pages. https://doi.org/10.1145/3359183

[6] Curley, A., O'Sullivan, D., Gordon, D., Tierney, B., Stavrakakis, I. (2021) The Design of a Framework for the Detection of Web-Based Dark Patterns", ICDS 2021: The 15th International Conference on Digital Society, Nice, France, 18th – 22nd, July 2021 (online).

[7] Melissa A. Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, andMichael S. Bernstein. 2017. Flash Organizations: Crowdsourcing Complex Work by Structuring Crowds As Organizations. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). Association for Computing Machinery, New York, NY, USA, 3523–3537. https://doi.org/10.1145/3025453.3025811

[8] Dean, B. (2021) Google Chrome Statistics for 2022, BackLink. Available at: https://backlinko.com/chrome-usersnumber-of-chrome-extensions.

[9] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2021. A Comparative Study of Dark Patterns Across Mobile and Web Modalities. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 377 (October 2021), 29 pages. https://doi.org/10.1145/3479521

[10] Arunesh Mathur, Jonathan Mayer, and Mihir Kshirsagar. 2021. What Makes a Dark Pattern... Dark?: Design Attributes, Normative Considerations, and Measurement Methods. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445610

[11] Aditi M. Bhoot, Mayuri A. Shinde, and Wricha P. Mishra. 2020. Towards the Identification of Dark Patterns: An Analysis Based on End-User Reactions. In IndiaHCI '20: Proceedings of the 11th Indian Conference on Human-Computer Interaction (IndiaHCI 2020). Association for Computing Machinery, New York, NY, USA, 24–33. https://doi.org/10.1145/3429290.3429293

[12] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In Designing Interactive Systems Conference 2021 (DIS '21). Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[13] Walter S. Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P. Bigham, and Michael S. Bernstein. 2015. Apparition: Crowdsourced User Interfaces that Come to Life as You Sketch Them. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 1925–1934. https://doi.org/10.1145/2702123.2702565

[14] Arpita Ghosh, Satyen Kale, and Preston McAfee. 2011. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In Proceedings of the 12th ACM conference on Electronic commerce (EC '11). Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/1993574.1993599

[15] Anon. Leveraging crowdsourcing for efficient malicious users detection in large-scale social networks. Retrieved October 13, 2022 from https://ieeexplore.ieee.org/stamp/stamp.jsptp=arnumber=7462258

[16] Warby, S., Wendt, S., Welinder, P. et al. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat Methods 11, 385–392 (2014). https://doi.org/10.1038/nmeth.2855

[17] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K. Thiruvathukal, and Ming Yin. 2020. Crowdsourcing Detection of Sampling Biases in Image Datasets. In Proceedings of The Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 2955–2961. https://doi.org/10.1145/3366423.3380063