Fighting Malicious Designs: Towards Visual Countermeasures Against Dark Patterns

René Schäfer
rschaefer@cs.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Paul Preuschoff
paul.preuschoff@rwth-aachen.de
RWTH Aachen University
Aachen, Germany

René Röpke
roepke@cs.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Sarah Sahabi
sahabi@cs.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Jan Borchers
borchers@cs.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

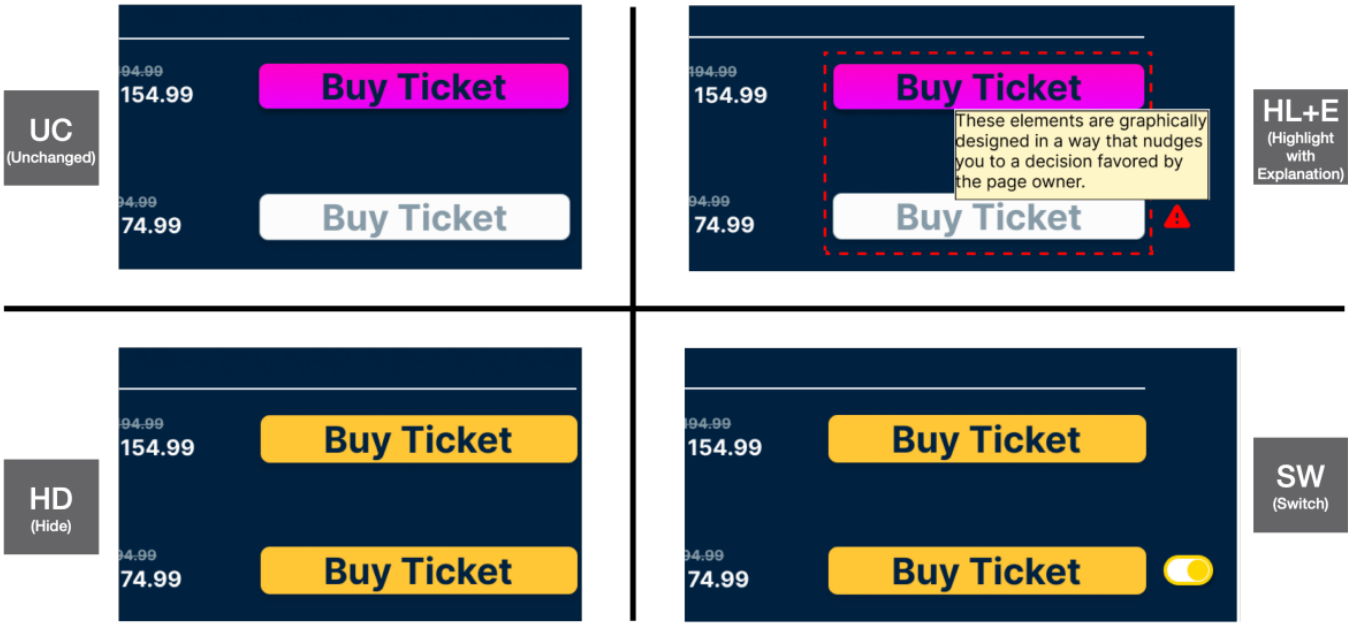

Figure 1: We investigated three countermeasures and tested them against 13 common dark patterns in a lab study. Participants were shown four variations of the same screen; one unchanged baseline (*UC*) and the three countermeasures *Highlight with Explanation* (*HL+E*), *Hide* (*HD*), and *Switch* (*SW*) that are based on the work by Schäfer et al. [23]. For each dark pattern, we asked participants to pick their favorite variation and justify their decision.

ABSTRACT

Dark patterns are malicious UI design strategies that nudge users towards decisions going against their best interests. To create technical countermeasures against them, dark patterns must be automatically detectable. While researchers have devised algorithms to detect some patterns automatically, there has only been little work to use obtained results to technically counter the effects of dark patterns when users face them on their devices.

To address this, we tested three visual countermeasures against 13 common dark patterns in an interactive lab study. The countermeasures we tested either (a) highlighted and explained the manipulation, (b) hid it from the user, or (c) let the user switch between the original view and the hidden version. From our data, we were able to extract multiple clusters of dark patterns where participants preferred specific countermeasures for similar reasons. To support creating effective countermeasures, we discuss our findings with a recent ontology of dark patterns.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642661

CCS CONCEPTS

• **Human-centered computing → User studies**; **Graphical user interfaces**.

KEYWORDS

dark patterns, deceptive design, visual countermeasures, lab study

ACM Reference Format:

René Schäfer, Paul Preuschoff, René Röpke, Sarah Sahabi, and Jan Borchers. 2024. Fighting Malicious Designs: Towards Visual Countermeasures Against Dark Patterns. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3613904.3642661

1 INTRODUCTION

Dark patterns, a term initially shaped by Brignull in 2010[1], are malicious design strategies that steer users towards making decisions that benefit the respective (online) service [17]. In recent years, they have received increasing attention in the HCI research community, as demonstrated by a CHI'21 workshop [16] with 18 position papers[2], a CHI'23 panel [11], an accompanying Special Interest Group (SIG) [14], and a recent book by Brignull [5]. Additionally, researchers started classifying dark patterns into taxonomies [12, 13, 18–20].

With the rise of dark patterns, researchers started calling for effective countermeasures [2, 17, 23]. Research fields for dark patterns that appear promising to allow countering them include legislation [6], user awareness [1, 2], and technical countermeasures [7]. While passing laws against dark patterns or building up a deep user awareness can take a substantial amount of time, technical countermeasures can instantly intervene when facing manipulative designs, requiring little to no prior knowledge of dark patterns by the user. However, while all approaches above rely on taxonomies, technical countermeasures additionally require automatic detection of such patterns. One example of automatic detection is the crawler by Mathur et al. [17] that automatically detects text-based dark patterns on websites. However, research on visualizing countermeasures for detected dark patterns is surprisingly sparse.

Recently, Schäfer et al. [23] took initial steps by investigating six visualization techniques as possible countermeasures against three common dark patterns in an online study. However, the authors used non-interactive screenshots to illustrate the effect of applying each countermeasure. In their study, one of the most favored approaches introduces additional elements and, thus, visual clutter to the screen. This indicates the need to test them in live interactions to understand better how well these countermeasures would perform in actual use.

We expand on their work with a new study on visual countermeasures against dark patterns to better understand how people actually perceive them in more realistic scenarios. As such, we tested three of the most relevant visual concepts with a more extensive set of 13 dark patterns from Mathur et al.'s established taxonomy [17]. Furthermore, we cluster this set of dark patterns based on user justifications for countermeasures and compare it with a recent ontology by Gray et al. [13] to support the creation of effective countermeasures against dark patterns. Overall, our work contributes first insights into actual user interactions with visual countermeasures against dark patterns. It thus strengthens, expands upon, and sharpens the results of Schäfer et al.'s proposed visualization techniques [23] and puts them into the context of the ontology by Gray et al. [13].

2 RELATED WORK

Given the prevalence and the strong effects of dark patterns on people [6], researchers emphasize the pressing need for effective countermeasures [2, 7]. While the current corpus of proposed countermeasures against dark patterns is limited, some general approaches and ideas exist [2, 3, 17]. For example, Bösch et al. [3] investigated privacy-related dark patterns and proposed countermeasures like *raised user awareness* through education. Other approaches include so-called *bright patterns* [10, 22] and automatic *dark pattern detection tools* [17]. We review related work exploring these approaches below.

2.1 Education & Awareness

Several researchers have investigated the relationship between users' awareness of dark patterns and their capabilities to detect and resist them. Di Geronimo et al. [9] discovered that the prevalence of dark patterns can result in *dark pattern-blindness*: Users have difficulties detecting patterns that have already become common in everyday interactions. However, they also found that users performed better at detecting dark patterns when they were more aware and knowledgeable about them. Nevertheless, Bongard-Blanchy et al. [2] observed that raising user awareness does not necessarily increase manipulation resistance when they are unaware of the risks. Consequently, creating awareness among users and educating them about possible risks and how to resist them is a promising approach to counteract dark patterns [2].

2.2 Bright & Fair Patterns

A recent approach to counteract dark patterns is using *bright patterns*. These are design strategies that, for example, nudge users towards privacy-friendly options [10] or prioritize user over company goals [22]. To effectively counteract dark patterns automatically, each pattern would need a bright version that allows manipulation in favor of the users while not requiring server-side changes. Therefore, while bright patterns provide interesting research directions, their current use as technical countermeasures against willingly placed dark patterns appears limited. Another approach in a similar direction uses so-called *fair patterns*[3] [21]. In contrast to bright patterns, fair patterns propose user interfaces without manipulations in any direction to enable users to make fair decisions.

2.3 Technical Countermeasures

Several projects have addressed automatic dark pattern detection. Mathur et al. [17] laid a foundation for this with their crawler that automatically detects certain text-based dark patterns on websites. Building upon this, a winning hackathon team[4] created the

[1] https://www.deceptive.design (former: https://www.darkpatterns.org) *Accessed: February 2024*

[2] https://darkpatternsindesign.com/position-papers/ *Accessed: February 2024*

[3] https://fairpatterns.com Accessed: February 2024

[4] TeenHacks LI, https://thli-fall-2019.devpost.com/ *Accessed: February 2024*

browser extension *Insite*[5] that detects and highlights dark pattern instances online. Similarly, researchers from the *Dark Pattern Detection Project*[6] created a plugin that detects certain dark patterns on websites using simple regular expressions[7]. Hausner and Gertz [15] propose detecting dark patterns in cookie banners by analyzing their CSS styles.

Curley et al. [8] discussed the technical applicability of automatic dark pattern detection. They proposed a framework for dark pattern detection consisting of the following three classes: 1) patterns that can be detected in an automated way, 2) patterns that can be detected manually, and 3) patterns that cannot be detected. The authors suggest that this classification can be used by detection tools that detect and remove class 1 patterns while highlighting potential class 2 patterns to warn the user.

2.4 Visual Countermeasures

Research on automatic dark pattern detection also raised the question of how to best communicate detected dark patterns to users to mitigate their influence. While Mathur et al. [17] proposed highlighting detected patterns and providing additional explanations to users, little research has explored different types of visual countermeasures and their effectiveness. Schäfer et al. [23] investigated such visual countermeasures in an online study with screenshots of possible countermeasure concepts. They applied six countermeasures to three common dark patterns. Results indicate that users wish to know why a design is classified as a dark pattern. Simply removing the manipulation created mixed impressions. Some participants liked not needing to deal with dark patterns anymore, while others feared that vital information might be hidden accidentally. Additionally, participants disliked that a program altered or removed content silently. Providing visual cues without further information was met with disapproval. Since that study only contained non-interactive screenshots of visual countermeasures against a limited amount of dark patterns, it remains unclear whether added visual clutter (e.g., when highlighting a dark pattern for the user) distracts or annoys users too much, especially when multiple dark patterns are present simultaneously.

In summary, numerous research projects have begun to address the prevalence of dark patterns. They agree that one promising step towards countermeasures is automatically detecting dark patterns through dedicated tools. However, comparing different methods to counter dark patterns in an interactive setup visually remains unexplored. Our study aims to fill this gap to further the field of visual countermeasures against dark patterns that can be applied on the user's end and thus under their control.

3 STUDY

Our study investigates 13 of the 15 dark patterns from Mathur et al.'s taxonomy [17] (see Figure 2 and Appendix B). We excluded *Hard to Cancel* as this usually includes a multitude of steps for users, and *Low-stock Message* because it is structurally similar to *High-demand Message*. Instead, we included it to familiarize participants with our countermeasures.

We based our study on Schäfer et al. [23] by adapting their visual countermeasures *Highlight with Explanation* (*HL+E*) and *Switch* (*SW*), which received the best overall user rankings. We also included *Hide* (*HD*) as the most controversial countermeasure and *Unchanged* (*UC*) as the baseline status quo. The countermeasures are defined as follows:

- ***Unchanged* (*UC*)**: The manipulative element is not changed in any way.
- ***Highlight with Explanation* (*HL+E*)**: A red dashed box is drawn around the manipulative element. A red warning sign is added, explaining why this content is marked on hover. This strategy was proposed by Mathur et al. [17].
- ***Hide* (*HD*)**: If possible, the manipulative element is visually altered, rephrased, or removed completely. With this, *HD* turns dark patterns into fair patterns [21] by allowing user decisions that are not manipulated in any way.
- ***Switch* (*SW*)**: The manipulative element is hidden. A switch next to it allows users to toggle back to the original content.

We use similar assumptions as Schäfer et al. [23]: (1) We can detect some dark patterns automatically, (2) we can only alter the visual appearance of dark pattern representations, but not any functionality, (3) content revealed after future interactions is unknown to our countermeasures, and (4) the countermeasure does not trigger actions on behalf of the user (like rejecting cookies or an extended warranty automatically).

3.1 Merit

As Schäfer et al. [23] tried to reveal a basic understanding of the applicability of their visualization techniques, they only tested them against three dark patterns. In an interactive lab study, we focus on testing a subset of their most salient techniques against 13 patterns. A key difference is that, compared to Schäfer et al. [23], our participants could actually interact with the prototypes. Additionally, our scenarios contain multiple simultaneous dark patterns, as it is a common and more realistic practice [6, 9]. Table 1 compares both works.

3.2 Study Scenarios

We created one mock-up screen to show how each countermeasure deals with a *Low-stock Message*, allowing participants to familiarize themselves with the countermeasures without priming them regarding the dark patterns in this study. This training screen and our two main shopping scenarios were created using Figma[8]. In the first scenario, participants were tasked to buy a smartphone; in the second scenario, they bought a ticket for a concert. Each scenario consisted of multiple screens and embedded a distinct subset of the 13 investigated dark patterns from the taxonomy we refer to [17]. Both scenarios contain multiple manipulative elements inside a single view, as this represents actual practice more closely [6]. The smartphone scenario contained patterns A–F, while the ticket scenario embedded patterns G–M (see Figure 2). The scenarios are visualized in Appendix B. Images of our Figma prototypes can be found in the supplementary materials to this paper.

[5] https://github.com/NicholasTung/dark-patterns-recognition *Accessed: February 2024*
[6] https://dapde.de/ *Accessed: February 2024*
[7] https://github.com/Dapde/Pattern-Highlighter/ *Accessed: February 2024*
[8] https://www.figma.com/ *Accessed: February 2024*

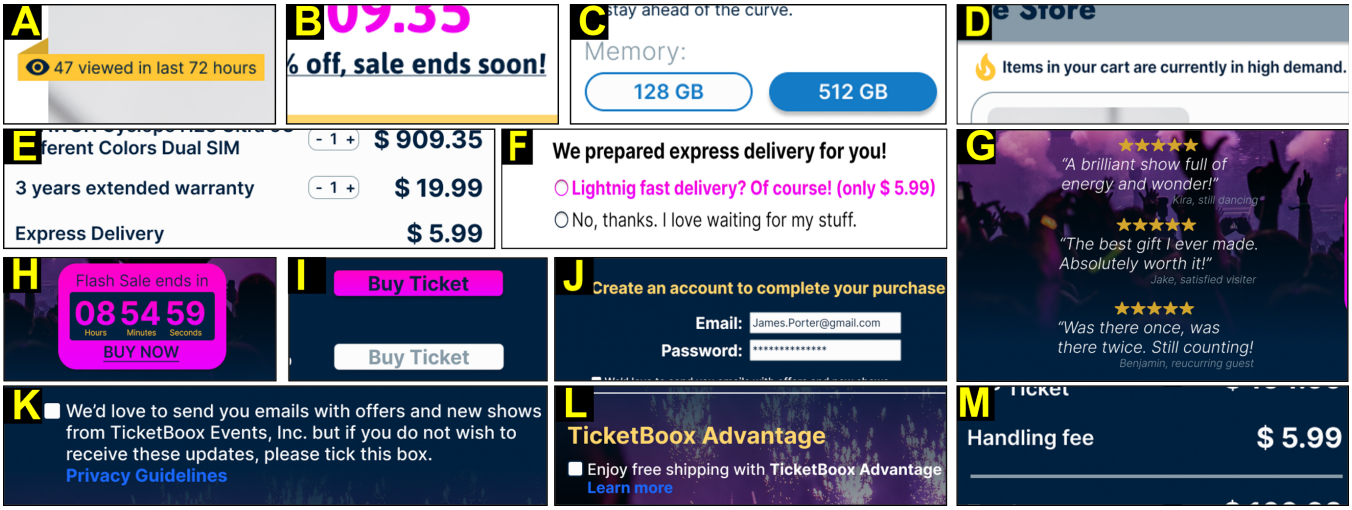

Figure 2: The 13 dark patterns we investigated in our study. (A)–(F) were used in the *smartphone shop scenario*, while (G)–(M) were used in the *ticket shop scenario*. The dark patterns are: (A) *Activity Message*, (B) *Limited-time Message*, (C) *Pressured Selling*, (D) *High-demand Message*, (E) *Sneak into Basket*, (F) *Confirmshaming*, (G) *Testimonials*, (H) *Countdown Timer*, (I) *Visual Interference*, (J) *Forced Enrollment*, (K) *Trick Questions*, (L) *Hidden Subscription*, and (M) *Hidden Costs*. *Confirmshaming* and *Visual Interference* were also used by Schäfer et al. [23].

Paper	Study	Countermeasures	Patterns	Patterns per Page	Screens
Schäfer et al. [23]	non-interactive online study (n=40)	6	3	1	1
This work	interactive lab study (n=20)	3	13	Up to 3	Multiple

Table 1: Comparing our approach to the most closely related work [23]. We selected the three most promising visual countermeasures from that work and evaluated them against a much larger collection of dark patterns in two studies. Our second study used interactive prototypes rather than screenshots. Our study also included multiple patterns on a page simultaneously, and they covered a multiple-screen interaction, for higher ecological validity. Overall, our approach allowed us to better understand the generalizability of these techniques.

3.3 Study Procedure

In our study, participants interacted with live prototypes of the three countermeasures. Since our institution does not have an ethics review panel, we followed the *ACM Code of Ethics*[9]. After we collected participants' demographics (Appendix A), we measured subjective dark pattern awareness using a 7-point Likert scale. Then, they familiarized themselves with the first countermeasure using the example screen with the *Low-stock Message* dark pattern for as long as they wanted before continuing with one of the two main scenarios (see Section 3.2). Afterward, they completed a post-task questionnaire to rate the given countermeasure on six categories adapted from [23] using 7-point semantic differential scales: *USABILITY*, *CLARITY*, *EFFICIENCY*, *SAFETY*, *HELPFULNESS*, and *FEELING*, i.e., whether the given countermeasure made the website feel better or worse (see Figure 3). Since *UC* did not change anything, it was not rated. Participants additionally rated *TRUSTWORTHINESS* of the pages presented for all countermeasures and *UC*.

For our study, we used two different scenarios with four variations each (3× countermeasures, 1× baseline) containing a total of 13 dark patterns (see Figure 2). Since showing each variant in combination with all four conditions would have taken too long, we let participants encounter all dark patterns twice using two countermeasures each. Using a single scenario would have resulted in learning effects regarding the implemented dark patterns, undermining our measures. Table 2 visualizes this process. Participants only rated *TRUSTWORTHINESS* when interacting with a scenario the first time to capture their initial impression.

After completing a task in a scenario for the first time, participants described their goal and decision-making. They then rated how easy it was to reach their self-stated goal and how confident they felt about it. Participants repeated this process for all countermeasures and the baseline. Finally, they provided an overall ranking. The order of countermeasures was counterbalanced using a Latin Square, and the scenarios alternated. In the second part of the study, participants received printouts of the scenarios in all variants, picked their favorite countermeasure, and justified their

[9]https://www.acm.org/code-of-ethics *Accessed: February 2024*

Scenario	Dark Patterns	HL+E	HD	SW	UC
Phone	(A) *Activity Message*, (B) *Limited-time Message*, (C) *Pressured Selling*, (D) *High-demand Message*, (E) *Sneak into Basket*, (F) *Confirmshaming*	X			X
Ticket	(G) *Testimonials*, (H) *Countdown Timer*, (I) *Visual Interference*, (J) *Forced Enrollment*, (K) *Trick Questions*, (L) *Hidden Subscription*, and (M) *Hidden Costs*		X	X	

Table 2: To encounter all dark patterns multiple times during the study, participants were given each scenario twice with different conditions. For example, as depicted here, a participant could see the phone scenario using *Highlight with Explanation* (*HL+E*) and *Unchanged* (*UC*). In this case, the participant would then encounter the ticket scenario with *Hide* (*HD*) and *Switch* (*SW*). The combinations of countermeasures and scenarios varied between participants.

decision for each dark pattern (see Appendix B). Finally, participants provided opinions on the strengths and weaknesses of each countermeasure.

3.4 Qualitative Coding

We coded qualitative responses and assigned anonymous IDs to all participants (e.g., P17). For the task where participants picked their favorite countermeasures, one researcher inductively created an initial codebook containing 42 codes following the thematic analysis approach [4]. Afterward, three other researchers familiarized themselves with the data, discussed the initial codebook with said researcher, and refined it accordingly. This codebook contained 22 codes after merging semantically similar codes. Two of the researchers then fully and independently coded the data again in two rounds using the updated codebook. To assess the inter-coder agreement, we computed Cohen's kappa and obtained a strong agreement ($\kappa = 0.90$). Finally, we discussed the largest deviations.

4 RESULTS

20 people participated in our study (21 to 31 years, M=25.80 years, SD=2.48 years, 11 male, and 9 female). Self-reported awareness of dark patterns was high (M=5.5, SD=1.63, on a scale from 1 *"Not aware at all"* to 7 *"Very aware"*), with only two participants reporting limited to no awareness of dark patterns (see Appendix A). Participants were recruited via convenience sampling and received no monetary compensation. Overall, the study took approximately 50–60 minutes.

4.1 Rankings

Overall, *HL+E* received very good rankings: 9×1^{st}, 6×2^{nd}, 4×3^{rd}, and 1×4^{th}. *SW* received similar positive results: 7×1^{st}, 10×2^{nd}, 2×3^{rd}, and 1×4^{th}. Participants' opinions on *HD* were mixed: 4×1^{st}, 2×2^{nd}, 8×3^{rd}, and 8×4^{th}. *UC* received low rankings: 0×1^{st}, 2×2^{nd}, 8×3^{rd}, and 10×4^{th}.

The semantic differential ratings contain trends and controversies. *HD* appears to be seen as more dangerous, slightly more unclear, and less helpful compared to both other countermeasures. At the same time, *HL+E* is rather controversial regarding whether it makes the overall page look better or worse.

All countermeasures received positive results for *EFFICIENCY* and *USABILITY*. *CLARITY*, *HELPFULNESS*, and *SAFETY* received similar ratings with *HL+E* and *SW* being liked, while *HD* was neutral. Surprisingly, participants thought that the overall pages looked best (*FEELING*) with *SW*, followed by *HD*, while *HL+E* received mixed results due to added visual clutter. The ratings are shown in Figure 3.

Both scenarios were rated similarly regarding their *TRUSTWORTHINESS*. *UC* was seen as being slightly dubious (M=3.1, SD=1.35). *HL+E* received worse ratings (M=2.8, SD=1.66) than *UC*, while *SW* (M=4.3, SD=1.81) caused a rather neutral trustworthiness among participants. *HD* received the best overall trustworthiness rating (M=4.6, SD=1.46).

4.2 Achieved Goals

Participants also described their goals and decision-making when first interacting with a scenario. If participants did not fully reach their self-set goals, e.g., because of mistakes or false assumptions, this counted as an unsuccessful attempt. Overall, there were 32 successful and 8 unsuccessful attempts (80% success). In the smartphone scenario, three participants failed their goal. With *HD*, one participant fell for *Sneak into Basket* by thinking that the added warranty had to be bought. Additionally, two participants also overlooked said warranty when using *SW*. Participants reported very high confidence (M>6) and easiness (M>5.5) for each method.

The ticket scenario had completely distinct results: Five participants failed their goal. For *UC*, two participants fell for *Visual Interference* by misinterpreting the button colors which resulted in them choosing a more expensive option. Another participant fell for *Trick Question* and accidentally subscribed to a newsletter. The same happened for two further participants with *HL+E*. For this scenario, *HL+E* and *SW* received higher confidence ratings (M>6.4) than *HD* and *UC* (M=5.4). *UC* had its lowest easiness rating (M=3.8) in the ticket scenario. All other attempts were considered successful.

4.3 Countermeasure Justifications

Participants picked their favored countermeasure for each dark pattern (Figure 4) and justified their decision. After coding, we excluded all responses from P09, since all coders agreed that they did not match the given task.

By far the most common justification for choosing *HD* was to *remove unnecessary content* (38×) which was particularly often used for *Activity Message* and *High-demand Message*. Other common reasons were: *allows unbiased decisions* (17×), *removes malicious*

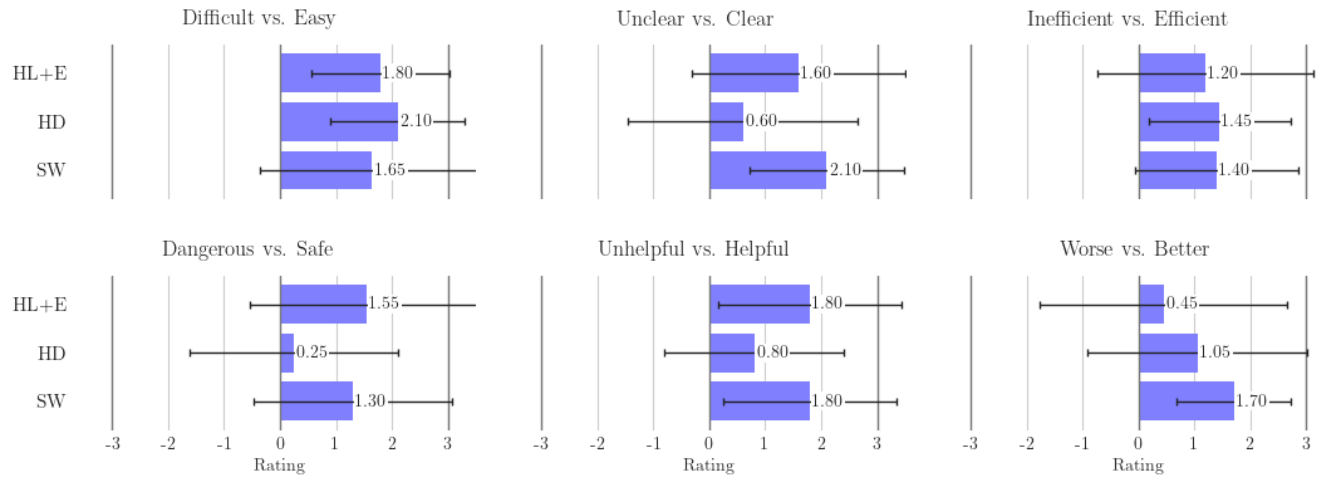

Figure 3: The semantic differential ratings of all countermeasures. Whiskers denote the standard deviation. Overall, all countermeasures received neutral to positive scores across all categories. Most ratings also exceed the same ratings from the online study [23].

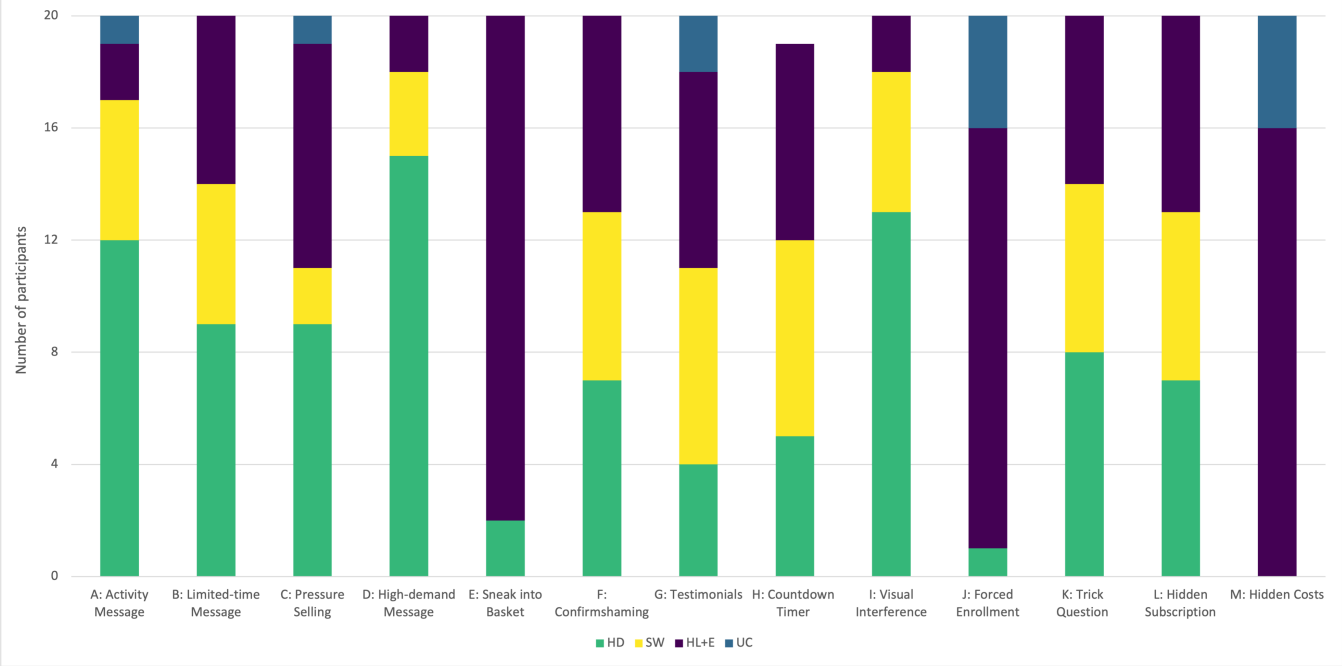

Figure 4: The frequencies of the preference of the four countermeasures (HD, SW, HL+E, UC) per dark pattern (A–L) from the last task of the lab study. *Hide* (HD) was the most controversial countermeasure, as it was highly preferred for some patterns (e.g., *High-demand Message*), while not selected by any participant for others (e.g., *Hidden Costs*). *Unchanged* (UC) was rarely preferred.

content (13×), *simplifies content* (12×), and *cleans the website visually* (11×). Frequent justifications for using *HL+E* were that it *alerts the user* (24×) and *provides additional information* (22×) or *explanations* (17×). These were the main reasons for *Sneak into Basket*, *Forced Enrollment*, and *Hidden Costs*. Other reasons were that it *highlights malicious content* (15×) and *preserves content* (13×) without losing information. Participants also chose it in cases where the dark pattern *could not be changed anyway* (9×), like for *Hidden Costs*. *SW* was mainly chosen as it *allows to see the original content* (33×), most frequently for *Testimonials* and *Countdown Timer* with five

times each and for *Confirmshaming* and *Limited-Time Message* with four times each. Participants also chose *SW* as it *indicates changes* on the website (16×) and *removes unnecessary content* (7×). The main reason for choosing the baseline *UC* was that participants stated that the given element *could not be changed anyway* (8×), like the *Forced Enrollment*, or that a *change was not necessary* (6×).

4.4 Strenghts and Weaknesses

At the end of the study, participants provided strengths and weaknesses regarding all variants. Regarding *HD*, participants particularly liked that it *removes present manipulation* (7×). This was also seen critically, as it *removes and alters information* without the user knowing about it (6×). Participants further stated that it was *easy to use* (3×), *simplified content* (2×), and *cleaned the website* (2×). Additionally, participants pointed out that the latter could make a *shady website more trustworthy* (2×). Frequently named strengths of *HL+E* were that it *highlights manipulation* (7×), that it *provides additional information* (5×), and that *users can learn from it* (5×). Additionally, participants appreciated the *explanations* (3×) and that users are still able to *decide on their own* (3×). A downside of *HL+E* was *visual clutter* (6×) and the fact that the *manipulation is still present* which might bias users (4×). Participants also remarked that *HL+E draws attention towards the manipulation*, which might be counter-productive (2×). For *SW*, participants liked that it allows users to *see the original content* (7×) and *check it for control* (4×). Furthermore, it *highlights manipulation* (3×), *provides additional information* (2×), and *educates users* (2×). However, participants also argued that *SW* introduces *visual clutter* (4×), *interactions may take longer* (4×), and it *does not provide explanations* (3×). Participants frequently suggested combining *SW* with *HL+E* (7×). The main strength of *UC* is that it does not alter any information (5×). Additionally, it preserves the content (2×) and allows users to decide freely on their own (2×). However, participants disliked the manipulation still being present (10×) and that some dark patterns might be overlooked (5×).

5 DARK PATTERN CLUSTERS

The choices for participants' favored countermeasures (Figure 4) split the dark patterns roughly into three groups. For *Activity Message*, *High-demand Message*, and *Visual Interference*, participants clearly favored *HD*. *Sneak into Basket*, *Forced Enrollment* and *Hidden Costs* resulted in participants picking *HL+E*. All other patterns did not have a clear majority for a specific countermeasure. By analyzing participants' justifications, we extracted six more nuanced groups of dark patterns.

5.1 Identified Clusters

All six clusters are visualized in Figure 5. In the following, numbers in parentheses show how many participants provided a given justification for the respective pattern.

5.1.1 Unnecessary Elements (A, D). This cluster contains (A) *Activity Message* and (D) *High-demand Message*. For both, half of our participants wanted to remove them as they were unnecessary (A:10, D:9). Additionally, they share similar ratings regarding *SW* where participants want to be able to see the original (A:3, D:3), have changes indicated (A:2, D:2), and have unnecessary content removed (A:3, D:2).

5.1.2 Unclear Authenticity (B, G, H). For (B) *Limited-time Message*, (G) *Testimonials*, and (H) *Countdown Timer*, participants chose *HD* because the elements were unnecessary (B:5, G:3, H:3), *SW* to see the original content (B:4, G:5, H:5), or *HL+E* to preserve the content (B:4, G:2, H:3). While (G) and (H) shared very similar justifications, (B) is also similar to our cluster of *Unnecessary Elements*. However, a key reason for placing it in this cluster is that participants shared a wish to preserve the content which was not the case for the other cluster. We received multiple comments for (B), (G), and (H) that countermeasures should remove these patterns if they represent false information. This indicates uncertainty of participants since they fear important information could be lost, even if it might be fake.

5.1.3 Dangerous Elements (E, M). For (E) *Sneak into Basket* and (M) *Hidden Costs*, the most common justification was that participants wanted to be alerted to them (E:7, M:6). We hypothesize that the main reason could be that both patterns result in a potential loss of money. Additionally, participants appreciated the highlighted manipulation (E:4, M:3) and given explanations (E:4, M:2). It was expected that participants would largely choose *HL+E* as *SW* and *HD* could not remove them.

5.1.4 Forced Action (J). While (J) *Forced Enrollment*, like *Sneak into Basket* and *Hidden Costs*, cannot be changed by *SW* and *HD*, the main justification for choosing *HL+E* differed compared to the other two. Here, participants were interested in additional information (J:7) while alerting was only secondary (J:4). This might be because (J) does not directly cause financial harm to users. With this, receiving additional information appeared more interesting, as this pattern cannot be overlooked.

5.1.5 Biased Decision (C, F, I). This cluster contains (C) *Pressured Selling*, (F) *Confirmshaming*, and (I) *Visual Interference*. Here, the most common reason was to remove them to allow an unbiased decision (C:4, F:4, I:6). While (C) and (F) also share justifications for *HL+E*, justifications for (F) and (I) are mainly focused on removing the elements. Additionally, for (F) and (I), participants sometimes argued for *SW* to access the original content (F:4, I:3).

5.1.6 Complex Elements (K, L). (K) *Trick Questions* and (L) *Hidden Subscription* were the only dark patterns where participants mainly chose *HD* to simplify the content (K:6, L:4), possibly because both are complex in different ways. While (K) is manipulatively formulated and also includes a negation, (L) hides crucial information behind a label.

5.2 Categories of Clusters

We can further group our previous clusters to a more abstract level where patterns share similar properties or where justifications target the same goal. *Dangerous Elements* and *Forced Action* contained elements which *SW* and *HD* could not alter. At the same time, *HL+E* provided additional information to users, highlighted the manipulation, and warned them, if necessary. Another group of similar patterns combines *Biased Decision* and *Complex Elements*. Both share a similar goal, although the main justifications for choosing

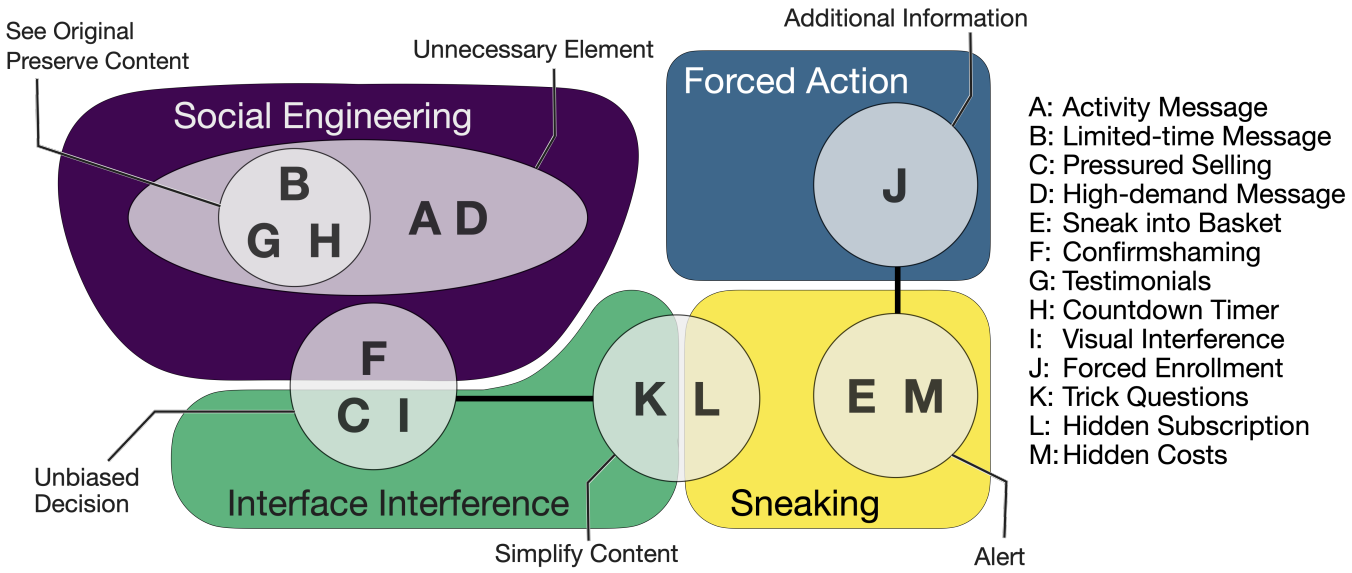

Figure 5: Our clusters of dark patterns fall under four of the six high-level patterns in the ontology by Gray et al. [13]: *Social Engineering*, *Interface Interference*, *Sneaking*, and *Forced Action*. Encircled surroundings resemble patterns with similar justifications while the groups within each of our higher cluster categories are connected with lines. Overall, only patterns from *Interface Interference* were grouped with patterns from other high-level patterns in the ontology.

specific countermeasures differed, e.g. allowing unbiased decisions vs. simplifying content. Simplifying might be a pre-condition for allowing users to make an informed and unbiased decision. With this, it could be a more specific cluster within *Biased Decision*. However, more empirical evidence is needed to support this finding.

6 DISCUSSION

In this section, we first compare our results with the findings of Schäfer et al. [23] as we adapted their countermeasures. Afterward, we compare our dark pattern clusters with a recent ontology of dark patterns by Gray et al. [13].

6.1 Screenshots vs. Interactive Prototpyes

Compared to Schäfer et al. [23], all three countermeasures received similar or better ratings across all categories in our lab study. This might be because participants could now interact with the countermeasures directly, instead of only having access to non-interactive screenshots. Only *HD* regarding *CLARITY* (M=0.60 vs. M=1.55) and *HL+E* regarding *FEELING* (M=0.45 vs. M=1.05) received clearly lower ratings. Unlike in their study, our scenarios contained screens with several simultaneous dark patterns and spanned multiple screens to make them more realistic. This increased visual clutter of *HL+E* could explain the lower *FEELING* ratings in both scenarios. A reason for the decreased *CLARITY* ratings for *HD* might be that *HD* could not successfully remove some dark patterns, which is an interesting finding on its own. Surprisingly, *SW* was rated much better compared to their study in every category, even though, like *HD*, it cannot alter certain patterns. On average, its *FEELING* was rated better than both other countermeasures. Participants probably felt less confident with *HD* after realizing this.

During the study, participants tended to be very confident that they could fully reach their goals. However, the success rate was only 80%, with several participants missing their goal. Surprisingly, both scenarios caused failures with a distinct subset of the investigated countermeasures. Since our sample size is rather small, this is a promising finding for future work.

Similarly to [23], opinions were split regarding *HD*: Four participants ranked it best, while eight participants put it last. However, this split shifted towards the lower ranks compared to their study. Another reason could be that participants learned that *HD* does not work on every dark pattern which also holds for *SW*, but does not seem to have affected respective user rankings. Interestingly, participants still chose *HD* frequently when picking their favorite countermeasure against specific dark patterns.

6.2 Matching Preferred Countermeasures to Dark Patterns

When designing countermeasures against dark patterns, knowing which patterns can be counteracted similarly is crucial. To understand whether our clusters already resemble common groupings of dark patterns, we put them into perspective with a recent ontology by Gray et al. [13], who expanded on common taxonomies and grouped dark patterns into six high-level patterns: *Nagging*, *Obstruction*, *Sneaking*, *Interface Interference*, *Forced Action*, and *Social Engineering*; each of which contains multiple meso and low-level patterns. Our study contained 6× *Social Engineering*, 3× *Interface Interference* and *Sneaking*, and 1× *Forced Action*. Figure 5 shows how our patterns align within the ontology. We need to note that *Hidden Subscription* was not classified in the ontology. To our understanding, it would be a low-level pattern within *Sneaking* under

the meso-level *Hiding Information*. Additionally, *Visual Interference* was also not present with this exact name. Our implementation (see Figure 2) best resembles the low-level pattern *Visual Prominence*.

In their ontology, Gray et al. [13] grouped eight low-level dark patterns into the high-level pattern *Social Engineering*. Six patterns from our study match respective low-level patterns from this group: (A) Activity Message, (B) Limited-Time Message, (D) High-Demand Message, (F) Confirmshaming, (G) Testimonials, and (H) Countdown Timer. Interestingly, there are two justifications that participants provided for every single of those patterns: using *HD* to remove them since they are unnecessary (A:10, B:5, D:9, F:2, G:3, H:3) and using *SW* to see the original content (A:3, B:4, D:3, F:4, G:5, H:5). While the first justification was the most frequent one for (A), (B), and (D), the latter was most frequent for (F), (G), and (H), and the second most frequent for (B). Overall, this indicates that *Social Engineering* creates uncertainty within participants on whether the content is fake or legit. With our clusters, we found two groups of dark patterns that were fully contained in *Social Engineering* and preferred countermeasures for different reasons (see Section 5).

The only pattern from *Social Engineering* that we could not match to the other two clusters was *Confirmshaming*. Instead, we grouped it with two *Interface Interference* patterns: *Pressured Selling* and *Visual Interference*. Here, participants wanted to form an unbiased decision. One reason for *Confirmshaming* falling under this cluster might be that it also used highlighting. The other cluster that contains a pattern from *Interface Interference* overlaps with *Sneaking*. Both *Trick Question* and *Hidden Subscription* made the interface more complex and, thus, caused our participants to favor removing the manipulation.

Even though *Hidden Subscription* is a form of *Sneaking*, participants first wanted the interface to be simplified. We hypothesize that this pattern would receive similar justifications as the other *Sneaking* patterns once the interface is simplified. This is because, as with *Sneak into Basket* and *Hidden Costs*, it directly causes financial harm to users. We grouped those two patterns, as participants largely favored *HL+E* and wanted to be alerted. This might also hold for other *Sneaking* patterns, as long as they cannot be automatically removed. While *Forced Enrollment* could also not be altered by *SW* and *HD*, we did not group it with the other *Sneaking* patterns. Here, participants mainly wanted additional information, while being alerted was only secondary. Overall, both groups preferred *HL+E* for different reasons.

Overall, there are parallels between our clusters and the ontology by Gray et al. [13]. Still, the individual clusters within the high-level patterns provide slightly different reasons for choosing certain countermeasures. Interestingly, only our clusters involving *Interface Interference* patterns overlap with other high-level patterns.

7 LIMITATIONS

Our user group was fairly homogeneous with a rather technical background, limiting generalizability. Also, *SW* used Figma overlays, which let participants only activate one switch at a time and required turning it off to activate another. This software limitation was communicated before such trials. Finally, our countermeasures modify the contents of a website or app locally, similarly to ad blockers. Such on-client manipulations could imply legal issues.

8 CONCLUSION

In this work, we investigated three visual countermeasures against 13 common dark patterns using a lab study with interactive mock-up prototypes (n=20). Our results indicate that users prefer having access to more information, even when accompanied by increased visual clutter. As a result, *Highlight with Explanation* and *Switch* received better ratings and rankings than *Hide*, which silently removes the manipulation from a page. Future research could address this issue by reducing visual clutter while still providing access to valuable information about dark patterns, for example, by combining *Highlight with Explanation* and *Switch*.

When participants were asked to assign their preferred countermeasures to specific scenarios with different dark patterns, we discovered that users often preferred *Highlight with Explanation* in situations in which they could incur unnoticed costs and *Hide* for manipulations that made one option seem superior to another. However, we also found a stronger mistrust of *Hide* among our study participants, as it removes elements unnoticed. Based on participants' justifications, we clustered all investigated dark patterns into groups where participants provided similar reasons for specific countermeasures and compared our findings with a recent ontology by Gray et al. [13]. Our clustering can support extending current taxonomies and creating effective countermeasures for dark patterns that could be countered in similar ways.

Overall, our work strengthens and supports the findings from a previous online study [23], but also provides new insights into the applicability of promising visual countermeasures against dark patterns. Functioning versions of visual countermeasures that can be embedded on the user's end are an important next step for future research. Additionally, designing respective countermeasures will require a deeper understanding of the large amount of different dark patterns.

ACKNOWLEDGMENTS

This work was funded in part by the German B-IT Foundation

REFERENCES

[1] Aditi M. Bhoot, Mayuri A. Shinde, and Wricha P. Mishra. 2021. Towards the Identification of Dark Patterns: An Analysis Based on End-User Reactions. In *Proceedings of the 11th Indian Conference on Human-Computer Interaction* (Online, India) *(IndiaHCI '20)*. Association for Computing Machinery, New York, NY, USA, 24–33. https://doi.org/10.1145/3429290.3429293

[2] 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* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[3] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* 2016, 4 (2016), 237–254. https://doi.org/10.1515/popets-2016-0038

[4] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. *Thematic Analysis*. Springer Singapore, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[5] Harry Brignull. 2023. *Deceptive Patterns: Exposing the Tricks Tech Companies Use to Control You*. Testimonium Limited.

[6] European Commission, Directorate-General for Justice, Consumers, F Lupiáñez-Villanueva, A Boluda, F Bogliacino, G Liva, L Lechardoy, and T Rodríguez de las Heras Ballell. 2022. *Behavioural study on unfair commercial practices in the digital environment : dark patterns and manipulative personalisation : final report*. Publications Office of the European Union, Brussels, Belgium. https://doi.org/10.2838/859030

[7] Gregory Conti and Edward Sobiesk. 2010. Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/1772690.1772719

[8] Andrea Curley, Dympna O'Sullivan, Damian Gordon, Brendan Tierney, and Ioannis Stavrakakis. 2021. The Design of a Framework for the Detection of Web-Based Dark Patterns. In *The Fifteenth International Conference on Digital Society (ICDS 2021)*. IARIA, Nice, France, 24–30. https://doi.org/10.21427/20g8-d176

[9] 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 *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[10] Paul Graßl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. 2021. Dark and Bright Patterns in Cookie Consent Requests. https://doi.org/10.31234/osf.io/gqs5h

[11] Colin M. Gray, Shruthi Sai Chivukula, Kerstin Bongard-Blanchy, Arunesh Mathur, Johanna T. Gunawan, and Brennan Schaffner. 2023. Emerging Transdisciplinary Perspectives to Confront Dark Patterns. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 522, 4 pages. https://doi.org/10.1145/3544549.3583745

[12] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[13] Colin M. Gray, Cristiana Santos, and Nataliia Bielova. 2023. Towards a Preliminary Ontology of Dark Patterns Knowledge. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 286, 9 pages. https://doi.org/10.1145/3544549.3585676

[14] Colin M. Gray, Cristiana Teixeira Santos, Nicole Tong, Thomas Mildner, Arianna Rossi, Johanna T. Gunawan, and Caroline Sinders. 2023. Dark Patterns and the Emerging Threats of Deceptive Design Practices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 510, 4 pages. https://doi.org/10.1145/3544549.3583173

[15] 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 (CHI '21). , 5 pages. https://dbs.ifi.uni-heidelberg.de/files/Team/phausner/publications/Hausner_Gertz_CHI2021.pdf

[16] Kai Lukoff, Alexis Hiniker, Colin M. Gray, Arunesh Mathur, and Shruthi Sai Chivukula. 2021. What Can CHI Do About Dark Patterns?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 102, 6 pages. https://doi.org/10.1145/3411763.3441360

[17] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 81 (nov 2019), 32 pages. https://doi.org/10.1145/3359183

[18] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[19] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 192, 15 pages. https://doi.org/10.1145/3544548.3580695

[20] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. 2023. Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 194, 19 pages. https://doi.org/10.1145/3544548.3580729

[21] Marie Potel-Saville and Mathilde Francois. 2023. From Dark Patterns to Fair Patterns? Usable Taxonomy to Contribute Solving the Issue with Countermeasures. In *Annual Privacy Forum* (Lyon, France). 24 pages. https://www.researchgate.net/publication/371314839_From_Dark_Patterns_to_Fair_Patterns_Usable_Taxonomy_to_Contribute_Solving_the_Issue_with_Countermeasures

[22] Hauke Sandhaus. 2023. Promoting Bright Patterns. Position Paper at the Workshop "Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice" at the CHI Conference on Human Factors in Computing Systems (CHI '23). , 9 pages. https://doi.org/10.48550/arXiv.2304.01157

[23] René Schäfer, Paul Miles Preuschoff, and Jan Borchers. 2023. Investigating Visual Countermeasures Against Dark Patterns in User Interfaces. In *Proceedings of Mensch Und Computer 2023* (Rapperswil, Switzerland) *(MuC '23)*. Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/3603555.3603563

A STUDY PARTICIPANTS

The following table contains anonymized information on our study participants.

ID	Age	Sex	Profession Field	Awareness
P01	26	female	ecology	7
P02	25	male	aviation	7
P03	22	female	computer science	5
P04	31	male	computer science	6
P05	23	female	human resources	6
P06	26	male	engineering	4
P07	25	male	computer science	6
P08	25	male	IT security	7
P09	27	male	electrical engineering	5
P10	25	female	computer science	6
P11	26	male	computer science	2
P12	26	female	computer science	6
P13	24	male	IT	5
P14	25	male	computer science	7
P15	28	female	computer science	1
P16	31	female	computer science	6
P17	26	female	computer science	6
P18	21	male	computer science	5
P19	25	male	chemistry	7
P20	25	female	computer science	7

Table 3: Demographics of our participants containing their anonymous ID, age, stated sex, field of profession, and their subjective awareness regarding the existance of dark patterns on websites (scale from 1 *"Not aware at all"* to 7 *"Very aware"*).

B DARK PATTERNS AND STUDY SCENARIOS

In this appendix, we provide a table containing all 15 dark patterns with their original description from Mathur et al. [17] and the last part of the questionnaire in our lab study, in which participants chose specific countermeasures against each given pattern.

Dark Pattern	Assigned Letter	Description by Mathur et al. [17]
Activity Message	A	Informing the user about the activity on the website (e.g., purchases, views, visits)
Limited-time Message	B	Indicating to users that a deal or sale will expire will expire soon without specifying a deadline
Pressured Selling	C	Pre-selecting more expensive variations of a product, or pressuring the user to accept the more expensive variations of a product and related products
High-demand Message	D	Indicating to users that a product is in high-demand and likely to sell out soon, increasing its desirability
Sneak into Basket	E	Adding additional products to users' shopping carts without their consent
Confirmshaming	F	Using language and emotion (shame) to steer users away from making a certain choice
Testimonials	G	Testimonials on a product page whose origin is unclear
Countdown Timer	H	Indicating to users that a deal or discount will expire using a counting-down timer
Visual Interference	I	Using style and visual presentation to steer users to or away from certain choices
Forced Enrollment	J	Coercing users to create accounts or share their information to complete their tasks
Trick Questions	K	Using confusing language to steer users into making certain choices
Hidden Subscription	L	Charging users a recurring fee under the pretense of a one-time fee or a free trial
Hidden Costs	M	Revealing previously undisclosed charges to users right before they make a purchase
Low-stock Message	-	Indicating to users that limited quantities of a product are available, increasing its de- sirability
Hard to Cancel	-	Making it easy for the user to sign up for a service but hard to cancel it

Table 4: The 15 patterns that were listed by Mathur et al. [17] with their original description. In our study, we used 13 out of all 15 patterns. The assigned letter for each pattern matches our implementation of that pattern for our study as shown in Figure 2.

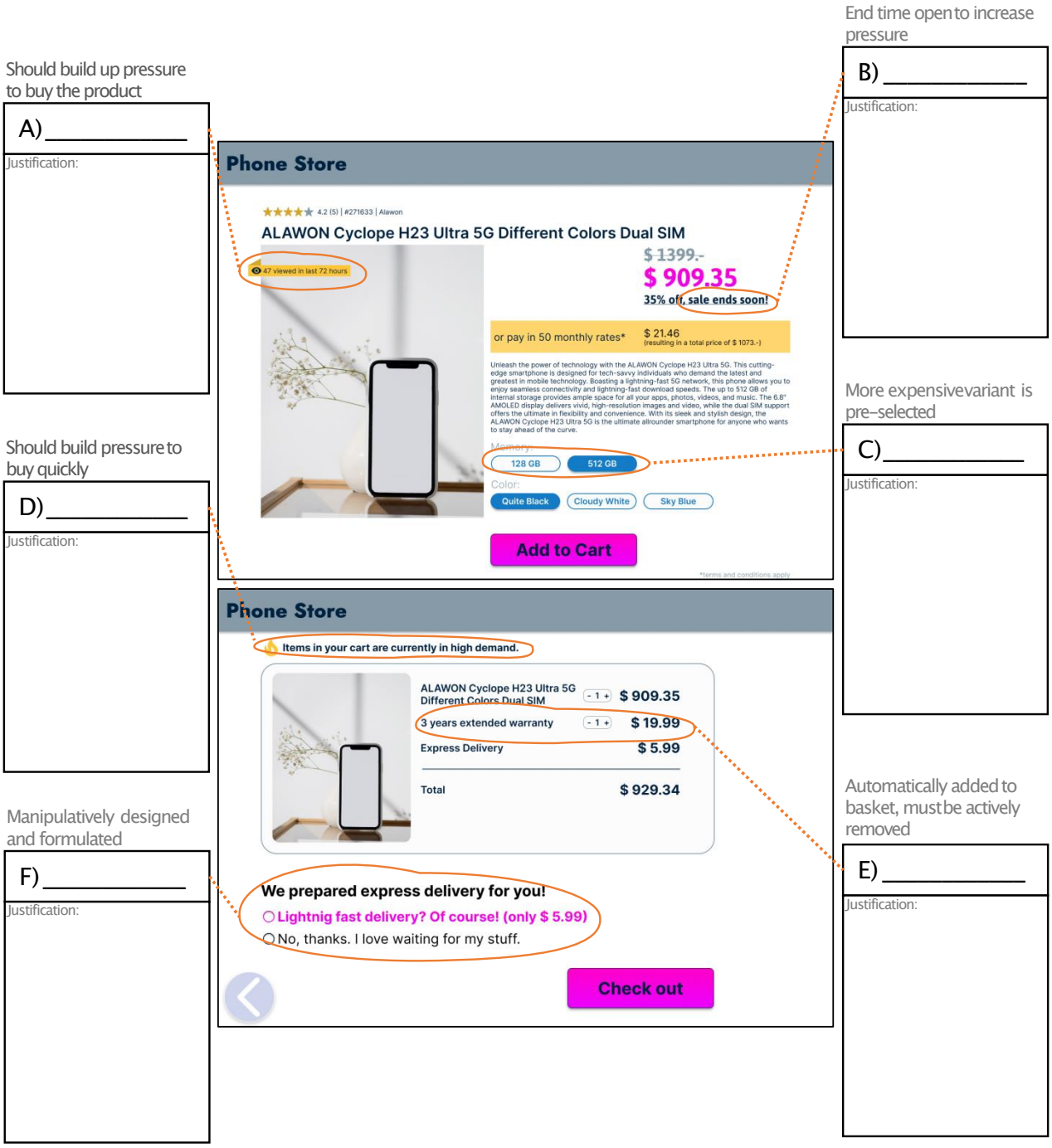

Figure 6: The phone scenario we used in our study. It contains the following dark patterns: (A) *Activity Message*, (B) *Limited-time Message*, (C) *Pressured Selling*, (D) *High-demand Message*, (E) *Sneak into Basket*, and (F) *Confirmshaming*.

Figure 7: The ticket scenario we used in our study. It contains the following dark patterns: (G) *Testimonials*, (H) *Countdown Timer*, (I) *Visual Interference*, (J) *Forced Enrollment*, (K) *Trick Questions*, (L) *Hidden Subscription*, and (M) *Hidden Costs*.