Investigating Visual Countermeasures Against Dark Patterns in User Interfaces

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

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

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

ABSTRACT

Dark patterns are malicious interface design strategies on the web and in apps that trick users into decisions that go against their best interests, costing them money, time, or private data. While there are approaches to classifying these patterns and investigating user awareness, there has been little work looking into visual countermeasures against dark patterns. In this work, we used an online survey to investigate concepts for six visual countermeasures against three common dark patterns: Confirmshaming, Low-stock Message, and Visual Interference. Our results indicate two opposing forces for users: On the one hand, users dislike systems actively making silent changes to their screen, preferring to be informed about the presence of dark patterns. On the other hand, they do not want applications to become visually cluttered, as this may impact their productivity. We found that different applications of dark patterns require different countermeasures, and that individual preferences vary strongly.

CCS CONCEPTS

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

KEYWORDS

deceptive patterns, dark patterns, visual countermeasures, online survey

ACM Reference Format:


1 INTRODUCTION

Interest in dark patterns within the research community has grown rapidly in recent years [19, 25]. The term Dark Pattern was introduced by Brignull on his website in 20101. It describes dark patterns as “tricks used in websites and apps that make you do things that you didn’t mean to, like buying or signing up for something”. There are many different ways to influence customers on websites [24] using, e.g., attention-grabbing dark patterns that try to maximize the time users spend within an application [23]. Of course, many aspects of webshops intend to influence the user without being dark patterns. For example, offering a limited-time discount is not a dark pattern, but including an attention-grabbing fake countdown would be. Without knowing the intentions behind a design, it can be quite tricky to be certain about whether a given design actually is a dark pattern. Such deceptive designs facilitate user frustration, confusion, and errors [16], so detecting and countering them is important. Researchers have worked to detect dark patterns [21], to identify and classify existing patterns [13, 21, 22], to show the effect of such patterns [26], to investigate users’ awareness of them [1, 2], and to propose possible countermeasures [3].

Looking at both the research literature and online practices reveals four main approaches to deal with dark patterns:

1. Raising the awareness of users [2, 9].
2. Putting pressure on companies by unveiling their practices (e.g., on social media)3.
3. Passing laws influencing the use of dark patterns [10, 17, 26, 30].
4. Introducing technological detection [21] and countermeasures [15].

While there is a relatively substantial body of research on dark pattern taxonomies and the influence of such patterns on users, research into possible countermeasures has been comparatively sparse. The concept of removing unwanted content from websites is already commonly applied in ad blockers. Assuming we can detect malicious design elements automatically (using approaches such as the one proposed in [21]), how should such programs then deal with dark patterns? Should they remove them quietly or highlight them with warning labels? How should they deal with dark patterns that might not be automatically detectable [7]? How should they handle those dark patterns that cannot be removed visually, like forced enrollment or the practice of displaying additional fees only on the last page right before checkout in online shopping? Currently, there is a large research gap regarding countermeasures for dark patterns that go beyond only detecting them.

In general, dealing with dark patterns on a visual level appears promising, as it can be done completely on the user’s side. So far, proposed visual countermeasures include drawing a box around the dark pattern and providing information to the user [21].

Our research thus focuses on this important next step of fighting dark patterns by introducing, evaluating, and comparing different techniques and concepts for visual countermeasures. We conducted


2like Reddit (/r/assholedesign) or Twitter (#darkpattern)
an online survey to gather first insights on different strategies to visually counteract three common types of dark patterns: Visual Interference, Confirmshaming, and Low-stock Messages (see Section 3.1). For this, we adapted a study design by Bongard-Blanchy et al. [2] and introduced six different visual countermeasures to deal with these malicious designs.

We found that generally, people liked receiving more information, while quietly removing dark patterns was seen as very controversial. Interestingly, the countermeasures introducing the most and least visual clutter were perceived as looking similarly good. Furthermore, our results indicate that dark patterns require different types of countermeasures and that preferences vary strongly between individuals.

With our work, we want to motivate the creation of tools capable of dealing with dark patterns on the user’s side. Overall, our contributions are:

1. Concepts for six visual countermeasures against dark patterns.
2. A study providing first insights into the applicability and usefulness of these countermeasures for three common dark patterns.

2 RELATED WORK

There are different approaches to countering the use of dark patterns, like introducing privacy-friendly (bright) patterns [2, 10, 28]. However, there will always be providers that consciously make the decision to use dark patterns. Following the four main approaches to dealing with dark patterns, we describe related work regarding all of the aforementioned ways to counter dark patterns.

2.1 Detection

To enable the automatic detection of dark patterns, it is crucial to be aware of the different types of dark patterns. For this, researchers already created taxonomies listing and classifying various types of dark patterns. Gray et al. [13] created a corpus of dark patterns using social media (e.g., Twitter). Their taxonomy contains over 15 dark patterns, which the authors clustered into five categories: nagging, obstruction, sneaking, interface interference, and forced action. Mathur et al. [21] automatically detect dark patterns and, with this, investigated around 11,000 shopping websites regarding the usage of dark patterns. They provide a lower bound of used dark patterns as their crawler only works text-based, meaning that only written components of dark patterns are detectable. Using the data from their crawler, the authors created a taxonomy containing 15 types of dark patterns and suggest future work on, inter alia, countermeasures against dark patterns. The authors later built upon this work by creating a more detailed taxonomy that summarized the research in the field of dark patterns [22].

Approaches to automatically detect dark patterns exist. The aforementioned work by Mathur et al. [21] is among the most prominent. More recently, Soe et al. [29] trained a machine learning model to detect dark patterns in cookie banners. The authors argue that there are many challenges to detecting dark patterns automatically. One example is that it is not trivial to deduce the intent behind a specific design. Curley et al. [7] discuss the detectability of different types of dark patterns and argue that some patterns are hard or even impossible to detect automatically due to the large variety in which they present themselves in applications. Hausner and Gertz [15] argue that dark pattern detection tools will always be reactive, as it is difficult to predict what new patterns may arise in the future. The authors also state that, besides user awareness and regulation laws, automatic detection of dark patterns is an important step to protect people from these malicious designs.

Overall, most existing approaches to detect dark patterns automatically work text-based.

2.2 Awareness

Once dark patterns are known and classified using, e.g., taxonomies [4, 9, 13, 21, 22], it becomes easier to raise public and user awareness regarding these malicious design practices. This is an important step in fighting dark patterns [2, 9], especially since users are already beginning to consider them normal when using the internet [1, 5, 9]. As a result, how to raise awareness has been investigated extensively over the last decade.

Bhoot et al. [1] found five variables that are important for users to identify dark patterns: frequency of occurrence, level of frustration, trustworthiness, misleading behavior, and physical appearance. In their self-administered questionnaire with 300 collected responses followed up with an experiment with ten participants, the use of a forced enrollment was identified most often (by around 88% of participants), while roach motel, a pattern trapping users into subscriptions, was only identified by less than 20% of participants. Trick questions were only identified by around 32% of participants. The authors state that participants were aware of the use of dark patterns by companies, but accepted them as a part of the internet experience.

Di Geronimo et al. [9] analyzed 240 popular mobile apps and found that 95% of those apps contained at least one dark pattern. On average, apps contained around seven dark patterns (SD=5) each. With being present in approximately 60% of all investigated apps, preselections and false hierarchy were the most used dark patterns across apps, while nagging was most frequently found overall. The authors also conducted an online survey with 589 users. In contrast to [1], most users did not recognize dark patterns. The authors argue that informing people about dark patterns helps them to better identify and recognize these malicious designs. Like Bhoot et al. [1], some participants stated that dark patterns belonged to the internet experience. Finally, the authors state that creating tools to educate people on dark patterns might be a good next step.

In contrast to Di Geronimo et al. [9], Bongard-Blanchy et al. [2] surveyed 406 people and found that people were generally aware of dark patterns. However, they were still manipulated by them. People were unsure about the potential harm that some dark patterns might inflict. In their experiment, they gave participants ten screenshots of interfaces (nine of which contained a dark pattern) and asked participants to spot dark patterns in the interfaces. 59% of the participants were able to detect five or more dark patterns overall. The authors stress that raising awareness and automatic detection are ways of countering dark patterns that should be explored further.

Like Bongard-Blanchy et al. [2], Keleher et al. [17] conducted a study asking participants to identify whether given images of
interfaces contained manipulative patterns. Overall, participants were able to correctly identify manipulations in around 50% of cases. The authors state that end users often did not recognize manipulative patterns and tended to be more positive when confronted with rather implicit dark patterns such as confusing wordings or countdowns.

Gray et al. [11] also investigated the end user perception of dark patterns. In their survey of 169 participants, 79% reported being aware of manipulations, and most blamed designers (n=97), other stakeholders (n=92), and developers (n=54) for manipulations. Interestingly, 34 people blamed themselves for being manipulated. Additionally, the authors report that most participants stated that they mistrusted digital products at least "sometimes", over 82% for smartphones and around 89% for websites.

In their study with 204 participants, Voigt et al. [31] found that people who were more familiar with technology did not identify significantly more dark patterns compared to other people. Apart from investigating user awareness using studies, it is also important to increase public awareness and interest in dark patterns. A CHI’21 workshop [19] with 18 position papers helped focus HCI research on the topic and demonstrated its relevance to the HCI community.

Public communities are also aware of dark patterns and share their findings on social platforms such as Reddit[4] or Twitter[5]. Here, Gray et al. [12] investigated the subreddit r/assholedesign. The authors analyzed 1002 posts on the platform and found many already-known dark patterns. They found that about 27% of the analyzed posts actively included the name of the company that used the dark pattern to shame them.

In addition to awareness of dark patterns on social media platforms, there are already small games showing the ridiculousness of some dark patterns. One example is Cookie Consent Speed.Run[6], which challenges visitors to decline all cookies as quickly as possible.

2.3 Laws and Regulations
Another factor that influences dark patterns is regulation laws. One example is the General Data Protection Regulation introduced in the EU in 2016, which increased the use of cookie banners on websites[8]. Several researcher groups subsequently investigated the influence of such laws on dark patterns in cookie banners and showed that the design of cookie banners highly influences the decision-making of users[10, 20, 26, 30]. Luguri and Strahilevitz[18] provided evidence that some dark patterns can be seen as unlawful, and Habib et al. [14] investigated the usability of consent interfaces.

In 2018, the Norwegian Consumer Council (NCC) published a report explaining how tech companies make use of dark patterns. They investigated the GDPR settings of Facebook, Google, and Windows 10. Results include that several dark patterns try to nudge users to pick privacy-intrusive settings using, inter alia, "bad" default values, misleading texts, and the illusion of control, which the NCC considered unethical. In 2022, an extensive report by the EU[5] also analyzed the use of dark patterns and emphasized their international relevance. According to the report, dark patterns often place themselves in a gray area between legitimate persuasion attempts and unlawfulness. The authors also state that a lack of user awareness regarding dark patterns exists, but emphasize that once people are aware, using dark patterns is seen negatively by them. The European Data Protection Board also released a report on best practices to recognize and avoid dark patterns on social media platforms.

2.4 Countermeasures
Automatically detecting dark patterns is not trivial and may sometimes even be impossible[7]. Even if it works, it leads to the question of how best to utilize this knowledge. Consequently, research in the field of dark patterns already suggests investigating possible countermeasures against dark patterns [2, 3, 21].

Bösch et al. [3] investigated various privacy-related dark patterns and describe possible countermeasures against them. Their proposed countermeasures mainly include raised user awareness, active user actions to work against dark patterns, and support from third party tools such as BugMeNot[10] to bypass forced enrollments.

Mathur et al. [21] suggest using their data set to create tools that can flag and annotate dark patterns with an explanation for the user. Their suggestion inspired one of the countermeasures we evaluate in our study. Building upon the work of Mathur et al. [21] to automatically detect text-based dark patterns, the TeenHacks LI (2019)[11] winning team created the browser extension Insite[12]. Once it finds a pattern, it highlights it for the user and provides a pop-up explaining the category of the present dark pattern.

Hausner and Gertz[15] try to detect dark patterns in cookie banners using graph neural networks on the DOM (Document Object Model) of a webpage. With this, they are capable of detecting Visual Interference by analyzing the CSS styles of buttons within the cookie banners. Their program then manipulates the DOM for users to visually detect and spot dark patterns more easily.

To our knowledge, there is no other published research that compared different methods to visually counter dark patterns. With this, our work is one of the first important steps in creating a deeper understanding regarding visual dark pattern countermeasures that are applicable on the user’s side.

3 STUDY
As seen in Section 2.1, detecting dark patterns is an area of active research. We were interested, however, in countermeasures as the logical next step. In particular, we wanted to focus on the potential of applying visual countermeasures on the user side. Since there are numerous ways to do this, we make the following assumptions to focus the scope of our work on a set of comparable countermeasures:

---

[3]using hashtags such as #darkpattern
We introduce and investigate seven strategies to deal with dark patterns. We build upon suggestions from the literature and extend them to provide short definitions of the three dark patterns:

- **Confirmshaming**: Uses emotions to make users feel guilty when, e.g., declining an offer. An example text on respective buttons could be “No thanks, I do not want the discount”, or “I do not care about losing my data”.
- **Low-stock Message**: False information about a product being nearly sold out that can urge people to buy it faster.
- **Visual Interference**: Creating visual imbalance to influence user decisions. An example is a common cookie banner where one button is intentionally made bigger and more colorful while the other option is displayed as simple text.

Since these dark patterns are very common, they provided a solid starting point to test a larger set of countermeasure concepts on and get first insights into their applicability.

### 3.2 Countermeasures

We introduce and investigate seven strategies to deal with dark patterns by altering the current screen visually. Since there has been only little work on visual countermeasures for dark patterns, we build upon suggestions from the literature and extend them with our own ideas.

- **Unchanged (UC)**: The manipulative element is not changed in any way. UC serves as the baseline to compare our visual countermeasures against.
- **Highlight with Explanation (HL+E)**: A red dotted box is drawn around the manipulative element on the page. We also add a small red warning sign that explains why this content is marked on hover. The explanation includes what exactly is manipulative and how it aims to affect the user. This strategy was proposed by Mathur et al. [21].
- **Highlight without Explanation (HL)**: Like HL+E, but without the warning sign and explanation. This still makes the user aware that a dark pattern may be present while reducing screen clutter.
- **Lowlight (LL)**: The manipulative element is made less prominent, by graying out text or buttons, for example. Since some dark patterns (e.g., Low-stock Messages and Countdowns) want to be seen, this technique counters this effect by making them fade into the background more without removing the information they convey. This strategy tries to not draw any attention toward the manipulative element.
- **Hide without Marking (HD)**: The manipulative element is visually altered and/or rephrased or removed completely. Simply removing dark patterns is one of the most obvious countermeasures.
- **Hide with Marking (HD+M)**: Like HD, but additionally a green dotted box is drawn around the area of the manipulative element to indicate its removal. We included this countermeasure because trust in a tool is important and we hypothesized that users would disapprove of a countermeasure that makes hidden changes. HD+M communicates that a certain area on the screen has been altered.
- **Switch (SW)**: The manipulative element is hidden. Next to it, a glowing light bulb is shown. When clicked, the bulb goes out and the manipulative element is shown instead, allowing the user to revisit the original content. This countermeasure tries to combine the benefit of simply removing the manipulation with increasing trust in the tool, as users can always toggle between the original and the altered screen.

### 3.3 Independent Variables

Our independent variables were COUNTERMEASURE with seven levels (see Section 3.2) and PATTERN with the three levels Confirmshaming, Visual Interference, and Low-stock Message.

### 3.4 Dependent Variables

We measured self-reported user preferences using six 7-point semantic differential scales about our proposed countermeasures. The scales were Usability (Hard to use vs. Easy to use), Clarity (Confusing vs. Clear), Safety (Dangerous vs. Safe), Efficiency (Inefficient vs. Efficient), Helpfulness (Unhelpful vs. Helpful), and Feeling (Whether the given countermeasure would make the website feel better or worse).

### 3.5 Study Design

For our study regarding visual countermeasures against dark patterns, we created three non-interactive, static screenshot prototypes of an online shopping scenario that we adapted from Bongard-Blanchy et al. [2] (Fig. 1). Each screenshot included one of three common dark patterns: Confirmshaming, Low-stock Message, and Visual Interference (see Section 3.1).

Our study consisted of two parts. First, participants were given screenshots showing the effect of each countermeasure applied to one dark pattern at a time. For this, the dark pattern in the shopping scenario (Fig. 1) was altered according to the respective countermeasures from Fig. 2.
Figure 1: We investigated countermeasures for three dark patterns on a shopping website mockup: (a) Low-stock Message, (b) Visual Interference, and (c) Confirmshaming. Only one pattern was used at a time. The buttons for Visual Interference and Confirmshaming replaced the default button. The Low-stock Message was placed as shown. The overall website design was adapted from Bongard-Blanchy et al. [2].

Participants then ranked all seven countermeasures for the given pattern. This procedure was repeated for the remaining patterns before providing an overall ranking of all countermeasures at the end of the survey. The second part focused on participants’ opinions regarding each individual countermeasure by asking for the six 7-point semantic differential ratings listed in Secton 3.4. Two optional free-text questions regarding strengths and weaknesses of the countermeasure were included to gather more qualitative insights. We used a within-subjects study design. The order of the three dark patterns was randomized, with the general ranking always being at the end of the survey. We did not change the order of the countermeasure screenshots between dark patterns to reduce potential confusion caused by the number of screenshots. Since all countermeasures for a dark pattern were shown on one page simultaneously, participants could think about all countermeasures before answering any questions, minimizing the effect of the unchanged order in this part of the study. In the second part, the order in which countermeasures were rated and semantic differential ratings were randomized. The study was conducted using the online survey tool SoSci Survey\(^{13}\). We invited our participants over private and professional connections and through posts on Reddit (r/SampleSize)\(^{14}\).

\(^{13}\)https://www.soscisurvey.de Accessed: March 2023
\(^{14}\)https://www.reddit.com/r/SampleSize/ Accessed: March 2023
4 RESULTS

Overall, 40 people participated in our study, aged 20 to 62 (M=30.45 years, SD=11.49 years, 12 women, 25 men, 2 non-binary, and 1 n/a). 32 of them reported professions we classified as a technical academic background. They reported an average daily computer browsing time of 5.58 hours (SD=3.24, min=0, max=12). Self-reported awareness of dark patterns was high (M=4.3, SD=0.91, on a scale from 1 “Not at all” to 5 “Very much”), with only two participants reporting limited experience with dark patterns.

4.1 Semantic Differential Scales

We analyzed the effects of COUNTERMEASURE on all six semantic differential scales (see Fig. 4). A Friedman test revealed significant effects of COUNTERMEASURE on USABILITY ($\chi^2(5)=24.11$, p<0.001), CLARITY ($\chi^2(5)=47.46$, p<0.001), SAFETY ($\chi^2(5)=26.20$, p<0.001), EFFICIENCY ($\chi^2(5)=22.17$, p<0.001), HELPFULNESS ($\chi^2(5)=31.77$, p<0.001), and FEELING ($\chi^2(5)=29.72$, p<0.001). We used Wilcoxon signed-rank tests with Holm corrections as post-hoc tests to further analyze all of these effects.

Overall, countermeasures were perceived as at least neutral on average for most ratings. HL+E and HD had the most significant differences from other countermeasures across all ratings.

4.1.1 USABILITY. Participants found HL+E significantly easier to use than HL. This shows that only highlighting a dark pattern without explaining it further was considered much less useful. Overall, our countermeasures were considered rather easy to use, with mostly similar ratings. Only HL+E and HD stand out slightly against the remaining countermeasures.

4.1.2 CLARITY. HL+E provided significantly more clarity to our participants than all other countermeasures except for HD. SW and HD were both perceived as significantly clearer than HD+M and HL, with HD being also significantly clearer than LL. While HL+E and HD were seen as very clear, HL and HD+M rather confused participants as they highlight areas without further information.

4.1.3 EFFICIENCY. As with CLARITY, participants found HL+E and HD significantly more efficient than HL. In addition, HD was perceived as significantly more efficient than LL. Overall, HL+E and HD were rated as very efficient and received similar scores.

4.1.4 SAFETY. The post-hoc tests revealed significant differences, with HL+E being rated significantly safer than all other countermeasures except for SW. Generally, our countermeasures were perceived as rather safe. However, some participants rated HD as very dangerous as the changes it made were hidden from them.

4.1.5 HELPFULNESS. HL+E was rated as significantly more helpful than any other countermeasure except for SW. Additionally, SW was only considered significantly more helpful than HL.

4.1.6 FEELING. Overall, participants felt that HL made the website look significantly worse than with every other countermeasure except for HD+M. Surprisingly, the countermeasures with the most and least visual clutter (HL+E and HD) received very similar, positive ratings.
4.2 Ranking

Participants ranked all countermeasures four times: once for each dark pattern and once overall, shown in Fig. 3. In this overall ranking, HL+E was ranked very highly, with 19 participants giving it the top rank. SW and LL were also rather liked, while HL and HD+M were rather disliked. UC was ranked last by 14 of our 40 participants, but was also rated best three times. HD was the most controversial countermeasure: It was ranked best 9 times but also worst 11 times.

Regarding the rankings per pattern, HL+E and HD were generally ranked highly, with HD receiving mixed results for Confirmshaming. LL was ranked well regarding Low-stock Message, and was generally distributed across all ranks. Still, few participants liked this countermeasure, with seven participants placing it on the second rank for Confirmshaming. SW and HD+M were placed mostly on the middle ranks. HL received mixed rankings for Confirmshaming and was placed more toward the lower ranks for the other dark patterns. Still, eight participants put it on the second rank for Visual Interference. UC was often placed in the last ranks. Visualizations of these rankings per pattern are included in Appendix A.
4.3 Participant Comments

We asked participants to comment on the strengths and weaknesses of each countermeasure. When citing participants, we corrected spelling and grammar and translated them when necessary. Cita-
tions are placed in quotation marks and followed by the partici-
 pant’s anonymized number (e.g., P17). Statements made by multiple partici-
pants are indicated with their frequency in parentheses.

Highlight with Explanation (HL+E). Many participants perceived this countermeasure as too visually cluttered (14). The warning sign was perceived as “patronizing” (P6). “It looks like a download button for a virus” (P32). Especially on websites with many dark patterns, “several such indicators may make the website somewhat confusing” (P19). It was perceived as “annoying over time” (P3), “daunting” (P17) or “confusing” (P18). P12 suggested adding the option to deactivate certain explanations for the future, to counteract this. Participants liked the fact that there was an explanation given (12), thus educating the user (9) and fostering informed decisions (6): “you get all the information you need and still get to make the final decision” (P1). This would also help inexperienced users (P2). Participants entertained the idea that this countermeasure would “frustrate” (P19) the manipulating parties as “clear call-outs of scumf**k behavior” (P26) (3). While two participants liked that the explanations only showed up on hover, there were concerns if they would be used since “nobody reads hidden information” (P29). P6 remarked that “The learning effect will be enormous, maybe even to the point where this can be turned off again”.

Highlight without Explanation (HL). The most frequent critique for HL was that it provided no additional explanation why something was highlighted (11): “Users might need an explanation to decide which is the right option to choose” (P6). Additionally, the highlighting with a red box was perceived as intimidating (2): “Not even I would dare to click any button” (P32). This is in line with the second frequent remark that HL clutters/worsens the website (7): “I don’t like the visual clutter” (P11). Participants were concerned that the red box could strengthen a dark pattern in certain situations (5), defying the purpose of the countermeasure: “This highlights the text ‘Hurry, limited quantity left’ so that the user might be even more confused” (P3). Participants liked about HL that it alerts users (5) since “it clearly shows the attempt” (P10), and facilitates user reflection (5) by “alarm[ing] you to check yourself what the website probably wants you to do” (P13). Participants appreciated the content of the website not being changed (3) so that it is “transparent” (P12) “like a helpful overlay” (P1), making it “very simple” (P22).

Lowlight (LL). With LL, the effect of manipulation is reduced (8), as “it robs a lot of common dark patterns of their most useful feature”. But manipulation is not stopped completely (8). This applies especially to manipulative text since “most biased phrasings are still more noticeable” (P32) and “the evil text might still be visible” (P3). Participants were concerned that LL makes the site harder to use (6), especially for visually impaired users (2). Applied changes might be part of the website design (5), which could result in shady websites looking misleadingly trustworthy (2) since “the manipulation is now even more subtle” (P6). Participants liked that LL is visually clear (4) and supports prioritizing information on a site (4) because “it does not visually interfere that much” (P1). Additionally, it exposes manipulation attempts (3) but does not change the content (3), making it “much safer and clearer than hiding things” (P13).

Hide without Marking (HD). Participants feared the removal of relevant content (11). Since it does not indicate changes in any way (10), this “could lead to really dangerous results [...]” (P9). If the countermeasure works flawlessly, users could not use the presence of dark patterns to recognize shady websites (2), since it “hides the villainous nature of the vendor” (P26). Additionally, this countermeasure “might lead to censoring” (P3) or “creators of the countermeasure could use it to manipulate in other ways” (P12). Participants appreciated that this countermeasure “disarms” (P10) or even “eliminates the possibility” (P9) for manipulation (11). This results in a clear website without extra content (9), reducing cognitive load for the user (8), and allowing them to “decide completely unbiased” (P2) (4). Therefore, it “could be a good solution for parents to protect their children” (P36).

Hide with Marking (HD+M). 12 participants stated that it “might be confusing if hidden information is highlighted” (P39). Participants were curious why something was hidden (8). Users “want to know what is hidden and why” (P23), because HD+M “could make you feel unsafe about a potentially harmless website through the fear of the unknown it creates” (40). Users suggested that it should be possible to see the original content (6) or that an explanation should be available (3). Note that the former resembles SW. As with HL, participants liked that it counteracts the manipulation effectively (7), but additionally shows that manipulation was present (5): “I am unbiased while still knowing that someone wanted to trick me” (P2). It is a simple countermeasure (3) that lowers cognitive load (2).

Switch (SW). All other ratings were in the midterm. Participants liked the option to see the original content (11), allowing them to reflect on it (4). They liked that it hides the dark pattern initially (6): “I am initially unbiased. It isn’t distracting. I can see the manipulation as it was intended if I want to.” (P2). “It protects the user from unaware manipulation, as the manipulation becomes only visible after being made aware” (P36). Participants think “it might be fun to hit the switch” (P40) (2), but there are concerns that it makes surfing inefficient (5) as it clutters the page (3) and users might “click it all the time to double-check whether the program decided right” (P13). However, the option to switch to the original content allows “to build up trust with the countermeasure tool: Whenever I still desire to know why the interface was altered, I have access to the original, but over time I will look at that less and less often as I see that the tool only does helpful alterations” (P12).

5 DISCUSSION

In the following, we discuss the results we obtained, and highlight interesting and surprising findings as well as open research gaps. We conclude with a summary of our most important findings.

5.1 Information is Power

While hiding dark patterns results in a clean website, users were concerned that relevant content could be lost. HD was rated as the easiest countermeasure, but it was also rated as the one that feels the second most dangerous. Users want to be able to reflect on
the decisions of the countermeasures, resulting in split opinions regarding HD. Some users liked that the dark patterns were simply removed since they do not have to bother with them at all. However, others disliked the idea that a program unobservably alters their current view. This controversy might explain half of all participants placing HD on the best or worst rank (see Fig. 3). Users could also be tricked by shady websites and apps that get to look more trustworthy, which would be a dangerous unintended side-effect of HD.

Overall, participants greatly appreciated explanations for why an element was classified as a dark pattern using HL+E which is a suggested countermeasure by Mathur et al. [21]. Thus, it could be interesting to combine HD (or HD+M) with explanations as used by HL+E. This could reduce visual clutter while the user can still access information regarding changes.

5.2 Visual Clutter
Displaying more information results in either more occupied screen space or more necessary interaction steps by the user to retrieve the information (e.g., hovering over an icon). Since HL+E was frequently placed on the top rank, a next step is to investigate whether the added visual elements cause too much annoyance among users to be constantly active, as this was a recurring remark by participants. Across all countermeasures, HL+E added the most visual elements, thus possibly resulting in visual clutter. Therefore, it is very surprising that participants perceived HL+E and HD (least visual clutter) to make the site look similarly good. This is especially interesting since 14 out of our 40 participants stated that HL+E creates too much visual clutter.

Adding visual cues may also result in relevant downsides. Similarly to HD that could make shady websites look more trustworthy, adding warning signs and labels to a website could make the website look shadier and could increase distrust for users—which is not necessarily bad. However, this becomes especially important when dealing with false positives (designs that were wrongly classified as a dark pattern by an algorithm), and could result in financial harm and lowered credibility for the owners of the site.

One way to cause less visual clutter without losing information is SW. However, this countermeasure is comparable to a curtain hiding the original content of the website. With this, it could easily distract users from their goal on the site by making them curious to see what content was hidden.

5.3 Countering Dark Patterns
Overall, our results indicate that a good countermeasure should react to the kind of dark pattern it deals with. For example, patterns that want to grab the user’s attention, such as Low-stock Messages, can be better countered using LL. LL also received good ratings for Low-stock Messages. Reducing the contrast of the manipulative design also allows for preventing information loss (as with HD) when dealing with false positives. These are hard to rule out for, e.g., Low-stock Messages, since it may be impossible to detect whether such messages are genuine or fake. While LL is useful for these attention-grabbing patterns, HL+E and HD appear to be useful when dealing with Confirmshaming and Visual Interference.

While we can already present insights on visual countermeasures for the three aforementioned dark patterns, it is not clear how well our findings transfer to other dark patterns. Additionally, two main factors emerged influencing the selection of a fitting countermeasure: the dark pattern and user preference.

User requirements for countermeasures may change as their expertise increases (see P12’s comment regarding SW in Section 4.3). With this, using countermeasures that include more explanations (like HL+E and SW) might be a good choice for novices. Once a user got used to recognizing dark patterns and the user’s trust in the countermeasures increased, explanations could become superfluous, leading to countermeasures such as HD. However, it is important to keep in mind that certain dark patterns might not be automatically detectable at all [7], which limits the applicability of automatic countermeasures.

Quietly removing possible dark patterns will always come with the danger of accidentally removing genuine content (false positives). In the end, user preference will be defined by a combination of educated knowledge of the domain, trust in the system, and expertise and efficiency in using the browser. With this, behavior customization of countermeasures becomes an interesting avenue for further research.

5.4 Summary of Findings
In the following, we summarize our most interesting findings that could help researchers and practitioners to further investigate and create automatic countermeasures for dark patterns.

- Users want to know why content is classified as a dark pattern. Providing visual cues without further information is not recommended.
- HD was the most controversial countermeasure. It provides the least visual clutter, but also performs hidden changes in the users’ views, which was seen as very dangerous.
- HD could make shady websites look inadvertently more trustworthy.
- HL+E could make websites look less trustworthy, or even dangerous, especially when false positives occur.
- HL+E creates the most visual clutter but received similarly good ratings on the look of the website as HD.
- LL received good ratings overall, while it received only mixed results for all individual rankings.
- Preferences vary strongly between individuals, indicating a likely need for countermeasures to be customizable. With this, increasing user expertise regarding dark patterns might require different or adaptive countermeasures.
- Trust in the countermeasures is very important, especially when content is hidden from the user. If users do not trust the countermeasures, HD and HD+M should not be used.

6 LIMITATIONS
Most of our participants had a technical background, which suggests that testing with a broader user base will help to verify that our findings are more generally applicable. We were also only able to test six countermeasures for three popular dark patterns. We limited this work to three dark patterns to be able to test and explore a larger variety of visual countermeasures, since this space
has not been explored much in research until now. Future research should use our findings on dark pattern countermeasures and, with this, be more exhaustive regarding tested patterns. Since there are many more dark patterns (see [4, 9, 13, 21, 22] for an overview), other countermeasures might be more appropriate for other patterns. Additionally, our countermeasures would modify content locally on the user’s side, similar to ad blockers, which may imply legal issues. Finally, our study did not include interacting with the site but was entirely based on static screenshots. Therefore, a next step to continue our work is to implement a set of interactive countermeasures and to observe whether future studies continue to confirm our findings.

7 CONCLUSION AND FUTURE WORK

In this research, we investigated six different visual countermeasures and an unchanged baseline and applied them to three common dark patterns embedded in a shopping website mockup. Our findings indicate that people want to be informed about automatic content changes. Thus, $\text{HL+L}$ performed well throughout our study. $\text{HD}$ was liked especially for dealing with Low-stock Messages and Visual Interface, but the participants’ final ranking was controversial since some considered a program dangerous that changed their view without informing them about applied changes. $\text{HL}$ and $\text{HD+M}$ did not perform well, as they highlight an area without providing explanatory cues which was rather disliked by our participants. $\text{SW}$ was generally appreciated, since it allowed users to toggle between original and changed versions. However, participants feared that adding interactive elements might make them less efficient as it would hide information behind a curtain.

With our work, we gained first valuable insights into the applicability and usefulness of different visual countermeasures for certain dark patterns. The next step is to investigate some of the most promising countermeasures, such as $\text{HL+L}$, in more depth using interactive prototypes. This would also allow us to investigate whether user efficiency and satisfaction is influenced by applied countermeasures. It may also be worth combining the situational advantages of different countermeasures by developing a mixed or factorial approach. Additionally, creating browser plugins that apply countermeasures is an interesting research direction [6, 21]. Such plugins could combine existing data sets [21] with our proposed visual countermeasures. Finally, a major factor for successful long-term usage of the countermeasures presented will be low user annoyance. With this, allowing users to customize the behavior and intrusiveness of applied countermeasures becomes an intriguing research direction.

Overall, lastingly approaches to handling dark patterns may be user awareness and exposing manipulative companies. User awareness helps people to be more cautious around the internet, but it only treats a symptom of dark patterns, and users are affected by dark patterns even when they are aware of them [2]. By exposing such practices to non-expert users (for example, with warnings in search results), the respective companies might experience financial losses, providing incentives to remove dark patterns in the long run. Companies have shown creativity in developing dark patterns, and they will likely find ways around new laws; and any automatic detection of dark patterns will include errors. Indeed, if users were to rely solely on technology to protect themselves against dark patterns, missed patterns might inflict even more harm than today. Nevertheless, countermeasures can support users in making informed decisions, and they can help reduce the influence of malicious designs, making it sensible to combine this approach with raising user awareness and introducing legislation.

With our efforts, we want to encourage future research on automatic dark pattern detection that goes beyond text-based manipulation, the exploration of dark pattern countermeasures, and the development of appropriate tools, to minimize the manipulative effects of dark patterns.

ACKNOWLEDGMENTS

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

REFERENCES


In this appendix, we provide visualizations of the individual rankings of all countermeasures regarding the three dark patterns: **Confirmation Shaming**, **Low-stock Message**, and **Visual Interference**. Throughout all rankings, there was a general preference for Highlight with Explanation and Hide without Marking. Unchange was disliked the most. Still, few participants preferred it over all other countermeasures. Countermeasures that only highlight areas without providing more information were generally disliked by our participants.
Figure 5: Ranking of all countermeasures for the dark pattern Confirmshaming. *Highlight with Explanation* was clearly preferred, with nearly half of our participants giving it the top rank.

Figure 6: Ranking of all countermeasures for the dark pattern Low-stock Message. Participants equally preferred *Highlight with Explanation* and *Hide without Marking*.

Figure 7: Ranking of all countermeasures for the dark pattern Visual Interference. Participants equally preferred *Highlight with Explanation* and *Hide without Marking* with a slight tendency towards *Hide without Marking*. 