"Deception Detected!" — A Serious Game About Detecting Dark Patterns

Kevin Fiedler[1 [0009−0008−3644−1488]], René Schäfer[1 [0000−0002−0078−1412]], Jan Borchers[1 [0000−0003−1509−3257]], and René Röpke[2 [0000−0003−0250−8521]]

[1] RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany
{kfiedler,rschaefer,borchers}@cs.rwth-aachen.de
[2] TU Wien, Favoritenstr. 9-11, 1040 Vienna, Austria
rene.roepke@tuwien.ac.at

Abstract. Dark patterns are malicious user interface design strategies that nudge users toward decisions that may be against their best interests. As countermeasures, related work has explored legislation, detecting and modifying web content with such patterns, and educating users about dark patterns. Serious games that teach detecting and classifying dark patterns, however, have received little attention so far. We present a web-based prototype of such a game and the results of a comprehensive user study investigating its effectiveness, user experience, and knowledge retention among participants. Our findings demonstrate that participants became better at detecting dark patterns, indicating that they became more aware of such design manipulations. While our analysis of participants' classification abilities yields mixed results, our game log data confirms that their detection skills increased.

Keywords: Dark Patterns · Deceptive Design · Serious Game.

1 Introduction

Dark patterns are malicious user interface design strategies widely used in apps and online services [5,13,18] that aim to manipulate users into specific actions and decisions that may go against their best interests, such as expensive purchases or hidden subscriptions [7,18]. To mitigate the manipulative effects of these patterns, researchers have called for effective countermeasures [1,8]. Examples of such countermeasures include legislation [7] and techniques to detect and counteract dark patterns in the user interface [1,4,23]. Educating users, however, presents a complementary approach to fighting dark patterns [1,2,17]. As related work in IT security education has demonstrated, serious games can enable learning about such topics in a risk-free and engaging learning environment [21].

Therefore, we introduce a web-based serious game about detecting and classifying dark patterns in online content (Sections 3 and 4). It lets users familiarize themselves with different manipulation techniques widely applied online and in apps. Our game adopts the preliminary ontology by Gray et al. [9]. After describing the design of our game, we present the results of a comprehensive user study

that evaluates the game's effectiveness, user experience, and knowledge retention (Section 5). We use a pre-/post-/re-test experimental study design. Lastly, we discuss our findings and possible limitations (Section 6) before concluding the paper and outlining future work (Section 7).

2 Related Work

To better understand how the concept of serious games may be applied to the topic of dark patterns, we discuss prior work in both areas.

2.1 Dark Patterns

The term *dark pattern*[3] was introduced by UX researcher Harry Brignull in 2010 on his website[4]. Nowadays, Brignull [3] refers to them as *deceptive patterns* and defines them 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" on his website[4].

Such manipulative design strategies can be found in mobile games [6], social media [19], shopping websites [18], and many other domains. Because of this, researchers have collected existing types of dark patterns [10,18] and call for effective countermeasures to mitigate the manipulative effects of these patterns [8]. Bongard-Blanchy et al. [1] propose an intervention space for such countermeasures. This space contains the fields *Regulatory* (e.g., [7]), *Technical* (e.g., [2,18]), *Design* (e.g., [11,20]), and *Educational* (e.g, [1]). While current research does not agree on the impact of user awareness on resilience [1,12], sensitizing people to dark patterns could help them detect patterns themselves and, thus, reduce the influence of prevalent manipulations. One approach to sensitizing people is to provide real-world examples combined with an explanation, as Brignull does on his website[5]. Another approach involves serious games, as they allow people to expose themselves to manipulative designs without the risk of harm [22].

2.2 Serious Games on Dark Patterns

Work on game-based learning regarding dark patterns is still sparse. "The Dark Pattern Game" is a board game focusing on dark patterns and privacy [24]. It targets adolescent players and can be played in groups of three to five people. Its objective is to set up a mobile phone with a selection of apps while trying to minimize data sharing for privacy reasons. Although playing the game had only limited impact on participants' knowledge about dark patterns, there was a significant influence on behavioral intention, i.e., students with higher prior dark pattern knowledge reported higher intentions to protect their personal data and

[3] While the ACM Ethics Board already considers the term to be problematic, the research community has not agreed on a different name yet. Therefore, we use the term to better connect with the existing literature.

[4] https://www.deceptive.design, last accessed July 17, 2024

[5] https://www.deceptive.design/hall-of-shame, last accessed July 17, 2024

privacy. Our own work intends to further this research through a web-based game prototype on the topic.

In a workshop at CHI'24 [8], two papers presented games in the context of dark patterns[6]: The first game [15] tackles dark patterns in a virtual 3D setting in which a narrator tries to keep players in a location as long as possible. To do this, the narrator uses 3D analogies of dark patterns, like manipulating the light sources in a room to create a prolonged path for the players.

The second game [16] is a decision-making game in which players always choose between two options and try to maximize their score. However, as they play, they are also confronted with several manipulative techniques. For example, players may choose to quickly walk to a low-scoring item or take a longer path to reach a higher-scoring item, which resembles the "Obstruction" dark pattern. With their game, the authors propose a new paradigm to better estimate the deceptive potential of dark patterns. Additionally, aspects like decision time or cognitive load can also be adjusted in this game by altering, e.g., task complexity, level of distraction, or time pressure.

3 Concept and Design

Our game design takes inspiration from serious games on anti-phishing education due to the similarity of subject matter: In both areas, users need to detect and understand manipulation and deception. However, research in anti-phishing games has also shown that detection alone is not sufficient to prepare users [22] since it does not offer much insight into players' decision-making processes and allows for guesswork. Therefore, using classification in addition to detection allows for deeper insights into players' understanding of the manipulation and helps to identify possible misconceptions.

As a first step, we evaluated three game mechanics on three different levels of Bloom's Revised Taxonomy [14] in a preliminary user study: (a) classifying a potential dark pattern, (b) finding and classifying a single dark pattern hidden on a full website, and (c) finding and classifying an unknown number of dark patterns on a website. Our results show that players prefer the latter game mechanic as it closely mimics the real-world situation and was perceived as the most fun and most satisfying game mechanic.

Consequently, our game prototype uses this as its core game mechanic. We use a level-based design with each level themed around a specific type of website (e.g., a hotel booking website, or a smartphone shop). In each level, the player is presented with a partially interactive website that they can search for dark patterns. To select a potential dark pattern, they activate a "magic wand" and click the element (see Fig. 1). This displays a list of dark pattern categories based

[6] All accepted papers to the workshop are available at https://ceur-ws.org/Vol-3720/, last accessed July 17, 2024

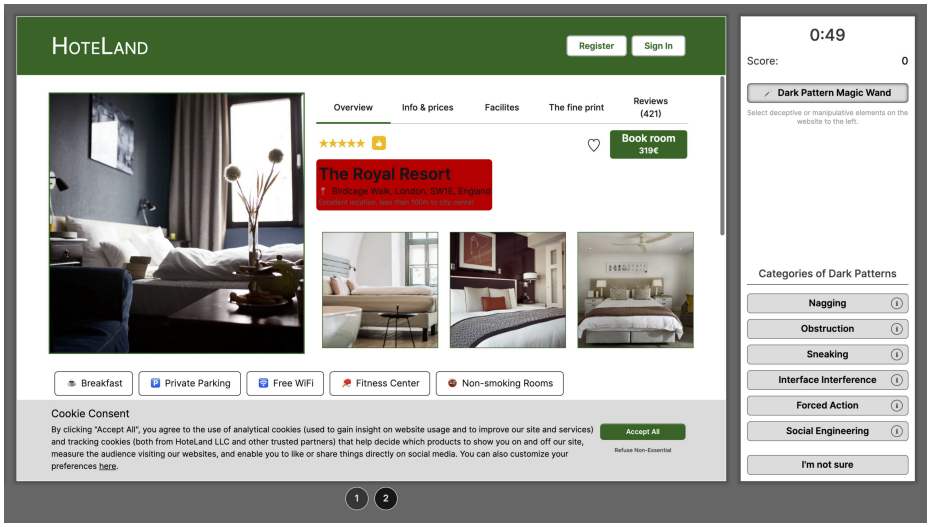

Fig. 1. In our game, players select elements from a website that they believe to be dark patterns. The active selection is highlighted in red. They then classify the element based on the high-level patterns by Gray et al. [9] shown on the right.

on the high-level patterns from Gray et al.'s ontology [9] for classification[7]. If the detection and classification are correct, players receive instant feedback with a short explanation and score for both aspects.

Additionally, we use storytelling elements to increase player engagement: Players assume the role of the newest recruit in the fictional "Dark Pattern Defense Force" and can climb a metaphorical career ladder to higher positions by completing levels and earning more points. The look of the game and its tutorial are themed to match this storytelling.

4 Implementation

The game is implemented as a web application to make it easily accessible and widely available. It is built with React[8] and Next.js[9]. The React framework allows for a very modular design with reusable components and offers a modern state-based implementation, making it easy to maintain and extend. We provide exemplary gameplay in an anonymized OSF repository[10].

[7] When conducting the study, the updated version of the ontology presented at CHI'24 [10] was not available yet. This is why our game still classified *Nagging* as a high-level pattern.

[8] https://react.dev/, last accessed July 21, 2024

[9] https://nextjs.org/, last accessed July 21, 2024

[10] https://osf.io/uq9zf/?view_only=cf79d804cbb24e9689d7ac7d607830ed

Currently, the game offers a tutorial to teach the categories for classification and six levels in three different themes: a hotel booking website, a smartphone shop, and a concert ticket shop. The levels themselves are built in a modular fashion with parameters to generate different versions (e.g., to include a particular dark pattern). For instance, we include a collection of different cookie banner designs that can be added to each level while matching its design.

Since the game itself and the game content (i.e., the levels) are written in the same language, exchanging data between them is straightforward, which enables the core game mechanic: selecting dark pattern elements on the website. Every HTML tag can be augmented with the class "darkPatternContainer" to make it selectable as a possible dark pattern. Additionally, the correct category of dark patterns and links to other elements (e.g., siblings) are modeled using HTML classes. Custom attributes let us specify explanations for each tag (with a list of pre-defined explanations for common dark patterns) and custom scoring information. That way, the level contents remain compatible with basic HTML.

Lastly, we implemented extensive gameplay logging to capture events like users selecting an element, choosing a category, or hovering a tooltip. The resulting log data can be used to analyze gameplay using game learning analytics or to discover user behaviors to improve countermeasures against dark patterns.

5 Evaluation

5.1 Study Design and Procedure

Our user study was designed as a two-part within-subjects experiment with pre-post-test (part 1) and re-test (part 2). Each player completed the tutorial and actual game between the pre- and post-test. The order of levels was randomized. As part of the pre-, post-, and re-test, participants were presented with images of potentially manipulative websites and had to decide if they were manipulative and why. We used the same ten sample websites from Bongard-Blanchy et al. [1] for the pre-test, plus twice an additional five websites in a similar style each for the post- and re-test to check for learning biases. For the re-test, four months later, players first classified the $10 + 5$ potentially manipulative images before playing two levels: one new level and one level from the previous session. No additional tutorial was offered, but explanation tooltips were available throughout the game if players needed to refresh their knowledge on the different types of dark patterns [9].

5.2 Participants

We conducted the study with 22 university students (12 female, 9 male, 1 diverse), all between 20 and 35 years old ($M = 25.0$, $SD = 3.86$). 15 participants had a technical background (computer science, engineering), seven a non-technical background (economics, musicology). Eleven participants were undergraduate students and held a high school diploma as their highest academic

degree, four completed their Bachelor's, and four a Master's degree. Participants reported spending an average of 4.86 hours/day online ($SD = 2.17$). For the re-test, 19 of these students participated again and as such, only the $n = 19$ participants who completed all steps of the study were considered for analysis.

5.3 Results

For performance improvements between pre- and post-test, we checked for learning bias first. We calculated the mean of correct answers (in %) for the pre-test items ($M_{\text{pre}} = .73$, $SD = .16$), the combined post-test items ($M_{\text{post}} = .87$, $SD = .06$), the post-test items that were part of the pre-test ($M_{\text{post-pre}} = .88$, $SD = .1$), and the new post-test items ($M_{\text{post-new}} = .86$, $SD = .11$). The descriptive results show that the improvements in the post-test are similarly higher for the pre-test images and the new images. Furthermore, both are substantially higher than in the pre-test. As such, potential learning bias should be negligible. Similarly, the results for the re-test suggest that the learning bias is negligible too ($M_{\text{re}} = .8$, $SD = .12$; $M_{\text{re-pre}} = .79$, $SD = .13$; $M_{\text{re-new}} = .82$, $SD = .17$).

To determine the influence on the participants' performance as learning effects, we performed paired t-tests with a Bonferroni correction for the correctly identified percentages between the pre-test and post-test and the pre-test and re-test, all on the data of $n = 19$ participants who completed all three tests. For assumption checking, we confirmed a normal distribution for the differences with a Shapiro-Wilk test ($p_{\text{pre-post}} > .48$, $p_{\text{pre-re}} > .54$). The results of the paired t-tests show significant differences both between the pre- and post-test results ($t(19) = -4.89$, $p = .0002$, $df = 18$, Cohen's $d = 1.12$) and between the pre- and re-test results ($t(19) = -4.02$, $p = .0016$, $df = 18$, Cohen's $d = .92$).

On average, participants' performance improved between the pre-test and the post-test by 13.64% ($SD = 12.6$, $CI = [8.03, 19.24]$). The gain in performance was less pronounced the better participants already were in the pre-test. In the re-test, participants' performance improved by 7.19% on average compared to the pre-test ($SD = 7.8$, $CI = [3.43, 10.95]$).

When analyzing gameplay log data, results show that player performance within the game also improved during the second play session (as part of the re-test) compared to the first play session. The average detection rate for a level during the first playthrough was 68.08% ($SD = 7.62$). In the re-test, players detected 13.83% ($SD = 6.48$) more dark patterns for levels that they had encountered four months prior. However, while the detection rate of dark patterns improved during the second playthrough, the correct classification score decreased by 3.44% ($SD = 13.15$). While the improvement in detection was quite uniform (all except one participant detected more dark patterns than in the first study), the successful classification varied greatly between participants, with the majority doing slightly worse and four participants each doing considerably worse and considerably better respectively.

6 Discussion and Limitations

Based on the data collected in our user study, we are able to discuss and reflect on the results, their interpretation and limitations applicable to our work.

6.1 Performance between pre-, post- and re-test

The improvement between the pre- and post-test shows that the game fulfilled one of its core purposes: imparting domain knowledge. It does so in two steps: first by teaching dark patterns terminology, and then by testing that knowledge in an interactive and playful way. The improvements were more pronounced for participants who detected less than half the manipulations during the pre-test, thus indicating its benefits for novices.

Even though participants' performance in the re-test was lower than in the post-test, it was still significantly higher than in the pre-test. This indicates that at least some knowledge was retained over the four months between the tests. However, we would still recommend that players continue to play the game from time to time to refresh their knowledge, similar to practices with regular compliance and safety trainings.

6.2 Performance during gameplay

The improvements between re- and pre-test are also reflected during gameplay. Players were better at detecting dark patterns not just in the replayed level but also in the new level. However, performance improvements in the replayed level were more pronounced. While this is true for the re-test in general, it also extends to recurring instances of dark patterns. Our log data also indicates that players appear more likely to detect further instances or variations of the same type of dark pattern after encountering it once. For instance, the initial instance of a *disguised ad* dark pattern was only detected by 27% of participants. Its second instance was detected and correctly classified by 41%. Additionally, participants were more decisive in their selection and classification of recurring dark patterns. These findings support our goal to teach detection and classification of different dark patterns so that users can remember their characteristics and make informed decisions.

Players were able to detect (and correctly classify) certain dark patterns better than others: Generally, they were best at detecting and classifying dark patterns from the *social engineering* category of the ontology [9], such as *low stock messages* (e.g., "only one room left") and *countdown timer*. Similarly, *false hierarchy* dark patterns were detected in 90% of all occurrences. Dark patterns from the *sneaking* category, which hide information, were detected least often. Our findings match the results of [1] on people's ability to detect certain dark patterns better than others.

Participants often mistook either *sneaking* or *forced action* for obstruction (less often vice versa) because the hidden information or required action was considered an unnecessary step and thus made the task unnecessarily harder.

Participants also frequently classified *nagging* as *forced action*. Interestingly, in the recent update of the ontology [10], *nagging* was re-classified as a sub-category of *forced action*. This seems to match our participants' mental model.

For a more comprehensive performance evaluation, we need to consider socio-demographical participant data. All participants were students in higher education or with a university degree. In the first part of our study, there were 15 participants with a technical and 7 with a non-technical background. Their respective detection and classification rates were very similar, suggesting that affinity to technology does not influence one's ability to resist dark patterns.

Furthermore, we analyzed performance in correlation to time spent online. Participants who spend less time online (1-3 hours per day) detected slightly more dark patterns, whereas participants who spend much time online (6 or more hours per day) were slightly better in classifying them. However, these differences were minimal, and the rather small sample size does not allow us to draw general conclusions from this.

6.3 Limitations

One limitation derives from our implementation of how the game handles multiple dark patterns. While it is possible for one element to contain multiple dark patterns (in this case selecting any one would be considered correct), a problem arises with nested dark patterns (i.e., an outer container being one dark pattern and an inner container being another one). During the study, it became apparent that participants often *meant* the right thing, but their selection was either too broad or too narrow.

Another limitation is that our participant sample is small and only consisted of university students. The perception of dark patterns might vary by age [1]. As such, generalizability of our results to other demographics is limited and requires further validation, e.g., in a larger study with a more diverse user group.

7 Conclusion and Future Work

In this paper, we presented the design, implementation, and first evaluation of a serious game about detecting and classifying dark patterns in sample websites. Through an empirical user study with pre-, post-, and re-test and two gameplay sessions, we evaluated possible learning effects by focusing on participants' performance improvements. Our results show that participants performed significantly better in detecting dark patterns after playing the game, both in the immediate post-test as well as in a re-test four months later. Furthermore, participants showed differences in classifying different types of dark patterns; e.g., *social engineering* was correctly classified more often than *sneaking*. This may allow further reasoning on participants' misconceptions and difficulties.

For future work, we plan to make our game publicly available to validate our study results with a larger and more diverse participant sample. Using detailed

game log data, we want to extract participants' mental models and possible misconceptions for specific instances of dark patterns, and use this data to improve countermeasures against dark patterns.

References

1. Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., Lenzini, G.: "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. pp. 763–776. ACM, Virtual Event, USA (2021). https://doi.org/10.1145/3461778.3462086
2. Bösch, C., Erb, B., Kargl, F., Kopp, H., Pfattheicher, S.: Tales from the dark side: Privacy dark strategies and privacy dark patterns. Privacy Enhancing Technologies **2016**(4), 237–254 (2016). https://doi.org/10.1515/popets-2016-0038
3. Brignull, H.: Deceptive Patterns: Exposing the Tricks Tech Companies Use to Control You. Testimonium Limited (2023)
4. Conti, G., Sobiesk, E.: Malicious interface design: exploiting the user. In: 19th Int. Conf. on World Wide Web. pp. 271–280. WWW '10, ACM, New York (2010). https://doi.org/10.1145/1772690.1772719
5. Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., Bacchelli, A.: 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. p. 1–14. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376600
6. Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., Bacchelli, A.: Ui dark patterns and where to find them: A study on mobile applications and user perception. In: CHI Conference on Human Factors in Computing Systems. p. 1–14. CHI '20, ACM, New York (2020). https://doi.org/10.1145/3313831.3376600
7. European Commission and Directorate-General for Justice and Consumers, Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F., Liva, G., Lechardoy, L., Rodríguez de las Heras Ballell, T.: Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation (final report). Publications Office of the European Union, Brussels (2022). https://doi.org/10.2838/859030
8. Gray, C.M., Gunawan, J.T., Schäfer, R., Bielova, N., Sanchez Chamorro, L., Seaborn, K., Mildner, T., Sandhaus, H.: Mobilizing research and regulatory action on dark patterns and deceptive design practices. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems. CHI EA '24, ACM, New York (2024). https://doi.org/10.1145/3613905.3636310
9. Gray, C.M., Santos, C., Bielova, N.: Towards a preliminary ontology of dark patterns knowledge. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. CHI EA '23, ACM, New York (2023). https://doi.org/10.1145/3544549.3585676
10. Gray, C.M., Santos, C.T., Bielova, N., Mildner, T.: An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In: CHI Conference on Human Factors in Computing Systems. CHI '24, ACM, New York (2024). https://doi.org/10.1145/3613904.3642436
11. Graßl, P., Schraffenberger, H., Zuiderveen Borgesius, F., Buijzen, M.: Dark and bright patterns in cookie consent requests. Journal of Digital Social Research **3**(1), 1–38 (2021). https://doi.org/10.33621/jdsr.v3i1.54

12. Grossklags, J., Acquisti, A.: When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In: WEIS. Citeseer (2007)
13. Gunawan, J., Pradeep, A., Choffnes, D., Hartzog, W., Wilson, C.: A comparative study of dark patterns across web and mobile modalities. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (Oct 2021). https://doi.org/10.1145/3479521
14. Krathwohl, D.R.: A revision of bloom's taxonomy: An overview. Theory into practice **41**(4), 212–218 (2002)
15. Kronhardt, K., Gerken, J.: Start playing around-serious & persuasive games as a viable counter-measure against deceptive patterns? In: Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices Workshop at CHI '24, Honolulu (2024), https://ceur-ws.org/Vol-3720/paper7.pdf
16. Löschner, D.M., Pannasch, S.: Measuring the deceptive potential of design patterns: A decision-making game. In: Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices Workshop at CHI '24, Honolulu (2024), https://ceur-ws.org/Vol-3720/paper9.pdf
17. Lu, Y., Zhang, C., Yang, Y., Yao, Y., Li, T.J.J.: From awareness to action: Exploring end-user empowerment interventions for dark patterns in ux. Proc. ACM Hum.-Comput. Interact. **8**(CSCW) (2024). https://doi.org/10.1145/3637336
18. Mathur, A., Acar, G., Friedman, M.J., Lucherini, E., Mayer, J., Chetty, M., Narayanan, A.: Dark patterns at scale: Findings from a crawl of 11k shopping websites. Proc. ACM Hum.-Comput. Interact. **3**(CSCW) (2019). https://doi.org/10.1145/3359183
19. Mildner, T., Freye, M., Savino, G.L., Doyle, P.R., Cowan, B.R., Malaka, R.: Defending against the dark arts: Recognising dark patterns in social media. In: 2023 ACM Designing Interactive Systems Conference. p. 2362–2374. DIS '23, ACM, New York (2023). https://doi.org/10.1145/3563657.3595964
20. Potel-Saville, M., Da Rocha, M.: From dark patterns to fair patterns? usable taxonomy to contribute solving the issue with countermeasures. In: Privacy Technologies and Policy. pp. 145–165. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-61089-9_7
21. Roepke, R., Schroeder, U.: The Problem with Teaching Defence against the Dark Arts: A Review of Game-based Learning Applications and Serious Games for Cyber Security Education. In: Lane, H., Zvacek, S., Uhomoibhi, J. (eds.) 11th Int. Conf. on Computer Supported Education. pp. 58–66. CSEDU '19, SciTePress, Heraklion, Greece (2019). https://doi.org/10.5220/0007706100580066
22. Röpke, R.: Extending Game-Based Anti-Phishing Education Using Personalization: Design and Implementation of a Framework for Personalized Learning Game Content in Anti-Phishing Learning Games. Doctoral Thesis, RWTH Aachen University, Aachen (2023). https://doi.org/10.18154/RWTH-2023-04991
23. Schäfer, R., Preuschoff, P.M., Röpke, R., Sahabi, S., Borchers, J.: Fighting malicious designs: Towards visual countermeasures against dark patterns. In: CHI Conference on Human Factors in Computing Systems. CHI '24, ACM, New York (2024). https://doi.org/10.1145/3613904.3642661
24. Tjostheim, I., Ayres-Pereira, V., Wales, C., Manna, A., Egenfeldt-Nielsen, S.: Dark Pattern: A Serious Game for Learning About the Dangers of Sharing Data. ECGBL **16**(1), 774–783 (2022). https://doi.org/10.34190/ecgbl.16.1.872