Chair for Computer Science 10 (Media Computing and Human-Computer Interaction)

RWTH AACHEN UNIVERSITY

Exploring Game Mechanics for the Design of Dark Pattern Learning Games

Master's Thesis
submitted to the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Kevin Fiedler

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr.-Ing. Ulrik Schroeder

Registration date: 20.11.2023
Submission date: 21.03.2024

Eidesstattliche Versicherung
Statutory Declaration in Lieu of an Oath

___________________________ ___________________________

Name, Vorname/Last Name, First Name

Matrikelnummer (freiwillige Angabe)
Matriculation No. (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel

I hereby declare in lieu of an oath that I have completed the present paper/Bachelor thesis/Master thesis* entitled

____

____

____

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

independently and without illegitimate assistance from third parties (such as academic ghostwriters). I have used no other than the specified sources and aids. In case that the thesis is additionally submitted in an electronic format, I declare that the written and electronic versions are fully identical. The thesis has not been submitted to any examination body in this, or similar, form.

___________________________ ___________________________

Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen

*Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

Para. 156 StGB (German Criminal Code): False Statutory Declarations

Whoever before a public authority competent to administer statutory declarations falsely makes such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment not exceeding three years or a fine.

§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

Para. 161 StGB (German Criminal Code): False Statutory Declarations Due to Negligence

(1) If a person commits one of the offences listed in sections 154 through 156 negligently the penalty shall be imprisonment not exceeding one year or a fine.

(2) The offender shall be exempt from liability if he or she corrects their false testimony in time. The provisions of section 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:

I have read and understood the above official notification:

___________________________ ___________________________

Ort, Datum/City, Date

Unterschrift/Signature

Contents

List of Figures

List of Tables

Abstract

Dark patterns are manipulative design elements that try to deceive or coerce the user into doing something that is in the best interest of the website or app and not necessarily of the user. They have become somewhat ubiquitous throughout the internet. While there are countermeasures from a regulatory direction (such as the GDPR in the European Union) and active research going into technical solutions against dark patterns, this thesis explores another direction: educating the user to better detect dark patterns and gain confidence in doing so. For this purpose, we explore a learning game against dark patterns. In a preliminary user study, we evaluate three game mechanics that combine finding and classifying dark patterns regarding their suitability and user preference. The results show that users prefer a game mechanic that is closest to the real world, where they have to find multiple dark patterns hidden on a website. We proceed by building a learning game prototype artifact based on the results of this preliminary study. We evaluate this prototype in a second user study and measure its learning effects in a pre-test/post-test design. The results show that there is a significant increase in performance in detecting dark patterns after playing the game. It further indicates that players gain more confidence when encountering dark patterns, especially recurring ones. We describe some opportunities to utilize the game to create a large dataset on susceptibility to certain dark patterns.

Überblick

Dark Patterns sind manipulative Design Elemente, die versuchen den Besucher einer Website oder App zu täuschen oder dazu zu bewegen etwas zu tun, was im Interesse des Betreibers und nicht notwendigerweise im Interesse des Besuchers ist. Solche Dark Patterns sind relativ weit verbreitet im Internet. Maßnahmen gegen Dark Patterns kommen einerseits von Aufsichtsbehörden (wie z.B. mit der GDPR in der Europäischen Union), doch auch seitens der Forschung, um technische Lösungen gegen Dark Patterns zu finden. In dieser Arbeit erkunden wir eine weitere Richtung: Die Aufklärung von Nutzern, damit sie Dark Patterns besser erkennen und selbstbewusster umgehen können. Wir überprüfen, ob ein Lernspiel dafür geeignet ist. In einer ersten Nutzerstudie testen wir drei verschiedene Spielmechaniken, die das Finden und Klassifizieren von Dark Patterns kombinieren. Wir prüfen deren Eignung und Nutzer-Präferenz. Unsere Ergebnisse zeigen, dass Nutzer die Spielmechanik bevorzugen, die am nähesten die Realität widerspiegelt und bei der mehrere Dark Patterns auf einer Website versteckt sind, die sie finden müssen. Basierend auf diesem Ergebnis entwickeln wir einen Prototyp für ein Lernspiel. Wir testen diesen Prototyp in einer zweiten Nutzerstudie, um dessen Lerneffekt zu messen. Dazu verwenden wir ein pre-test/post-test Studiendesign. Die Ergebnisse zeigen, dass durch das Spielen unseres Prototyps ein signifikanter Lerneffekt zu erkennen ist. Darüber hinaus deuten die Ergebnisse darauf hin, dass Nutzer durch das Spiel mehr Selbstbewusstsein gewinnen. Im Weiteren beschreiben wir das Potential des Spiels, damit große Datenmengen zu sammeln, die zur weiteren Forschung, gerade was Anfälligkeiten gegenüber spezifischen Dark Patterns angeht, verwendet werden können.

Acknowledgements

First of all, I want to thank Prof. Dr. Jan Borchers and Prof. Dr.-Ing. Ulrik Schroeder for examining this thesis.

Secondly, I would like to thank my thesis advisers René Schäfer and René Röpke. Your continuous support, feedback and motivation was invaluable and made my work on this thesis a joy. You have certainly taught me a lot over the last couple of month and I really appreciate all the time and energy that you put into it.

This totally extends to all of you at i10. I could always count on each and everyone of you no matter the question or problem. Whether it was an introduction to MAXQDA or a crash course in R or just general advice, I feel like I didn't even have to ask. So thanks to all of you.

Furthermore, thank you to all my friends and family who supported me during this time in one way or another.

Finally, a big thanks to everybody who generously sacrificed their time to participate in my user studies. Without your help, this thesis would not have been possible.

Conventions

Throughout this thesis we use the following conventions.

Text conventions

Definitions of technical terms or short excursus are set off in colored boxes.

EXCURSUS:
Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:
Excursus

Source code and implementation symbols are written in typewriter-style text.

`myClass`

The whole thesis is written in American English. We use the plural form for the first person. For unidentified third persons, we use the pronouns they/their.

Download links are set off in colored boxes.

File: myFile[a]

[a]http://hci.rwth-aachen.de/public/folder/file_number.file

Chapter 1

Introduction

The five most visited websites in the world are Google Search, YouTube, Facebook, Instagram, and X, which have combined more than 149 billion visits per month [Similarweb, 2023]. There is a good chance that many people will visit at least some of those websites fairly regularly. For instance, Urman and Makhortykh [2023] showed that people in Germany and Switzerland use Google Search on average 8.8 times per day. Similarly, a survey on social media usage by Auxier and Anderson [2021] showed that about 70% of adults in the U.S. visit Facebook at least once a day.

These websites are not only the most visited websites. They and their respective parent companies are also prominently listed by Brignull et al. [2023] on their "Hall of Shame" section of their website https://www.deceptive.design/[1], where the authors showcase and publicly shame websites that use dark patterns.

Many popular websites use dark patterns.

Dark Patterns:
Dark patterns (or deceptive patterns) are often malicious elements or tricks used on websites and in apps to make the users do something they did not intend to and which may be against their best interest. [Brignull et al., 2023]

Definition: *Dark Patterns*

[1] https://www.deceptive.design/hall-of-shame [Accessed: Dec. 2, 2023]

Dark patterns are a common occurrence throughout the internet.

However, Brignull et al.'s "Hall of Shame" is not limited to just popular websites or big corporations. On the contrary, they present over 400 examples of dark patterns throughout the internet. This shows the prevalence of dark patterns. For instance, Mathur et al. [2019] showed in the context of shopping websites, that at least 11% of their inspected websites (~11K) used dark patterns. Furthermore, popular shopping websites were more likely to use dark patterns than less popular ones. Unfortunately from an end-user perspective, companies have a monetary incentive to use dark patterns [Maier and Harr, 2020]. The usage of mild (i.e., subtle) and even aggressive dark patterns is effective in steering and manipulating customers [Luguri and Strahilevitz, 2021] while at the same time risking to upset said customers and harming the brand [Voigt et al., 2021].

Falling for dark patterns can have serious consequences.

The consequences of being tricked by a dark pattern can vary in severity. At best, they are just an annoyance and cause frustration [Conti and Sobiesk, 2010, Bhoot et al., 2021]. However, many dark patterns are designed to coerce people into sharing more personal data than they intended [Gunawan et al., 2022, Brignull, 2023, Norwegian Consumer Council, 2018]. There are also many reported cases where dark patterns have led to financial harm [Brignull, 2023] and they have consequently been investigated by lawmakers (e.g., by the Federal Trade Commision (FTC)[2]).

Visual countermeasures are possible but have limitations.

Because dark patterns are potentially harmful, researchers are already exploring countermeasures. Mathur et al. [2019] proposed visual highlighting of dark patterns. Schäfer et al. [2023] further investigated which visual countermeasures, such as *highlighting with explanation* or *hiding*, are preferred by users for which kind of dark pattern. There are also some artifact browser extensions that highlight dark patterns on websites (e.g., insite[3] or dapde[4]). However, reliably detecting dark patterns is a complex problem (the two examples of *insite* and *dapde* use machine learning and regular expressions, respectively) and work mostly

[2] https://www.ftc.gov/news-events/news/press-releases/2021/07/lendingclub-agrees-pay-18-million-settle-ftc-charges/ [Accessed: Dec. 2, 2023]
[3] https://devpost.com/software/insite-qfpjcd [Accessed: Dec. 2, 2023
[4] https://dapde.de/de/ [Accessed: Dec. 2, 2023]

text-based or on a limited set of dark patterns [Mansur et al., 2023]. For some types of dark patterns, it may not even be possible to detect them with an automatic approach [Curley et al., 2021]. Even if there is some automated approach to detect dark patterns, it is still possible that dark pattern detection browser extensions will face similar challenges as ad blockers and have to constantly evolve in order to keep up with the newest changes from advertisers [Storey et al., 2017].

We aim to educate people on dark patterns.

Apart from the *regulatory* interventions mentioned before (e.g., by the European Union or the FTC) and the *technical* countermeasures just described, Bongard-Blanchy et al. [2021] propose another user-directed intervention measure focusing on *education*. They suggest something like a *"spot the dark pattern"*-game. In this thesis, we will explore whether learning games are suitable to educate people against dark patterns.

Definition: *Learning Games*

LEARNING GAMES:
Serious games are games that have at least one additional goal besides entertainment. Learning games describe a type of serious games that focus on learning for educational purposes. [Dörner et al., 2016, Plass et al., 2015]

We want people to be more confident when encountering dark patterns.

The idea is that people don't have to rely on a browser extension or other technical or regulatory solution but instead gain the required knowledge and confidence to detect and properly circumnavigate dark patterns on their own. This is especially relevant because even being aware of dark patterns does not necessarily mean people can resist them [Bongard-Blanchy et al., 2021]. Yet people performed better in detecting dark patterns once they had been informed about them [Di Geronimo et al., 2020].

Games are a well-suited method for learning.

Games have been shown to be a very effective tool for learning if they are designed for a specific problem or to teach a certain skill [Griffiths, 2002]. They are well suited to reach a large demographic and offer opportunities for measurements and research. They can be more motivating and engaging than traditional educational methods [Tang et al., 2009].

Learning games allow people to engage with and learn about dark patterns in a consequence-free environment.

Learning about dark patterns within a game has additional advantages: While a learning game has certain rules and presents some sort of artificial conflict, it also provides a risk-free, consequence-free virtual environment where players can freely experiment and explore dark patterns within the confinements of the learning game and any mistakes do not have real-world ramifications [Röpke, 2023]. On the contrary, making mistakes within a learning game can even be desirable as it can be a necessary step towards learning [Plass et al., 2015].

1.1 Motivation & Aim

Research has focused on technical solutions against dark patterns.

We have already outlined the different intervention measures against dark patterns proposed by Bongard-Blanchy et al. [2021]. *Regulatory* interventions, like the General Data Protection Regulation (GDPR) in the European Union, only had limited success[5]. *Technical* and *design* intervention measures are both being currently researched (e.g., by Schäfer et al. [2023]) and, though they have their own limits and challenges, being applied (e.g., with *dapde* and *insite*).

There are games about dark patterns with limited scope.

However, there is less research going into the *educational* direction of intervention measures. There is a browser game called Cookie Consent Speed.Run[6]. While it is fun to play, it is more a reductio ad absurdum of how convoluted some cookie consents have become rather than an actual learning experience[7]. Another game is called *The Dark Pattern Game* [Tjøstheim et al., 2022] and focuses on preserving privacy and the dangers of sharing too much data.

Yet, to the best of our knowledge, there is no dark pattern learning game that focuses on the general detection of dark patterns (i.e., the suggested *"spot the dark pattern"*-game by Bongard-Blanchy et al.).

[5] https://netzpolitik.org/2022/manipulative-cookie-banner-viele-beschwerden-wenige-strafen/ [Accessed: Mar. 4, 2024]
[6] https://cookieconsentspeed.run [Accessed: Mar. 4, 2024]
[7] It is still great to raise awareness for the issue.

With this thesis, we aim to fill this gap. The focus is to explore which game mechanics are suitable for such a dark pattern learning game and match our learning goals. Those are the improved ability to generally detect dark patterns and increased confidence in doing so.

Furthermore, we aim to evaluate the effectiveness of such a dark pattern learning game by measuring its immediate and long-term learning effects.

1.2 Outline

In Chapter 2 "Related Work", we first introduce the domain of dark patterns for our own learning game by looking into and comparing existing taxonomies. We explore other domains where learning games have already been successfully applied to and provide an overview of the taxonomies of educational goals.

In Chapter 3 "Exploring Game Mechanics", we present our preliminary study where we investigate which game mechanics are suitable for dark pattern learning games. We test three game mechanics and evaluate them to get a user preference and an indication of their suitability.

In Chapter 4 "A Dark Pattern Learning Game Prototype", we describe the design and implementation of the dark pattern learning game prototype artifact that is based on the results of our preliminary study.

We evaluate the learning game prototype in Chapter 5 "Dark Pattern Learning Game User Study & Evaluation" with a second user study to measure the learning effects and benefits of the game.

We conclude our research in Chapter 6 "Summary and Future Work". We summarize our findings along with their limitations and suggest some future research opportunities.

Chapter 2

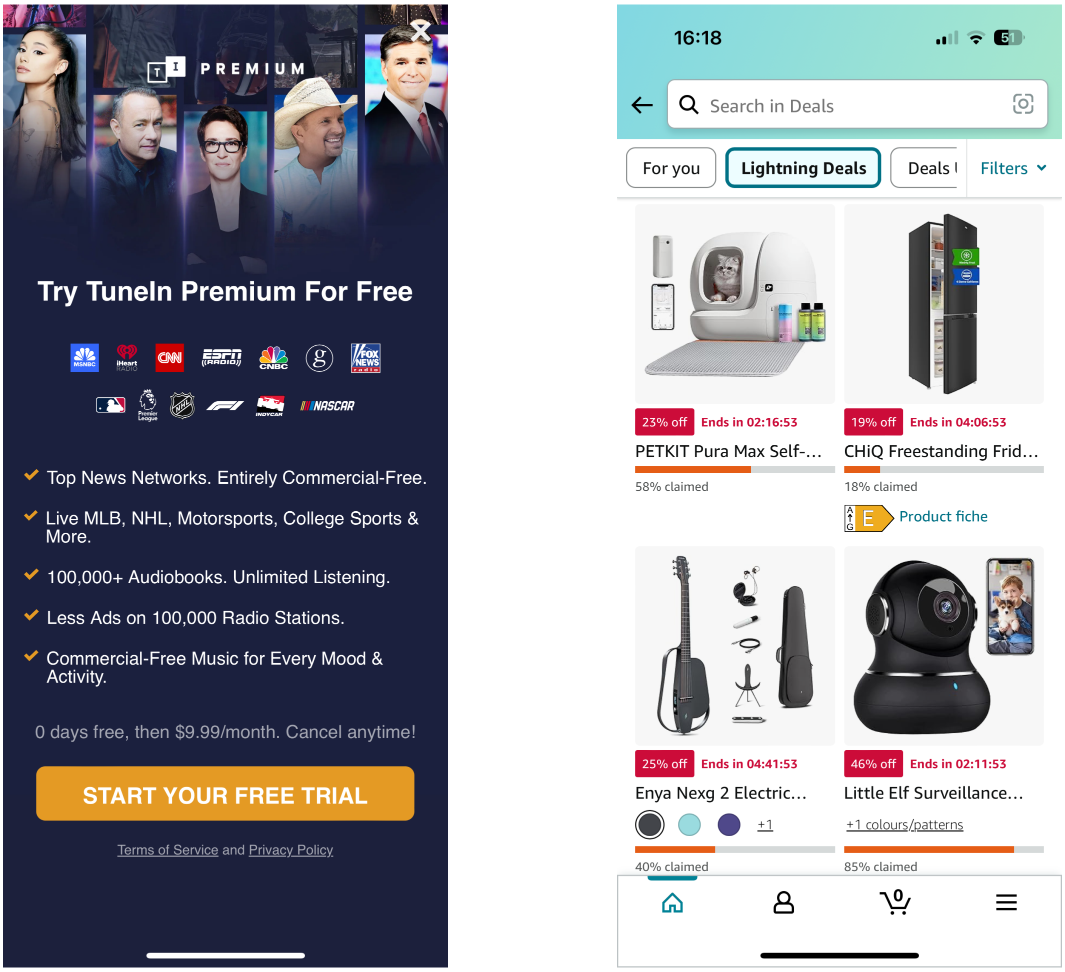
Related Work

In this chapter, we look at related work from two domains. Firstly, in Chapter 2.1, we explore the area of dark patterns. Specifically, we look into existing taxonomies and categorization of dark patterns. We further discuss the prevalence of dark patterns and their effect on users. Secondly, in Chapter 2.2, we look into learning games and the theory of educational objectives and present learning games from other domains.

2.1 Dark Patterns

There is not one clear definition of *"dark patterns"*. Its exact meaning and categories vary based on the context and domain [European Commission et al., 2022]. There are more generic categories for dark patterns (e.g., by Gray et al. [2018]) and domain-specific dark patterns (e.g., by Mathur et al. [2019] or Mildner et al. [2023]). However, a *dark pattern* generally refers to malicious user interface elements that are meant to manipulate or deceive users [Brignull et al., 2023, Gray et al., 2018]. Figure 2.1 shows two examples of dark patterns: Figure 2.1a uses *interface interference* to make the subscription button the obvious choice and hide the close button. Figure 2.1b attempts to rush the user into making a purchase by using *scarcity* and *urgency*.

Dark patterns are malicious design elements to manipulate or deceive users.



(a) Screenshot from the initial screen when a user opens the TuneIn app

(b) Screenshot from the Lightning Deals section in the Amazon app.

Figure 2.1: Two examples of dark patterns in the real world: (a) shows an example of *interface interference* [Gray et al., 2018], putting a strong emphasis on the subscription button while almost hiding the close button in the top right corner. (b) shows an example of *urgency* (limited time deal) and *scarcity* (limited stock), two dark patterns that Mathur et al. [2019] found in the context of shopping websites.

Dark patterns describe established solutions to exploit or deceive users.

The term *"pattern"* generally refers to a successful solution to a recurring problem [Borchers, 2000]. It originated from the domain of architecture [Alexander, 1977], where the authors listed established designs for cities and buildings in a hierarchy. It was first adapted for computer science and software development [Beck, 1987], and later for the domain of HCI [Borchers, 2000]. However, while *design patterns* generally describe successful good practices, a *dark pattern* describes an established solution to exploit or deceive users [Bösch et al., 2016].

2.1.1 Existing Taxonomies

Early work in the area of dark patterns and building a taxonomy for them dates back to 2010. Brignull [2010] first used the term *"dark pattern"*, which they defined as "user interfaces that have been designed to trick users into doing things they wouldn't otherwise have done, or to prevent users from doing the things they want to do". They collected and showcased websites that used dark patterns and identified and categorized recurring dark patterns. Their original taxonomy consisted of eleven types of specific dark patterns such as *bait and switch*, *sneak into basket*, *privacy zuckering*, or *roach motel*. Though their work is still relevant today, it has since then been updated and expanded upon. In their newest version, Brignull et al. [2023] list 16 types of dark patterns (which they rebranded *"deceptive patterns"* in 2022), including those added by Gray et al. [2018] and Mathur et al. [2019].

Brignull [2010] provided an early taxonomy for dark patterns.

In 2010, Conti and Sobiesk [2010] provided another taxonomy independent of the works of Brignull. They analyzed websites, desktop software, and interfaces in a 12-month study to find malicious interface design techniques. They combined data that were gathered using an automated approach and data that was manually sought out, as well as data collected from group discussions at a hacker conference (Hackers of Planet Earth Conference). Their taxonomy describes eleven high-level techniques of how malicious interfaces work, such as *confusion*, *distraction*, *exploiting errors*, and *obfuscation*.

Conti and Sobiesk [2010] created a taxonomy that describes high-level malicious interface design techniques.

Gray et al. [2018] analyzed the topic of dark patterns from an ethical perspective. They considered the strategic decisions that designers made for each dark pattern to come up with five high-level categories to place existing dark patterns into: *nagging*, *obstruction*, *sneaking*, *interface interference*, and *forced action*. These categories are shown in

Gray et al. [2018] expanded on Brignulls taxonomy and introduced five superordinate categories for dark patterns.

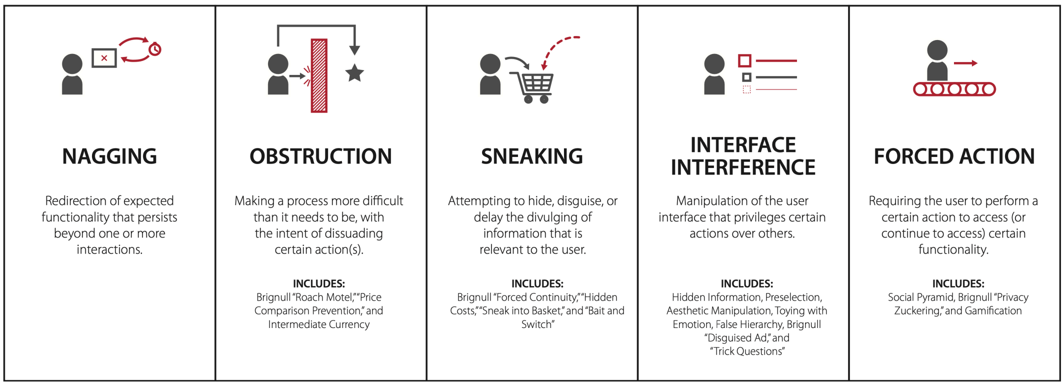


Figure 2.2: The five high-level categories by Gray et al. are *nagging*, *obstruction*, *sneaking*, *interface interference*, and *forced action*. Additionally, they indicate what specific dark patterns fall into each category. Figure taken from [Gray et al., 2018]

Figure 2.2. Gray et al. list what types of dark patterns by Brignull fall into each category, but they also expanded the existing list of dark patterns. For instance, they included *nagging* as a completely new category or added *preselection* as a new type of dark pattern.

Gray et al. [2023] provided a draft for an ontology for dark patterns.

Additionally, Gray et al. [2023] provided a draft for an ontology of dark patterns as a foundation work for future research. They aggregated data from their previous work as well as from other sources such as Brignull et al. [2023] or Mathur et al. [2019] and from regulatory reports (e.g., the European Union Commission or the FTC). They then divided these data into a three-tier hierarchy: high-level patterns that describe general strategies of how to manipulate people, in-between *meso-level* patterns that describe an angle of attack, and low-level patterns that describe a specific means of execution. Most noteworthy, they added a sixth high-level category, *"social engineering"*, where many of the domain-specific dark patterns by Mathur et al. fall into.

There is an update to the ontology.

A paper with updates to the ontology has since then been accepted at CHI 2024[1]. In this version, Gray et al. [2024] updated the ontology to five high-level patterns (moving *nagging* to the meso-level of *forced action*), 25 meso-level

[1]The paper has already been published on the website of the author.

patterns, and 35 low-level patterns[2]. They provide further details and guidance on how to integrate future dark patterns or dark patterns from other domains into the existing ontology.

Whereas the previous works attempt to categorize and classify dark patterns in general, there is further research looking into dark patterns in specific domains. We present three domains and their specific taxonomies that cover a wide range of popular use cases: shopping, social media, and gaming.

Further research investigates specific domains for dark patterns.

Mathur et al. [2019] used an automated approach to collect and classify dark patterns on ~11K shopping sites. They discovered 1818 dark patterns on over 1254 websites (~11.1% of all investigated websites). They identified 15 different types of dark patterns that they placed in 7 superordinate categories: *sneaking*, *urgency*, *misdirection*, *social proof*, *scarcity*, *obstruction*, and *forced action*. For each dark pattern, they named the cognitive bias that the dark pattern exploits (*anchoring effect*, *bandwagon effect*, *default effect*, *framing effect*, *scarcity bias*, and *sunk cost fallacy*). Furthermore, they provide higher-level attributes or characteristics to describe dark patterns: *asymmetric*, *covert*, *deceptive*, *hides information*, and *restrictive*. They added *disparate treatment* (disadvantage one group of users) as a sixth attribute in a follow-up work [Mathur et al., 2021].

Mathur et al. [2019] classified dark patterns in the context of shopping websites into seven categories.

They classified dark patterns based on six characteristics and the cognitive bias they exploit.

In the same way that Mathur et al. looked specifically into shopping websites, Mildner et al. [2023] investigated dark patterns in the context of social media (Facebook, Instagram, TikTok, and Twitter). There, they found 44 instances of dark patterns that they coded into two high-level strategies and five social-media-specific dark patterns: The first strategy is *engaging* the user that they are entertained longer and thusly interact with the service for longer. This includes the dark patterns *interactive hooks* (e.g., addictive design or gamification) and *social brokering* (e.g., social connections or reappearing popular content). The second strategy is *governing* the user, which includes *decision uncertainty*,

Mildner et al. [2023] looked into dark patterns in the context of social media.

[2]Because this paper was only published recently, we use the original draft of the ontology throughout most of this thesis. In the end, we will draw conclusions on how these recent updates match our results.

labyrinthine navigation, and *redirective conditions*. They also characterized each of their dark patterns using the same characteristics as Mathur et al. did, thus expanding on this taxonomy. However, the authors also note that dark patterns in social media are less effective because their goal is to engage the user for longer. Therefore, they need to satisfy their users rather than aggravate them, which is why the two strategies have to be carefully balanced.

Dark patterns in social media are less harmful.

While previous domains looked into dark patterns in the context of websites or apps, Zagal et al. [2013] looked into dark patterns in the context of gaming by analyzing strategies of game designers, observation, and player reactions. They discovered seven game-specific dark patterns in three categories: Firstly, *temporal* dark patterns that are designed to waste the players' time with *grinding* or *playing by appointment*. Secondly, *monetary* dark patterns so that the player spends more money. This includes *pay to skip* (often combined with *grinding*), *pre-delivered content* (i.e., paying to unlock all content shipped with the game) and *monetized rivalries* (also known as "pay-to-win"). Finally, there are *social capital-based* dark patterns that exploit the social standing component of games. There are *social pyramid schemes* (inviting friends to the game) and *impersonation* (i.e., making it seem as if a real player performed an action they didn't do). However, the authors also note that not every occurrence of these elements is necessarily a dark pattern. *Grinding* or *playing by appointment* can also be part of regular gameplay (e.g., for optional goals or rare achievements[3]).

Zagal et al. [2013] found seven dark patterns in three categories in the context of games.

2.1.2 Prevalence and Susceptibility

Some of the previous papers have already given an indication of the prevalence of dark patterns. Mathur et al. [2019] found dark patterns on at least 11.1% of shopping websites. However, this is just a lower bound since their automated crawler approach could only detect text-based dark patterns. Additionally, more popular shopping websites were more likely to use dark patterns.

[3] Zagal et al. name the respawn timer in *World of Warcraft* as an example of *playing by appointment* that is not a dark pattern.

The European Commission issued a report on dark patterns [European Commission et al., 2022] where they also investigated the prevalence of dark patterns. They compiled a list of the 30 most visited websites and 30 most used apps across the European Union, plus 15 additional popular national websites. Their results show that 73 of the 75 investigated websites and apps (97%) used at least one form of dark pattern. It further indicates that specific dark patterns are far more common than others: *preselection*, *hidden information / false hierarchy*, *nagging*, *roach motel*, and *forced registration* were the most five common dark patterns while *bait and switch*, *confirmshaming*, *price comparison prevention*, *sneak into basket*, and *trick questions* were the five least common ones. There is also no real difference in dark pattern use based on the category of the website or app.

A report on dark patterns in the EU showed that 97% of investigated websites and apps use dark patterns.

Di Geronimo et al. [2020] looked specifically at the prevalence of dark patterns in mobile apps. They looked at the top 30 trending apps in the eight categories of the Google PlayStore. In order to find certain dark patterns, the authors did not only analyze screenshots but interacted with each app performing tasks for 10 minutes and then analyzed the recordings. They found 1787 dark patterns in the 240 inspected apps, with 95% of them containing at least one dark pattern. Furthermore, almost half of the apps contained seven or more dark patterns. Similar to the EU report [European Commission et al., 2022], the most common dark patterns were *nagging*, *false hierarchy*, and *preselection*. Di Geronimo et al. performed a second study with 589 participants, where they investigated how well people detect dark patterns in apps. The results show that the majority of users are unable to detect the dark patterns or are unsure about them. Some participants stated that some dark patterns are so ubiquitous and common that they have become part of the normal interaction with apps.

Di Geronimo et al. [2020] found dark patterns in 95% of popular smartphone apps.

Some dark patterns are so common that people don't recognize them as such.

However, Bhoot et al. [2021] present a somewhat contradictory result based on the results of an online questionnaire (Figure 2.3). There, they tested with twelve dark patterns if people recognized them. The results varied based on the individual dark patterns, with *forced action* being recognized most often and *roach motel* being recognized least often. Bhoot et al. assume that this might be based on

Bhoot et al. [2021] investigated which dark patterns users are more susceptible to and how it affects their perception of the website.

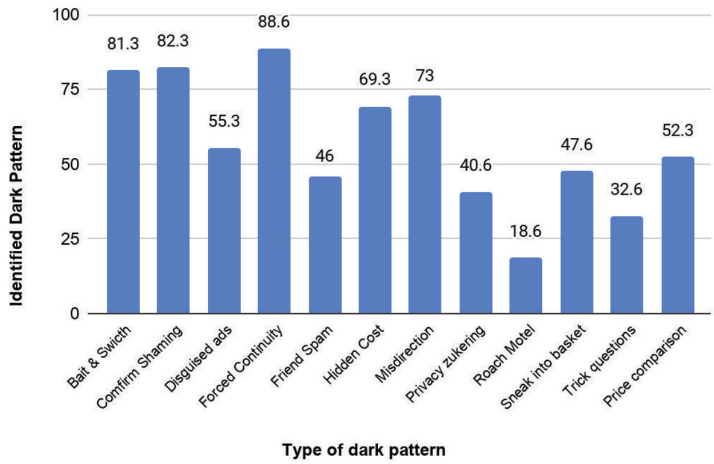


Figure 2.3: The results of a questionnaire by Bhoot et al. [2021] show how well users recognized different dark patterns. *Forced continuity* was recognized most often, *roach motel* least often. Figure taken from [Bhoot et al., 2021]

familiarity and frequent occurrences of certain dark patterns. They further investigated how the usage of dark patterns affects user perception of the website: users generally get frustrated when they encounter a dark pattern, and the usage of dark patterns reduces the trust in the website. This matches the results of Voigt et al. [2021] that using dark patterns increases annoyance and hurts brand trust. Despite that, people tolerate the use of dark patterns by large companies and services they depend on because the benefits that these services offer often outweigh the presence and manipulation of dark patterns [Maier and Harr, 2020].

Dark patterns annoy users, but the benefits that services offer outweigh the manipulation.

In an online experiment van Nimwegen and de Wit [2022] compared the differences in dark pattern detection based on the user platform (desktop vs. mobile) with three concrete dark patterns (*sneak into basket* together with *toying with emotions*, and *trick question*). Their results show that there is a significant difference for *sneak into basket* between desktop and mobile, with users being tricked more often on mobile. Possible explanations by the authors are that mobile users typically spend less time inspecting each element and that the smaller screen and layout change may

People on mobile devices are more likely to be tricked by certain dark patterns.

have contributed to the result. While *trick question* was also more effective on mobile, the differences between the two platforms were smaller.

Bongard-Blanchy et al. [2021] investigated the awareness of people regarding dark patterns and their ability to detect them. Their results show that people are aware that websites can manipulate them. They are less worried about being manipulated or the potential harm that dark patterns might cause. In a study with ten user interfaces (9 containing dark patterns and one benign) 59% of participants detected at least five dark patterns. However, some dark patterns were recognized better than others. For instance, *high-demand* or *limited-time messages* and *confirmshaming* were detected more often than *trick questions* or *pre-selection*.

Bongard-Blanchy et al. [2021] investigated the awareness of people regarding dark patterns.

Luguri and Strahilevitz [2021] further investigated the effects of dark pattern usage with an extensive online survey. They tested the differences between *mild* and *aggressive* dark patterns by trying to sell a subscription service. Using mild dark patterns doubled the acceptance rate compared to the control group (no dark patterns) (11.3% for no dark patterns, 25.8% for mild dark patterns) without causing much backlash. Aggressive dark patterns almost quadrupled the acceptance rate (41.9%) but did cause more backlash in doing so. Furthermore, less educated people were more susceptible to mild dark patterns. In a second study, Luguri and Strahilevitz investigated which dark patterns were more effective in steering people toward a decision. Those were *hidden information*, *trick questions*, and *obstruction*. *Urgency*, however, had no effect toward nudging people.

Luguri and Strahilevitz [2021] show that mild and aggressive dark patterns are effective, but the latter annoys customers.

Similarly, Voigt et al. [2021] performed an online experiment with 204 participants where they tested annoyance and brand trust for two webshops: one without dark patterns and one with five dark patterns (one of each of Gray et al. [2018] categories). Their results show significantly higher annoyance and lower brand trust for the dark pattern version, with *nagging* being described as particularly annoying. The authors further note that there is no connection between the users' affinity for technology and their detection rate of dark patterns.

There is no connection between affinity for technology and dark pattern detection rate.

2.2 Learning Games

Games are suitable for learning and offer measurements for research.

Griffiths [2002] describe the advantages of using games for educational purposes from a general point of view and a research perspective. Games are good at grabbing attention for younger players but are also well suited for a larger demographic (age, gender, educational background). They are more engaging and elements such as interactivity, feedback, and challenges can further stimulate learning. Moreover, games also provide many benefits for researchers. They allow for measurements and comparable (or even standardized) results. Plass et al. [2015] name another advantage of learning games compared to other methods of learning: games can be very adaptive and tailor the learning content to each player's knowledge and skill level.

Learning games possess certain characteristics such as transferable knowledge and clear educational objectives.

Whereas regular games are primarily designed for entertainment, *learning games* (or educational games) have the primary goal of imparting knowledge or teaching certain skills while still being entertaining [Dörner et al., 2016, Tang et al., 2009]. Tang et al. [2009] list characteristics from learning theory that learning games typically possess: they are supposed to be motivating and engaging, embark knowledge that is transferable to the real world, provide feedback and assessment, and have clear educational objectives.

Bloom's Revised Taxonomy (BRT) organizes educational objectives in a two-dimensional array.

A widely used taxonomy for educational objectives is the revision of Bloom's taxonomy by Krathwohl [2002], commonly referred to as *Bloom's Revised Taxonomy* (BRT). Educational objectives are organized into a two-dimensional array. The first dimension is the *knowledge dimension* and includes *factual knowledge*, *conceptual knowledge*, *procedural knowledge*, and *metacognitive knowledge*. The second dimension is the *cognitive process dimension*. This dimension can be considered hierarchical because it scales from simple to complex, and more complex levels include the simpler ones. This dimension includes - from simple to complex - *remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create*. The resulting two-dimensional array, including a description for each educational objective, is visualized in Figure 2.4.

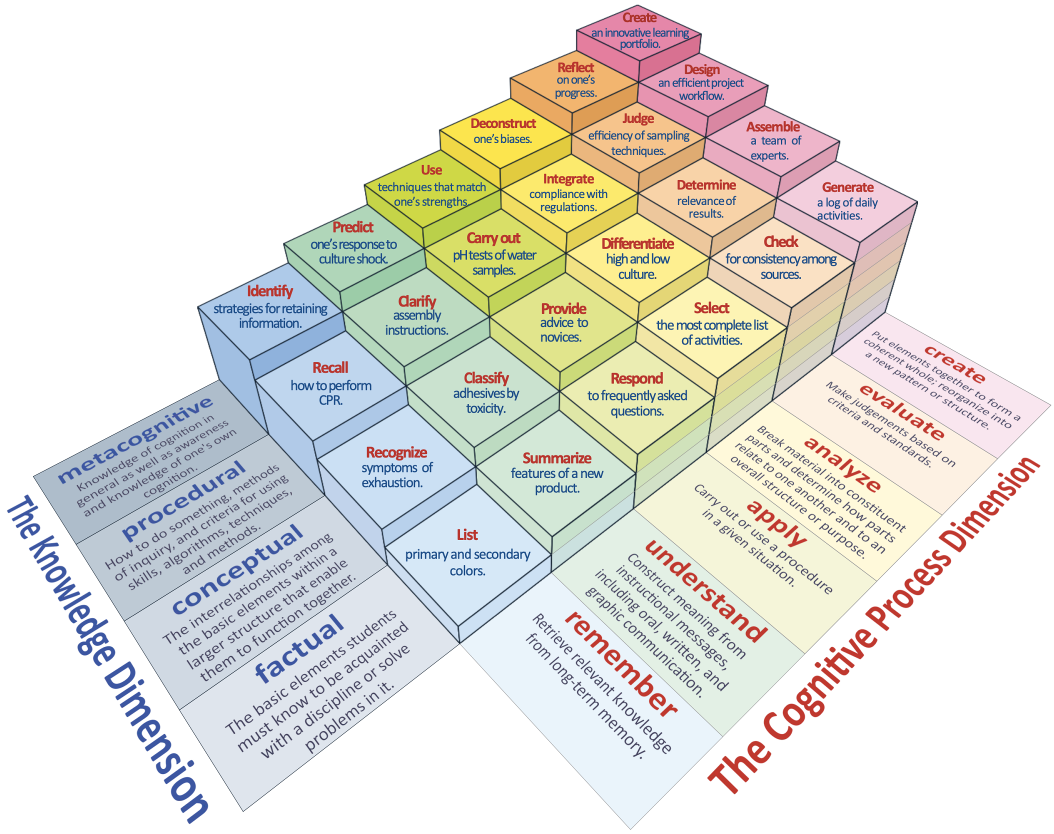


Figure 2.4: A visualization of learning objectives and the two dimensions of Bloom's Revised Taxonomy based on the works of Krathwohl [2002] by Rex Heer, Center for Excellence in Learning and Teaching, Iowa State University (licensed under a Creative Commons Attribution-ShareAlike 4.0 International License[a]).

[a]https://creativecommons.org/licenses/by-sa/4.0/

One early example of a learning game in cyber security is *Anti-Phishing Phil* [Sheng et al., 2007], an educational anti-phishing game. The authors describe the process of specifying the learning goals and educational objectives (*conceptual* and *procedural*). They use a level-based game mechanic, where players are presented with a fixed number of URLs (phishing and benign) and have to decide which is which. This includes a scoring system for correct answers, false positives, and false negatives. The whole game is wrapped in a story and includes a tutorial, a post-level summary, and between-level coaching. The game uses an instant feedback system as well as hints. The results of a user study show

Learning games in phishing improved recognition performance compared to traditional learning methods.

that people who play the game perform better than those who use other learning materials.

Learning games can also improve performance for non-novice users.

Similarly, CJ et al. [2018] designed an anti-phishing game *Phishy* and evaluated it with tech-savvy users. They also use a binary decision game mechanic where players have to decide if a URL is phishing or not. They use a story-based concept, a progressive level design, learning tips, and feedback to enhance learning further. Their evaluation with tech-savvy enterprise users shows that the game improves phishing identification (in correctness and confidence) and is more engaging than other learning methods.

Röpke et al. [2022] show that game mechanics on higher levels in BRT offer more insights.

Both examples of *Anti-Phishing Phil* and *Phishy* are located on the lower end of the *cognitive process dimension* (i.e., simpler) in Bloom's Revised Taxonomy (*remember*). While *recognition* is a sufficient educational goal for them, Röpke et al. [2022] explored if higher levels on the *cognitive process dimension* in Bloom's Revised Taxonomy can further improve performance. They tested two game mechanics on higher levels: one on an *analyze* level and one on a *create* level. Their results show that while there is no significant improvement in performance, both game mechanics on the higher levels offer more insights into the decision-making process and, thus, more feedback and information for the researchers.

Narratives have little to no effect towards increased learning gains.

Another aspect of learning games is the usage of a story or narrative. Clark et al. performed a systematic review and meta-analysis of digital learning games to analyze the impact of various factors on learning. This included the relevance and depth of a narrative. A story that is closely tied to the game mechanics is *relevant*, whereas a story that wraps around and does not affect the gameplay is *irrelevant*. Their results show that the use of irrelevant stories can lead to greater learning gains (significant if looked at individually but insignificant if combined with other factors such as realism and variety of game mechanics). Regarding the depth of a story, no story or a thin story (i.e., setting, scenery, or context) has a significant effect on learning gains. In contrast, medium stories (i.e., stories evolving over time) show a slight adverse effect. Clark et al. note that this might be because a too rich story might distract from learning. They further note that their results only consider learning outcomes. Stories make the game more engaging. This was similarly confirmed by Jemmali et al. [2018] that stories do not positively affect learning outcomes but improve motivation and engagement.

Narratives improve motivation and engagement.

There is also a board game that focuses on learning about dark patterns with a focus on privacy called *The Dark Pattern Game* [Tjøstheim et al., 2022] based on the work of Nyvoll [2020]. This game is targeted towards older teenagers and played in groups of three to five people. The goal is to set up a new phone with a list of apps while trying to share as little data as possible. However, their evaluation shows that playing the game had only partial success in players being more knowledgeable or understanding of dark patterns. Yet, players with higher previous knowledge performed better in the game.

A dark pattern learning board game showed little improvements in increasing knowledge.

Chapter 3

Exploring Game Mechanics

Because we are creating a learning game, we had to carefully plan the game mechanics and find the right balance between game elements and learning goals. Therefore, to understand what game mechanics are suitable for a dark pattern learning game, we decided to begin with a preliminary user study to investigate the strengths and weaknesses of different game mechanics and which are preferred by users.

Before we can create a dark pattern learning game, we have to test which game mechanics are well suited.

This chapter presents the design and evaluation of this preliminary user study. In Chapter 3.1, we describe the design of the study and the general considerations and decisions of the study design. In Chapter 3.2, we present the results of the study as well as an evaluation of them and discuss what these results mean for the creation of our learning game.

3.1 Methodology

This section describes the design and procedure of the preliminary user study. This includes an explication of the selection of the game mechanics tested within this study.

3.1.1 General Considerations

We defined the goal of our learning game that players gain confidence in recognizing dark patterns.

As a first step, we defined the goals and learning objectives for our dark pattern learning game. The primary goal of previous examples (e.g., *Anti-Phishing Phil* [Sheng et al., 2007] or *Phishy* [CJ et al., 2018]) is to raise awareness. However, we determined that raising awareness would not be the primary goal of a dark pattern learning game. A study by Bongard-Blanchy et al. [2021] shows that even though people might be aware of a *"darkness"* in an interface, they still might not be able to resist the dark pattern. Furthermore, given the prevalence of dark patterns, their media coverage, and attempts by UX practitioners to raise public awareness [Brignull et al., 2023, Fansher et al., 2018], we assume that most people who would play a dark pattern learning game are already at least somewhat aware of their existence. Instead, we determined that the goal of our learning game should be for players to learn to better recognize and, consequently, circumnavigate dark patterns. Thus, they should gain increased confidence when interacting with potentially manipulative websites.

Binary decision games don't offer enough feedback and are susceptible to guessing.

In Bloom's Revised Taxonomy (BRT) [Krathwohl, 2002] *recognition* as a learning objective is on the lowest level of the *cognitive process dimension* and the *conceptual*-level of the *knowledge dimension*. A simple game mechanic for this can be that users are presented with images of potential dark patterns and have to decide whether or not they are. However, as Röpke et al. [2022] have shown with learning games in phishing, this binary decision is not very well suited for a learning game. This is mainly for two reasons:

1. It doesn't offer enough insights into the player's decision-making process. Thus, it is difficult to detect and correct any misconceptions.
2. With only a Yes-or-No decision, there is also a high possibility of players just guessing correctly.

Instead, Röpke et al. explored other game mechanics that provide more feedback and use a higher *cognitive process dimension*. The first one is an *analyze game*, where players not only have to decide if something is phishing but also what type of phishing it is. The second one is a *create game*, where players have to construct their own phishing URLs.

Game mechanics on a higher cognitive level offer more feedback and insights.

We determined that the *analyze game* might also be well suited for dark patterns. Furthermore, we already have taxonomies that include high-level categories for dark patterns [Gray et al., 2018, Mathur et al., 2019] which we can use for players to sort dark patterns into.

However, we didn't want to use a *create game* for two reasons: Firstly, the task of creating dark patterns does not offer much more insight into the player's decision-making process while at the same time increasing the complexity of the game. Secondly, we don't want to teach people how to create dark patterns[1]. While they might also be able to recognize dark patterns better, the knowledge of how to create them is not required.

We didn't want users to create dark patterns.

Instead, we considered game mechanics on an *evaluate* level on Bloom's Revised Taxonomy. We determined that a game mechanic where players would have to find a dark pattern on a website (in addition to categorizing it) would be on the *evaluate* level. Additionally, we considered a game mechanic where the player is once again presented with a website and has to find *all* dark patterns that are used by the website. This is also on the *evaluate* level on the *cognitive process dimension*, where the player would additionally have to *judge* if they found all dark patterns. This game mechanic also quite closely represents a real-world scenario.

We came up with two game mechanics on an *evaluate* level where users have to find dark patterns.

Based on these considerations, we decided to use the following three game mechanics in the study:

[1] We confirmed that this was indeed the correct decision when we had to design our own websites with dark patterns for the preliminary study and the learning game prototype. It did not increase our understanding of dark patterns, but coming up with dark pattern designs certainly brought out an evil side.

We came up with three game mechanics to test in the study: CLASSIFY, SINGLE-SPOT, and MULTI-SPOT.

1. CLASSIFY GAME[2]: Participants are presented with an image that may or may not be a dark pattern. They have to decide what category it belongs to. This is identical to the *analyze*-game by Röpke et al. in phishing.

2. SINGLE-SPOT GAME: Participants are presented with a complete website that contains precisely one dark pattern. Participants have to find and categorize it.

3. MULTI-SPOT GAME: Participants are presented with a complete website that contains zero to n patterns. The number of dark patterns is unknown to the participants. They have to find and categorize all of them.

We use Mathur et al.'s seven high-level categories for the study.

Another consideration was what categories to choose for our learning game. Both Gray et al. [2018] and Mathur et al. [2019] provide superordinate categories for dark patterns that are suitable for this study. We decided to use Mathur et al.'s seven categories for this study, mainly because their categories were created in the context of shopping websites. Since there is a higher potential for financial harm, preparing users specifically for these categories seems useful. Furthermore, those categories more specific to shopping websites (*urgency*, *social proof*, and *scarcity*) were the most common occurrences of dark patterns in Mathur et al.'s dataset.

With this study, we aim to answer the following research questions:

RQ-1: Are users able to differentiate between the three game mechanics?

RQ-2: How do users change their approach for each of the game mechanics?

RQ-3: Is there a user preference for a certain game mechanic?

[2]This game mechanic is possibly named somewhat ambiguous considering that *classify* is also the name of an educational objective in BRT on a *recognize*-level. However, this game mechanic is, in fact, on the *analyze* level with a *determine*-objective.

3.1.2 Study Design

We designed the study as a within-subjects experiment. Every participant played each of the three game mechanics. We used a balanced 3x3 Latin square to counterbalance any order effects to determine the order of three game mechanics [Lazar et al., 2017]. Since we had an odd number of conditions, this yielded six different orders [Bradley, 1958].

We designed a within-subjects study counterbalanced by using a Latin square.

Design of Dark Pattern Mock-ups

We decided to create custom made-up websites for this user study to avoid influencing participants because they might recognize a particular website or brand and hence might be prejudiced.

We created custom mock-ups for the study.

The CLASSIFY game consisted of 14 images of dark patterns (two each for every category from Mathur et al.) and two images of elements that were not dark patterns. As described in Chapter 3.1.1, the CLASSIFY images were just images of the dark pattern itself and did not include any surrounding website or context. The two images without dark patterns were a clean cookie consent banner conforming to the GDPR and an image of a countdown that indicated the beginning of a sale and should, therefore, not be *urgency*.

Both CLASSIFY and SINGLE-SPOT tested each dark pattern category twice. *Classify* included two additional images without dark patterns.

Similarly, we created 14 images of complete websites for the SINGLE-SPOT game that contained precisely one dark pattern (again, we used two images for every category from Mathur et al.). We created four major themes for this: a hotel booking website, an online shop for tech products, a food delivery service, and an event ticket shop.

We created nine images of complete websites for the MULTI-SPOT game. They used the four themes from the SINGLE-SPOT game and two additional website designs (a museum ticket shop and another online shop with a more dubious design). Like the other two game mechanics, we included every dark pattern at least twice and varied the number of dark patterns per website: two websites used no dark patterns, one used one, two used two, one used

For MULTI-SPOT, we included each category of dark patterns at least twice and varied the number of dark patterns per mock-up between 0 and 5.

three, and one used four. The other two designs included one element each that we considered a gray area as mentioned in Mathur et al.'s limitations (those had either one or two dark patterns and four or five dark patterns).

Figure 3.1 shows an example each for the CLASSIFY game design of images of dark patterns without the entire website and MULTI-SPOT game of full website designs.

Questionnaire

In this section, we describe the considerations and decisions that went into creating the questionnaire for this user study. The complete questionnaire is available in Appendix A.1.

The questionnaire consists of three parts: one before the study, one after each condition, and one at the end of the study.

We divided the questionnaire into three parts. The first part consists of the informed consent form and a demographics questionnaire that participants filled out before the study. The second part consists of questions regarding each game mechanic and was filled out immediately after playing the respective game mechanic. The third and final part is a ranking across all three game mechanics that participants filled out at the end of the study.

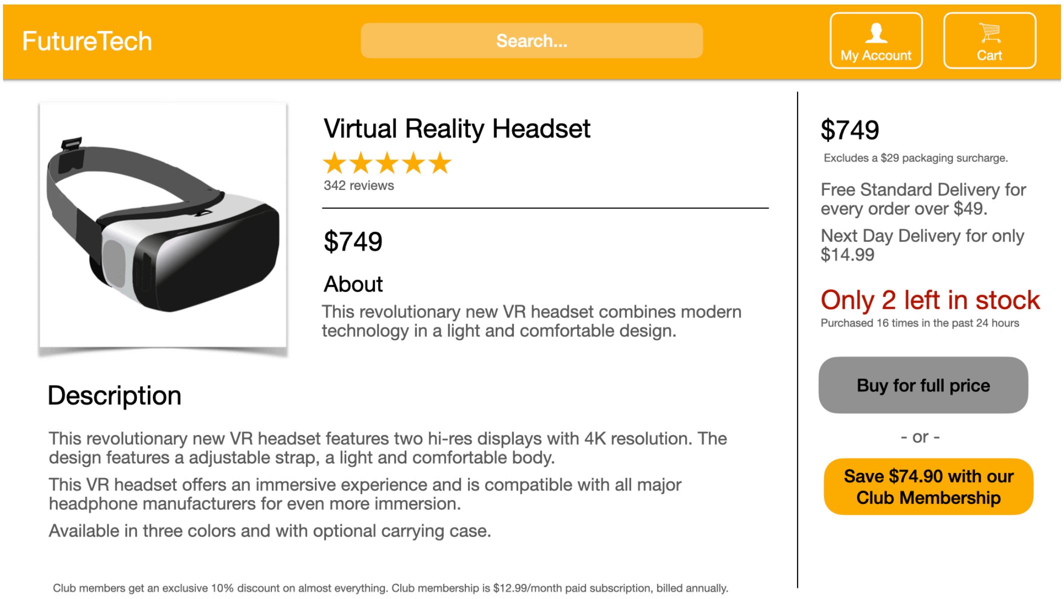
The pre-study questionnaire gathers demographics and self-reported experience and confidence with dark patterns.

The first part of the demographics questionnaire consists of questions about the general demographic background of the participants. This includes age, gender, occupation or field of study, and the highest achieved academic degree. Additionally, it includes a question about their daily internet and social media usage. The second part of this demographics questionnaire consists of two questions to assess their experience and confidence in spotting dark patterns. While those two self-reported metrics are less reliable or comparable among participants, we nevertheless included them for a pre-study and post-study comparison.

The second part of the questionnaire gathers usability and game mechanic related information.

To evaluate each game mechanic, we started by looking at literature in that area. There are questionnaires specifically designed to evaluate games [Calvillo Gamez, 2009]. However, their scope goes well beyond what we were looking

(a) Example of a CLASSIFY game design for *misdirection*. It only shows the dark pattern without the rest of the website. Participants have to decide which category it belongs to.



(b) Example of a full website design based on an online shop used in MULTI-SPOT. This design includes *sneaking*, *scarcity*, and *misdirection*. Participants have to circle and classify all dark patterns that they can find.

Figure 3.1: Two examples of designs that we used during the study: (a) shows only a part of a website whereas (b) shows the complete website. All image assets are royalty-free from Pixabay[a] with a full list and credits available in Appendix A.3.

[a]https://pixabay.com

for during this study (e.g., they evaluate graphics, sound design, immersion, etc.). Instead, we followed the recommendations by Olsen et al. [2011] for usability testing in early game development: Firstly, they recommend keeping the questionnaire as short as possible while still collecting all necessary data to not unnecessarily prolong the study. Secondly, they suggest using a mixture of questions from the SUS [Brooke, 1996] and the QUIS [Chin et al., 1988] to evaluate usability in the early stages of game development. Therefore, we combined game mechanic related questions from Calvillo Gamez [2009] with general usability questions for this part of the questionnaire. We also included text boxes to write down the strengths and weaknesses of the game mechanic as well as any additional thoughts or suggestions.

The post-study questionnaire contains a ranking across the three game mechanics and a self-evaluation of the performance during the study.

The final part of the questionnaire consists of a preference ranking across all three game mechanics specifically designed to answer RQ-3. The ranking is across similar questions as in the second part (*challenge*, *fun*, *frustration*, and *satisfaction*) [Calvillo Gamez, 2009], but also an *overall* question. We also included a text field for participants to explain the main reason behind their number one overall pick. Additionally, there is a single-choice picker for what game mechanic participants believe to be best suited to teach them about dark patterns with a justification text field. Finally, there is the post-study question regarding their confidence. Here, the questionnaire mirrors the pre-study question from the demographics part on their confidence in spotting dark patterns, but we also added a question regarding their confidence that they have correctly identified the dark patterns.

Procedure

Once our designed mock-ups and questionnaire were completed, we prepared a step-by-step outline of our study setup and procedure and performed a test run of it. Based on the results and feedback from this test run, we slightly adjusted the study like this:

The study setup consisted of the following parts:

The apparatus for the preliminary study.

- The printed-out informed consent form and questionnaire for participants to fill out.
- Printed out example images for each category of dark patterns (that we did not use in images during the three games).
- A printed-out list of all categories of dark patterns with a short explanation of each that participants could look at during all three games, similar to how the categories were always visible in the *classify-game* by Röpke et al. in their phishing game. We also included the special category *"It is a dark pattern, but I'm not sure which one"* in this list.
- An iPad Pro that we used to present the images of our dark pattern designs and where participants could freely draw with an Apple Pencil.
- An overhead camera that we used to film the interactions on the iPad (so that we could later measure some timing-related data as to where participants hesitated) and record audio that we could later transcribe.
- A laptop to take notes on during the study for a semi-structured interview at the end.

Based on our test run, we estimated that the duration of the study would be between 40 to 60 minutes.

Before we started with the study, we presented participants with the informed consent form and the first half of the demographics questionnaire. Once they had filled out both, we started the recording on our camera.

Participants consented to a recording of their actions during the study.

We then asked them whether they had any idea what dark patterns are and if they could explain them to us. We would then either correct any misconceptions, if there were any, or explain dark patterns to them. We used this technique to get all participants to the same base level of understanding of what dark patterns are.

We explained the taxonomy used in the study with one example each.

Afterward, we asked participants to fill out the second half of the demographics questionnaire with the self-reported dark pattern related questions. We would then continue explaining the different categories of dark patterns we used throughout the study with our example images and invite them to ask any questions. Once we had showcased all categories and there were no more questions, we would replace the example images with our list of categories instead to avoid any pattern matching with the study images.

In the next phase of the study, we played each of the three game mechanics with the participants in the order specified by our Latin square. The following steps are repeated for each of the game mechanics:

We explained how each game mechanic is supposed to be played.

Firstly, we explained the rules of the game. We explained to the participants what they were supposed to do. We specified that due to the study design, we would not provide any feedback on whether their spotting of dark patterns or classification was correct. We stressed, however, that this would be part of a future game and that we did this because we wanted to minimize learning effects throughout the study.

We did not use think-aloud to avoid the additional cognitive burden.

Then we let participants play the game at their own pace. We specifically did not ask them to do think-aloud since we wanted to avoid the additional cognitive burden [Zhang and Zhang, 2019, Olsen et al., 2011]. However, we also did not stop them if they did so on their own. Furthermore, if we took note of something during the study where we wanted to understand the decision-making process, we would follow up on that after the study during a short interview where we asked the participants to explain their thought process in a retrospective think-aloud.

Once they had completed the game, we asked them to fill out the questionnaire for the respective game mechanic and then continued with the next game mechanic.

After participants had completed all three game mechanics, we asked them to fill out the final part of the questionnaire, where they should rank the three game mechanics and pick the game mechanics that they believed to be best suited to teach people about dark patterns.

Finally, we concluded the study with a short, semi-structured interview where we specifically asked questions to answer our RQ-1 and RQ-2: if participants were able to differentiate between the three game mechanics and how they changed their approach. But we also followed up on any questions that arose during the study.

We concluded the study with a semi-structured interview to gather more qualitative data.

3.2 Results & Evaluation

We conducted our user study over the course of two weeks. We then digitized the questionnaires, transcribed the audio recordings, and combined this with the qualitative data from the questionnaires and our notes during the study. We coded the qualitative data using MAXQDA[3]. In the following section, we present the results of our study and evaluate those results.

3.2.1 Demographics

The Latin square yielded six permutations for our three game mechanics that we each tested twice. We conducted the study with 12 university students (7 female, 4 male, 1 divers). The participants' age ranged from 19 to 27 years (M=23.8, SD=2.4).

We conducted the study with 12 university students.

Seven of our participants had a technical background (computer science, engineering) as their current field of study. Five participants had a non-technical background (economics, musicology). Six participants had a high school diploma, four a Bachelor of Arts, and two a Bachelor of Science as their highest academic degree.

Participants were a mix of majors and Bachelor's and Master's students.

58% said that they spend between 3 and 5 hours online or on social media every day. 25% reported more than five

[3] https://www.maxqda.com [Accessed: Dec. 17, 2023]

hours per day and 17% between 1 and 3 hours per day. The self-reported experience with dark patterns on a 5-point Likert scale was M=3.1 (SD=0.8, on a scale from 1 to 5). Likewise, the self-reported confidence that they would be able to spot dark patterns on a 5-point Likert scale was rated at M=3.75 (SD=0.75).

3.2.2 Data Analysis

We gathered quantitative and qualitative data during the study.

We gathered a variety of different data types. The questionnaire itself already included both quantitative and qualitative data. We also gathered additional qualitative data from our notes during the study, the interviews afterward, and the transcribed audio. Furthermore, we measured timings and error rates for each game mechanic and every category of dark patterns as more quantitative data. While they were less relevant for the evaluation of the game mechanics, those data were essential for refining our learning goals and fine-tuning our learning game prototype as described in Chapter 4.

In the following sections, we analyze both quantitative and qualitative data individually before drawing overall conclusions.

Quantitative Analysis

For the quantitative analysis, we used Numbers[4] spreadsheets for descriptive analysis and Python[5] for further analysis.

We start by analyzing the quantitative data for each game mechanic. Figure 3.2 shows the results for the nine questions from the questionnaire where participants were asked to rate aspects of the game mechanic on a 5-point Likert scale on a scale from 1 (strongly disagree) to 5 (strongly agree). The complete data are available in Table A.1.

[4]https://www.apple.com/numbers/ [Accessed: Dec. 17, 2023]
[5]https://www.python.org [Accessed: Dec. 17, 2023]

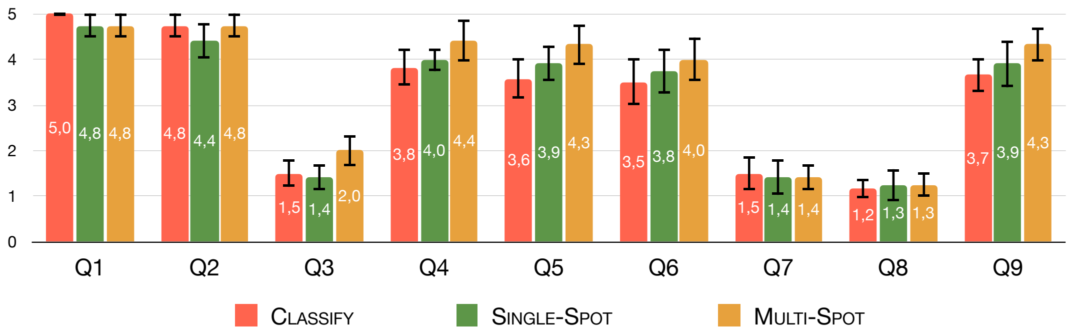


Figure 3.2: This figure shows the results of the questions for each game mechanic of our questionnaire. It shows the arithmetic mean and 95% confidence interval for every question and game mechanic. While there are only little differences across all questions, the results show a tendency that MULTI-SPOT was generally rated best in enjoyment (Q4), engaging (Q5), varying (Q6), and suitability for a dark pattern learning game (Q9).

Participants rated that the game had a clear goal (Q1) and that it was intuitive (Q2) very highly (M≥4.4 for each game mechanic in both questions). None of the game mechanics was rated too challenging or difficult (Q3), though MULTI-SPOT was rated slightly more challenging than the other two game mechanics. Participants rated the perceived enjoyment while playing the game (Q4), engaging gameplay (Q5), and varying and interesting tasks (Q6) as high for each game mechanic. However, while the differences in means are small, SINGLE-SPOT was rated slightly better than CLASSIFY, and MULTI-SPOT was rated even better than SINGLE-SPOT for all three categories. Frustration during the gameplay (Q7) and after completing the game (Q8) was rated very low, with minimal differences between the three game mechanics. The last question was whether participants believed that the game mechanic is suitable to learn about dark patterns (Q9). While all three game mechanics were rated positively in this regard, MULTI-SPOT (M=4.3, SD=0.65) was rated slightly better than SINGLE-SPOT (M=3.9, SD=0.9), which in turn was rated slightly better than CLASSIFY (M=3.67, SD=0.65).

MULTI-SPOT was generally rated best in enjoyment (Q4), engaging (Q5), varying (Q6), and suitability for a dark pattern learning game (Q9).

In the first part of the *Overall* questionnaire at the end of the user study, participants ranked the three game mechanics in the categories *challenge* (from 1 = most challenging to 3 = least challenging), *fun* (from 1 = most fun to 3 = least

Participants ranked all three game mechanics in five different aspects.

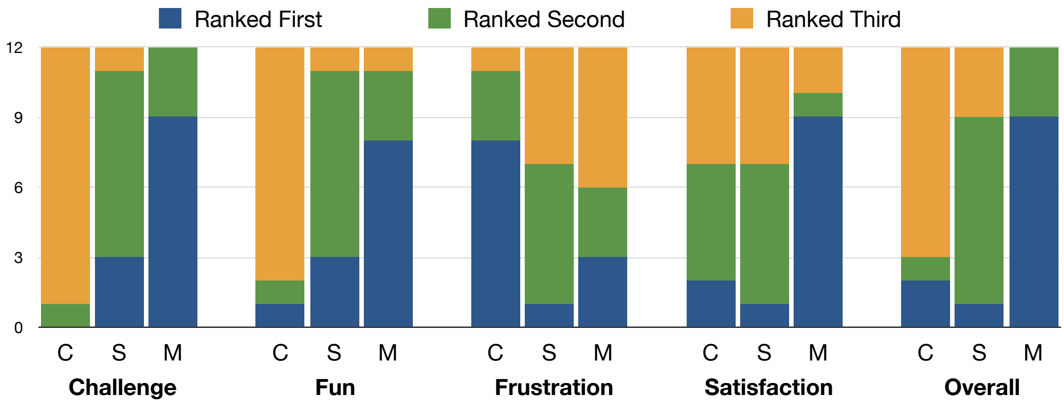


Figure 3.3: We asked participants to give a preference ranking across all three game mechanics (CLASSIFY, SINGLE-SPOT, MULTI-SPOT) in the categories *"challenge"*, *"fun"*, *"frustration"*, *"satisfaction"*, and *"overall"*.

fun), *frustration* (from 1 = least frustrating to 3 = most frustrating), *satisfaction* (from 1 = most satisfying to 3 = least most satisfying), and *overall* (from 1 = best to 3 = least best). The results of this question are visualized in Figure 3.3.

For *challenge*, *fun*, and *overall*, there is a clear indication for the first choice and the last choice. There is a clear first choice for *frustration* and *satisfaction*, but the second and third choices are more ambiguous.

MULTI-SPOT was ranked best in all categories except frustration.

75% of participants rated MULTI-SPOT as the most *challenging* and 92% rated CLASSIFY the least *challenging*. Similarly, 67% rated MULTI-SPOT the most *fun* and 83% rated CLASSIFY the least *fun*. 83% rated CLASSIFY the least *frustrating*, while 50% rated MULTI-SPOT and 42% SINGLE-SPOT the most *frustrating*. Likewise, 75% rated MULTI-SPOT the most *satisfying* while 42% each rated CLASSIFY and SINGLE-SPOT the least most *satisfying*. For the *overall* best game mechanic, 75% rated MULTI-SPOT the best and 75% rated CLASSIFY the least best.

MULTI-SPOT is considered best suited for learning about dark patterns.

Furthermore, participants were asked which game mechanic they believed could teach them the most about dark patterns. 10 of our 12 participants picked MULTI-SPOT and one participant each picked CLASSIFY and SINGLE-SPOT.

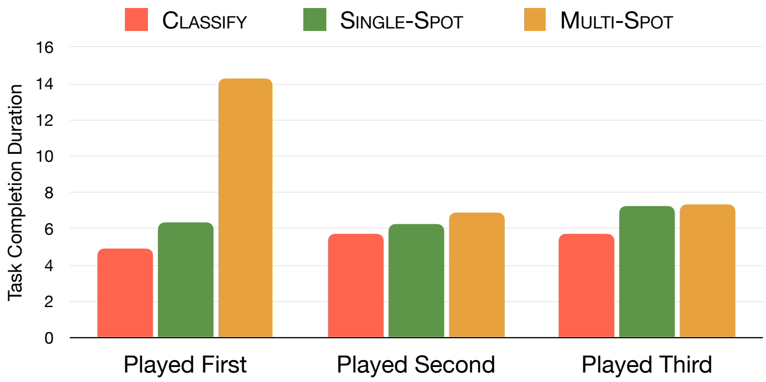


Figure 3.4: When we analyzed the times to complete each task, we observed noticeable differences for MULTI-SPOT. Therefore, we considered the position that the game mechanics were played at and found out that MULTI-SPOT took approximately twice as long when played first.

CLASSIFY was completed the quickest whereas MULTI-SPOT took the longest.

During the user study, when specifying the tasks, we told our participants to perform the tasks at their own pace. We did so because we didn't want to induce additional stress or time pressure. However, we did measure the time that participants took per game mechanic. The CLASSIFY game was completed the fastest (M=5.5, SD=1.3), followed by SINGLE-SPOT (M=6.6, SD=0.7) and MULTI-SPOT (M=9.7, SD=4.4). However, the comparably high standard deviation in MULTI-SPOT was interesting.

Participants spent more time with MULTI-SPOT, especially if it was the first condition.

Therefore, we compared the times for the three game mechanics based on the order in which they were played. The results (see Figure 3.4) show that MULTI-SPOT took considerably longer to complete if it was the first game mechanic that the participants played. More precisely, it took approximately twice as long compared to when it was played as the second or third game mechanic. This is especially interesting as neither CLASSIFY nor SINGLE-SPOT had a comparable difference between conditions. On the contrary, CLASSIFY even took longer when played second or third. We will try explaining this in Chapter 3.2.3 and draw consequences from this for our game design.

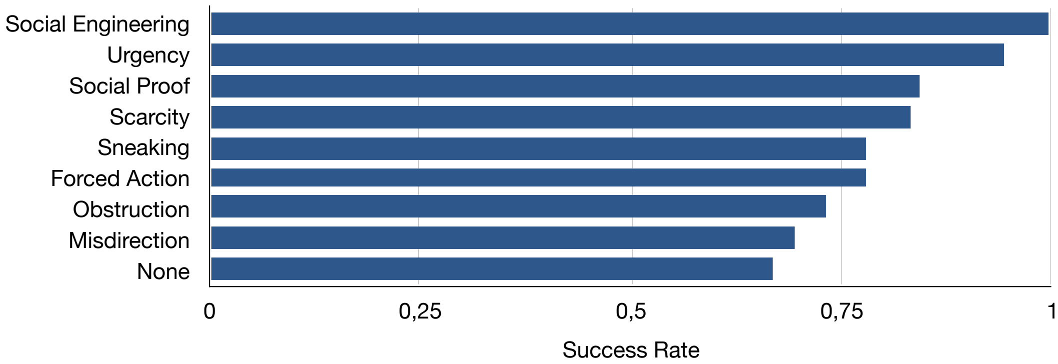


Figure 3.5: The success rates for each category of dark patterns in the preliminary study. Because many participants mistook *scarcity* or *social proof* for *urgency*, which Gray et al. [2023] combined in one superordinate category (*social engineering*), we also included this as the combined success rate for all three.

The difference in detection success rate between the three game mechanics is marginal.

Finally, we analyzed how successful participants were in correctly identifying the different categories of dark patterns. Firstly, there is little difference between the three game mechanics. CLASSIFY was the best with 79.9% successful identification, followed by MULTI-SPOT with 77.2% and SINGLE-SPOT with 74.6%.

Social engineering dark patterns were detected most accurately.

There are, however, more noticeable differences between the categories of dark patterns. This reflects the results from the questionnaire of Bhoot et al. [2021] that users recognized certain dark patterns more easily than others. The results of our study are visualized in Figure 3.5. Our results show that *urgency* (94%), *social proof* (84%) and *scarcity* (83%) were identified most successfully. This effect is even stronger when we consider that most participants who made mistakes in one of those categories mistook either *scarcity* or *social proof* for *urgency*. Since Gray et al. [2023] combine these three categories into one (*social engineering*) in their ontology, we also analyzed the success rate for them combined, which results in 99%.

There is a high number of false positives.

Both *sneaking* and *forced action* were correctly identified 77% of the time, followed by *obstruction* (73%) and *misdirection* (69%). The designs with no dark patterns, however, were only correctly identified at 66%. This high number of false positives can be related to participants expecting dark patterns in a dark pattern user study and thus trying to find a matching dark pattern even if none existed.

Qualitative Analysis

For the qualitative analysis, we followed the approach by Braun and Clarke [2006] and Saldaña [2021]. We used MAXQDA to aggregate the data from the questionnaires, the study notes, and the semi-structured interviews. We then coded the data and performed a thematic analysis on it [Braun et al., 2019]. The codebook is available in Appendix A.4.

We coded the qualitative data with MAXQDA and performed a thematic analysis.

One aspect was criticized across all three game mechanics. Even though we specifically informed participants that there would be no feedback due to the design of the study, this was commonly criticized both during the study and in the questionnaires.

Lack of feedback was most often criticized.

Concerning the individual game mechanics, the data from the questionnaire match what we had witnessed during the study. Participants liked CLASSIFY for its faster pace and as a way to get many examples of dark patterns. However, they criticized that it lacks realism, and the missing context makes it harder to correctly judge the intent of certain elements as to whether or not they are indeed manipulative.

CLASSIFY was praised for its faster pace but criticized for its lack of context.

While SINGLE-SPOT fixed the missing context issue, participants still criticized a lack of realism. Furthermore, having precisely one dark pattern made it more predictable. However, they appreciated that it allowed them to really focus on individual elements.

SINGLE-SPOT provides context but still lacks realism.

Since MULTI-SPOT was clearly ranked best overall and as the best-suited game mechanic for learning about dark patterns (see Chapter 3.2.2), we will describe its feedback in more detail:

> *"The longer you look at a website, the more blurry it gets what may be a dark pattern or not."*
>
> *—Participant 10*

MULTI-SPOT offers realism and more challenges.

Participants treated and considered MULTI-SPOT more like a detective game, searching and scanning for clues for dark patterns. They liked the realistic look and gameplay of MULTI-SPOT as it closely matches what they experience in real life.

> *"You have to do more at the same time."*
>
> *—Participant 2*

MULTI-SPOT allows players to reflect on intentions behind UI elements.

Another aspect that participants liked in MULTI-SPOT is that the context of a game, while still closely resembling the real world, allowed them to engage longer with the website and hence allowed them to reflect more deeply on the intentions behind individual elements and consider whether or not they are deceptive. Some participants were aware that this comes at the risk of more false positives.

> *"There is a cookie banner! They are always manipulative!"*
>
> *—Participant 5*

MULTI-SPOT has more false positives.

However, about one third was less worried about false positives and indeed picked everything that could be considered mildly manipulative (e.g., a strikethrough previous price next to the discounted price[6]) or not manipulative at all (e.g., the clean cookie banner).

The uncertainty of MULTI-SPOT is more controversial.

The open-ended aspect of MULTI-SPOT is more controversial. About 66% of participants liked the uncertainty of not knowing if they were finished. The other third of participants stated that they would have liked some sort of indication when there were no more dark patterns left.

[6] This is not something considered a dark pattern in Mathur et al.'s dataset.

3.2.3 Discussion

In this section, we discuss the results from our preliminary study and attempt to answer our research questions. We also discuss the limitations of the study and its consequences for our learning game prototype.

User Perception of Differences between Game Mechanics

The first research question was whether users are able to differentiate between the three game mechanics and how they perceive the differences between them.

We picked the three game mechanics to be on three different levels of Bloom's Revised Taxonomy with higher cognitive demands for SINGLE-SPOT and MULTI-SPOT, respectively. While participants were generally able to name differences between the game mechanics in the post-study interview, no one named differences in cognitive demand.

Participants did not perceive the differences in cognitive demand.

Interestingly, participants focused more on differences in the "categorize" part of the game mechanic (that each game mechanic offers) and did not consider the process of finding the dark patterns (in SINGLE-SPOT and MULTI-SPOT) as a real difference.

Participants focused on differences in categories.

Instead, they often stated that CLASSIFY and MULTI-SPOT are more similar because both have the option that something is not a dark pattern (CLASSIFY) or does not contain any dark patterns (MULTI-SPOT). On the other hand, SINGLE-SPOT always had one dark pattern, which made it easier in the perception of our participants.

Changes in Gameplay

Our second research question was how users change their approach for each game mechanics. Since the core concept of CLASSIFY is part of the other two game mechanics, we omit to discuss this individually.

The clear exit in SINGLE-SPOT is an advantage and disadvantage.

One observation during the study was that SINGLE-SPOT offers a clear exit. Except for one participant[7], everybody stopped looking at the website once they found one dark pattern. While this makes sense from a "goal achieved" point of view, problems arise if their choice was incorrect. It reduces the time that participants spend thinking about potentially manipulative elements or reflecting on the website in general, especially compared to MULTI-SPOT.

Participants tended to second-guess more elements in MULTI-SPOT.

As we already described in Chapter 3.2.2, MULTI-SPOT was the complete opposite. The majority of our participants pondered over almost every element trying to decide if it could be a dark pattern. We believe that it is certainly preferable from a learning game perspective that players spend more time and possibly overthink and second-guess certain elements instead of skipping over half the page. But at the same time, it will be useful to at least limit the amount of containers that players can select so they don't get carried away with their selection.

User Preferences & Suitability

The third research question was whether there is a user preference for a certain game mechanic. Both the quantitative results from the questionnaire and the qualitative results from the interviews show that all three game mechanics are generally well suited for a dark pattern learning game.

MULTI-SPOT was generally preferred by our participants.

However, there is a clear preference towards MULTI-SPOT. It was rated the highest on the 5-point Likert scale that participants filled out for each game mechanic regarding its suitability. MULTI-SPOT was also ranked as the overall best game mechanic by 75% of our participants.

The higher difficulty makes MULTI-SPOT more frustrating but also more satisfying.

While MULTI-SPOT was also rated the most frustrating game mechanic, it was at the same time rated the most satisfying and the most fun. This may seem contradictory. However, the individual rating for frustration was

[7]This participant believed to have found two dark patterns which confused them.

still pretty low, and it was only ranked the most frustrating compared to SINGLE-SPOT and CLASSIFY. We believe that this might be caused by the slightly higher difficulty and less guidance, which simultaneously explains the higher rated satisfaction and fun with a greater feeling of achievement.

Limitations

Missing interactivity was an issue for designing the mock-ups and for participants during the study.

One shortcoming in this study already became apparent during the design of the mock-ups and was also mentioned by some of the participants during the study. Static images lack the interactivity of a real website, which makes it harder for participants to judge certain elements. However, this goes beyond just missing context or reduced realism for the user. The designs of the dark pattern mock-ups are static as well, which limits them in their expressiveness for certain dark patterns (e.g., *sneaking* or *obstruction* often use a multi-step approach that can't be replicated with a single image). Chen et al. [2023] use the terms *static* and *dynamic* dark patterns to differentiate between the two. And while the former certainly works well in the design of our study, the latter was more challenging to convey properly. Thus, a learning game for dark patterns should offer at least some level of interactivity to reflect dark patterns that exploit multiple steps properly.

The number of complaints about lack of feedback - though intentional for the study - makes it clear how important it is to users.

The second big issue during the study was the lack of feedback. Even though we had told participants before the study that there would be no feedback due to the study design (with-in groups, learning effects), it was still criticized by the majority of participants. While we had always planned to use some sort of feedback system (which is also successfully used by, e.g., *Anti-phishing Phil* [Sheng et al., 2007], and *Phishy* [CJ et al., 2018]), the comments during and after the study made it clear that participants prefer instant feedback to know whether or not their decisions were indeed correct.

Feedback should include an explanation along with the correct answer.

An instant feedback system would also address the issue of the higher number of false positives in MULTI-SPOT. For this reason, feedback should include an explanation as to why an element is a dark pattern or not. For instance, for the cookie banner misconception, exemplary feedback could be: *"This cookie banner puts equal emphasis on the option to accept or to decline and conforms with the GDPR requirements. Therefore, it is not a dark pattern."*

MULTI-SPOT can be more overwhelming when played without prior knowledge.

Another issue is the task completion time as shown in Figure 3.3. When MULTI-SPOT was the first condition, participants took considerably longer compared to it being the second or third condition. Based on our observations, participants were more thorough and examined almost every element carefully before making a decision. Furthermore, when MULTI-SPOT was the first condition, participants spent more time once they (unknowing to them) had found all dark patterns to ensure they had not missed something. This, too, was something that was less prominent when MULTI-SPOT was the second or third condition.

This is not necessarily a bad thing. However, it also shows that there is a considerable learning effect early on. Once participants were familiar with the categories of dark patterns and first examples, they were faster and more decisive even when facing the uncertainty of MULTI-SPOT.

CLASSIFY or SINGLE-SPOT could be used as a tutorial.

Four participants even suggested that CLASSIFY (3 participants) or SINGLE-SPOT (1 participant) would make an excellent tutorial for MULTI-SPOT. Advantages would be that players would get a lot of examples early on, and the faster pace would help them learn the taxonomy and prepare players for MULTI-SPOT. This would also mitigate the timing issue discussed before and is certainly something we will consider for the learning game prototype.

Chapter 4

A Dark Pattern Learning Game Prototype

The previous chapter focused on suitable game mechanics for a dark pattern learning game. In this chapter, we continue with a functional and playable prototype of said learning game. We describe the requirements, design decisions, and core mechanics of the learning game. The resulting game prototype is available under this link[1].

Since MULTI-SPOT was the preferred game mechanic in our preliminary user study, we will use it as the underlying game mechanic for our learning game prototype.

4.1 Requirements

The requirements for the game are primarily defined by our choice of MULTI-SPOT as the underlying game mechanic.

Players need to be able to select parts of the website and categorize them.

At least, the game needs to display a website that players should investigate for dark patterns. Furthermore, there needs to be a way for players to select areas of the website

[1]https://git.rwth-aachen.de/i10/thesis/thesis-kevin-fiedler-dark-pattern-game-mechanics

that are a potential dark pattern. Finally, there must be a way for players to categorize their selection.

Feedback and interactivity are further requirements.

Additional requirements based on the results from the preliminary study (see Chapter 3.2.3) are a solid feedback system and some level of interactivity to support dynamic dark patterns [Chen et al., 2023].

4.2 Platform & Engine

We built a cross-platform game to increase availability.

One of the more important aspects of our learning game is availability. With the goal of educating people about dark patterns, it is certainly desirable to reach as many people as possible. As such, a *cross-platform* game was a logical choice. Here, we considered different options such as Godot[2], a cross-platform game engine suitable for 2D and 3D games. Or a web app using Next.js[3] and React[4] that is playable from any modern browser.

We use React for the game UI and easy content creation and integration.

Because part of our game already is a website, it made sense to also develop the surrounding game with React and, by doing so, directly expose the dark pattern elements from the website to the game. The idea is to mark elements in the dark pattern website with certain classes to indicate that they are a dark pattern. This makes content creation and integration very easy, which is beneficial for extending the game with additional levels in the future.

Furthermore, with React we can build modular elements as building blocks to recreate different versions of the same website for increased replayability. For instance, we have created multiple versions of cookie banners, either with or without dark patterns, that can be reused for every website. Furthermore, existing building blocks like this allow for the creation of new content even faster.

[2]https://godotengine.org [Accessed: Jan. 2, 2024]
[3]https://nextjs.org [Accessed: Jan. 2, 2024]
[4]https://react.dev [Accessed: Jan. 2, 2024]

4.3 Narrative

The role of a story or narrative in learning games has been studied quite extensively [Jemmali et al., 2018]. Even though stories in learning games have not been shown to increase learning performance [Adams et al., 2012], Jemmali et al. still recommend the use of narrative elements to increase engagement with the game, which may result in improved learning.

Story-based elements are recommended to increase engagement.

We decided to theme the game around the *"Dark Pattern Defense Force"* (DPDF), a fictitious agency dedicated to fighting against dark patterns. The player assumes the role of the newest recruit to that agency, which is expressed during the tutorial in the style of text-based communication with their new boss.

The player assumes the role of newest recruit in the fictitious *Dark Pattern Defense Force*.

Over the progression of the game, the player will climb the metaphorical *career ladder* from *DPDF Trainee* all the way up to *DPDF Director* based on their performance in the game. This is an additional incentive system to further increase motivation and engagement [Plass et al., 2015].

For game progression, the player climbs up a *career ladder*.

Furthermore, the styling of the user interface supports these story elements. For instance, players work on *contracts* (i.e., levels), and task descriptions are also written in the style of a work directive.

The design of the UI fits the narrative.

4.4 Game Mechanics

The game is divided into two segments: the tutorial and regular levels. The tutorial has to be completed in order to unlock the regular levels, which contain sub-tasks and can be played in any order.

4.4.1 Tutorial

We use Gray et al.'s six high-level patterns as categories.

The tutorial itself consists of two tasks. The first one introduces the topic of dark patterns and their taxonomy. Based on the results of the preliminary study (especially regarding the error rate in Chapter 3.2.2), we decided to use the six high-level patterns from the ontology by Gray et al. [2023][5].

The tutorial explains each category with a brief example.

Each high-level pattern is explained with one example and lists further specific cases (i.e., meso-level or low-level patterns). Limiting this part to one example was a deliberate choice not to make the tutorial too long and tedious.

This first task in the tutorial also includes the story element described in Chapter 4.3 and instructions on how to play the upcoming levels.

We use the CLASSIFY game mechanic to test the player's knowledge.

The second task in the tutorial uses the CLASSIFY game mechanic from the preliminary study to give new players the opportunity to test their knowledge and to familiarize themselves with the taxonomy and the gameplay. We included this based on the feedback from the preliminary study (see Chapter 3.2.3).

The game UI is kept minimalistic so players can focus on the content.

The design of the CLASSIFY game is shown in Figure 4.1. We kept the game user interface as minimal as possible to allow players to focus on the content of the game. The main view shows a single page element with a potential dark pattern. Once the player selects a category at the bottom of the sidebar, the correct answer and an explanation will be shown. It also unlocks the next example at the bottom of the page.

Besides the list of categories, the sidebar also shows the elapsed time and the current score. The scoring system will be explained in more detail in Chapter 4.4.4. Similarly to the preliminary study, the categories include the options *"None"* (exclusive for the tutorial) and *"I'm not sure"* (available in the tutorial and the regular game). The category

[5] At the time of building the learning game and the user study, only the draft version of the ontology was available. Therefore, any updates to the taxonomy, such as *nagging* being moved from a high-level pattern to the meso-level of *forced action* are not considered hereinafter.

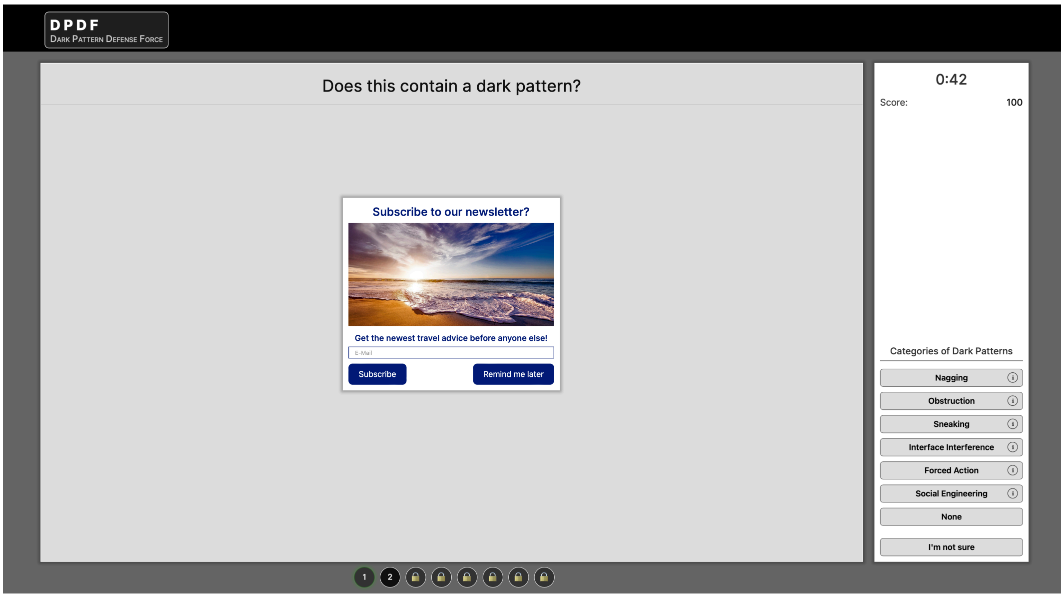


Figure 4.1: The design of the tutorial based on the CLASSIFY game mechanic. Players are presented with a series of potentially manipulative website elements and have to choose what category of dark pattern they belong to in the sidebar.

buttons further include little info icons that display a tooltip containing a short explanation about the sort of dark patterns they describe when hovered.

After the player completes all examples in the CLASSIFY game, they unlock the main game but can also revisit the tutorial anytime.

4.4.2 Main Game

The main game uses the MULTI-SPOT game mechanic as discussed in Chapter 3.2.3. The game user interface is very similar to the one from the tutorial game (see Figure 4.2). However, the sidebar now shows a button to enable the *"Dark Pattern Magic Wand"*, which is key to the gameplay. When enabled, the cursor changes to a magic wand icon and when moving over the website in the center, it highlights the currently hovered container in red.

Regular levels extent the tutorial gameplay with a *magic wand* that highlights elements on the website.

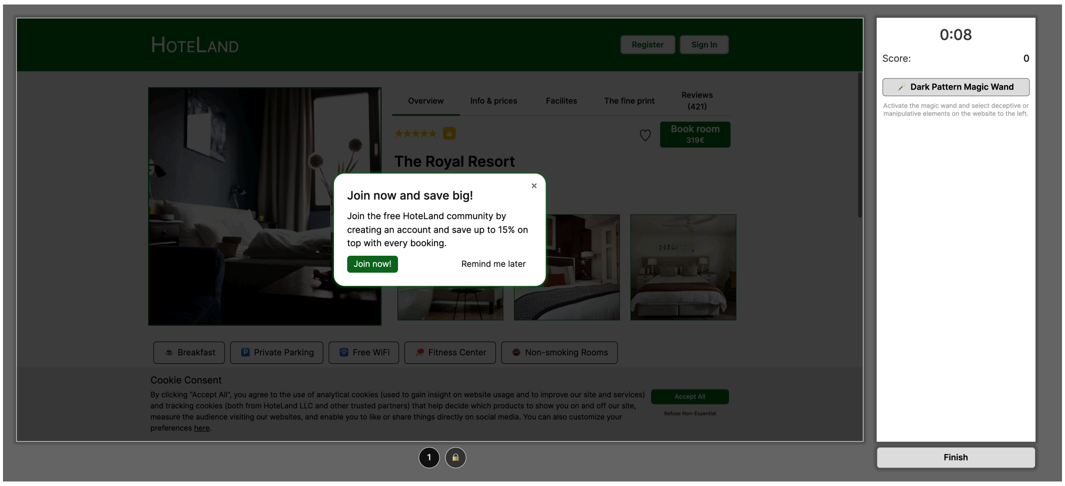


Figure 4.2: The design of the main game is based on the MULTI-SPOT game mechanic but intentionally similar to the one used in the tutorial (Figure 4.1). The sidebar now offers a way to enable the *Dark Pattern Magic Wand* that players use to select manipulative elements on the website on the left.

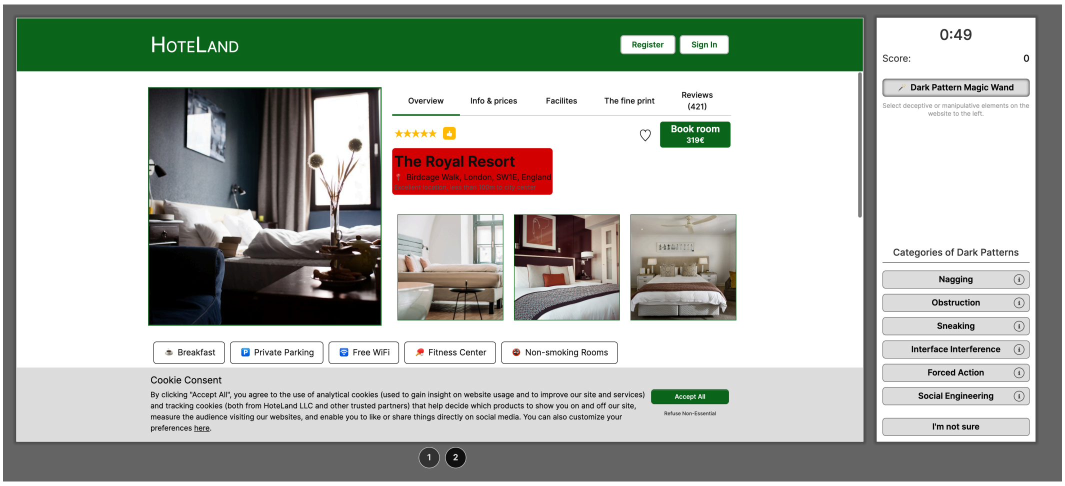


Figure 4.3: In this stage, the player has selected a potentially manipulative element on the website on the left and the category buttons have appeared in the sidebar.

When the player clicks on the highlighted container, the category buttons appear (see Figure 4.3) and they can pick a category for the dark pattern. The categories include the *"I'm not sure"* answer but omit *"None"* as an answer since players can simply unselect the container if they change their minds.

To address the issue of dynamic dark patterns (Chen et al. [2023]) that only become apparent over multiple steps or pages (which was one shortcoming of the static images used in the preliminary study), the game supports multiple pages where players can freely navigate between at the bottom of the screen. Initially, all but the first page are locked and players need to unlock them by finding the corresponding element on the website. When they hover over such an element, the cursor changes to an unlocked lock icon and the page is permanently unlocked. For example, in Figures 4.2 and 4.3, this is used to dismiss the alert. Actually dismissing the alert would prevent players from interacting with it again if it included a dark pattern they wanted to select. With the multi-page support, they can go back and revisit the alert at any time.

Dynamic dark patterns are supported with multiple pages within a level.

For this prototype, we included two groups of levels with two and three levels each. One is the order process of a smartphone online shop in two designs. The first level is free of dark patterns, and the second level (Figure 4.4a) uses *interface interference*, *sneaking*, and *social engineering* in an attempt to upsell.

Two levels are themed around a smartphone online shop.

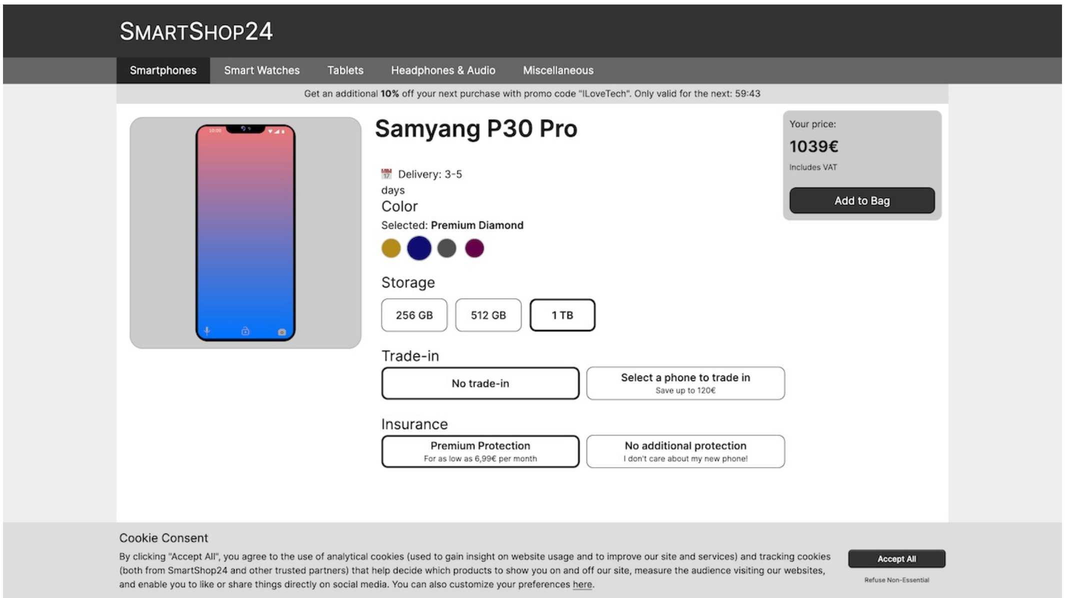
The other levels are three different pages of a hotel booking site: a listing page (Figure 4.4b) that heavily uses *social engineering* and *sneaking*, a booking page (as shown in Figures 4.2 and 4.3) that uses *nagging*, *interface interference*, *sneaking*, and *social engineering* and a newsletter unsubscribe page that uses *obstruction* and *interface interference* to discourage the user from unsubscribing.

The other three levels are themed around a hotel booking site.

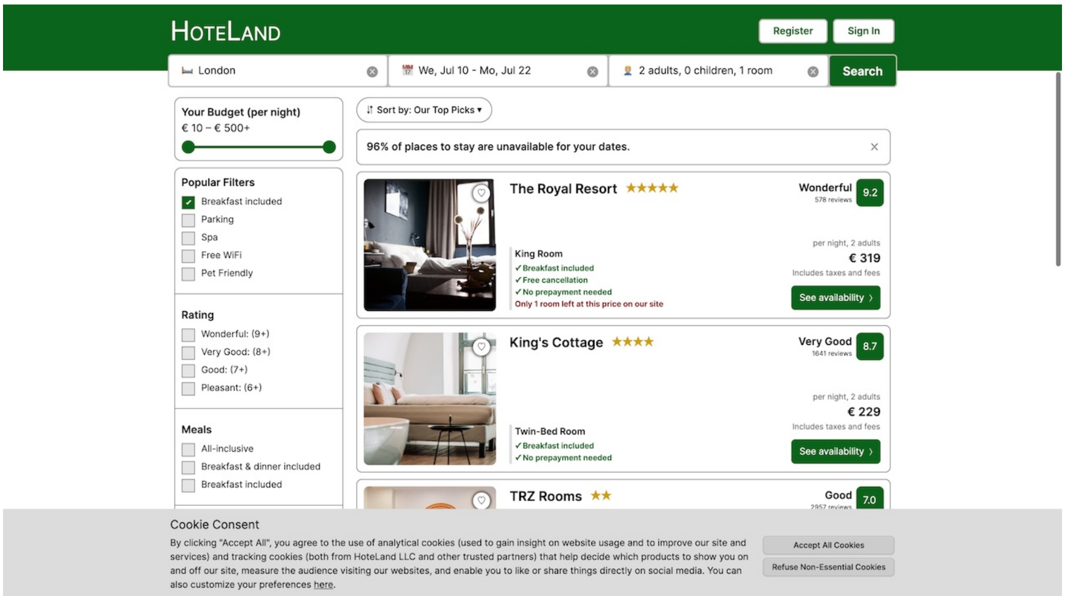
4.4.3 Feedback System

The results of the preliminary study made it very clear that players wanted and needed proper feedback. Therefore, we carefully designed the feedback system to be informative and yet not hinder gameplay. When a player selects a category of dark patterns, the game presents an overlay as instant feedback showing whether their choice was correct, along with a short explanation. If the answer is incorrect, it also names the correct category.

The game provides instant feedback and explanations.



(a) One design of an online smartphone shop that tries to upsell the phone offered. Amongst others, it uses the most expensive options (as preselection) and a hidden subscription for premium insurance.



(b) The design of a hotel listing site that uses urgency and scarcity to rush the visitor into making a booking.

Figure 4.4: Examples of both level groups: (a) shows an online smartphone shop. (b) shows the listing page of a hotel booking site.

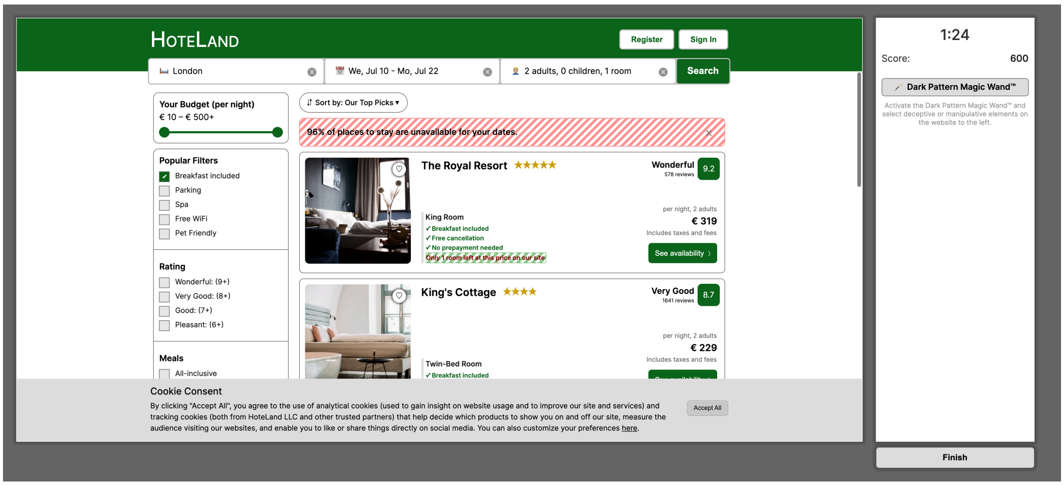


Figure 4.5: When the user categorizes an element, the game visually highlights them: green for correct categories, yellow for *"I'm not sure"* answers, and red for wrong categories and false positives.

Furthermore, the game colors the element background based on the answer (see Figure 4.5). Correctly categorized elements are shown in green, and wrongly categorized elements and false positives are shown in red. When the player selects the *"I'm not sure"* answer, the element is colored yellow.

Categorized elements are highlighted visually.

When the player finishes a level, they are presented with a list of all the dark patterns that they have missed. The list only names the category and not the specific element. This was a deliberate choice to present the player with two options: showing them what they have missed (in this case, all the missed elements are colored red and yellow) or attempting the level again and trying to find the missed elements on their own. We believe that this should be a choice that the player gets to make and not forced by the game.

Players can choose to view their missed dark patterns or try again.

If the player chooses to view their mistakes, they can hover over a wrong or missed element, and an explanation will be shown in a tooltip overlay.

4.4.4 Scoring System

Players are scored based on their actions.

Along with the feedback, players are scored during gameplay. The overall level score is always visible in the sidebar. The level overview also shows the high score for each level[6].

The instant feedback overlay, which is shown after the player categorizes an element, and the summary at the end of each level also includes how their answer affected the score.

Scoring considers spotting dark patterns, false positives, false negatives, and correct categorization.

The scoring system covers these cases:

- Finding a dark pattern element (omitted for the tutorial CLASSIFY game): 100 points
- Missing a dark pattern element (false negative): -100 points
- Selecting an element that is not a dark pattern (false positive): -50 points
- Correctly categorizing an element: 100 points
- Wrongly categorizing an element: -100 points
- Selecting the *I'm not sure* answer: -50 points

Points can be adjusted for difficult or easy dark patterns.

The points awarded for the first two cases are fixed. However, the points for the other four cases can be customized per element. For instance, a harder-to-find or more difficult-to-categorize element can be rewarded with more points, or a very mild or ambiguous element can score fewer points.

Additionally, players are rewarded 200 extra points at the end of a level if they found all dark patterns. They receive an additional 300 points if they made no mistakes (i.e., no false positives or wrong answers).

[6] During the user study, we reset the high score for each participant. Thus, it only showed their own high score.

The achieved score also updates the player's *career level* (see Chapter 4.3). However, while levels can be played repeatedly, the updates to the career level score are only applied the first time that a level is played. That way, the career level is a good representation of how well a player's initial performance was.

The initial score is used for the narrative career progression.

4.4.5 Logs

While the primary objective of our dark pattern learning game is to educate players to detect dark patterns better and increase their confidence when encountering them, the game also offers good opportunities for further research. To this end, we built an extensive logging system into the game.

The game logs data for further research opportunities.

Especially when hosted and publicly available, this can be used as a tool to collect large datasets regarding susceptibility towards certain dark patterns. It can show what dark patterns people have trouble recognizing and what dark patterns pose little problems.

It can be used to create a large data set regarding susceptibility.

Among others, the game logs what elements players select, the category they choose, the dark patterns they have missed, when and how they switch between pages, and when and for how long they look at the help tooltips for the categories. All data is logged with a timestamp and associated with a random user ID.

In its current version, log files are created locally, but we designed the logging system in such a way that it can be easily exchanged for a database to run unsupervised on a server and support multiple users.

Chapter 5

Dark Pattern Learning Game User Study & Evaluation

In this chapter, we describe the second user study to evaluate our dark pattern learning game prototype that we presented in Chapter 4. In Chapter 5.1, we describe the design of the study. We present the results of the user study and a discussion of these results in Chapter 5.2.

5.1 Methodology

The overall goal of this study is to evaluate the usability of the dark pattern learning game prototype and determine if it is suitable as a learning game for dark patterns. More precisely, we aim to answer the following research questions:

This study aims to test usability and measure learning effects and gains in players' confidence.

RQ-1: Does player performance improve after playing the game? Is there a measurable learning effect from playing the game?

RQ-2: Do players gain confidence in detecting dark patterns from playing the game?

RQ-3: Is playing the game once sufficient? How pronounced is the knowledge retention?

Measuring long-term learning effects is beyond the scope of this thesis.

RQ-1 and RQ-2 can be answered with a single user study. However, to measure any long-term learning effects for RQ-3, we designed our study to include a voluntary follow-up user study a couple of months later that measures knowledge retention. We outline a proposed study design for this follow-up study in Chapter 5.2, but due to the time constraints of this thesis, the execution and results of the follow-up study are beyond the scope of this work.

5.1.1 Study Design

We use a within-subjects study design and randomized the order of levels.

This chapter describes the design of the first study. The study was designed as a within-subjects experiment where each participant played the complete game. The order of the levels was randomized. Half the participants started with the hotel-themed group of levels, and the other half started with the smartphone-shop-themed group of levels. However, the order of levels inside each group remained the same.

Questionnaire

The questionnaire is integrated into the learning game prototype.

The questionnaire is divided into two parts with the informed consent form and a demographics questionnaire at the beginning of the study and a two-part questionnaire assessing the tutorial and game after the study. We integrated the questionnaires into the learning game so participants could fill them out within the browser window. Screenshots of all parts of the questionnaire are available in Appendix B.1.

The demographics questionnaire consists of questions regarding age, gender, and current occupation or field of study. It also includes the time spent online in hourly increments. Additionally, it includes the choice to be contacted for the voluntary follow-up study.

The post-study questionnaire for the tutorial was designed to check if the information content in the tutorial was adequate for playing the game. It consists of three 5-point Likert scale questions (from ”strongly disagree” to ”strongly agree”) and three free-text fields to express where participants would have wanted more information, less information, and leave further comments.

The post-study tutorial questionnaire checks if the amount of information is adequate.

The final part of the post-study questionnaire focuses on the game itself. Participants rated the same set of questions that we had used in the preliminary study, such as the clear goal, fun, or frustration, on a 5-point Likert scale (from ”strongly disagree” to ”strongly agree”) as suggested by Olsen et al. [2011]. Additionally, we added three more questions: that playing the game increased their knowledge about dark patterns, their confidence in detecting dark patterns, and their confidence in handling dark patterns. We provided four free-text fields for strengths and weaknesses of the game, suggestions for improvements, and further comments.

The post-study game questionnaire checks general usability issues and changes to the players' confidence about dark patterns.

Pre-test/post-test

To measure if there are learning effects from playing the game, we designed the user study with a pre-test/post-test design. This is an experimental approach where the same measurements are taken before and after applying some treatment [Campbell and Stanley, 2015, Knapp, 2016]. Before playing the dark pattern learning game, we tested how well participants performed when presented with potentially manipulative websites. Without telling them the results, they would then play the game, and afterward, we performed the same test as in the beginning with some additional websites to assess if there was any learning effect from playing the game.

We use a pre-test/post-test design to measure learning effects from playing the game.

For the pre-test/post-test, we used the same approach as Bongard-Blanchy et al. [2021] in the second part of their user study. We showed participants the image of the website for a fixed, limited amount of time and asked them afterward if they noticed any manipulations. Participants

Participants were presented with a picture of a website and had to decide if and why it is manipulative.

selected a yes-or-no radio button and justified why they believed the website was manipulative.

We use ten images from an existing study and five of our own.

For the pre-test, we used the same ten websites that Bongard-Blanchy et al. [2021] used in their study. We randomized the order of images for each participant. For the post-test, we tested the same ten websites in a different order again, combined with an additional five websites using a style and design similar to the initial ten websites. These additional websites are used to check for learning bias.

Procedure

We integrated the study setup into the game artifact.

In order to make the study and the evaluation more convenient, we integrated the demographics, questionnaire, and pre-test and post-test into the game prototype. Apart from signing the informed consent form, participants could perform the whole study within a single browser window.

We added an option for players to justify their decisions.

We then performed a test run of the study and adjusted the game prototype slightly based on the results. Most importantly, we included a text field to the feedback overlay during regular gameplay. This text field is only visible for incorrect answers and allows players to justify their thought process. Apart from that, we made the unlocking of pages more explicit and slightly adjusted the instruction texts to be more precise.

Participants played the game at their own pace.

The study was conducted in person, with participants performing the study on a computer with the screen mirrored to observe their actions. The study took between 60 and 90 minutes per participant. We did not limit or restrict their time, nor did we prevent participants from replaying a level if they chose to. In that case, we only considered their first attempt at the level for the evaluation.

The study starts with the demographics questionnaire and pre-test.

The study began by signing the informed consent form and filling out the demographics questionnaire. From there on, the game would guide the participants through each step: Firstly, participants performed the pre-test with the ten images of potentially manipulative websites.

Once completed, they proceeded to the game, where only the tutorial was unlocked. Upon completing both parts of the tutorial, the two groups of levels would unlock. Here, we asked them to play them in the order that they were shown, which was randomized for each participant.

Participants play the tutorial and all levels.

Once the participants had completed all five levels, we initiated the post-study questionnaire about the tutorial and the game. After finishing it, the game guided the participant to the post-test, which mirrored the pre-test but with 15 images of potentially manipulative websites.

The study concludes with the second questionnaire and post-test.

As we did in the preliminary study, we did not ask our participants to do think-aloud to avoid the additional cognitive burden [Zhang and Zhang, 2019, Olsen et al., 2011]. If there was anything that we took note of during the study and wanted to clarify, we asked the participants afterward to explain their actions.

5.1.2 Follow-up Study

The informed consent form asked participants if they would be willing to be contacted for a voluntary follow-up study a couple of months later to check for long-term knowledge retention (RQ-3). Twenty participants agreed to this.

20 participants agreed to participate in the follow-up study.

Our proposed design for this follow-up study is similar to the one used by Röpke [2023] in the domain of phishing. Our plan is for participants to perform a third version of the pre-test/post-test setup. They will be presented with images of potentially manipulative websites and have to classify them and justify their decision. To be comparable to the pre-test/post-test, it will include the same images plus some additional images to check for learning bias (like in the post-test).

Participants will perform a third version of the pre-test/post-test.

A questionnaire will assess changes in their behavior since the first study.

Additionally, we plan to include a short questionnaire in the follow-up study on how playing the game the first time influenced the participants in their everyday lives. For instance, if they recognized more dark patterns since then[1].

It might also be interesting to let participants play some additional levels of the game to see if their game performance also improved.

This version of a follow-up study wouldn't even have to be conducted in person. In this case, the logging system (see Chapter 4.4.5) needs to be adjusted to support multiple users and sessions.

5.2 Results & Evaluation

5.2.1 Demographics

22 university students participated in the study.

We conducted the study with 22 university students (12 female, 9 male, 1 divers). The participants were between 20 and 35 years old (M=25.0, SD=3.86).

15 participants had a technical background (computer science, engineering), and 7 had a non-technical background (economics, musicology). Eleven participants had a high school diploma as their highest academic degree, four had a Bachelor of Science, three had a Bachelor of Arts, and four had a Master of Science degree.

Participants reported spending an average of 4.86 hours online (SD=2.17).

[1] This is certainly something that we observed with friends and colleagues who regularly come to us to tell us about the newest dark patterns that they have spotted.

5.2.2 Data Analysis

Similar to the preliminary study, we aggregated the quantitative and qualitative data from different sources. This includes the questionnaires, log data from the learning game, and our notes during the study.

First, we will analyze the quantitative data from the questionnaire and log files described in Chapter 4.4.5. Afterward, we will analyze the qualitative data from the logs and study notes.

Quantitative Analysis

For the quantitative analysis, we used Numbers spreadsheets for descriptive analysis and R[2] to test for significance.

Pre-test/post-test Firstly, we look at the performance of players from the pre-test/post-test. Here, we compare the percentage of correct answers in the pre-test to the percentage of correct answers in the post-test.

The learning bias is negligible.

First, we checked for learning bias. We calculated the mean of correct answers for the pre-test answers (M_{pre}=0.73, SD=0.16), the combined (i.e., for all 15 images) post-test answers (M_{post}=0.87, SD=0.06), the post-test answers that are part of the pre-test ($M_{post\text{-}pre}$=0.88, SD=0.1), and the new post-test answers ($M_{post\text{-}new}$=0.86, SD=0.11). The results (see Figure 5.1) 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, the learning bias should be negligible.

There is a significant learning effect in the performance between the pre-test and post-test.

To determine if there are any learning effects, we performed a paired t-test for the correctly identified percentages between the pre-test and the post-test. We checked for a normal distribution for the differences

[2]https://www.r-project.org [Accessed: Feb. 9, 2024]

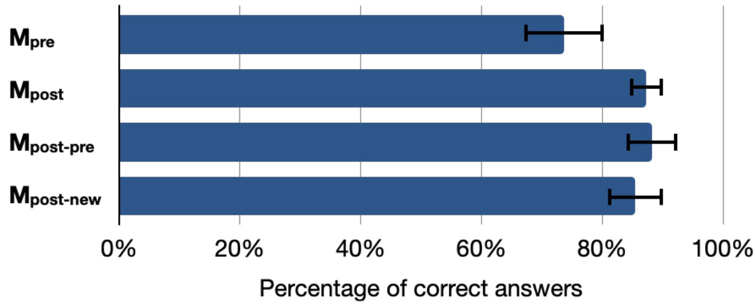


Figure 5.1: This figure shows the mean percentage of correct answers and the 95% confidence interval for the pre-test (M_{pre}), post-test (M_{post}), pre-test images used again in the post-test ($M_{post\text{-}pre}$), and new post-test images ($M_{post\text{-}new}$).

between the pre-test and post-test with a Shapiro-Wilk test (p-value > 0.28). The results of the paired t-test show a significant difference between the pre-test and post-test results ($t(22) = -5.22$, p-value = 0.00005, degrees of freedom (df) = 21, Cohen's d = 1.08).

Participants improved on average over 13%.

On average, the participants' performance improved in 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. Figure 5.2 shows these diminishing learning effects based on the number of correct answers in the pre-test.

Participants detected ∼2/3 of dark patterns and categorized more than 3/4 of them correctly.

Game performance Secondly, we have a look at the performance during the game. There are a total of 28 dark patterns divided across four levels[3]. On average, participants missed 33.12% of dark patterns (M=9.27, SD=2.05) and found 66.88%. They categorized 77.67% correctly (M=14.55, SD=3.32) and 22.33% incorrectly (M=4.18, SD=2.32). The average number of false positives was 2.23 (SD=1.19).

[3] One of the five levels includes no dark patterns.

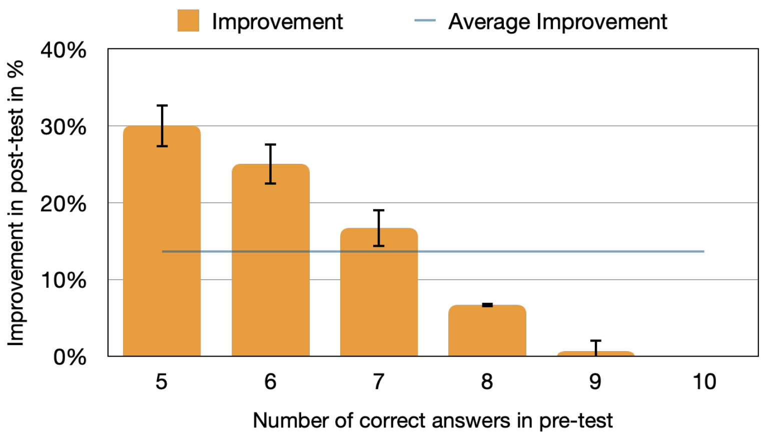


Figure 5.2: This figure shows the average improvement and 95% confidence interval of participants between the pre-test and post-test based on the number of correct answers in the pre-test.

Apart from two exceptions, performance in the individual levels is similar.

The performance for each level is available in Table 5.1. The hotel-themed levels took longer to complete than the smartphone-shop-themed ones. However, the hotel-themed levels are also more content-rich, and some span over multiple pages. The percentage of dark patterns found is similar for three levels (M>67%) and only the hotel listing page is noticeably lower (M=55.3%). The number of false positives is similar for three levels. The hotel newsletter unsubscribe level had no false positives at all (but used a more minimalistic design with fewer elements to select) and the second smartphone shop design had a very low value (M=0.09, SD=0.29). However, this level is very similar to the previous one with different pre-selections.

	Mean Time (SD)	**DP**	**Found**	**FP (SD)**
Hotel Listing Page	4.22 m (SD=2.4)	6	3.32 (55.3%)	0.68 (0.84)
Hotel Booking Page	5.09 m (SD=1.65)	8	5.41 (67.6%)	0.73 (0.94)
Hotel Newsletter	4.33 m (SD=1.7)	8	5.59 (69.9%)	0 (0)
SmartShop Design A	1.79 m (SD=0.98)	0	n/a	0.73 (0.63)
SmartShop Design B	2.31 m (SD=0.73)	6	4.41 (73.5%)	0.09 (0.29)

Table 5.1: This shows the mean time participants took, the number of dark patterns (DP), the mean total number and percentage of found dark patterns, and the total number of false positives (FP) per level.

Differences between participants with technical and non-technical backgrounds are minimal.

There are only minimal differences between participants with technical and non-technical backgrounds. Participants with a non-technical background missed slightly more dark patterns (0.64, about 2% difference). The exact differences are available in Table 5.2. This matches the results by Voigt et al. [2021] that there is no connection between a person's affinity for technology and their detection rate of dark patterns. However, our sample size of participants with a non-technical background is only seven. Therefore, we can't confidently draw any conclusions if this is generally applicable.

	F+C	**F+NC**	**NF**	**FP**
Technical Background	14.33 (3.49)	4.60 (2.5)	9.07 (1.91)	2.33 (1.35)
Non-technical Background	15.00 (3.11)	3.29 (1.7)	9.71 (2.43)	2.00 (0.82)

Table 5.2: This shows the differences in game performance between participants with technical (*n=15*) and non-technical (*n=7*) backgrounds. It lists the *found and correctly classified* (F+C), *found and incorrectly classified* (F+NC), and *missed* (NF) dark patterns, and the *false positives* (FP). Values are mean values (out of 28 dark patterns total) with standard deviation in parentheses.

Time spent online had little effect on performance.

Similarly, there were only small differences in the detection and classification rate based on the hours that participants spent online. Participants who spent less time online (1-3 hours) found slightly more dark patterns. However, participants who spent six or more hours online performed slightly better in categorizing the dark patterns. The differences in performance are shown in Table 5.3.

	n	**F+C**	**F+NC**	**NF**	**FP**
1 - 3 hours	7	14.58 (3.59)	4.54 (2.76)	8.88 (2.31)	2.00 (1.15)
4 - 5 hours	9	14.29 (3.55)	4.50 (2.28)	9.21 (2.06)	1.89 (0.93)
6 or more hours	6	15.22 (2.95)	2.89 (1.67)	9.89 (2.07)	2.44 (1.63)

Table 5.3: This shows the differences in game performance between participants based on their time spent online. It lists the *found and correctly classified* (F+C), *found and incorrectly classified* (F+NC), and *missed* (NF) dark patterns, and the *false positives* (FP). Values are mean values (out of 28 dark patterns total) with standard deviation in parentheses.

When considering the detection rate for individual dark patterns, the results vary greatly based on the different categories.

Three particular dark patterns were found by all of our participants: the "only x rooms remaining" *low stock* message (that appears in two levels) and a fake *countdown timer* for urgency. Generally, dark patterns from the *social engineering* category were found most often except for *previous price*, which was only found 31.82% of the time. *False hierarchy* was also found in 88.64% of all occurences.

Social engineering and *false hierarchy* dark patterns were detected most precisely.

Dark patterns that were found least often are from the *sneaking* category: *disguised ads* were only found by 40.9% of participants, and a prominent *"Book Now"* button combining *sneaking* and *interface interference* was only found by three participants (13.64%).

Sneaking dark patterns were detected least often.

A complete list of all 28 dark patterns and their respective detection and classification rates is available in Appendix B.2.

Questionnaire results Finally, we present the quantitative data from the tutorial and game questionnaires. All the questions were answered on a 5-point Likert scale from -2 (strongly disagree) to 2 (strongly agree).

The tutorial questionnaire focuses on the amount of information and whether it is adequate for the game. Participants rated that *the tutorial explained everything they needed to play the game* (Q1) at M=1.55 (SD=0.59). They rated that *the information in the tutorial helped them understand the concepts of dark patterns* (Q2) at M=1.64 (SD=0.4) and that *the amount of information in the tutorial was sufficient* (Q3) at M=1.36 (SD=0.66). The results are also visualized in Figure 5.3.

The amount of information in the tutorial was sufficient.

The overall results of the game-related questionnaire are generally positive. Figure 5.4 shows the results of the game questionnaire questions. Participants strongly agreed that the game had a clear goal (Q1, M=1.91, SD=0.29) and agreed that the game was intuitive (Q2, M=1.27, SD=0.55).

The game was rated positively and not too challenging.

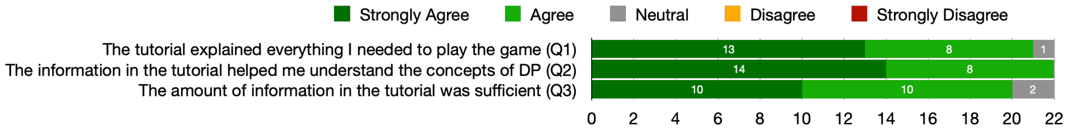


Figure 5.3: This shows the results of the tutorial questionnaire on a 5-point Likert scale from -2 (strongly disagree) to 2 (strongly agree). Participants rated the amount of information in all three aspects very highly.

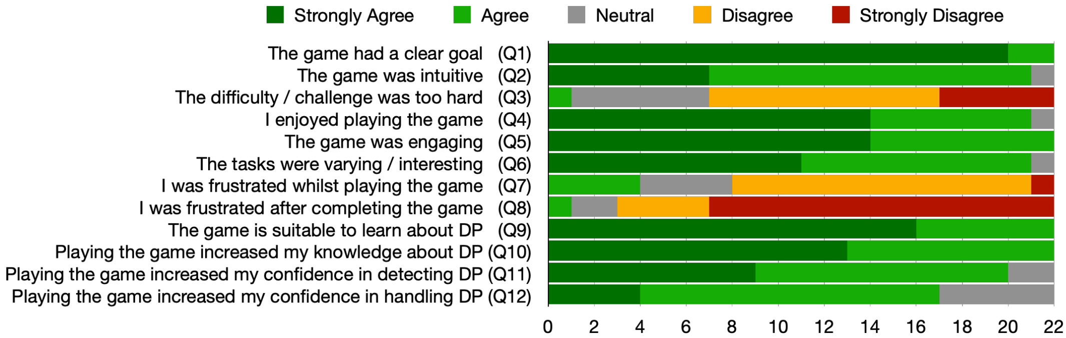


Figure 5.4: This shows the results of the game questionnaire on a 5-point Likert scale from -2 (strongly disagree) to 2 (strongly agree). Participants rated the same nine questions as in the preliminary study. We added three questions regarding knowledge and confidence.

The game was not too challenging (Q3, M=-0.86, SD=0.83) and participants enjoyed playing the game (Q4, M=1.59, SD=0.59). They found the game to be engaging (Q5, M=1.64, SD=0.49) and the tasks to be varying/interesting (Q6, M=1.45, SD=0.6). The frustration during the game was low (Q7, M=0.5, SD=0.86) and lower after completing the game (Q8, M=-1.5, SD=0.86).

The game is suitable for learning about dark patterns and boosts knowledge and confidence.

The participants agreed that the game is suitable for learning about dark patterns (Q9, M=1.73, SD=0.46) and that playing the game increased their knowledge about dark patterns (Q10, M=1.59, SD=0.5). Playing the game increased their confidence in detecting dark patterns (Q11, M=1.32, SD=0.65) and, to a lesser extent, their confidence in handling dark patterns (Q12, M=0.95, SD=0.65).

Qualitative Analysis

For the qualitative analysis, we coded the responses to the free-text questions in both questionnaires and our notes during the study in MAXQDA and performed a thematic analysis on them.

> *"It makes learning about dark patterns fun!"*
>
> *—Participant 14*

Participants liked the gameplay of spotting dark patterns.

Feedback from our participants shows that they really enjoyed playing the game. They liked looking for dark patterns on the website and the *dark pattern magic wand* as the tool to select dark patterns. They also liked the realistic designs of the websites[4] and that they could interact with them.

Participants liked the setting of the game as well. Although we used narratives sparingly, it helped to set the tone of the game.

Participants liked the career ladder as motivation.

The *career ladder* was also very well received. Participants liked the motivation to reach a higher career level (*"I need to get 600 more points in the last level to make it." - Participant 5*) and we also observed participants afterward comparing scores. Some participants suggested extending it further with in-game notifications if they reached a new career level or a generated certificate of their achieved score to share on social media.

Some participants disliked categorizing the dark patterns.

While the task of finding dark patterns was generally enjoyed, many participants disliked categorizing dark patterns either in general (*"Deciding the category is [...] annoying." - Participant 21*) or because they felt that their choices were correct but the game said otherwise (*"I can be a bit frustrating to have a different view on what pattern applies to the situation." - Participant 19*).

[4]They are, after all, inspired by real websites.

"By showing missed dark patterns and always providing explanations, it makes it easier to understand why certain things are (not) dark patterns"

—Participant 18

Participants liked the instant feedback.

Participants liked the feedback system a lot. Although they did not always agree with the results, they liked the instant feedback and the explanation, which helped them understand the concepts better.

"A lot of information to take in at the beginning"

—Participant 9

The tutorial can be further improved.

About a quarter of our participants felt that the tutorial was a bit too daunting. The amount of information was either too much to comprehend in a short time, or the delivery of the information was too boring. Suggested improvements were references to the tooltips so that new players will know that they don't have to remember everything by heart and more interactivity, such as multiple-choice questions after each section of the tutorial to check one's knowledge.

5.2.3 Discussion

In this section, we discuss the results of the study to answer our research questions.

Effects of the game on player's performance

Participants' performance increased after playing the game.

Firstly, we want to discuss the changes in performance upon playing the game for RQ-1. As described in Chapter 5.2.2, there is a significant improvement in performance between the pre-test and post-test, with an average increase in performance of 13.64%. However, the increase in performance was more pronounced the worse the participants performed in the pre-test. It makes sense that someone who was already quite good at spotting dark patterns would gain less performance from playing the game than someone who had more trouble spotting dark patterns.

This improvement was also noticeable within the game. In the tutorial, we intentionally included some of the more difficult-to-find dark patterns, such as *disguised ads*. When players first encountered this dark pattern in the CLASSIFY-tutorial, only 27% correctly identified it. During the regular gameplay, 41% found and correctly identified it.

Participants' performance improved during the game.

When we consider that most of our participants played the game for only about half an hour, this increase in performance is very promising. Additionally, performance further increased for recurring instances of dark patterns (e.g., *interface interference* in the cookie banner). Therefore, more levels that showcase even more variations of dark patterns could further boost performance and help players if they encounter any of these dark patterns in the real world.

Recurring dark patterns were recognized better.

Effects of the game on player's confidence

To assess the confidence changes of players after playing the dark pattern learning game for RQ-2, we have the self-reported metrics from the post-game questionnaire as well as our observations during the study.

The results from the questionnaire show that players felt that playing the game increased their confidence in detecting dark patterns (M=1.32, SD=0.65, on a scale from -2 to 2). This certainly matches what we witnessed during the study. Especially with some of the recurring dark patterns (such as *false hierarchy* or *low stock messages*), players did not only select them much quicker. They also picked the categories more confidently (often without checking the tooltips again).

Participants' confidence increased from playing the game.

Participants used more concise terminology after the game.

This confidence and knowledge gain was also apparent in the post-test answers. We already presented the improvements in Chapter 5.2.2. In addition, about 2/3 of our participants were more concise in their answers and often even used the proper terminology (e.g., calling it *confirmshaming* or *interface interference*) in their justifications.

Observations during gameplay

One participant didn't want to view their mistakes.

Differences in Playstyle When we designed the learning game prototype and the feedback system, we considered two different ways that players can finish a level: either viewing what they had missed or trying again without viewing their mistakes (see Chapter 4.4.3). During the study, only one participant opted for the latter. When we asked them about it, they answered that they wouldn't learn from it if the game just told them the missing elements and that they wanted to figure it out on their own.

Some participants relied less on tooltips than others.

Another interesting observation regarding the playstyle was about the use of tooltips. While three participants did not find the tooltips at all[5], all other participants used the tooltips to various extents. Most used them to check a certain category if they weren't sure. They would read the tooltip, sometimes compare it to another tooltip, and make their selection. The longer they played the game, the less frequently they checked the tooltips.

However, four participants relied more heavily on the tooltips. They would check the tooltips for each category for every dark pattern they selected and check which one matched best. It was especially interesting that they would, most of the time, begin with the category that was correct. So their instincts were correct, but they didn't want to make a hasty decision.

The order of levels had little effect on the performance.

Order of levels The order of levels seems to have little effect. We assumed that there would be more of a learning effect early on, but the results don't reflect that. On

[5]This is something we plan to address in the future in the tutorial.

the contrary, the participants performed slightly better in the early levels. Yet the difference between starting with the hotel-themed levels or the smartphone shop levels was less than half a dark pattern for each level. The only exception was the second smartphone shop level, where participants playing it first performed worse (-1.5 dark patterns compared to playing it second). However, the missed dark patterns were those using interactive elements (expensive pre-selections) that players may have missed because they did not yet realize that they could freely interact with those elements. Thus, we attribute this difference more to a shortcoming of the game and less to a learning effect.

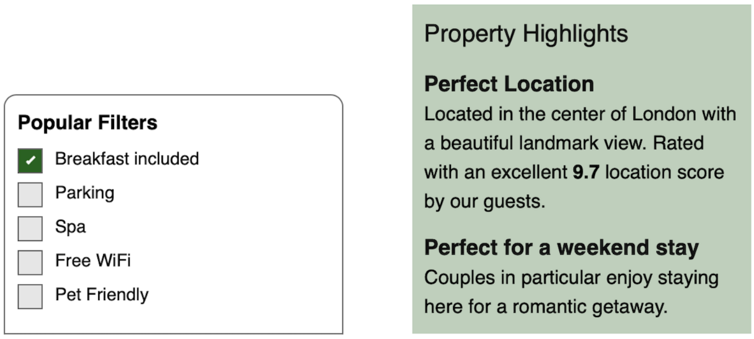
Social engineering and *false hierarchy* were detected most often.

Susceptibility As described in Chapter 4.4.5, the logs during gameplay offer great insights into susceptibility towards certain dark patterns. Even though we performed the study with a relatively small group of participants (*n=22*), it already shows some trends. In the following, we discuss the susceptibility towards particular dark patterns in more detail. *Low stock messages* (e.g., "only one room left") and *countdown timer* were correctly detected and categorized by all our participants. Similarly, *false hierarchy* dark patterns were found in 90% of all occurrences.

The dark pattern that was least often found in our game was a *"Book Now"* button that included the pre-selection for the most expensive hotel room available. Only three participants found this instance of a dark pattern.

There were few false positives.

Overall, there were few false positives across all five levels (M=2.23, SD=1.19). There were three elements that multiple participants picked as dark patterns: One is a preselected checkbox in the sidebar of the hotel listing website for included breakfast (Figure 5.5a). We did not mark this as *interface interference* because it is a pretty common request and not really to the advantage of the website or the disadvantage of the user. However, we can understand why some participants picked this as a dark pattern and might adjust it accordingly. The second frequently picked false positive was a box with information about property highlights on the hotel booking website (Figure 5.5b). Some



(a) The sidebar had one popular filter *breakfast included* pre-selected.

(b) An info box at the side listing only positive things about the property.

Figure 5.5: The two most frequently picked elements that we had not marked as dark patterns.

participants mistook it for social engineering (i.e., listing the positive things about the property).

We actually expected more false positives in the one clean level without any dark patterns (especially since this was the first level for half of our participants). While some participants selected the trade-in option for various reasons (e.g., for the wording "save *up to* 120 €"), the number of false positives was not higher compared to other levels (see Table 5.1). Furthermore, there was no difference in the number of false positives whether participants played this as the first or fourth level.

Changes to the ontology better reflect our participants' conceptions.

Misconceptions When we designed the learning game prototype and the user study, the ontology by Gray et al. [2023], which we used for our categories, was only available as a draft version. In the meantime, a paper presenting an updated ontology [Gray et al., 2024] has been accepted at CHI 2024 and has already been made available. The most noteworthy change is that the high-level category *nagging* has been moved to the meso-level of *forced action*. This is an interesting change as it reflects what we observed during the study. If participants wrongly categorized our *nagging*-alert, they often chose *forced action* instead. Hence, this change to the ontology seems to match the mental model of our participants.

There were two more common misconceptions: Firstly, many participants picked *obstruction* instead of *interface interference* in multiple instances. For some *bad defaults* preselections, they picked *obstruction* because they had to deselect checkboxes manually (i.e., imposing additional steps). Similarly, for *trick questions*, they also picked *obstruction* because they first had to parse the true meaning of the question. Secondly, *pressured selling* (*interface interference*) was commonly categorized as *sneaking* because participants believed that this was hiding information from them.

Some *interface interference* elements were considered *obstruction* or *sneaking*.

Differences to the preliminary study The post-study questionnaire about the game included the same nine questions that we used in the preliminary study for each game mechanic. Since our game prototype builds upon MULTISPOT, we looked at the differences in rating between the preliminary study and the learning game.

Generally, the results are similar with three exceptions: In the preliminary study, participants rated higher that the game was intuitive (0.48 difference). In the game prototype, participants rated that tasks were varying/interesting higher (0.45 difference). In the game prototype, participants rated frustration during the game higher (1.08 difference).

The ratings from the game prototype only differ slightly from those of the preliminary study.

The difference in intuitive gameplay might be explained by the setup of the preliminary study. Participants could just circle any arbitrary element on a picture. In the learning game prototype, they had to hover over containers and could only select the ones that were highlighted. This could also explain the difference in frustration. Furthermore, the preliminary study did not include the results of the categorization. Based on our observations, this was a contributing factor to frustration, especially if participants were undecided between two categories and picked the wrong one.

Selection and categorization changed from the preliminary study.

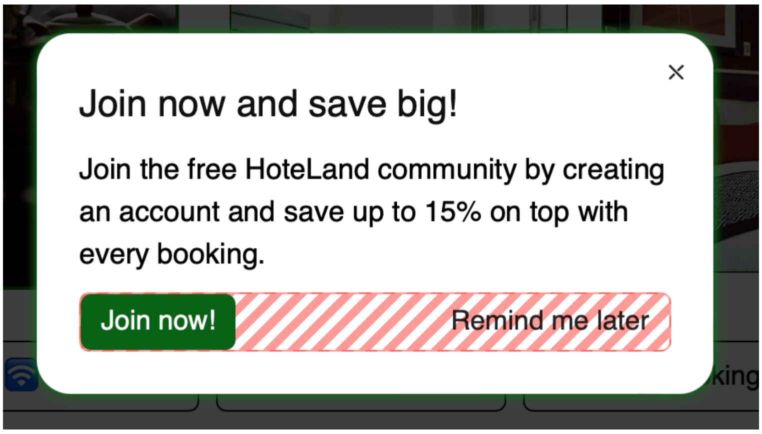


Figure 5.6: One example of nested dark patterns. The inner one (marked in red) is *interface interference*; the outer one is *nagging*.

The fact that the game prototype was rated higher regarding the varying/interesting tasks might have to do with the different levels and added interactivity.

Limitations & Roadmap

Nested dark patterns are not yet handled by our prototype.

Nested dark patterns Probably the biggest limitation of the game results from nested dark patterns. For example, the *nagging* alert illustrated in Figure 5.6 has two buttons: *"Sign up"* and *"Remind me later"*. In addition to it being *nagging*, the buttons are designed differently, with the *"Sign up"* button more prominent, making it *interface interference*, too.

The way the game is designed, this scenario is straightforward: the outer container of the alert is *nagging*, and the inner container containing both buttons is *interface interference*. However, in this example (as used in the second hotel level), six participants picked the inner container as *nagging* as it contained the *"Remind me later"*-text. They had identified the dark pattern correctly, but the container system was not flexible enough to handle this. To make matters worse, they could no longer select the buttons as *interface* *interference* (assuming they wanted to) because the game already marked the inner container as (wrongly) categorized.

This certainly needs to be adjusted so that the game can handle correct intentions even if the selection wasn't completely precise. The game already includes a mechanism to support the same nested dark pattern (e.g., the *"sponsored"* text is linked to the table cell that is sponsored content; selecting one automatically selects the other one, too). Extending this mechanism to support different nested dark patterns should also be possible.

Tutorial We received some suggestions to improve the tutorial. For instance, multiple-choice questions during the taxonomy should be included to directly quiz players on what they have just learned. This would make the initial amount of information more interactive and less daunting.

The tutorial could be more interactive.

A bigger issue resulted from the game explanation. This was part of the initial tutorial and explained how to play the game and how to use the *dark pattern magic wand*. However, the next thing participants had to play was the CLASSIFY-tutorial, which looks just like the regular game but without the *magic wand*. This confused all our participants, who first spent some time looking for the missing button. We plan to adjust the tutorial to explain certain elements gradually. In this case, it will only explain the gameplay necessary for CLASSIFY and afterward for the regular game.

The structure of the tutorial could be improved.

The last improvement to the tutorial is based on the performance in the first hotel-themed level for participants starting with this level group. Participants missed more dark patterns, especially the more subtle ones. Therefore, we plan to include a regular MULTI-SPOT level in the tutorial so players can practice the game once and see how strict the scoring of the game is.

Suggested Improvements Another common feature requested by our participants was more levels. This is certainly a good sign that participants enjoyed playing the

Participants want more levels.

game and wanted to keep playing. Especially since we have seen increased detection rates for recurring dark patterns, more levels that showcase even more instances of dark patterns should be very beneficial.

The scoring system could be upgraded to show individual strengths and weaknesses.

There were two more suggested improvements that we feel, although only suggested ones each, are worth considering to be implemented. Firstly, one suggestion was to split the score into multiple scores. For instance, one score is for detecting dark patterns, and another score is for categorizing them. That way, it is more obvious where the player's strengths and weaknesses are. This could even be extended to individual categories to show which categories the player can better recognize.

Secondly, one suggestion was that upon completing the game, it should generate a certificate with the final score and career level. This could then be shared on social media to help further increase awareness and motivate more people to participate in the game.

Chapter 6

Summary and Future Work

In this chapter, we conclude our research by providing a summary of our work and present opportunities for future work.

6.1 Summary and Contributions

This thesis explored intervention measures against dark patterns from the direction of user education. We presented the process of designing a learning game against dark patterns in three steps:

Firstly, we provided the groundwork by exploring suitable game mechanics for a dark pattern learning game. We brainstormed three game mechanics on different levels of cognitive difficulty on Bloom's Revised Taxonomy and evaluated them in a user study. The results show that while all three of them were suitable and liked, the most realistic one, where players had to find multiple dark patterns on a website and categorize them, was preferred by our participants.

We first explored which game mechanics are suitable for a dark pattern learning game.

We presented the design process of our prototype.

Secondly, we described the design of a learning game prototype based on the results of the preliminary study. We described the decisions that went into creating the individual components of the game.

We evaluated our prototype in a user study and measured its learning effects.

Lastly, we evaluated the learning game prototype in a second user study. We measured the learning effect from playing the game with a pre-test/post-test design. The results show that participants performed significantly better in detecting dark patterns after playing the game. Participants gained increased confidence by playing the game and were more decisive in their actions. Furthermore, people enjoyed finding the dark patterns more than categorizing them.

6.2 Limitations & Future Work

We only tested the game with university students.

One limitation is that we only evaluated the learning game prototype with university students. There might be differences in the perception of dark patterns across different age groups. Bongard-Blanchy et al. [2021] noted in their study that older generations are less aware of possible manipulations and, thus, more likely to fall for dark patterns. Similarly, Sahabi [2023] showed that the mental model of children regarding manipulative designs differs from those of adults. There might also be changes in the perception of manipulative designs for people of different demographics. Furthermore, the majority of our participants had a technical background. While their results only differ marginally from those with a non-technical background, this should be checked with a larger sample size to verify if it is generally applicable.

We focused on the underlying game mechanics and did not create a full game.

Another limitation is that we focused our learning game prototype on the core game mechanics from the preliminary user study. We added some other game mechanics, such as timers and high scores, and some immersive elements like the narrative. However, some other game aspects and learning aspects still need to be included. For instance, a learning game would be more effective if it adapted to the player's abilities and focused more on where the player made errors [Plass et al., 2015].

Furthermore, we developed the learning game prototype primarily to measure its learning effects. It certainly could be more polished. Its current version is not yet optimized for smaller screen sizes or touchscreens. Both would be useful in extending its reach.

The prototype could be more polished.

Therefore, one part of future work could focus on improvements to the learning game prototype itself. We have outlined suggested improvements and technical changes in Chapter 5.2.3 that would be beneficial to implement. More importantly, it would be good to add more levels and more instances of different dark patterns, especially since participants performed better in recognizing recurring dark patterns. In its current version, the game prototype consists only of two groups of levels within the same kind of context: a desktop website. Additional levels could further explore other contexts: mobile websites, smartphone apps, or games. This would offer opportunities to study dark patterns across different modalities and build upon the works of van Nimwegen and de Wit [2022].

The game prototype can be further improved.

Additional levels could cover other modalities.

Moreover, deploying the game online will offer even more opportunities for future work. Firstly, this would allow to verify the accuracy of our collected data with a larger sample size and more diverse demographics. Secondly, future work could build upon the existing logging system to collect large amounts of data in regard to the susceptibility of dark patterns.

Making the game publicly available offers opportunities for large-scale data collection.

Appendix A

Game Mechanics Study

In the following, we include additional material used in the preliminary user study on the suitability of the three game mechanics. Furthermore, we present more of the raw data used in the evaluation.

A.1 Questionnaire

The following questionnaire was used in the preliminary user study. We folded the first page in half so participants would only fill out the first half until they were introduced to the topic of dark patterns.

Participant: ________

Demographics

Age: ___________

Gender: ________________

Current Occupation / Field of Study: ___________________________

Last achieved academic degree (e.g., high school, Bachelor's): _____________________

How much time do you typically spend browsing the web / social media each day?

O less than 1 hour
O 1 - 3 hours
O 3 - 5 hours
O more than 5 hours

On a scale from 1 to 5, how experienced are you with Dark Patterns?

O 1 (no experience at all)
O 2
O 3
O 4
O 5 (very experienced)

On a scale from 1 to 5, how confident are you that you can spot Dark Patterns?

O 1 (not confident at all)
O 2
O 3
O 4
O 5 (very confident)

Figure A.1: Preliminary user study questionnaire. We collected some demographics at the start of the user study as well as self-reported metrics on daily internet and social media usage and expertise regarding dark patterns.

Participant: ________

{Game Mechanic} Game

Please rate the following statements on a scale from 1 (strongly disagree) to 5 (strongly agree).

	1 (strongly disagree)	2	3 (neutral)	4	5 (strongly agree)
The game had a clear goal	○	○	○	○	○
The game was intuitive	○	○	○	○	○
The difficulty / challenge was too hard	○	○	○	○	○
I enjoyed playing the game	○	○	○	○	○
The game was engaging	○	○	○	○	○
The tasks were varying / interesting	○	○	○	○	○
I was frustrated whilst playing the game	○	○	○	○	○
I was frustrated after completing the game	○	○	○	○	○
The game is suitable to learn about Dark Patterns	○	○	○	○	○

What are the strengths of this game?

What are the weaknesses of this game?

Further Comments

Figure A.1: Preliminary user study questionnaire [continued]. Participants filled out this questionnaire after playing each game mechanic (CLASSIFY, SINGLE-SPOT, and MULTI-SPOT).

Participant: ________

Overall

Please rank all three games.

	Classify	Single-Spot	Multi-Spot
Challenge (1 = most challenging, 3 = least challenging)	◯	◯	◯
Fun (1 = most fun, 3 = least fun)	◯	◯	◯
Frustration (1 = least frustrating, 3 = most frustrating)	◯	◯	◯
Satisfaction (1 = most satisfying, 3 = least most satisfying)	◯	◯	◯
Overall (1 = best, 3 = least best)	◯	◯	◯

What was the main reason why you ranked your overall best choice this way?

(Assuming feedback during gameplay,) which game do you believe can teach you the most?

- O Classify
- O Single-Spot
- O Multi-Spot

Justify your answer.

Figure A.1: Preliminary user study questionnaire [continued]. After completing all three game mechanics, participants ranked them in five categories and picked the one best suited to teach them about dark patterns.

Participant: ________

On a scale from 1 to 5, how confident are you that you have correctly spotted the Dark Patterns?

- O 1 (not confident at all)
- O 2
- O 3
- O 4
- O 5 (very confident)

On a scale from 1 to 5, how confident are you that you have correctly identified the Dark Patterns?

- O 1 (not confident at all)
- O 2
- O 3
- O 4
- O 5 (very confident)

Figure A.1: Preliminary user study questionnaire [continued]. Additionally, they rated their self-perceived performance during the games.

A.2 Quantitative Results

In Table A.1, we present the quantitative results from the questionnaire of the preliminary study. It shows the arithmetic mean and the standard deviation for every question.

	CLASSIFY	SINGLE-SPOT	MULTI-SPOT
The game had a clear goal	5.00 (SD=0)	4.75 (SD=0.45)	4.75 (SD=0.45)
The game was intuitive	4.75 (SD=0.45)	4.42 (SD=0.67)	4.75 (SD=0.45)
The difficulty / challenge was too hard	1.50 (SD=0.52)	1.42 (SD=0.51)	2.00 (SD=0.6)
I enjoyed playing the game	3.83 (SD=0.72)	4.00 (SD=0.43)	4.42 (SD=0.79)
The game was engaging	3.58 (SD=0.79)	3.92 (SD=0.67)	4.33 (SD=0.78)
The tasks were varying / interesting	3.50 (SD=0.9)	3.75 (SD=0.87)	4.00 (SD=0.85)
I was frustrated whilst playing the game	1.50 (SD=0.67)	1.42 (SD=0.67)	1.42 (SD=0.51)
I was frustrated after completing the game	1.67 (SD=0.39)	1.25 (SD=0.62)	1.25 (SD=0.45)
The game is suitable to learn about Dark Patterns	3.67 (SD=0.65)	3.92 (SD=0.9)	4.33 (SD=0.46)

Table A.1: Quantitative data from the questionnaire for each game mechanic. Every question was rated on a 5-point Likert scale from ″strongly disagree″ (1) to ″strongly agree″ (5). Displayed are the mean value and the standard deviation.

A.3 Image Assets

We used the following royalty-free images from Pixabay to create our study designs. While their license allows us to use the images free of charge and without giving credit, the study designs only truly came to life thanks to those assets. And we want to credit each of the artists for their work:

- hawkHD: Smartwatch[a]
- lequangutc89: Hotel Room[b]
- ManuelaJaeger: Hotel Room[c]
- peterweideman: Hotel Room[d]
- Pfüderi: Cabin by the Lake[e]
- Pixaline: VR Headset[f]
- thekaleidoscope: Concert[g]
- vlaaitje: Cute Dog[h]
- 3534679: Hotel Room[i]

[a] https://pixabay.com/photos/smart-watch-smartwatch-fitness-889639/
[b] https://pixabay.com/photos/bedroom-indoors-interior-design-bed-6577523/
[c] https://pixabay.com/photos/hotel-hotel-rooms-home-decoration-1749602/
[d] https://pixabay.com/photos/bedroom-interior-design-bed-room-5664221/
[e] https://pixabay.com/photos/lake-mountains-hut-mountain-lake-1681485/
[f] https://pixabay.com/vectors/virtual-reality-game-glasses-2055227/
[g] https://pixabay.com/photos/concert-live-audience-people-crowd-3387324/
[h] https://pixabay.com/photos/puppy-dogs-collie-cute-pet-sweet-2298832/
[i] https://pixabay.com/photos/bedroom-indoors-interior-design-3475656/

A.4 Codebook

The following section contains the codebook [Braun et al., 2019] used for the thematic analysis of the qualitative data in the preliminary study.

Category	Code	CLASSIFY	SINGLE-SPOT	MULTI-SPOT
Strength	Beginner-friendly	8	3	0
	Certainty	3	5	0
	Examples	7	0	0
	Fast pace	8	0	0
	Focus on elements	5	1	0
	Realistic	0	1	14
	Scrutinize over details	0	2	9
	Uncertainty	0	0	11
Weakness	Clear exit[a]	1	6	0
	False positives	0	0	6
	Hard to categorize	1	2	1
	Lack of context	11	0	0
	Overthinking	0	0	3
	Too easy	5	1	0
	Unrealistic	1	3	0

Table A.2: This shows the codes we developed from the preliminary user study. The numbers signify how often the respective code came up for each game mechanic. We grouped the codes into categories for strengths and weaknesses for further analysis.

[a]This was also considered a strength by some participants.

Appendix B

Learning Game Prototype User Study

In the following, we include additional material used in the second user study evaluating the dark pattern learning game prototype. Furthermore, we present more of the raw data used in the evaluation.

B.1 Questionnaire

The following three questionnaires were used in the second user study to evaluate the learning game prototype.

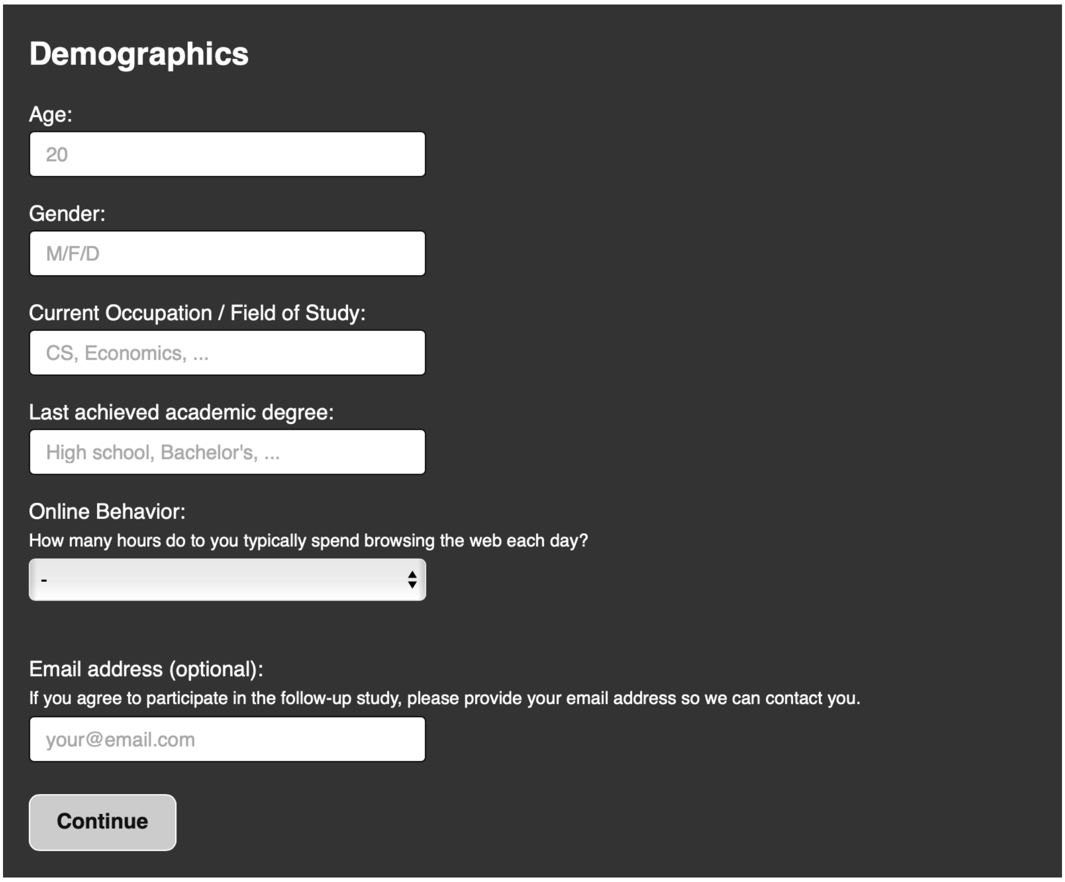
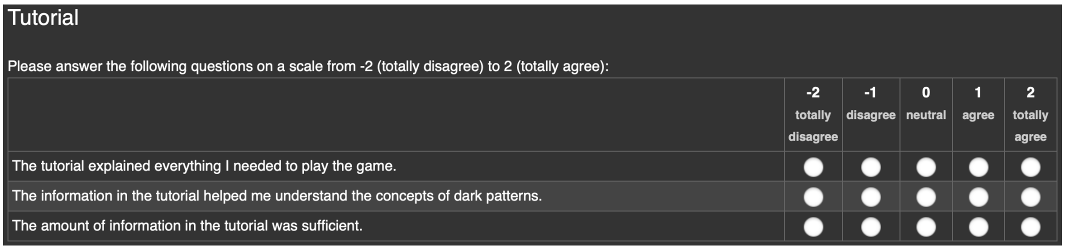


Figure B.1: Participants filled out this first part of the questionnaire before starting the game. It gathers basic demographics and the option to volunteer for the follow-up retention study.


Tutorial

Please answer the following questions on a scale from -2 (totally disagree) to 2 (totally agree):

	-2 totally disagree	-1 disagree	0 neutral	1 agree	2 totally agree
The tutorial explained everything I needed to play the game.	○	○	○	○	○
The information in the tutorial helped me understand the concepts of dark patterns.	○	○	○	○	○
The amount of information in the tutorial was sufficient.	○	○	○	○	○

Figure B.2: Participants filled out this tutorial-focused questionnaire after completing the whole game.

Game

Please answer the following questions on a scale from -2 (totally disagree) to 2 (totally agree):

	-2 totally disagree	-1 disagree	0 neutral	1 agree	2 totally agree
The game had a clear goal					
The game was intuitive					
The difficulty / challenge was too hard					
I enjoyed playing the game					
The game was engaging					
The tasks were varying / interesting					
I was frustrated whilst playing the game					
I was frustrated after completing the game					
The game is suitable to learn about dark patterns					
Playing the game increased my knowledge about dark patterns					
Playing the game increased my confidence in detecting dark patterns					
Playing the game increased my confidence in handling dark patterns					

Figure B.3: Participants filled out this game-focused questionnaire after completing the whole game.

B.2 Dark Pattern Detection and Classification Rate

The following table lists the detection and classification rates for all 28 dark patterns used within the learning game prototype.

Lvl	Dark Pattern	F+C	F+NC	NF
H1	Disguised ad	8	1	13
H1	Expert testimonial	2	6	14
H1	Hidden costs	14	0	8
H1	Previous price	6	1	15
H1	Scarcity 1	12	1	9
H1	Scarcity 2	22	0	0
H2	False hierarchy 1	10	6	6
H2	False hierarchy 2	19	2	1
H2	Hidden information	0	7	15
H2	Nagging alert	10	7	5
H2	Pressured selling	10	2	10
H2	Scarcity 1	22	0	0
H2	Scarcity 2	21	0	1
H2	Sneaking button	2	1	19
H3	Bad defaults	6	11	5
H3	Confirmshaming	12	0	10
H3	False hierarchy	19	2	1
H3	Obstruction 1	7	9	6
H3	Obstruction 2	7	0	15
H3	Obstruction 3	11	7	4
H3	Obstruction 4	16	3	3
H3	Trick question	4	9	9
S2	Confirmshaming	10	0	12
S2	Countdown timer	21	1	0
S2	False hierarchy	18	2	2
S2	Hidden subscription	16	3	3
S2	Pressured selling 1	7	4	11
S2	Pressured selling 2	8	7	7

Table B.1: This shows the detection and classification rates for all dark patterns used in the game. They are grouped by level (H1 = Hotel Listing, H2 = Hotel Booking, H3 = Hotel Newsletter, S2 = Smartshop Design B) and list how often they were *found and correctly categorized* (F+C), *found and incorrectly categorized* (F+NC), and *missed* (NF)

B.3 Image Assets

In addition to the images already used in the preliminary study (see Appendix A.3), we used the following images for the learning game prototype and the pre-test/post-test images.

- Alexas_Fotos: Plush Toy[a], Plush Toy[b]
- Annamos: Hotel Room[c]
- JanClaus: Hotel Room[d]
- Olichel: Hotel Room[e]
- Pexels: Plush Toy[f]
- Pezibear: Plush Toy[g]
- stevepb: Hotel Room[h]
- 12019: Sunset[i]
- 5132824: Workout[j]

[a]https://pixabay.com/photos/stuffed-animal-lion-fun-cute-5202849/
[b]https://pixabay.com/photos/teddy-bear-glitter-eyes-861048/
[c]https://pixabay.com/photos/beach-house-interior-palmetto-coasts-1505461/
[d]https://pixabay.com/photos/to-travel-hotel-room-hotel-room-1677347/
[e]https://pixabay.com/photos/hotel-room-bed-pillows-room-hotel-1447201/
[f]https://pixabay.com/photos/happy-smiling-cuddly-toy-toy-smile-1281590/
[g]https://pixabay.com/photos/teddy-bear-bear-plush-toy-524251/
[h]https://pixabay.com/photos/bedroom-bed-bedside-lamp-490779/
[i]https://pixabay.com/photos/beach-sea-sunset-sun-sunlight-1751455/
[j]https://pixabay.com/photos/woman-crunches-sport-training-2250970/

Bibliography

Deanne M Adams, Richard E Mayer, Andrew MacNamara, Alan Koenig, and Richard Wainess. Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of educational psychology*, 104(1):235, 2012.

Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

Brooke Auxier and Monica Anderson. Social media use in 2021. *Pew Research Center*, 1:1–4, 2021.

Kent Beck. Using pattern languages for object-oriented programs. In *OOPSLA-87 workshop on the Specification and Design for Object-Oriented Programming*, 1987.

Aditi M. Bhoot, Mayuri A. Shinde, and Wricha P. Mishra. Towards the identification of dark patterns: An analysis based on end-user reactions. In *Proceedings of the 11th Indian Conference on Human-Computer Interaction*, IndiaHCI '20, page 24–33, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389440. doi: 10.1145/3429290.3429293. URL `https://doi.org/10.1145/3429290.3429293`.

Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. "i am definitely manipulated, even when i am aware of it. it's ridiculous!" - dark patterns from the end-user perspective. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, page 763–776, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3462086. URL `https://doi.org/10.1145/3461778.3462086`.

Jan O Borchers. A pattern approach to interaction design. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 369–378, 2000.

Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: privacy dark strategies and privacy dark patterns. *Proc. Priv. Enhancing Technol.*, 2016(4):237–254, 2016.

James V Bradley. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528, 1958.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77, 2006.

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

H Brignull, M Leiser, C Santos, and K Doshi. Deceptive patterns – user interfaces designed to trick you, 4 2023. URL `https://www.deceptive.design/`.

Harry Brignull. Dark Patterns, 2010. URL `http://darkpatterns.org/`. original website accessed via https://web.archive.org/.

Harry Brignull. *Deceptive patterns - exposing the tricks that tech companies use to control you*. Testimonium Ltd, 2023. URL `https://www.deceptive.design/book`.

John Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3):189–194, 1996.

EH Calvillo Gamez. *On the core elements of the experience of playing video games*. PhD thesis, UCL (University College London), 2009.

Donald T Campbell and Julian C Stanley. *Experimental and quasi-experimental designs for research*. Ravenio books, 2015.

Jieshan Chen, Jiamou Sun, Sidong Feng, Zhenchang Xing, Qinghua Lu, Xiwei Xu, and Chunyang Chen. Unveiling the tricks: Automated detection of dark patterns in mobile applications. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606783. URL https://doi.org/10.1145/3586183.3606783.

John P Chin, Virginia A Diehl, and Kent L Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–218, 1988.

Gokul CJ, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. Phishy - a serious game to train enterprise users on phishing awareness. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts, page 169–181, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359689. doi: 10.1145/3270316.3273042. URL https://doi.org/10.1145/3270316.3273042.

Douglas B Clark, Emily E Tanner-Smith, and Stephen S Killingsworth. Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, 86(1):79–122, 2016.

Gregory Conti and Edward Sobiesk. Malicious interface design: Exploiting the user. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 271–280, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772719. URL https://doi.org/10.1145/1772690.1772719.

Andrea Curley, Dympna O'Sullivan, Damian Gordon, Brendan Tierney, and Ioannis Stavrakakis. The design of a framework for the detection of web-based dark patterns. 2021.

Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 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*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376600. URL `https://doi.org/10.1145/3313831.3376600`.

Ralf Dörner, Stefan Göbel, Wolfgang Effelsberg, and Josef Wiemeyer. *Serious games*. Springer, 2016.

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

Madison Fansher, Shruthi Sai Chivukula, and Colin M. Gray. darkpatterns: Ux practitioner conversations about ethical design. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356213. doi: 10.1145/3170427.3188553. URL `https://doi.org/10.1145/3170427.3188553`.

Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. The dark (patterns) side of ux design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174108. URL `https://doi.org/10.1145/3173574.3174108`.

Colin M. Gray, Cristiana Santos, and Nataliia Bielova. 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, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/

3544549.3585676. URL `https://doi.org/10.1145/3544549.3585676`.

Colin M. Gray, Nataliia Bielova, Cristiana Santos, and Thomas Mildner. An ontology of dark patterns: Foundations, definitions, and a structure for transdisciplinary action. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3613904.3642436. URL `https://doi.org/10.1145/3613904.3642436`.

Mark D Griffiths. The educational benefits of videogames. *Education and health*, 20(3):47–51, 2002.

Johanna Gunawan, Cristiana Santos, and Irene Kamara. Redress for dark patterns privacy harms? a case study on consent interactions. In *Proceedings of the 2022 Symposium on Computer Science and Law*, CSLAW '22, page 181–194, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392341. doi: 10.1145/3511265.3550448. URL `https://doi.org/10.1145/3511265.3550448`.

Chaima Jemmali, Sara Bunian, Andrea Mambretti, and Magy Seif El-Nasr. Educational game design: an empirical study of the effects of narrative. In *Proceedings of the 13th international conference on the foundations of digital games*, pages 1–10, 2018.

Thomas R Knapp. Why is the one-group pretest–posttest design still used?, 2016.

David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

Jamie Luguri and Lior Jacob Strahilevitz. Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1):43–109, 2021.

Maximilian Maier and Rikard Harr. Dark design patterns: An end-user perspective. *Human Technology*, 16(2):170–199, 2020.

S M Hasan Mansur, Sabiha Salma, Damilola Awofisayo, and Kevin Moran. Aidui: Toward automated recognition of dark patterns in user interfaces. In *Proceedings of the 45th International Conference on Software Engineering*, ICSE '23, page 1958–1970. IEEE Press, 2023. ISBN 9781665457019. doi: 10.1109/ICSE48619.2023.00166. URL `https://doi.org/10.1109/ICSE48619.2023.00166`.

Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), nov 2019. doi: 10.1145/3359183. URL `https://doi.org/10.1145/3359183`.

Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445610. URL `https://doi.org/10.1145/3411764.3445610`.

Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 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*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580695. URL `https://doi.org/10.1145/3544548.3580695`.

Norwegian Consumer Council. Deceived by design, how tech companies use dark patterns to discourage us from exercising our rights to privacy. *Norwegian Consumer Council Report*, 2018.

Karen Maria Nyvoll. Serious interactive board games: Increasing awareness of dark patterns in teenagers. Master's thesis, NTNU, 2020.

Tanner Olsen, Katelyn Procci, and Clint Bowers. Serious games usability testing: How to ensure proper usability, playability, and effectiveness. In Aaron Marcus, editor, *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, pages 625–634, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21708-1.

Jan L Plass, Bruce D Homer, and Charles K Kinzer. Foundations of game-based learning. *Educational psychologist*, 50(4):258–283, 2015.

René Röpke, Vincent Drury, Ulrike Meyer, and Ulrik Schroeder. Exploring and evaluating different game mechanics for anti-phishing learning games. *International Journal of Serious Games*, 9(3):23–41, 2022.

René Christian Röpke. *Extending game-based anti-phishing education using personalization: design and implementation of a framework for personalized learning game content in anti-phishing learning games*. PhD thesis, Dissertation, RWTH Aachen University, 2023, 2023.

Sarah Sahabi. Prevalent dark patterns: Investigating children's mental models of malicious designs. Master's thesis, RWTH Aachen University, Aachen, July 2023.

Johnny Saldaña. *The coding manual for qualitative researchers*. sage, 2021.

René Schäfer, Paul Miles Preuschoff, and Jan Borchers. Investigating visual countermeasures against dark patterns in user interfaces. In *Mensch Und Computer 2023*, MuC '23, page 161–172. Association for Computing Machinery, September 2023. ISBN 9798400707711. doi: 10.1145/3603555.3603563. URL `https://doi.org/10.1145/3603555.3603563`.

Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, page 88–99, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938015. doi: 10.1145/

1280680.1280692. URL https://doi.org/10.1145/1280680.1280692.

Similarweb, 2023. URL https://www.similarweb.com/top-websites/. Accessed on Dec. 02, 2023.

Grant Storey, Dillon Reisman, Jonathan Mayer, and Arvind Narayanan. The future of ad blocking: An analytical framework and new techniques. *arXiv preprint arXiv:1705.08568*, 2017.

Stephen Tang, Martin Hanneghan, and Abdennour El Rhalibi. Introduction to games-based learning. In *Games-based learning advancements for multi-sensory human computer interfaces: Techniques and effective practices*, pages 1–17. IGI Global, 2009.

Ingvar Tjøstheim, Vanessa Ayres Pereira, Chris Wales, Simon Egenfeldt-Nielsen, and Angela Manna. Dark pattern: A serious game for learning about the dangers of sharing data. 2022.

Aleksandra Urman and Mykola Makhortykh. You are how (and where) you search? comparative analysis of web search behavior using web tracking data. *Journal of Computational Social Science*, pages 1–16, 2023.

Christof van Nimwegen and Jesse de Wit. Shopping in the dark: Effects of platform choice on dark pattern recognition. In *International Conference on Human-Computer Interaction*, pages 462–475. Springer, 2022.

Christian Voigt, Stephan Schlögl, and Aleksander Groth. Dark patterns in online shopping: Of sneaky tricks, perceived annoyance and respective brand trust. In *International conference on human-computer interaction*, pages 143–155. Springer, 2021.

José P Zagal, Staffan Björk, and Chris Lewis. Dark patterns in the design of games. In *Foundations of Digital Games 2013*, 2013.

Lawrence Jun Zhang and Donglan Zhang. Think-aloud protocols. In *The Routledge handbook of research methods in applied linguistics*, pages 302–311. Routledge, 2019.

Index

Typeset March 20, 2024