

Bachelor's Programme in Science and Technology

# Into the Dark

The Fault of Machine Learning Methods in Detecting UX Dark Patterns

---

Duc Chu

# Into the Dark

The Fault of Machine Learning Methods  
in Detecting UX Dark Patterns

**Duc Chu**

Thesis submitted in partial fulfillment of the requirements for  
the degree of Bachelor of Science in Technology.  
Otaniemi, 05 May 2025

Supervisor: Professor Maarit Korpi-Lagg  
Advisor: D.Sc. (Tech) Jaakko Sahsten

**Aalto University**  
**School of Science**  
**Bachelor's Programme in Science and Technology**

**Author**

Duc Chu

**Title**

Into the Dark: The Fault of Machine Learning Methods in Detecting UX Dark Patterns

**School** School of Science**Degree programme** Bachelor's Programme in Science and Technology**Major** Data Science**Code** SCI3095**Supervisor** Professor Maarit Korpi-Lagg**Advisor** D.Sc. (Tech) Jaakko Sahsten**Level** Bachelor's thesis    **Date** 08 May 2025    **Pages** 30    **Language** English**Abstract**

The world of design and the world of science has inextricably been intertwined, but it has rarely been truly studied. Especially in the smaller field of UX Design, there is one specific subject that should be discussed: dark patterns. Dark patterns are tricks used in digital spaces such as websites and apps that make the user do something they did not intend to do. Some examples of dark patterns are forcing users to give personal information to access free content, i.e., forced continuity, and tricking users into expecting certain outcomes but producing a different undesirable outcome, i.e., bait and switch. The consequences of these deceptive design tactics can range from loss of customer trust to financial loss to leakage of confidential personal information. There have been many serious attempts at categorizing dark patterns into taxonomies and building machine learning methods to detect them automatically. However, research into machine learning methods to detect these patterns is still in its infancy, requiring much more delicate insight and analysis.

In this thesis, a state-of-the-art literature review is conducted on dark pattern taxonomies and the machine learning models used to detect these patterns. The shortcomings of the dark pattern taxonomies are their low comprehensiveness, narrow scope, potential of missing future updates and low usability for detection purposes. The shortcomings of the machine learning models consist of the usage of vague taxonomies, failure to detect dynamic dark patterns, failure to detect non-textual dark patterns, overfitting and the lack of fully autonomous models. Other dark pattern detection models that does not use machine learning are also analyzed to provide insights for future machine learning models. Some future directions are then suggested for both dark pattern taxonomies and machine learning models. The paper contributes to the dark pattern detection field by not only deepening the understanding of dark pattern taxonomies and detection tools, but also providing insights for future models and research.

**Keywords** Automatic detection methods, Dark patterns, Deceptive Patterns, Machine learning, UX design**urn** <https://aaltodoc.aalto.fi>

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>Symbols and abbreviations</b> . . . . .	v
<b>1. Introduction</b>	<b>1</b>
<b>2. Background</b>	<b>3</b>
<b>2.1 On UX/UI Design</b> . . . . .	3
<b>2.2 On Dark Patterns</b> . . . . .	3
<b>2.3 On Dark Pattern Taxonomies</b> . . . . .	4
<b>2.4 On Current Machine Learning Abilities to Detect Dark Patterns</b> . . . . .	5
<b>3. Pattern Taxonomies and Their Fault</b>	<b>7</b>
<b>3.1 Analysis Criteria</b> . . . . .	7
<b>3.2 Low comprehensiveness</b> . . . . .	10
<b>3.3 Narrow Scope</b> . . . . .	11
<b>3.4 Potential of missing future updates</b> . . . . .	12
<b>3.5 Low Usability for Detection Purposes</b> . . . . .	13
<b>4. Machine Learning Models and Their Fault</b>	<b>15</b>
<b>4.1 Machine Learning Key Formulas</b> . . . . .	15
<b>4.2 Vague and Unusable Taxonomies</b> . . . . .	16
<b>4.3 Detection of Dynamic Dark Patterns</b> . . . . .	18
<b>4.4 Detection of Non-textual Dark Patterns</b> . . . . .	20
<b>4.5 Overfitting</b> . . . . .	22
<b>4.6 Fully Autonomous Models</b> . . . . .	24
<b>4.7 Other non-ML dark pattern detection models</b> . . . . .	25

<b>5. Limitations and Future Work</b>	<b>28</b>
<b>5.1 Limitations</b> . . . . .	28
<b>5.2 Future Work</b> . . . . .	28
<b>6. Conclusions</b>	<b>30</b>
<b>Bibliography</b>	<b>31</b>

# Symbols and abbreviations

## Abbreviations

UX	user experience
UI	user interface
ML	machine learning
EU	European Union
USA	United States of America
TP	true positives
TN	true negatives
FP	false positives
FN	false negatives

# 1. Introduction

User Experience (UX) Design aims to create interfaces that ensure a smooth and satisfactory user experience. However, some designers attempt to create interfaces that ignore the user's best interests. Ever since 2010 when Brignull coined the term "dark patterns" [1], the research into these patterns has been quite limited. However, the interest in the field has been substantially increasing in recent years [2]. An example of a dark pattern is the "Roach Motel" pattern. In this design pattern, the website encourages the user to perform an action that traps them in an undesirable state that is difficult to escape. For instance, the website may make it easy to subscribe to a newsletter, but challenging to unsubscribe.

The prevalence of dark patterns has led to many detection tools, ranging from manual to machine learning to advanced algorithms, to be created by UX researchers. Some of these researchers have produced detailed taxonomies of different types of dark patterns and the best method to detect them [2]. However, a shortage of serious insight into the shortcomings of these detection methods is detrimental to the UX Design field, whether it is generalizability or insufficient datasets. An investigation into the fault of machine learning detection methods is needed because without it, the damage is of great concern to the users and the companies producing the digital products. Users could be mentally tricked into losing privacy or even developing an addiction in the case of social media. Companies could be losing a significant number of their customer base for a slight increase in profit from dark patterns, which are likely to be illegal in the first place.

Thus, the aim of this state-of-the-art literature review is to provide an empirical analysis of machine learning tools in detecting dark patterns by presenting a deep dive into the current abilities and inabilities of the tools. It will be structured as follows: Chapter 1 is the introduction to the thesis, Chapter 2 provides background on basic key terms as well

as the history of taxonomies and detection tools. Chapter 3 discusses in depth the shortcomings of current dark patterns taxonomies. Chapter 4 will discuss the shortcomings of machine learning detection methods. Chapter 5 discusses the limitations and future directions after this paper. Finally, Chapter 6 concludes the thesis with a hopeful tone for the future of dark pattern detection.

## **2. Background**

This chapter will discuss the relevant key terms in this thesis: UX/UI design, dark patterns, the current state of dark patterns taxonomies, and dark patterns detection methods.

### **2.1 On UX/UI Design**

User experience (UX) design is the field that focuses on creating products, mainly digital ones, that facilitate an easy and intuitive experience for users of a product. The aim of UX Designers is ensuring a smooth and satisfactory user experience by utilizing user psychology, design principles, and a user-centered approach. The term "user experience" was coined in the 1990s by Don Norman, who was a cognitive scientist at Apple Inc. [3].

User interface (UI) design is the field that focuses on designing the look and feel of the digital product, focusing more on the aesthetics of the product. The product should be both functional and visually appealing. As a result, UI Designers work with the visual aspects of a product which includes typography, imagery, layout and colors.

### **2.2 On Dark Patterns**

Dark patterns, as defined by Harry Brignull, are "tricks used in websites and apps that make you do things that you didn't mean to" [1]. These design strategies capitalize on the cognitive biases and habitual behaviors of individuals to achieve specific commercial or other objectives. Dark patterns can appear in many different forms such as hidden costs, disguised ads and complex unsubscribe procedures. Even though they might give the owners of the websites additional benefits such as income, in the long run, they erode the users' trust and possibly leading to legal liabilities [4].

It is interesting here to note the legality of dark patterns. According to the website created by Harry Brignull himself about dark patterns, there has been legal precedence in the European Union (EU) and the United States of America (USA). The Digital Services Act, Data Act, and AI Act in the European Union have all taken part in regulating deceptive design [5]. However, there is currently no common definition for dark patterns in the EU legislation. Among USA regulatory bodies, however, dark patterns have been more heavily regulated. The Federal Trade Commission and the California Privacy Protection Agency have issued regulations on dark patterns recognition and the penalties associated with them [6],[7]. However, the terms "dark patterns" are still not legally recognized in many countries [8]. Therefore, it is of great importance that better tools and methods are proposed to categorize and detect dark patterns.

### **2.3 On Dark Pattern Taxonomies**

Firstly, the reason taxonomies are necessary in the first place is because they allow researchers to "form a solid knowledge base of our detection system" [9]. They provide researchers with a clear-cut benchmark on the definition of a dark pattern and a method to recognize it. Secondly, a taxonomy is important because it allows professionals in the field to communicate to each other more easily. An analogous example of this from another field is the DSM-V, or Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. Therapists and psychologists have utilized it all over the world to diagnose and treat patients. Although it is not perfect, surveys of clinicians about their attitude towards the DSM consistently show that they believe it is useful. For example, in a survey with more than 1,000 psychiatry practitioners and other psychology professionals, 95 percent reported a "willingness to continue to use the DSM even if it was not required" [10].

In the field of UX/UI design, there have been many efforts to provide a complete and definitive taxonomy of dark patterns. Gray et al. [11] have provided a five-class taxonomy of dark patterns that categorize dark patterns into 5 types: interface interference, nagging, forced action, obstruction, and sneaking. Harry Brignull [1] has published a taxonomy that recognizes 16 types of dark patterns. Nie et al. [2] conducted a comprehensive study on dark patterns and produced a 64-item taxonomy that is derived from 76 papers, which is arguably the most extensive taxonomy to

date. It took into account both Gray's and Brignull's taxonomies and did extensive research and testing to produce a final "optimized taxonomy". This taxonomy is comprised of six main categories: nagging, obstruction, sneaking, interface interference, forced action, and social engineering.

## 2.4 On Current Machine Learning Abilities to Detect Dark Patterns

According to Nie et al. [2] who conducted a comprehensive study in 2024 on dark patterns and tools to detect them, only six of the tools emerged as satisfiable solutions. Their criteria were that the tools must have relevant verbs in their descriptions, have automated capabilities, have a clear documentation, and provide clear guidance on what types of dark patterns they can detect. Among those tools, only one [12] was executable as the others either did not provide the source code or had a non-functional source code. Another important finding is that across all of these tools, the most successful tool was able to detect only 50 percent of dark patterns

It is important to note here the evolution of tools in the past decade or two, ranging from manual exploration to text and visual analysis. It used to be that dark patterns were manually detected either by domain experts [1], [13] or normal users [14]. Mathur et al. [15] then introduced semi-automated techniques, which utilize clustering techniques to divide dark patterns into related groups, facilitating the detection process's efficiency. Afterwards, Soe et al. [16] experimented with supervised and unsupervised machine learning techniques with unsatisfying results due to the use of limited labeled training data and the lack of consensus regarding the definition of "dark patterns" in legislative documents. Building on this model, Kirkman et al. [17] introduced DarkDialogs with a detection accuracy of at least 98 percent. Others have then published prototypes of automated software tools [18], a dialog box to notify users of dark patterns [19], and a detection technique using cluster analysis with an intended addition of a machine learning method in the future [20]. The latest detection tools introduced image processing as a key method, for example, AidUI [12] and UIGuard [9]. AidUI has a major advantage in the sense that it can be operated in different software domains as it relies on pixel-based data, unlike those models intended only for cookie banners and dialogs.

Even though these detection tools have gained some success at detecting dark patterns, they have many disadvantages, ranging from non-functional source code to lack of ability to detect all recognized dark patterns. For

example, DarkDialogs , which had a 98 percent accuracy rate, was only able to detect 10 dark patterns and also only considered consent dialogs, not the whole website. The shortcomings of current detection methods can be summarized below:

- Usage of vague and unusable taxonomies;
- Inability to detect dynamic dark patterns;
- Inability to detect non-textual dark patterns;
- Overfitting;
- Inability to be fully automated.

A final note should be made about dark pattern datasets. The quality of a dataset is important to ensure the quality of a machine learning model, because if the dataset is biased and unreliable, the model will be too. Many datasets consist of unequally divided samples, meaning that the number of some dark pattern samples is much higher than others [21]. This is the phenomenon called "class imbalance", where there are more instances of a class in a ML classification problem than the other class. If there is high class imbalance, model performance can be poor with the class that has fewer samples. Another common flaw of dataset is that they are not annotated and labelled properly. In other words, the dark pattern samples are not labelled with "predictors" [20] that can help the model in categorizing dark pattern samples. One example is the dataset by Soe et al [16], which uses Gray et al's dataset, which is too general and unfit for ML classification. Better labels are needed for automatic detection, which propels the need for better taxonomies. Better labelled and more detailed taxonomies lead to better labelled datasets. This is the topic of the next chapter discussed below.

### 3. Pattern Taxonomies and Their Fault

This chapter discusses in detail current pattern taxonomies and their shortcomings. The analysis criteria is provided, followed by four different shortcomings with illustrated examples for each.

#### 3.1 Analysis Criteria

In order to assess a taxonomy, it is valuable to note the standards for a well-rounded taxonomy. According to [22], the following criteria are important in evaluating taxonomies:

- Comprehensiveness: "ability to classify all known objects for the domain it was developed for";
- Robustness: "ability to differentiate objects of interest, determined by the degree to which a taxonomy's categories and characteristics represent distinct concepts";
- Conciseness: "ability to classify objects of interest with the least possible amount of dimensions, categories and characteristics";
- Extensibility: "ability to allow for changes in its structure, that is adding, modifying or deleting dimensions, categories or characteristics";
- Explanatory: "ability to enable the user to locate an object in the taxonomy based on its characteristics, or to deduce from the location of an object what characteristics it has";
- Mutual Exclusiveness: "ability to identify an object uniquely, that is that no object exists in the same dimension under different categories";
- Reliability: "ability to support consistent classification decisions".

The authors of this paper have provided a more detailed table which provides measurement methods, as seen in Table 3.1. The "Nature" column

describes the way the attribute is measured, either internally (measured by observing the product) or externally (measured by observing its relation to its environment). The "Perspective" column describes whether the attribute is objective or subjective. The "Mapping" column means that the mapping of an attribute to its measurement can be direct or derived, meaning the measurement can involve no other attributes or multiple attributes, respectively. Finally, the "Scale" column describes the type of data the attribute refers to: nominal, ordinal, interval, ratio or absolute data.

Attribute	Nature	Measurement Name	Nature	Perspective	Mapping	Scale
Comprehensiveness	External	Design process quality	Internal	Subjective	Direct	Nominal
		Count of unclassified objects	External	Subjective	Direct	Absolute
Robustness	External	Count of taxonomy constructs	Internal	Objective	Direct	Absolute
		Semantic proximity and distance	Internal	Objective	Derived	Ratio
		Count of misclassified objects	External	Subjective	Direct	Absolute
Conciseness	External	Human memory heuristic	Internal	Objective	Direct	Absolute
		Amount and depth of constructs	Internal	Objective	Derived	Ratio
		Count of unused constructs	External	Subjective	Direct	Absolute
Extensibility	Internal	Documented extension points	Internal	Subjective	Direct	Nominal
		Rate of change	Internal	Objective	Derived	Ratio
Explanatory	External	Location support	Internal	Subjective	Direct	Nominal
		Orientation efficiency and effectiveness	External	Objective	Derived	Ratio
Mutual exclusiveness	External	Classification constraint	Internal	Subjective	Direct	Nominal
		Ambiguous objects	External	Objective	Derived	Ratio
Reliability	External	Inter-rater reliability	External	Objective	Derived	Ratio
		Intra-rater reliability	External	Objective	Derived	Ratio

**Table 3.1.** A table showing the measurements and their characteristics of each evaluation criteria in [22]

Based on the measurement methods, an initial analysis was conducted for six taxonomies, five of which are the most cited in Google Scholar (as of May 5th 2025) in the UX/UI field: Gray et al.'s [11], Mathur et al.'s [15], Luguri et al.'s [23], Bösch et al.'s [24], and Brignull et al.'s [1]. The sixth one is Nie et al.'s, which is relatively new, thus not cited very often. However, it is included due to its claim being "the most comprehensive taxonomy to date" (and recently published in July 2024). The findings of the analysis is found in Table 3.2.

The equations used to calculate the scores of some of the measurements are taken from the paper. First, the design process quality is on a scale from 0 to 4 because there were four different subcriteria in this criteria (diverse set of data sources, diverse background of the creators, diverse data analysis methods and external expert evaluation). Next, the count of taxonomy constructs were the sum of all items of each dimension of the taxonomy. For example, if the taxonomy has "Severity" and "Common Scenarios" as columns, that counts as 2 items. If it has 18 different dark pattern types, that counts as 18 items. Therefore, the example taxonomy would have 20 items. Count of misclassified objects were subjectively evaluated by the author, who concluded that no items were mistakenly categorized in all the papers.

The next criteria, which is Conciseness, includes two measurements M(t) and C(T). M(t) are calculated based on the number of pieces of information needed at three decision point: choice of dimension, choice of category at level [1...n-1] and choice of characteristic at level n. The equation for C(T) is presented below.

$$C(T) = \frac{1}{1 + \ln(\sum_{i=1}^{ncat} (\frac{1}{d(Cat_i)}) + \sum_{i=1}^{nchar} (\frac{1}{d(Char_j)}) - 1)}$$

where T stands for taxonomy, ncat is the number of categories, nchar is the number of the characteristics, and d is the depth.

The next criteria, which is Extensibility, has two measurements, which were on a scale of 0 to 4 and 0 to 5 respectively. The first measurement ("Documented extension points") has four sub-measurements, while the second one ("Rate of change") has an equation:

$$RoC_{now} = \frac{n}{t_{now} - t_{initialrelease}}$$

where n is the number of changes, the denominator is the number of days between the first release and the current date. The range is originally [0...1], but it was scaled to [0...5] in this thesis for easier comprehension.

The rest of the measurements were either non-calculable or scaled on [0...2] or [0...5], based on the sub-measurements that were provided in [22].

Criteria	Measurements	Brignull et al's (2010)	Gray et al's (2018)	Mathur et al's (2019)	Bösch et al's (2016)	Luguri et al's (2021)	Nie et al's (2024)
Comprehensiveness	Design process quality (4 points)	4	4	3	1	3	4
	Count of unclassified objects	NaN	NaN	NaN	NaN	NaN	NaN
	Count of taxonomy constructs	16	18	20	17	27	58
Robustness	Semantic proximity and distance	NaN	NaN	NaN	NaN	NaN	NaN
	Count of misclassified objects (5 points)	0/16 objects -> 5	0/18 -> 5	0/15 -> 5	0/7 -> 5	0/27 -> 5	0/50 objects -> 5
	Human memory heuristic (t=7) (3 points)	M(t)=0 -> 3 C(T) = 0.269	M(t)=0 -> 3 C(T) = 0.28	M(t)=0 -> 3 C(T) = 0.277	M(t)=0 -> 3 C(T) = 0.294	M(t)=0 -> 3 C(T) = 0.265	M(t)=0 -> 3 C(T) = 0.257
Conciseness	Amount and depth of Constructs	NaN	NaN	NaN	NaN	NaN	NaN
	Count of unused constructs	NaN	NaN	NaN	NaN	NaN	NaN
Extensibility	Documented Extension Points (4 points)	0	0	4	3	3	4
	Rate of change (5 points)	0	0	0	NaN	0	NaN
Explanatory	Location support (2 points)	0	0	2	2	1	2
	Orientation efficiency and effectiveness	NaN	NaN	NaN	NaN	NaN	NaN
	Classification constraint (2 points)	0	0	0	0	0	0
Mutual Exclusiveness	Ambiguous objects (5 points)	0/16 objects -> 5	0/18 -> 5	0/15 -> 5	0/7 -> 5	0/27 -> 5	0/50 -> 5
Reliability	Inter-rater reliability	NaN	NaN	NaN	NaN	NaN	NaN
	Intra-rater reliability	NaN	NaN	NaN	NaN	NaN	NaN
	<b>Total points</b>	31.27	35.28	42.28	36.29	44.26	81.26

**Table 3.2.** Initial Analysis of six dark pattern taxonomies

A few caveats should be mentioned in this initial analysis. Firstly, the point system given to some of the measurements was based on the author's own interpretation of the criteria given in [22]. However, they were often appropriately guided by the number of characteristics in each of the criteria. Secondly, there were some measurements that could not be determined due to the time restriction of a Bachelor's thesis limiting the analysis. Thirdly, the scores, such as C(T) and M(t), that were calculated are not externally validated by other professionals. Therefore, the initial analysis's findings are not as accurate as it can be. However, it serves as a benchmark to

determine the areas in which these taxonomies lack. In addition, the year next to the authors' name in the table is the year of publication, not the year of evaluation (which would all be 2025). Finally, the taxonomies that are discussed in the next sections include taxonomies not mentioned in Table 3.2 which allows for a wider scope of analysis.

After conducting the initial analysis and synthesizing findings from previous research, the four shortcomings of current taxonomies are deducted to be low comprehensiveness, narrow scope, potential of missing future updates and low usability for detection purposes.

### **3.2 Low comprehensiveness**

The first fault of the existing taxonomies is that none of these taxonomies are definitive or extensive. This fault refers to the first criteria of a well-rounded taxonomy provided by [22], which is comprehensiveness, or the "ability to classify all known objects for the domain it was developed for". The taxonomies either overlap without any of them being extensive or vary in their level of details.

One example is Brignull et al.'s taxonomy [1], which can be considered the first dark pattern taxonomy to exist. It was published in 2010. Ever since, little improvements have been made to the taxonomy, except for one removal, two additions and some denomination changes [25]. There are many cases in which this taxonomy failed to include a dark pattern. For example, "cuteness" was found to be a dark pattern by Lacey and Caudwell [26], but was not included in the taxonomy. There is, however, "confirmshaming", which is described as triggering uncomfortable emotions to persuade the user into unintended actions. Even though both "confirmshaming" and "cuteness" deal with emotions, they are not the same. Therefore, Brignull et al.'s taxonomy cannot be considered to be extensive.

Another example is Bösch et al.'s taxonomy [24]. This taxonomy was published in 2016. However, the provided categorization in this taxonomy is not sufficient as they mainly focus on user accounts. For example, forced registration, bad defaults and immortal accounts, which are almost half of the proposed categories, focus on the complications of creating an account. Dark patterns such as sneak into basket or comparison prevention [1] are not mentioned in this taxonomy, making it incomprehensive.

When discussing comprehensiveness, the 64 Dark Pattern Taxonomy

(renamed for better readability) [2] needs to be mentioned. The 64 Dark Pattern taxonomy is claimed to be "optimized" by the authors. Furthermore, the title of the paper is "Shadows in the Interface: A Comprehensive Study on Dark Patterns" and the paper was published in July 2024. Therefore, it is reasonable to assume that this taxonomy would be the most comprehensive one to date. After conducting the initial analysis, it is clear from the points given that it is the most comprehensive, as the design process quality and count of taxonomy constructs had the highest points. However, since the initial analysis is subjective and only two out of four measurements in the Comprehensiveness category were taken, it cannot be concluded that this taxonomy is comprehensive. Nevertheless, the author conducted a survey for 173 industry experts and more than 92 percent believe it to have completeness. It is also noteworthy to mention that, as stated in their Threats to Validity section, the methods they use to evaluate current detection tools may not cover all dark patterns possible, thus affecting the completeness of their evaluation.

### **3.3 Narrow Scope**

The second fault of dark pattern taxonomies is that they are oftentimes focused on a single domain. Their narrow scope prevents researchers from using it as a standardized taxonomy for fields other than the field it was intended for. The main four fields in the existing taxonomies are: user experience ethics [27], digital privacy [24], regulatory bodies [6], [23], and human-computer interaction [11]. Aside from these major areas, there are taxonomies that deal with even more specific subfields such as e-commerce [15].

It is noteworthy to distinguish the first and second fault of taxonomies. Low comprehensiveness refers to the inability of taxonomies to include all possible known instances of the intended domain such as e-commerce or digital privacy. Narrow scope, however, refers to the inability of taxonomies to sufficiently cover all areas related to the subject matter, which is dark patterns in this case. In other words, a taxonomy could be comprehensive because it has listed all known dark patterns in the e-commerce area, but it does not have a wide scope because it only focuses on that specific area.

One notable taxonomy is the taxonomy by Mathur et al[15]. The author studied 11 000 shopping websites and distilled the findings into a taxonomy that has 7 main categories divided into 15 types of dark patterns. Even

though the data sources were plentiful, they were homogenous in the sense that they only come from shopping websites. In addition, the authors later stated in the conclusion section that they only took into account text-based dark patterns, which left out other types that are visual by nature. Therefore, even though the depth of insights into dark patterns were significant in this study, the breadth was missing. This is not to say that lack of scope is necessarily always a negative aspect. In fact, these taxonomies that are focused on only one sub-area of dark patterns can reveal many more insights than those who aim to be more wider in scope. Oftentimes, they discover new types of dark patterns, which only helps the construction of other taxonomies [2].

### **3.4 Potential of missing future updates**

The third fault of current taxonomies is that they, in many cases, do not provide documentation or guidance on the update process. In other words, the way to update the taxonomies in case of new dark patterns or merging old ones is not present in the documentation of these taxonomies. This fault directly correlates to the extensibility criteria. As new technologies are introduced, new types of dark patterns keep emerging. Thus, these taxonomies are prone to being outdated and therefore unusable in real-life settings.

One critical example of this fault is the Gray et al.'s taxonomy. Given its citation counts, the authors need to improve their documentation so that it includes an update process, indicating the people and the methods needed to update the taxonomy. Other authors have built on Gray et al.'s taxonomy to make their own taxonomies [2], [23]. The authors in [2] built on Gray et al.'s taxonomy and added 20 new important dark patterns, while the authors in [23] included the taxonomy in their summary of existing dark pattern taxonomies.

An example of a taxonomy that provided documentation on the taxonomy update procedure (how a taxonomy can be updated) is the original taxonomy by Brignull et al. [1]. In the website, a user can easily locate the place to submit new dark patterns in the menu section. The form is easy to understand and does not take much effort from the user to submit a new dark pattern. Brignull himself also set up an X account called @dark-patterns where people can submit undiscovered patterns. Even though Brignull et al.'s change process is documented, the actual taxonomy itself

has not been updated much. Ever since 2010, approximately five or six minor modifications have been made to the taxonomy [25]. The reason can be that either people find the updating process too cumbersome or that the team behind Brignull's website are not active in updating the taxonomy. However, this cannot be determined in the scope of this thesis.

### 3.5 Low Usability for Detection Purposes

The final major fault of dark pattern taxonomies is its low usability when utilized for dark pattern detection, namely when using machine learning methods. The taxonomies are either too complex and vague to apply to a machine learning context or too narrow in their scope, the latter of which has been discussed in section 4.2.

Gray et al's taxonomy [11] is a typical example of this fault. In the taxonomy, there are five different categories: nagging, obstruction, sneaking, interface interference and forced action. Obstruction, for example, is described as "making a process more difficult than it needs to be, with the intent of dissuading certain actions(s)". Even though the authors did give some examples of this dark pattern, they did not give specific characteristic clue on the context this dark pattern can be found or some textual or visual signs that the pattern is existent. Even though specific features can be left undefined for ML training, the problem is that the categories themselves are not well-defined, which can cause worse ML performance.

A good example of a taxonomy that has high usability for detection settings is the Potel-Saville and Mathilde's taxonomy, or the Fair Patterns Taxonomy for better readability [28]. The authors in this paper proposed a two-part taxonomy that detailed both dark patterns and its counterpart "fair patterns". It aims not to create an extensive taxonomy but an usable one. Usability is, according to ISO norm 9241-11 [29], "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". For detection purposes, the taxonomy assists the researchers to understand the type of dark pattern present easily with self-explanatory names such as "maze" or "missing information". With sufficient dataset, the taxonomy could be an efficient, though not extensive, tool to categorize dark patterns. Dark pattern dataset will be discussed in more detail in the next two chapters about machine learning methods. Moreover, the "fair patterns", which are essentially proposed solutions to the dark patterns,

help the researchers to take a more active role. They can suggest better ways to design the same content that is fair and user-friendly. However, the taxonomy could still be more usable by including more context-specific and detailed signs of dark patterns. This is due to the nature of dark patterns being elusive and oftentimes controversial among researchers. In addition, dark patterns usually appear in various forms and need detailed descriptions to be able to be captured.

## 4. Machine Learning Models and Their Fault

In this chapter, the shortcomings of current machine learning models for dark pattern detection will be discussed. Four main shortcomings of the current methods will be discussed through detailed analysis and comparisons. Other non-ML dark pattern detection frameworks will be also discussed, because there are some insights from those framework that can be of great assistance to improve current machine learning methods.

### 4.1 Machine Learning Key Formulas

First and foremost, it is important to note the key formulas when discussing machine learning methods. There are four main formulas: accuracy, precision, recall and F1.

**Accuracy** is a metric that measures how often a machine learning model correctly predicts the outcome. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

in which TP means True Positives, TN means True Negatives, FP means False Positives and FN means False Negatives.

**Precision** is a metric that measures how often a positive prediction is actually correct. The formula for precision is:

$$Precision = \frac{TP}{TP + FP},$$

,in which TP means True Positives, TN means False Positives.

**Recall** is a metric that measures how often a machine leaning model correctly identifies positive instances out of all actual positive instances in the dataset. The formula for recall is:

$$Recall = \frac{TP}{TP + FN},$$

,in which TP means True Positives, FN means False Negatives.

**F1** is a metric that computes the average of precision and recall to better evaluate a machine learning model's performance. It provides a balanced view of a model's ability to correctly identify positive samples and minimize false positives. The formula for the F1 score is:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

However, the models in this thesis deal with multiclass classification, which tries to assign objects in several predefined categories. It is different from binary classification, which tries to assign objects in two categories. Therefore, the metrics that are used are different. Multiclass classification still uses the same four metrics mentioned above, but it adds a macro-averaged or micro-averaged precision or recall scores. Only the macro-averaged scores' formula will be mentioned below, because micro-averaged scores cannot be manually calculated in this case (since it needs the specific numbers of True Positives, False Negatives and False Positives, which are not mentioned in the papers).

**The macro-averaged precision** formula is:

$$Precision \text{ (macro-averaged)} = \frac{\text{Sum of precision scores of each class}}{\text{The number of classes}}, \quad (4.1)$$

**The macro-averaged recall** formula is:

$$Recall \text{ (macro-averaged)} = \frac{\text{Sum of recall scores of each class}}{\text{The number of classes}}. \quad (4.2)$$

## 4.2 Vague and Unusable Taxonomies

The first fault of current machine learning methods relates back to the previous chapter about taxonomies. The current taxonomies available are not specific and detailed enough to get an accurate classification from the machine learning models. In other words, even if the machine learning models were to be able to detect dark patterns, putting them in a category

proved to be challenging due to the vagueness of the taxonomies.

An example is Mathur et al's unsupervised machine learning model that was released in 2019 [15]. It aimed to present automated methods that allow professionals to categorize dark patterns on a large sample of websites in one certain category. They used a Logistic Regression classifier for data collection and performed data analysis using a Hierarchical Density-Based Spatial Clustering of Applications with Noise algorithm. The results were that they found dark patterns in 11.1 percent of the 11000 shopping websites that were examined, which they noted were a "lower bound" for the actual number of dark patterns. In the limitation section, they informed the readers that the method that they used had left out inherently visual or other non-text-based dark patterns. And they also used many existing taxonomies when deciding which dark patterns were present in the shopping websites. And as discussed in Chapter 3, the existing taxonomies are incomprehensive and unfit for machine learning purposes. A better taxonomy would be a context-specific and narrowly defined one with at least a few examples for each of the dark patterns.

A machine learning model that tried to use this approach of giving a few examples per dark pattern is Sazid et al's model [30]. They used a GPT-3.5 Turbo model, and examined three models. The first model provided no example of any dark patterns, essentially a "blind" machine learning model (zero-shot learning). The second model provided one example of each dark pattern (one-shot learning). And the third model provided two examples for each dark pattern (few-shot learning). They examined seven dark patterns in total: Misdirection, Urgency, Scarcity, Social Proof, Obstruction, Forced Action and Sneaking. The text that did not fall into any categories were put into the "Not Dark Pattern" category. The results showed that few-shot learning worked best, with the accuracy score being above 92 percent for most dark patterns, except for Misdirection and Sneaking, with 65.8% accuracy and 0% accuracy, respectively. The macro-averaged scores were calculated to be 70.14% for precision and 79.92% for recall. The few-shot learning model was tested on 2,342 texts. The reason the score for the Sneaking dark pattern was suboptimal was because: firstly, it was a dark pattern that used website logic, not textual patterns, to deceive users (which will be discussed in section 4.3); secondly, the categorizations of dark patterns are not detailed and specific enough. It is noteworthy to mention that the model used in-context learning, which helped focus on the contextual information of dark pattern. However, it still suffered from

overfitting as the accuracy score when using another dataset plummeted from 92.2 percent to 58.67 percent with few-shot prompting. This highlights the necessity of having a clear and defined taxonomy that omits any gray area, and the necessity of having multiple varied datasets.

Overall, it is easy to see that the taxonomy that these machine learning models use is of great importance. Even if an author had access to a balanced and wide dataset with equal number of each dark patterns, which is needed for optimal performance, the model would still be suboptimal if the taxonomy is not good enough. Therefore, a detailed and comprehensive taxonomy must be created in accordance with the standards of a good taxonomy [22].

### 4.3 Detection of Dynamic Dark Patterns

The second fault of current machine learning models is that they cannot detect dark patterns that span multiple screens or involve other non-UI-related considerations such as user's intent. Chen et al. [9] had invented a name for these types of dark patterns: "dynamic dark patterns". A dynamic dark pattern is a dark pattern that requires additional contextual information to determine if it is actually deceptive or not, such as user's intent or transition animations. For example, the Bait and Switch dark pattern depends on the context in which the user was using the interface. The user can have varying levels of expectations and intents when using the interfaces, which can make a design pattern non-deceptive or deceptive. Another example is the Roach Motel pattern, which needs the information of all other user interfaces in the application to determine if there is an option to log out or unsubscribe.

An example of a machine learning model here is AidUI [12]. The model uses an almost fully autonomous process, using computer vision and natural language template matching techniques. It is very rare that the ML model is fully autonomous. In addition, it also uses deep learning based object detection techniques, such as Faster R-CNN [31]. They invented CONTEXTDP, a dataset that consists of 501 labelled screenshots. Regarding the taxonomy used, they created a new taxonomy, which is merged from existing taxonomy, that consisted of 7 dark pattern categories. The final scores were: 0.66 on precision, 0.67 on recall and 0.65 on F1. The score is not optimal due to various dynamic dark patterns. For example, the Default Choice dark pattern, as illustrated in Figure 4.1. This dark

The image shows a mobile application sign-up screen. At the top, there is a red header with a back arrow and the text "Sign Up". Below the header, there are several input fields: "FIRST NAME" with the value "Grace", "LAST NAME" with the value "Chan", "EMAIL" with the value "appcrawler6@gmail.com", "PASSWORD" (masked with dots) with a note "8 TO 12 CHARACTERS" and a "SHOW" link, "SECURITY QUESTION" with the value "City of birth?" and a dropdown arrow, and "SECURITY ANSWER" with the value "San Francisco" and a note "3 OR MORE CHARACTERS". Below these fields is a link "Link to existing loyalty card (optional)". At the bottom, there are two checked checkboxes: "Receive email offers" and "I agree to the Terms of Use". A red "Create Account" button is positioned below the checkboxes. The bottom of the screen shows the standard Android navigation bar.

**Figure 4.1.** An example of a dynamic dark pattern [12]. The text "Receive email offers" lies outside the vocabulary of textual patterns that AidUI recognizes, therefore making it ambiguous and harder to recognize as a dark pattern.

pattern uses texts that are semantically similar to deceptive ones ("Receive email offers"). It is reasonable to state that the "Receive email offers" text is not too complex. However, it is complex in the sense that it requires external considerations such as the user's intents.

Another example is the model UIGuard by Chen et al. [9], who invented the term "dynamic dark patterns". For many of the dynamic dark patterns that they examined in their taxonomy, the outcomes of the proposed model were not encouraging. For example, the "Nag to rate" and "Nag to upgrade" dark patterns were "in-between" dark patterns, meaning they were hard to categorize even with contextual UIs, which makes them dynamic dark patterns. The results were that the first pattern received a precision score of 0.10 and an F1 score of 0.18, while the second pattern received a precision score of 0.21 and a F1 score of 0.34. For further context, the macro-averaged precision was 0.83 and the macro-averaged recall was 0.82. The scores are quite good overall but low for some dynamic dark patterns, which means it is possible that the model is not complex enough

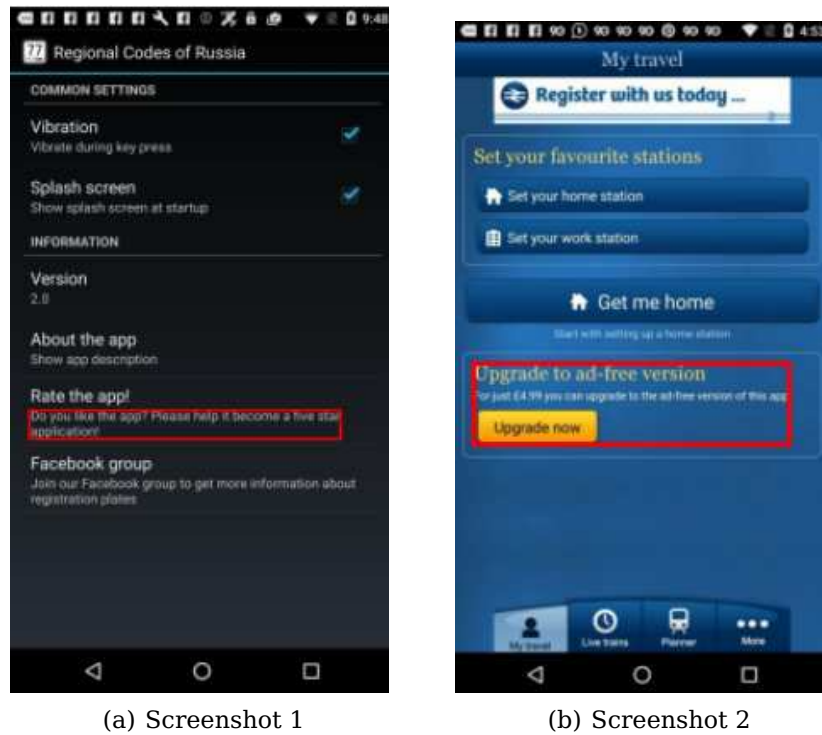
to detect such patterns.

The two models mentioned above are both machine learning models that utilize not only textual or visual analysis. They are advanced models, meaning that the authors use more than one feature of the dark patterns in training the models. Specifically, UIGuard considers six key characteristics: element coordinates, element type, text content, icon semantic, checkbox status, color and element relationship. AidUI, on the other hand, considers textual information as well as icon, color and spatial information. However, they still perform poorly with dynamic dark patterns. Evidently, the two machine learning models are not optimal in detecting more complex patterns. Since dynamic dark patterns involve user's intents in many cases, it is recommended to develop a more personalized detector model to warn the users if they want to be alerted about the potential dark pattern. In fact, UIGuard has recommended a similar future direction for their model. Additionally, since dynamic dark patterns also involve previous UI screens to determine their deceptiveness, machine learning models could benefit from a dataset that does not only consist of pictures of a single screen, but also videos of user flows from the start to the end of completing a task. For example, a video could show the user logging in the app, scrolling its features, buying a product, and then trying to find a way to delete the account (possibly because the product is too expensive). Consequently, if there is no screen in the video that has the option "Log out", the dark pattern Roach Motel (Hard to Cancel) is present [1].

#### **4.4 Detection of Non-textual Dark Patterns**

The third fault of current machine learning models is their lack of proficiency in detecting non-textual dark patterns. Aside from the two machine learning models mentioned in the previous section, many current machine learning models only focus on inherently textual dark patterns such as Urgency (giving user a limited time to buy a product) or Forced Action (forcing user to do something they cannot opt out) [30].

The first machine learning model to be discussed is the one that uses GPT-3.5 Turbo model, developed by Sazid et al. [30]. Their overall approach has been described in Section 4.2. The results showed that the Sneaking pattern were not recognized at all, earning a 0 percent score for accuracy, precision, recall and F1 score for each of the three models (no examples, one example and two examples). Their reasoning for the failure to capture



**Figure 4.2.** Examples of dynamic dark patterns in UIGuard model [9]. They are marked as dark patterns but they are not because they do not appear as a pop-up window

this dark pattern was that Sneaking "mostly leverages website logic and policy, rather than individual webpage texts". This is only one of the many dark patterns that are inherently non-textual. One example is Disguised Ad, which are ads that are designed to look like normal content. The textual information may be an advertisement, but the sizing and color of the advertisement makes it a dark pattern. However, the performance scores of the machine learning model are encouraging, with 92.57 percent overall accuracy (the results are considerably lower when testing with new dataset, which suggests overfitting and will be discussed in Section 4.5). One advantage of this textual-based machine learning model is that it focuses on textual dark patterns, whereas advanced machine learning models such as AidUI [12] focus on every single dark patterns. Since each dark pattern tends to have different variations, it is beneficial to focus on a single type of dark pattern before adding others. Therefore, Sazid et al.'s model is still an advancement in dark pattern detection, even with this shortcoming.

Another example is Yada et al.'s model, which was released in 2022 [32]. This is a textual-only ML model which similarly showed great scores when assessing performance. In fact, their model showed better scores than Sazid et al.'s. The authors chose two types of machine learning methods:

classical NLP (Natural Language Processing) methods and transformer-based pre-trained language models. The accuracy and F1 scores of all the classical NLP methods were above 0.95, and the Precision score were approximately 0.98. The transformer-based pre-trained language models performed even better, with the best being RoBERTa-large with an accuracy score of 0.975 [33]. Although they only focused on texts on e-commerce websites, similar to Mathur et al. [15], they have created a machine learning model that is well documented and can be used for further development to detect non-textual dark patterns as well.

The two models above all had great performance scores when testing against a textual dark pattern dataset, even better than the advanced ML models mentioned in Section 4.3. Therefore, even though they are flawed, they should still be considered to be an advancement in dark pattern detection. One suggestion could be that they combined the transformer-based pre-trained model RoBERTa-large with the UIGuard or AidUI model. Another suggestion would be to build a ML model specifically for detecting visual-based dark patterns, such as Disguised Ad, and ensure it has the highest performance score possible. Afterwards, it can be combined with the RoBERTa-large model to make a more well-rounded model.

#### **4.5 Overfitting**

The fourth fault of machine learning methods is the classic weakness in many ML models: overfitting. Overfitting is the phenomenon that a machine learning model performs worse on a new dataset that it has not learned before, due to the fact that it has "overfitted", or fit too closely, with the trained dataset. Many dark pattern detection ML models have this shortcoming, either due to lack of data or the ML method itself.

The Sazid et al's model [30] is one example. The model, when used for the original test dataset, detected 92 percent of the dark pattern texts. However, when used for a new dataset that consists of only 30 e-commerce websites, it only detected at most 58.67 percent of the dark pattern texts available (using the few-shot prompting, meaning giving two examples to the ML model). It can also be argued that the test data sources are different, leading to different results (the first dataset was merged from two datasets, the second dataset was manually annotated by the authors). Even though the performance score is not encouraging, the self-annotated dataset is a new direction that other UX experts can consider. The lan-

guages of the dataset might also have affected the results, as the first dataset was completely in English, while the second dataset were comprised of Bangladesh e-commerce websites. The authors did not mention the language of the second dataset, therefore language may or may not have been a confounding factor. Therefore, the Sazid et al.'s model is prone to overfitting, even if the in-context learning was claimed to "reduce the risk of overfitting". However, using in-context learning for machine learning models seems to have a good effect on the outcome, even if the outcome is still not optimal. The authors compared their results with another model [32] on the new dataset, and got a slightly better results (58.67 percent and 42.8 percent, respectively). Therefore, using in-context learning is a good insight that other researchers and machine learning experts need to take into consideration when building new models.

The work of Kodamdam et al [21] showed a positive example of a model that did not suffer from overfitting. They focused on detecting dark pattern on advertisements for blind screen-reader users. They built their own advertisement dataset and utilized a multi-modal classification model. Firstly, they derived text using Tesseract OCR [34], which is pre-trained using BERT model [35]. Secondly, they derived imaged using a pre-trained Vision Transformers model. All of the texts and images are then fed to a neural pipeline. Finally, they also included a "handcrafted" feature, which can include web page summary and ad text content. The training dataset consisted of 500 samples, 20% of which was used for testing. 20 annotated web pages were used to validate the model. They were able to achieve impressive scores. The model that used the most "handcrafted" features usually had the higher scores, with the highest score in accuracy being 0.862 on the test data. However, since the validation test is quite small with 20 samples, the metrics that were retrieved are not very accurate. Nevertheless, the model should still be considered as an advancement in the dark pattern detection field.

The Kodamradam et al. and Sazid et al.'s models have given the field of dark pattern detection many interesting insights. First of all, the inclusion of in-context learning in building ML models seems to be a reasonable direction to follow. As stated in Section 4.2, using context-specific taxonomies will help improve the performance of ML models. The same can be applied to the ML models themselves; the more in-context they are, the better they are at noticing smaller details about dark patterns and avoid overfitting. Secondly, Sazid et al.'s model's inclusion of "handcrafted"

features can be an encouraging step forward. "Handcrafted" features allow personalization for the ML models. In other words, they allow the models to better understand the unique features that some dark patterns have while other patterns do not. In addition, they allow the authors to manually label dark patterns. It has been shown that using deep learning only, which lets the models figure out the signs of dark patterns themselves, leads to suboptimal results [12]. Therefore, using in-context learning and creating personalized or "handcrafted" features are two good directions in which future machine learning models could follow.

#### **4.6 Fully Autonomous Models**

The fifth and final fault of current machine learning models is that they are not yet fully autonomous. Full autonomy would allow models to work with large dataset without the manual labelling by humans, and allow models to consistently update themselves with the newest dark patterns available. However, full autonomy is a very challenging task as dark patterns have varying forms of appearance [12]. Very few autonomous models are available [9], and the ones that aim to reach autonomy or semi-autonomy will be presented below.

Mathur et al.'s model [15] needs to be mentioned. The techniques they used were semi-automated, as they adopted clustering to group related UI patterns, then manually check each cluster for further evaluation. While clustering is a good option to facilitate data analysis, the manual exploration that accompanies it makes the model time-consuming. In addition, it is not conducive to adapting to new UIs. Soe et al.'s model [16] aims to be an advancement towards autonomy, but they got unsatisfactory results. In addition, the model developed by them is still semi-automated as they require manual labelling. Yada et al [32] provided a semi-automated model, specifically RoBERTa-large, that achieved the highest performance scores of all available machine learning models. However, their weakness is that they only focus on textual-based dark patterns in e-commerce websites, which are only a fraction of all dark patterns. In addition, they did not categorize the dark patterns into categories, but instead gave a binary classification (whether a webpage consisted of a dark pattern or not).

Despite the challenge, some notable models that are close to full autonomy have been created. For example, the aforementioned AidUI [12] is close to being fully autonomous. It used a synthetic dataset generation pipeline

(Faster R-CNN), which avoids manual labelling and can scale easily. When running, the model does not need labeled data and makes predictions from UI inputs, which is a key trait of autonomy. It uses a multi-phase detection pipeline that performs icon detection, text/graphical UI element parsing, heuristic-based pattern recognition, segment-level voting and UI-level decision aggregation. However, it uses heuristic pattern matching, which is human-made and static. Therefore, the model is not so conducive to changes in the patterns.

In general, it can be seen that full autonomy is still a challenge for UX researchers and machine learning experts to solve. However, a point must be noted about the nature of full autonomy. Given that dark patterns appear in various forms, full autonomy might not ever be the solution to dark pattern detection, simply because some dark patterns are only recognizable by a human [16]. For example, Nagging is a dark pattern that continuously "nags" or convinces the user to do some specific unwanted action during when a desired task is being performed. This dark pattern involves the user's intents, which can only be recognizable by a human. Therefore, even though autonomy is not achieved, a more optimal solution to dark pattern detection may lie in a mixed approach, or a semi-automated model.

#### **4.7 Other non-ML dark pattern detection models**

In light of current machine learning methods' shortcomings, it is beneficial to also consider the insights that non-ML detection methods can provide. Over the years, multiple frameworks and conceptual designs have been published with the shared goal of detecting dark patterns. The insights they can provide to improve machine learning detection methods will be discussed below.

First, one valuable lesson is that non-ML frameworks are generally better at creating usable and detailed taxonomies. Kirkman et al. created DarkDialogs in 2023 using Selenium and the CSS programming language. DarkDialogs automatically extracts consent dialogs and check for the presence of dark patterns. It uses a taxonomy that is comprised of 10 dark patterns with the category name, the impact the pattern has on the user, its legal implications and the criteria for detection. The last component of the taxonomy makes the taxonomy much more usable for machine learning purposes. Kocyigit et al. [36] created similar criteria for dark pattern

detection, which they called "measurable HCI features". They provided six dark pattern attributes that can be converted into code: Asymmetry, Restrictive, Information Hiding, Covert, Deceptive, and Disparate Treatment. In their "What makes a dark pattern... dark?" article, Mathur et al. [37] analyzed dark patterns under four lenses: individual welfare, collective welfare, regulatory objectives and individual autonomy. All of these frameworks have created specific and measurable criteria to detect dark patterns, which machine learning methods can learn from.

Another insight that comes from non-ML methods is that there are many dark patterns that are complicated and need better detection criteria. This reinforces the idea of "dynamic dark patterns" presented in Section 4.3. Kocyigit et al. [36] concluded that some dark patterns cannot be characterized by a single attribute. They added the attribute "complexity" to the existing six HCI measures, because dark patterns such as "Choice Overload" cannot be explained by any of the existing criteria. Thus, they suspected that more attributes can be added to account for the variety of interfaces. Kirkman et al. [17] concluded that human judgment was needed for some dark patterns such as Obstruction and Sneaking. Curley et al. [18] explored if a certain dark pattern can be automatically detected, manually detected or unable to be detected at all. The undetectable patterns include Bait and Switch (the user sets out to do one thing but an unwanted different thing happens) or Confirmshaming (guilting the user into opting for something). Their reason was that there is too much variance in the way these patterns appear. Thus, the common obstacle is that there are complex patterns that may not be detected automatically. This is true across the literature about dark pattern detection methods. Solutions have been proposed to deal with these seemingly undetectable patterns, which is linked to the final lesson from non-ML methods.

The final lesson from non-ML methods is that for every model or framework, an accompanied interface would support the effectiveness of it. The reason is that it allows for external input and feedback from other people. This is the solution to the undetectable dark patterns. In fact, Curley et al. has added two elements to their proposed system, which are Reporting Feature and Educational Feature, which is to allow users to report suspected presence of dark patterns and to educate them on dark patterns in general. However, their proposed system is not yet available to the public; only an initial prototype is mentioned in the paper with no link to it. DarkDialogs' authors, on the other hand, provided the source code of the

system as well as the installation and usage instructions. However, it seems that the intended audience is not the general public, but UX professionals, as the installation requires Python and other programming knowledge. A suggestion would be to make these applications into Chrome extensions or mobile applications, which are more user-friendly.

## 5. Limitations and Future Work

### 5.1 Limitations

The thesis's state-of-the-art literature review has some limitations that need to be considered. Firstly, the list of machine learning models as well as the taxonomies analyzed in the thesis is not comprehensive. There are some models that could have been analyzed, but due to time limits, they were omitted. However, the included models and taxonomies are a representative sample of the current state of dark pattern detection. Secondly, the initial analysis in Table 3.2 is a mix of subjective and objective evaluation, given the nature of some of the measurements. Additional evaluation by field experts is needed, but not possible given the scope of the thesis. Thirdly, a better evaluation would be to compare all of the machine learning models on the same dataset, which would require public access to the models. However, for many cases, this was not an option. Fourth, dark pattern taxonomies from legal bodies were largely ignored, due to the nature of the thesis being mainly about machine learning methods. A more comprehensive state-of-the-art review would also include legal taxonomies and frameworks.

### 5.2 Future Work

In light of the limitations, some future directions are described for other UX specialists to consider. This section will consist of directions for both taxonomies and machine learning models, as the two are inexplicably intertwined.

Firstly, future taxonomies need to be more detailed and context-specific. As discussed in Chapter 3, the current taxonomies are not optimal for

machine learning detection purposes. Features that could be added to taxonomies include "handcrafted" features [21] (personalized features for each of the dark patterns), severity, common scenarios [2] and UI-related features (a textbox is bigger than others, for example). Another improvement to taxonomies is the inclusion of guiding questions to support the user in categorizing dark pattern types [22]. Since there are a lot of disagreements among scholars on the classification of dark patterns, some questions that facilitates the categorization would be beneficial. One more improvement to taxonomies is to always include extensibility rules. In other words, the documentation of taxonomies should always include instructions on updating the taxonomies, whether it be through tweets [14] or contact forms. This ensures that the taxonomy stays updated with the new dark patterns that emerge as new technologies get introduced.

Secondly, future machine learning experts need to be more aware of current shortcomings and build models that address these issues. The future models need to utilize a more detailed and context-specific taxonomy. They also need to consider dynamic dark patterns, not just textual ones. In addition, full autonomy might not be a good direction because of the nature of dark patterns. Instead, a mixed approach should be prioritized. ML models should also be publicly available with an accompanied user interface, as user feedback is incredibly important in the changing field of dark pattern detection. As new dark patterns emerge, a constant feedback loop between the web or mobile users and the ML developers needs to be developed and sustained.

## 6. Conclusions

In summary, the thesis presents a state-of-the-art literature review about the shortcomings of machine learning methods in detecting dark patterns. Through meticulous analysis and research, a representative picture has been painted about the current dark pattern detection abilities of machine learning methods. The shortcomings of dark pattern taxonomies were also discussed, with a conclusion that taxonomies need to be more detailed and context-specific to improve the performance of these machine learning methods.

The future of dark pattern detection is bright, and there are many positive directions that dark pattern detection can take. The changing landscape of technology will always introduce new dark patterns, but it is the job of UX/UI designers, legal authorities, and internet security specialists to detect these deceptive patterns, and to develop new methods, especially machine learning methods, to counteract this issue. This thesis was a journey into the dark, but hopefully it is the start of the journey into the light too.

# Bibliography

- [1] H. Brignull, M. Leiser, C. Santos, and K. Doshi, "Deceptive patterns - user interfaces designed to trick you," April 2023. Accessed: Apr. 20th, 2025. [Online.]. Available: <https://www.deceptive.design/>.
- [2] L. Nie, Y. Zhao, C. Li, X. Luo, and Y. Liu, "Shadows in the interface: A comprehensive study on dark patterns," *Proceedings of the ACM on Software Engineering Volume 1, Issue FSE*, vol. 1, no. FSE, 2024.
- [3] Interaction Design Foundation, "Who is don norman?," 2021. Accessed: April 23th 2025. [Online.]. Available at: <https://www.interaction-design.org/literature/topics/don-norman>.
- [4] H. Brignull, M. Leiser, C. Santos, and K. Doshi, "Deceptive patterns - user interfaces designed to trick you," April 2023. Accessed: Apr. 20th, 2025. [Online.]. Available: <https://www.deceptive.design/laws>.
- [5] L. Moore, "Dark patterns: Not a new concept but will now be heavily regulated." Accessed: 2010-09-30. [Online.]. Available at: <https://www.williamfry.com/knowledge/dark-patterns-not-a-new-concept-but-will-now-be-heavily-regulated/>.
- [6] FTC, "Bringing dark patterns to light," 2022. Accessed: April 30th, 2025. [Online.] Available at: <https://www.ftc.gov/reports/bringing-dark-patterns-light>.
- [7] Loeb & Loeb LLP, "If you don't read this article about dark patterns, you're missing the opportunity of a lifetime," 2022. Accessed: Apr. 30th, 2025. [Online.]. Available at: <https://www.loeb.com/en/insights/publications/2022/11/if-you-dont-read-this-article-about-dark-patterns>.
- [8] C. M. Gray, C. Santos, and N. Bielova, "Towards a preliminary ontology of dark patterns knowledge," in *Extended Abstracts 2023 CHI Conf. Human Factors in Computing Systems*, CHI EA '23, (New York, NY, USA), Association for Computing Machinery, 2023.
- [9] J. Chen, J. Sun, S. Feng, Z. Xing, Q. Lu, X. Xu, and C. Chen, "Unveiling the Tricks: Automated Detection of Dark Patterns in Mobile Applications," Nov. 2024. arXiv:2308.05898 [cs].
- [10] V. Jampala, M. Zimmerman, F. S. Sierles, and M. A. Taylor, "Consumers' attitudes toward dsm-iii and dsm-iii-r: A 1989 survey of psychiatric educators, researchers, practitioners, and senior residents," *Comprehensive Psychiatry*, vol. 33, no. 3, pp. 180-185, 1992.

- [11] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of ux design," in *Proc. 2018 CHI Conf. Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), p. 1-14, Association for Computing Machinery, 2018.
- [12] S. M. Hasan Mansur, S. Salma, D. Awofisayo, and K. Moran, "AidUI: Toward Automated Recognition of Dark Patterns in User Interfaces," *2023 IEEE/ACM 45th International Conf. Software Engineering (ICSE)*, pp. 1958-1970, May 2023. ISBN: 9781665457019 Place: Melbourne, Australia.
- [13] L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, and A. Bacchelli, "Ui dark patterns and where to find them: A study on mobile applications and user perception," in *Proc. 2020 CHI Conf. Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), p. 1-14, Association for Computing Machinery, 2020.
- [14] H. Brignull, "Dark Patterns Twitter Account." Accessed: Apr. 15, 2025. [Online.] Available at: <https://x.com/darkpatterns>.
- [15] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites," *Proc. ACM Human-Computer Interaction*, vol. 3, pp. 1-32, Nov. 2019.
- [16] T. H. Soe, C. T. Santos, and M. Slavkovik, "Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way," 2022. arXiv:2204.11836v1 [cs.LG].
- [17] D. Kirkman, K. Vaniea, and D. W. Woods, "DarkDialogs: Automated detection of 10 dark patterns on cookie dialogs," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 847-867, July 2023.
- [18] A. Curley, D. O'Sullivan, D. Gordon, B. Tierney, and I. Stavrakakis, "The Design of a Framework for the Detection of Web-Based Dark Patterns," *ICDS 2021: The 15th Int. Conf. Digital Society*, 2021.
- [19] H. R. Sangaraju, S. Waris, A. Salina, D. Chandra Jadala, and G. Rao, "Smart Dark Pattern Detection: Making Aware of Misleading Patterns Through the Intended App," in *Advances in Intelligent Systems and Computing*, pp. 933-947, Hrushikesava Raju Sangaraju, Jan. 2022.
- [20] D. Nazarov and B. Yerkebulan, "Clustering of Dark Patterns in the User Interfaces of Websites and Online Trading Portals (E-Commerce)," *Mathematics*, vol. 10, p. 3219, Sept. 2022.
- [21] S. Kodandaram, "Detecting Deceptive Dark-Pattern Web Advertisements for Blind Screen-Reader Users," *Journal of imaging*, 9(11), 239., 2023.
- [22] M. Unterkalmsteiner and W. Abdeen, "A compendium and evaluation of taxonomy quality attributes," *Expert systems (Print)*, vol. 40, no. 1e13098, 2023.
- [23] J. Luguri and L. J. Strahilevitz, "Shining a light on dark patterns," *Journal of Legal Analysis*, vol. 13, no. 1, pp. 43-109, 2021.
- [24] C. Bösch, B. Erb, F. Kargl, H. Kopp, and S. Pfattheicher, "Tales from the dark side: Privacy dark strategies and privacy dark patterns," *Proc. Privacy Enhancing Technologies*, 2016.

- [25] C. Roşca, *Chapter 3. The dark (pattern) ages*. Maastricht University Press, 1 ed., 2024. Accessed: April 10th 2025. [Online]. Available at: <https://pubpub.maastrichtuniversitypress.nl/pub/chapter-3-the-dark-pattern-ages>.
- [26] C. Lacey and C. Caudwell, "Cuteness as a 'dark pattern' in home robots," in *Proc. 14th ACM/IEEE Int. Conf. Human-Robot Interaction, HRI '19*, p. 374-381, IEEE Press, 2020.
- [27] G. Conti and E. Sobiesk, "Malicious interface design: exploiting the user," in *Proc. 19th Int. Conf. World wide web, WWW '10*, (New York, NY, USA), pp. 271-280, Association for Computing Machinery, Apr. 2010.
- [28] M. Potel-Saville and M. Da Rocha, "From dark patterns to fair patterns? usable taxonomy to contribute solving the issue with countermeasures," in *Privacy Technologies and Policy* (K. Rannenberg, P. Drogkaris, and C. Lauradoux, eds.), (Cham), pp. 145-165, Springer Nature Switzerland, 2024.
- [29] ISO, "International standard iso 9241-11," 2018. Accessed: April 11th 2025. [Online]. Available at: <https://cdn.standards.iteh.ai/samples/63500/33c267a5a7564f298f02bbd65721a181/ISO-9241-11-2018.pdf>.
- [30] Y. Sazid, M. M. Nafis Fuad, and K. Sakib, "Automated Detection of Dark Patterns Using In-Context Learning Capabilities of GPT-3," in *2023 30th Asia-Pacific Software Engineering Conf. (APSEC)*, pp. 569-573, Dec. 2023. ISSN: 2640-0715.
- [31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, June 2015.
- [32] Y. Yada, J. Feng, T. Matsumoto, N. Fukushima, F. Kido, and H. Yamana, "Dark patterns in e-commerce: a dataset and its baseline evaluations," Nov. 2022. arXiv:2211.06543 [cs].
- [33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," July 2019. arXiv:1907.11692 [cs].
- [34] R. Smith, "An overview of the tesseract ocr engine," in *9th Int. Conf. Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 629-633, 2007.
- [35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. arXiv:1810.04805v2 [cs.CL].
- [36] E. Kocyigit, A. Rossi, and G. Lenzini, "A Systematic Approach for A Reliable Detection of Deceptive Design Patterns Through Measurable HCI Features," in *Proc. 2024 European Symposium on Usable Security, EuroUSEC '24*, (New York, NY, USA), pp. 290-308, Association for Computing Machinery, Nov. 2024.
- [37] A. Mathur, M. Kshirsagar, and J. Mayer, "What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods," in *Proc. 2021 CHI Conf. Human Factors in Computing Systems, CHI '21*,

(New York, NY, USA), pp. 1-18, Association for Computing Machinery, May 2021.