



PhD-FSTM-2026-016

The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 10 February 2026 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN INFORMATIQUE

by

EMRE KOÇYİĞİT

Born on 14 September 1993 in Seydişehir (Türkiye)

MITIGATING DECEPTIVE DESIGN: AN INTERDISCIPLINARY APPROACH TO AI-DRIVEN DETECTION

Dissertation defence committee

Prof. Dr Gabriele LENZINI, Dissertation Supervisor

Université du Luxembourg, Luxembourg

Prof. Dr Domenico BIANCULLI, Chairman

Université du Luxembourg, Luxembourg

Dr Arianna ROSSI, Member

Scuola Superiore Sant'Anna, Italy

Prof. Dr Gilbert FRIDGEN, Member

Université du Luxembourg, Luxembourg

Dr Qiang TANG, Member

Luxembourg Institute of Science and Technology, Luxembourg

Affidavit

I declare that this thesis:

- is the result of my own work. Any contribution from any other party, and any use of generative artificial intelligence technologies have been duly cited and acknowledged;
- is not substantially the same as any other that I have submitted, and;
- is not being concurrently submitted for a degree, diploma or other qualification at the University of Luxembourg or any other University or similar institution except as specified in the text.

With my approval I furthermore confirm the following:

- I have adhered to the rules set out in the University of Luxembourg's Code of Conduct and the Doctoral Education Agreement (DEA), in particular with regard to Research Integrity.
- I have documented all methods, data, and processes truthfully and fully.
- I have mentioned all the significant contributors to the work.
- I am aware that the work may be screened electronically for originality.

I acknowledge that if any issues are raised regarding good research practices based on the review of the thesis, the examination may be postponed pending the outcome of any investigation of such issues. If a degree was conferred, any such subsequently discovered issues may result in the cancellation of the degree.

Luxembourg, 2026-01-30

Emre KOÇYİĞİT

“Saatin kendisi mekân, yürüyüşü zaman, ayarı insandır.”

— A.H. Tanpınar

Sevgili eşim, saatimin ayarı Rana'ya.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Gabriele Lenzini. The atmosphere he fostered, the vision he set, his guidance, and his open-minded attitude played a vital role throughout my doctoral journey. Our conversations—sometimes on intellectual topics, sometimes on high-level research questions, and sometimes at the very technical level of code and algorithms—were not only enjoyable but also immensely enriching.

I extend special thanks to Arianna Rossi for her guidance, positive and constructive attitude, thorough reviews, and detailed feedback, all of which were instrumental throughout my research. This thesis would not have reached its current level without her invaluable contribution.

I am grateful to Davide Liga and Anastasia Sergeeva, whose collaboration, support, and generous sharing of expertise added great depth to my work.

My sincere thanks go to the members of my dissertation defense committee for agreeing to serve on my committee, for their time, and for their thought-provoking and valuable questions and comments.

I would also like to thank my master's thesis advisor, Banu Diri, who guided my very first steps on the path that led to this doctorate, as well as Özgür Koray Şahingöz, whose guidance, support, and vision have always been by my side.

I am grateful to my colleagues at the University of Luxembourg, with whom I shared the highs and lows of this journey. In particular, I wish to acknowledge the current and former members of the IRiSC research group at SnT, whose feedback, inspiration, encouragement, and support over the years meant more than I can express. All of these people brought smiles, encouragement, and a healthy work environment—something I believe is truly critical.

I would also like to express my gratitude to the DECEPTICON project, which was a cornerstone of this journey, contributing not only to the completion of this PhD but also to my multifaceted development. I am thankful to my colleagues within the project, who enabled me to conduct research with an interdisciplinary and holistic perspective, and of course to the FNR for funding this endeavor.

Beyond the university, I thank my friends and relatives for their unwavering support throughout this journey. I have always felt your presence behind me.

My father, M, and my mother, C, have always stood behind me and supported me unconditionally. Every achievement I have earned during and before this doctorate bears their imprint, and I can never fully repay what they have given me. *Teşekkür ederim anne, baba, hakkınızı ödeyemem.*

Finally, my beloved wife—thank you for your constant support and selfless sacrifices. Without you, this doctorate simply would not have happened. While I was pursuing a PhD, you were earning your own PhD in Patience. Your understanding and support were my greatest anchor in the most difficult moments. My son Rutkay, you contributed in your own way by graciously giving up some of your playtime on those exhausting days. And my daughter İnci, who joined our family toward the end of this journey—you too left your mark on this thesis with the joy and energy you brought, along with, of course, those sleepless nights. You have all always made me feel like the luckiest person in the world. Looking back, this journey has been a truly unique and invaluable experience—one that I will carry with me always.

Abstract

Deceptive designs (i. e., dark patterns) are design strategies that manipulate or force users to make decisions against their best interests. Since their impacts caused various harms such as privacy violations, financial loss, etc., and raise legal or ethical concerns, effective dark pattern detection strategies are essential to mitigate the risks they pose. Nevertheless, there are several challenges that hinder effective dark pattern detection, which can be grouped into three problem domains. First, although the diverse nature of dark patterns has been systematically categorized and described in various taxonomies, the assessment of their presence lacks objective criteria and measurement instruments. Second, dataset limitations, such as limited size and diversity, as well as labeling inconsistencies or incompatibilities, create challenges in the development and training of robust detection tools. They also invalidate a reliable evaluation of the effectiveness of the tools. Third, traditional detection approaches, whether rule-based or single-model AI methods, struggle with the evolving, diverse, and multimodal nature of dark patterns. Additionally, the role of detection tools needs to be clarified, as current solutions offer limited effectiveness, transparency, and explainability.

This thesis addresses these challenges through several contributions, adopting an interdisciplinary approach that integrates computer science, artificial intelligence, human-computer interaction, law, and user experience design to ensure both technical rigor and practical usability. To mitigate subjectivity in dark pattern detection, measurable features are proposed within a systematic and structured framework for one of the most common dark pattern use cases, namely cookie consent processes that allow a more objective assessment. This thesis further examines existing dark pattern datasets and identify critical quality issues, such as limited representativeness and noisy labeling, which have not been systematically presented and analyzed before. In response, a benchmark dataset is built, annotated by dark pattern experts and aligned with a recent unifying taxonomy to mitigate the aforementioned issues. To contribute to the data scarcity problem, a multi-agent framework employing Large Language Models for dark pattern data augmentation is developed, which is validated by showing improved performance of fine-tuned models in the task of dark pattern detection. Finally, the work focuses

on Multimodal Large Language Model-based approaches and efficient detection tools are built by leveraging recent techniques such as Retrieval Augmented Generation (RAG) and Chain-of-Thought (CoT), while integrating the concept of measurable features, which are proposed to address the subjectivity problem, to enhance both accuracy and explainability of the model. The proposed detection strategies were evaluated through quantitative analyses, expert interviews, and empirical comparisons. Furthermore, the study extends to open-source Multimodal Large Language Models, conducting the first empirical evaluations of their performance in the dark pattern detection task, using various prompt engineering strategies and comparing their effectiveness with proprietary models. The thesis concludes with a discussion of the remaining challenges, limitations, and open problems from an interdisciplinary perspective.

Index

1	Introduction	1
1.1	Understanding Dark Patterns	4
1.2	Research Gaps and Objectives	7
1.3	Research Questions	9
1.4	Thesis Overview	11
1.5	Scientific Contributions	14
2	Background and Related Work	19
2.1	Taxonomies and Descriptions of Dark Patterns	20
2.2	Attributes of Dark Patterns	23
2.3	Applications and Use Cases	26
2.4	Automated Detection of Dark Patterns	27
3	Assessing Features of Deceptive Designs	31
3.1	Introduction	31
3.2	Cookie Consent Process Background	33
3.2.1	Human-Computer and Machine-to-Machine Interaction Layers of the Cookie Consent Process	34
3.3	Methodology	35
3.4	Features in Cookie Consent Processes	38
3.4.1	Human-Computer Interaction-based Features	39
3.4.2	Machine-to-Machine Interaction-based Quality Features	41
3.5	Discussions	43
3.6	Limitations and Future Research Directions	44
3.7	Wrap-up and Perspectives	46
4	Reliable Detection of Deceptive Design Patterns	47
4.1	Introduction	48

4.2	Terminology	51
4.2.1	Type	51
4.2.2	Attribute	51
4.2.3	Measurable HCI Feature	52
4.3	Methodology	53
4.3.1	Identifying Dark Patterns in Cookie Consent Processes	53
4.3.2	Attributes of Deceptive Design Patterns in the Target Use Case	55
4.3.3	Identifying Measurable Features of Cookie Consent Processes	56
4.3.4	Describing Deceptive Design Pattern Attributes Based on the Features	57
4.4	Identifying, Mapping and Describing Deceptive Design Patterns	57
4.4.1	Identifying the Attributes Characterizing Deceptive Design Patterns Types	59
4.4.2	Defining Measurable Features of Dark Patterns	61
4.4.3	Measurable Feature-based Descriptions of the Dark Pattern Attributes	61
4.4.4	Applied Example of Feature-Based Dark Pattern Detection	63
4.5	Discussion	64
4.5.1	Identification of Dark Pattern Types in Cookie Consent Processes	64
4.5.2	Attributes Characterizing Dark Patterns	64
4.5.3	Measurable Features	67
4.5.4	Relevance	67
4.5.5	Future Improvements	69
4.5.6	Limitations	69
4.6	Wrap-up and Perspectives	70
5	Deceptive Design Datasets	71
5.1	Introduction	72
5.2	Background	73
5.3	Research Gap	76
5.4	Methodology	78
5.4.1	Evaluation	82
5.5	Results	82
5.6	Limitations	86
5.7	Closing Remarks	86
6	Dark Pattern Data Augmentation	87
6.1	Introduction	87

6.2	Background	89
6.2.1	Dark Patterns	89
6.2.2	Source Dataset	90
6.3	Methodology	90
6.3.1	Pre-processing	90
6.3.2	Augmentation	92
6.3.3	Evaluation	94
6.4	Results	97
6.4.1	Detection Performance of GPT-4 and Llama3 on Source Data	97
6.4.2	Zero-shot Prompting	99
6.4.3	Few-shot Prompting	99
6.4.4	Chain of Thought (CoT) Prompting	99
6.4.5	Agreement Among GPT-4 and Llama-3 While Classifying the Augmented Data	100
6.4.6	Human Evaluation of Augmented Dark Pattern Data	100
6.4.7	Parameter-Efficient Fine-Tuning Low-Rank Adaptation Over the Aug- mented Data	101
6.5	Discussion	104
6.6	Wrap-up and Perspectives	104
7	DeceptiLens: Deceptive Pattern Detection	107
7.1	Introduction	108
7.2	Methods	110
7.2.1	Research Gaps & Research Questions	110
7.2.2	DP Dataset	111
7.2.3	Design	112
7.2.4	Expert Evaluation	113
7.2.5	Ethical and Legal Aspects	118
7.3	Results	118
7.3.1	Performance of DeceptiLens Compared to Experts	118
7.3.2	Clarity, Correctness, Completeness and Verifiability of the DeceptiLens’ Explanations	120
7.3.3	Qualitative Feedback on the Explanations	121
7.4	Discussions	124
7.4.1	Overall System Performance	124

7.4.2	How Might the Presence of Dark Patterns Reliably Assessed by Human Beings and by MM-LLMs?	124
7.4.3	How Might We Make the Explanations More Reliable While Avoiding Overreliance?	126
7.4.4	How Might We Enhance the Approach?	127
7.5	Compliance With the AI Act	129
7.6	Limitations	129
7.7	Future Work	130
7.8	Wrap-up and Perspectives	130
8	Evaluating Dark Patterns in GUI	133
8.1	Introduction	134
8.2	Research Gap	136
8.3	Methodology	137
8.3.1	The Task of Dark Pattern Detection	137
8.3.2	Model Selection	138
8.3.3	Prompt Engineering Techniques	138
8.3.4	Evaluation	139
8.4	Results	140
8.4.1	Open Source Models	140
8.4.2	Comparison	145
8.5	Discussion	147
8.6	Limitations	149
8.7	Wrap-up and Perspectives	150
9	Discussions and Closing Remarks with Future Work	153
9.1	Why Detecting Dark Patterns Matter?	153
9.2	Challenges of Detecting Dark Patterns	156
9.3	Construction and Validation of Dark Pattern Datasets	159
9.4	Developing An Automated Dark Pattern Detection Tool	162
9.5	Limitations, Open Problems and Future Work	165
9.5.1	Moving Beyond Static Visual Analysis: Developing Agentic AI-based Detection Systems	165
9.5.2	Need for Objective Criteria: Towards Interdisciplinary Formalization	166
9.5.3	Performance Enhancement Strategies for LLMs	166

9.5.4	Open versus Proprietary LLMs for Dark Pattern Detection	167
9.5.5	Dark Pattern Dataset: Establishing a Collaborative Platform for Dataset Expansion and Expert Validation	168
9.5.6	Where and How Dark Pattern Detection Tools Should Be Used	168
9.6	Closing Remarks	169
A	Overview of Dark Pattern Taxonomies	197
B	Cookie Consent Scenarios, Measurable Features, and Source Websites	199
C	Feature Association and Measurable Feature Taxonomy	203
D	Dark Pattern Categories and Sources of the Curated Dataset	207
E	Detailed Classification Reports: Comparative Analysis of Prompting Strategies	209
F	Semi-Structured Expert Interview Protocol	211

List of Figures

1.1	Interest in the topic of "Dark Patterns" over time based on Google Trends	2
1.2	Yearly number of Scopus-indexed publications containing "Dark Patterns" in the title	3
1.3	Examples of common dark patterns: (a) False Hierarchy, (b) Disguised Ads, (c) Sneak into Basket.	6
2.1	An example of cookie consent process illustrating asymmetric design with click counts for "accept all" and "reject all."	25
3.1	Cookie consent flow phases	33
3.2	User Action and User Interface layers of cookie consent flow of website 2	34
3.3	User Action and User Interface layers of cookie consent flow of website 3	36
3.4	Multi-layer interaction flow (human-computer and machine-machine) of the cookie consent process	37
3.5	Extracted machine-to-machine interaction features from the cookie consent process.	44
4.1	Methodology workflow exhibiting the steps and corresponding research questions.	53
4.2	Examples of visual, linguistic, and process-based feature extraction from a cookie consent process.	58
4.3	Association between dark pattern attributes and types in cookie consent processes, classified as required, optional, or not required	60
5.1	Distribution of dark pattern categories after removing empty and duplicate entries in [19]	74
5.2	Dark pattern category - type distribution of UI-based dataset by Mansur et al. [72]	75
5.3	Examples labeled as dark pattern in existing datasets	77
5.4	A true positive example of <i>Countdown Timer</i> dark pattern	78
5.5	A false positive example of <i>Countdown Timer</i> dark pattern	79

5.6	Overview of dataset creation process	81
5.7	Case study: dark pattern expert opinion regarding UDS labeling.	84
5.8	Performance of open-source models across datasets (UDS vs our benchmark dataset)	85
6.1	Distribution of dark pattern categories of DS1 (%)	91
6.2	Flow of the dark pattern text data augmentation by multiple agents	93
6.3	The prompt to generate augmented dark pattern text data (CoT part zoomed).	94
6.4	The CoT prompt for the controller model.	95
6.5	Pipeline of the PEFT approach.	96
6.6	Classification prompts for the binary and multiclass classifications.	97
6.7	Number of agreed and disagreed augmented text data according to the binary classification of the evaluator models for each prompting technique	100
6.8	Number of agreed and disagreed augmented text data according to the multiple classification of the evaluator models for each prompting technique	101
6.9	Mean F1-score results of the PEFT using the various augmentations, considering both the binary (top) and the multiclass (bottom) scenarios.	102
6.10	Confusion matrix for human evaluators' annotations on augmented data	103
7.1	A true positive example of <i>Countdown Timer</i> dark pattern	111
7.3	Overview of the deceptive pattern detection and reporting approach design.	112
7.2	A false positive example of <i>Countdown Timer</i> dark pattern	112
7.4	A prompt for the MM-LLM and <code>target-dp</code> refers to the targeted DP in the UI, and its description is stored in <code>target-dp-description</code> variable.	114
7.5	An example question for the experts that was asked in the Task 1.	115
7.6	An example question for the experts that was asked in the Task 2.	116
8.1	A simulated example of <i>False Hierarchy</i> dark pattern.	134
8.2	A simulated example of <i>Confirmshaming</i> dark pattern.	134
8.3	Baseline performance scores of open-source models with prompt strategy (i)	141
8.4	Performance scores of open-source models with prompt strategy (ii): basic + contrastive learning	142
8.5	Performance scores of open-source models with prompt strategy (iii): basic + chain-of-thought	143
8.6	Performance scores of open-source models with prompt strategy (iv): basic + contrastive + chain-of-thought	144

8.7	Performance scores of open-source models with prompt strategy (v): category-specific expert chain-of-thought	145
8.8	Comparison of the models' best performing scenarios	146
8.9	GPT-4o performance on both datasets	147
8.10	Dark pattern category-based sensitivity status of best performing model - GPT-4o	148
B.1	Cookie consent scenarios based on user actions in the first two phases	200
C.1	Comparison of first- and second-round expert mappings of dark pattern attributes.	204

List of Tables

1.1	The attributes of dark patterns [18]	7
1.2	Chapters and the corresponding research questions.	10
4.1	Dark pattern attributes with their descriptions from Mathur et al. [18]	52
4.2	Dark pattern types used in cookie consent processes	59
5.1	Samples from dark pattern text dataset [19]	73
5.2	Example annotations from a family–category game app video. FA: Forced Action, DA: Deceptive Advertising, NG: Nagging	74
5.3	Dark Pattern (DP) categories in our expanded dataset. The newly added DP categories are shown as bold	80
5.4	Descriptions of the newly added DP categories [29] and our criteria determined for their labeling.	83
5.5	Agreement (%) between UI Guard labeling and open-source MM-LLMs on overall, non-dark-pattern (Non-DP), and dark-pattern (DP) samples.	85
6.1	Dark pattern categories and subcategories in <i>DS1</i> [19]	90
6.2	Classification report of the GPT-4 for the source data	98
6.3	Classification report of the Llama-3 for the source data	98
6.4	Classification report of agreed decisions of the both models for the source data	99
7.1	Target deceptive design pattern categories	111
7.2	Demographics of the interviewed participants	117
7.4	Example input and output of the DeceptiLens	119
7.3	DeceptiLens’s performance considering expert opinions	120
7.5	Evaluation results of the explanations by experts on a 1-5 Scale. (1: strongly disagree, 5: strongly agree)	121

A.1	Chronological Overview of Foundational and Influential Dark Pattern Taxonomies, Illustrating the Evolving and Expanding Landscape	198
B.1	Source websites that are selected to inspect cookies consent processes	199
B.2	Features of cookie consent process relevant for dark pattern detection	201
C.1	Description of measurable features and their design categories	205
D.1	Sources of the curated dataset	207
D.2	Target deceptive design pattern categories	208
E.1	Binary classification report of the generator model that employs zero-shot prompting	209
E.2	Dark Pattern Category Classification Report of the generator model that employs zero-shot prompting	209
E.3	Binary classification report of the generator model that employs few-shot prompting	210
E.4	Dark pattern category classification report of the generator model that employs few-shot prompting	210
E.5	Binary classification report of the generator model that employs Chain of Thought prompting	210
E.6	Dark pattern category classification report of the generator model that employs Chain of Thought prompting	210

Acronyms

AI Artificial Intelligence

NLP Natural Language Processing

GUI Graphical User Interface

UI User Interface

UX User Experience

API Application Programming Interface

UML Unified Modeling Language

LLM Large Language Model

EU European Union

EDPB European Data Protection Board

GDPR General Data Protection Regulation

EC European Commission

DSA Digital Service Act

CoT Chain-of-Thought

RAG Retrieval Augmented Generation

RLHF Reinforcement Learning with Human Feedback

HTML Hypertext Markup Language

CSS Cascading Style Sheets

Chapter 1

Introduction

Dark patterns are manipulative digital designs that mislead, coerce, or deceive users against their interests, often causing harm such as privacy infringement, and diminish user autonomy [1]. In 2010, user experience (UX) expert Harry Brignull launched a website, darkpatterns.org, which gave a name to this unethical online practice.¹ Researchers also began exploring dark patterns, and three years after the website's launch, a conceptualization of dark patterns, particularly in game design, was published by Zagal et al. [2]. They examined non-player-centric game designs that are potentially unethical and have dubious objectives. In the following years, scholars from various fields, including privacy, user experience design, and media studies [3] examined distinct aspects of this phenomenon, such as ethical concerns, the effectiveness of dark patterns in a psychological context [4], and described its characteristics [5].

Over time, dark patterns have also attracted increasing attention from regulators [6, 7, 8], as they may lead to illegal practices. In particular, the enactment of privacy-related regulations, such as GDPR (2018) and CCPA (2020), has led to a significant increase in scholarly and industry interest in the topic of dark patterns. As a result of collective efforts, dark patterns have been through both guidance and binding legislation. For example, at the end of 2021, they were identified as "unfair commercial practices" in the Commission Notice on Directive 2005/29/EC, which included examples, e.g., "visual obscuring" or "trick questions" to manipulate users, and references to breached articles. The following year, the European Data Protection Board issued guidelines [9] to prevent dark patterns that violate GDPR requirements, e.g., freely given consent Article 7. And nearly one year from the Commission Notice, binding legislation was enacted in the form of the Digital Services Act[10]. Efforts continued and converted into comprehensive and evidence-based reports [11], and lastly, the Digital Fairness Act is being prepared as a recent

¹Note: The term dark pattern has been replaced by "deceptive design patterns" or "deceptive pattern". While there is a tendency to use "deceptive design", e.g., website darkpattern.org has been changed as deceptive.design, we used the terms interchangeably.

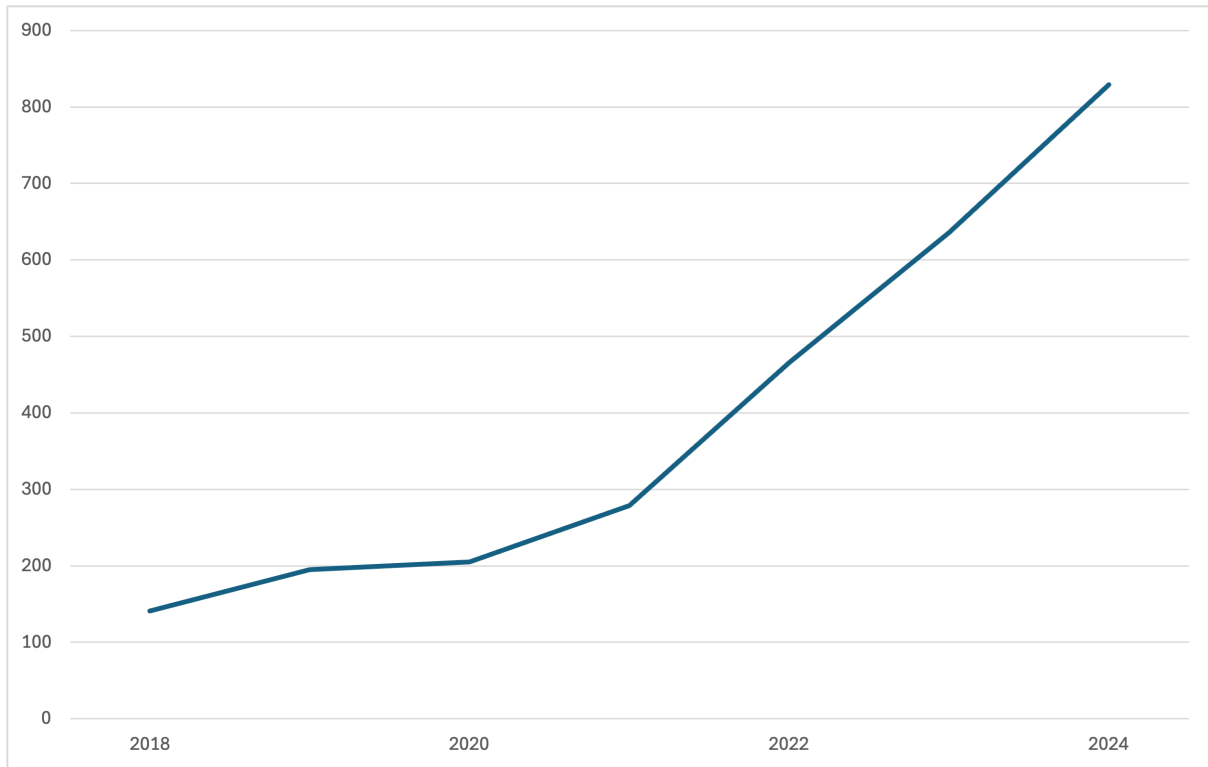


Figure 1.1: Interest in the topic of "Dark Patterns" over time based on Google Trends

complementary initiative, focusing on dark patterns and other problematic digital practices to ensure fairness [12].

Today, interest in dark patterns remains strong and its popularity continues to grow among the general public (see Figure 1.1), and academic scholars (see Figure 1.2). Furthermore, media outlets have reported on dark patterns emphasizing their impact on users and business [13, 14].

Dark patterns can cause a range of negative consequences, such as loss of autonomy, erosion of trust, financial loss, psychological (emotional) distress, and even physical harm to users, as well as collective harm, such as loss of trust in market [15]. Beyond raising ethical concerns, dark patterns also result in explicit legal infringements, such as privacy violation [16] and legal requirement breach, e.g. the requirement of freely given consent under Articles 4(11) and 7(4) of the GDPR [17]. Therefore, developing methods for the **detection of dark patterns** is both crucial and necessary to mitigate their harms.

Existing taxonomies, guidelines, and definitions of dark patterns have made valuable contributions, such as facilitating understanding of the phenomenon, raising awareness, supporting early classification efforts, and informing regulatory discussions. However, they largely remain descriptive, high-level, and non-actionable [18]. For example, Mathur et al. [19] describe asymmetric design that is one of the characteristics of dark patterns, as imposing an "unequal user burden", yet the notion of burden is ambiguous and context-dependent. To objectively deter-

mine the presence of dark patterns, it is necessary to establish **objective criteria that are measurable and operational** rather than abstract or interpretive. Such objective criteria not only enable consistent assessment, but also play a crucial role in explaining and justifying detection decision.

Detecting dark patterns poses significant challenges because the deceptive and manipulative characteristics of their design are often difficult to detect or quantify, unlike traditional cybersecurity threats such as phishing, where harmful content can be clearly identified.

Dark patterns can be likened to insidious and adaptive pathogens, which are harmful entities capable of altering form and spreading across diverse environments. This analogy is grounded in three key considerations. First, although dark patterns are recognized as harmful by regulatory authorities [8, 7] and scholars [20, 21], they reside within legitimate and mainstream digital systems, such as e-commerce websites [19, 22, 23] or online games [2, 24, 25]. Second, they can emerge in numerous contexts and use cases, including online video games [26], mobile applications and e-commerce websites[19], cookie consent processes [27], healthcare, and social media [28]. For instance, a dark pattern may camouflage advertisements within a user interface to induce inadvertent clicks, namely *Disguised Ads* [29], while another may use emotionally manipulative language to pressure users to subscribe to a newsletter, i.e., *Confirmshaming* [29]. Third, dark patterns manifest through different modalities, including visual, textual, or process-

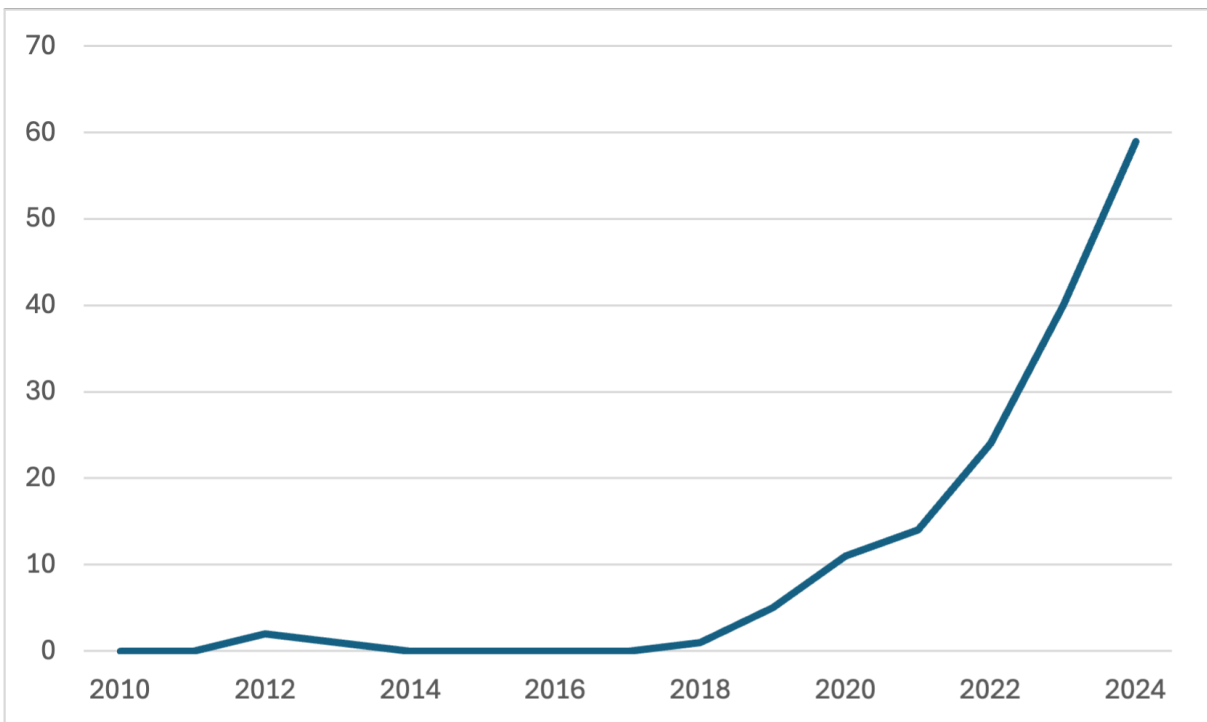


Figure 1.2: Yearly number of Scopus-indexed publications containing "Dark Patterns" in the title

based. The hidden advertisement example relies on visual deception and therefore requires image-processing capabilities for detection, whereas emotionally manipulative prompts necessitate semantic understanding of text. Other patterns can be embedded within interaction flows of website, revealing themselves only through user engagement, and some even extend to alternative interfaces such as voice assistants of home-type robotic devices [30]. The ubiquity of dark patterns across diverse digital environments makes their potential harms particularly concerning and underscore the need for effective and scalable **automated detection** solutions.

In the context of automated detection, different paradigms can be identified, such as intrusion detection and phishing detection. While the first relies primarily on unsupervised anomaly detection methods that identify deviations or outliers from a baseline [31], the second requires supervised learning approaches with explicitly labeled datasets [32]. Automated dark pattern detection aligns more closely with phishing detection paradigm since there is not standard or normal design to benchmark against. Instead, the goal is to identify manipulative or deceptive design elements within legitimate interfaces. Both of them can be formulated as classification problems. However, they are not identical: the distinction between phishing and dark pattern detection does not lie in the need for labeled dataset, but in the nature and complexity of the labeling process itself. When a website or email is identified as phishing, it is typically illegitimate, and even non-experts can label it accurately. In contrast, (i) dark patterns exist within legitimate websites, mobile applications, or online video games, and (ii) while phishing aims primarily to steal information or money, dark patterns serve diverse purposes, such as getting more personal data, inducing unnecessary purchases, or manipulating user consent. Therefore, the first essential step in dark pattern detection is to **define objective and operational labeling criteria**, which can then serve as the foundation for constructing reliable datasets. **Structured and labeled datasets** are crucial for both evaluating detection methods, tools, or models and training, particularly artificial intelligence-based detection models.

Although objective and operational labeling criteria and labeled datasets are crucial for automated dark pattern detection, existing research has not fully addressed these issues (see Section 1.2). After presenting additional background on dark patterns with illustrative examples, the following sections outline the research gaps, objectives, and questions and explain how this thesis contributes to addressing the aforementioned issues.

1.1 Understanding Dark Patterns

Dark patterns have been described through multiple definitions over the years. For instance, Brignull, who first coined the term, defined them on his website `deceptive.design` as "tricks

used in websites and apps that make you do things that you didn't mean to, like buying or signing up for something" [33]. In academia, the most cited publication on dark patterns defines them as "instances where designers use their knowledge of human behavior (e.g. psychology) and the desires of end users to implement deception functionality that is not in the user's best interest" [5]. While numerous definitions of dark patterns exist, Mathur et al. compiled 19 different definitions from academic and governmental sources and examined them across four dimensions: user interface (UI) properties, effects on users, designers' duty, and resulting benefits or harms [18]. From this analysis, common components emerge as subverting user intent or preferences, tricking, deceiving, manipulating, misleading, or coercing users, and ultimately benefiting the service provider and/or harming the user. Figure 1.3 presents examples of common dark patterns, e.g., *False Hierarchy*, *Disguised Ads* and *Sneak into Basket*.^{2 3 4}

As research on dark patterns expanded, both academic and industry efforts began developing taxonomies to categorize them, providing a structured understanding of their diverse dimensions. The dark pattern taxonomies created by both academic scholars and governmental associations are chronologically listed in Table A.1, in appendix. While each taxonomy contributes to describing and conceptualizing dark patterns, differences in classification priorities and methodological approaches create challenges for comparison and synthesis. For instance, some taxonomies describe similar patterns using different labels, while others concentrate on specific contexts and therefore do not address all types.⁵ Furthermore, the dark pattern types' level of abstraction varies across taxonomies, i.e., granularity of dark pattern types is various, from general classifications to specific low-level types. The growing number of taxonomies has not necessarily simplified the work of researchers; rather, it has led them to question how a given dark pattern type in one taxonomy corresponds to those in another. Gray et al. [29] addressed this challenge by producing a mapping table that integrates both academic and regulatory taxonomies. In this thesis, we refer to this ontology as it is both current and comprehensive.

²**False Hierarchy** refers to a dark pattern where an option, which usually does not serve the user's interest, is made more visually prominent than the other option [29].

³**Disguised Ads** refer to a dark pattern in which the elements of an advertisement are not clearly marked or styled [29].

⁴**Sneak into Basket** refers to a dark pattern that leads to unintentional purchase by adding unwanted items without user consent or explicit action [29].

⁵The term of 'type' refers to the dark pattern categories such as 'Forced Action', 'Nagging', 'Confirmshaming', etc.

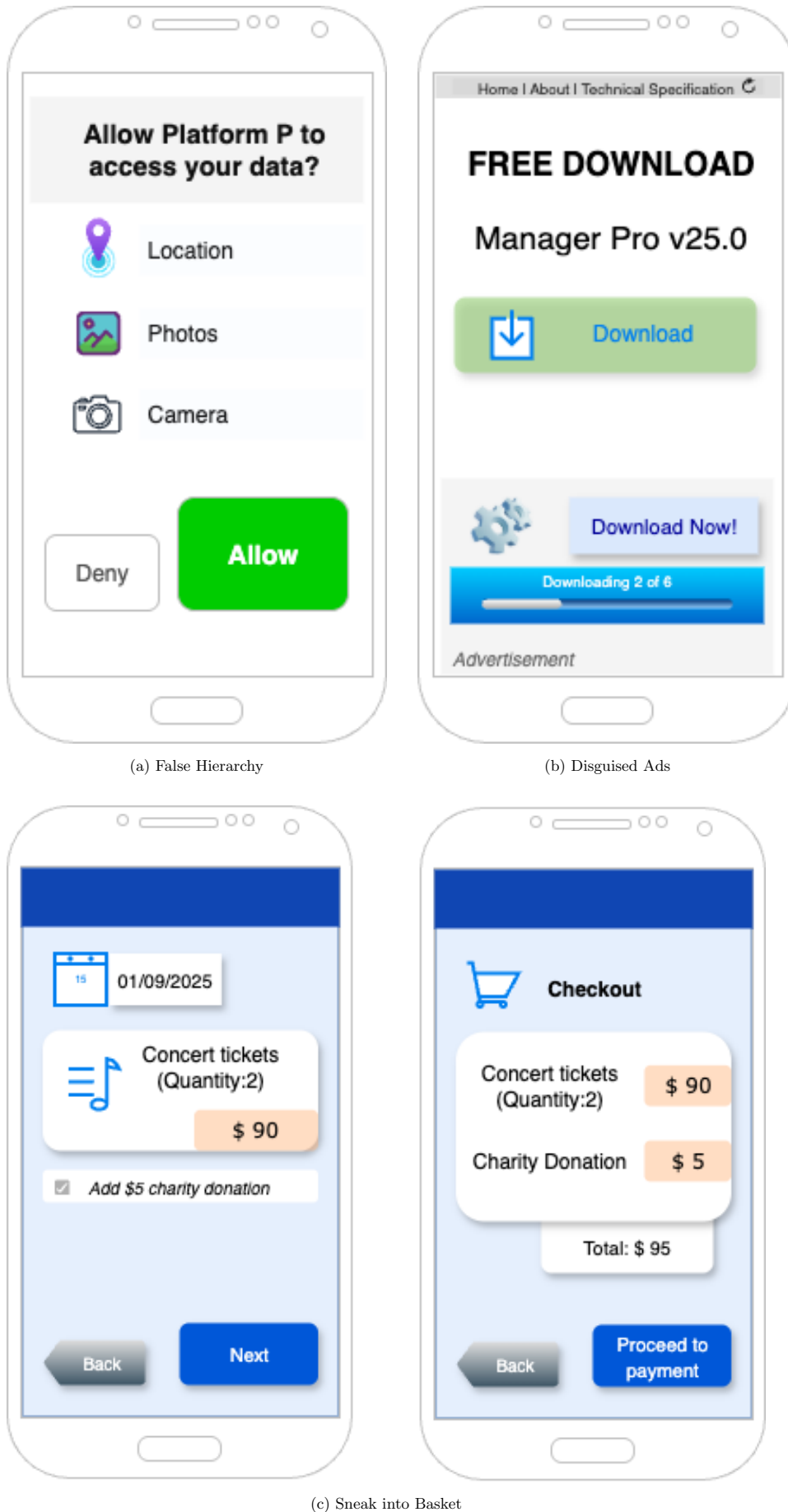


Figure 1.3: Examples of common dark patterns: (a) False Hierarchy, (b) Disguised Ads, (c) Sneak into Basket.

The extensive spectrum of dark pattern types exhibits several common attributes. While not every attribute applies to every dark pattern type, many types contain one or more of the following attributes: "asymmetry, restriction, disparate treatment, covert, deception, and information hiding" [18]. These attributes, along with their descriptions, are given in Table 1.1.

Table 1.1: The attributes of dark patterns [18]

Attribute	Description
Asymmetry	Unequal burdens on choices available to the user.
Restriction	Eliminating certain choices that should be available to users.
Covert	Hiding the influence mechanism from users.
Deception	Inducing false beliefs in users either through affirmative misstatements, misleading statements, or omissions.
Disparate Treatment	Disadvantaging and treating one group of users differently from another.
Information Hiding	Obscuring or delaying the presentation of necessary information to users.

1.2 Research Gaps and Objectives

In the previous sections, it is discussed that why and how dark patterns pose threats for digital user journey, and emphasized the importance of detecting them in an effective and robust manner. At this point, the first challenge emerges due to the fact that definitions of dark pattern types often contain vague, ambiguous expressions that allow for subjective interpretation. It becomes difficult to conduct an objective identification or assessment. For example, one of the most common and well-known dark patterns, *Disguised Ads*, is defined as being "not clearly marked as an advertisement" [29]. The term "clearly marked" is not expressed in measurable or observable terms, which means that experts may classify the same instance differently. Although some *Disguised Ads* cases are straightforward, others present ambiguity. For example, some experts regard adequately contrasted advertisement content as a clear indicator and thus refrain from it as a dark pattern, whereas others expect explicit textual or visual cue identifying it as an advertisement [34]. While such definitions are valuable for descriptive purposes and for understanding the concept, they are insufficient for objective evaluation. Objective evaluation is critical, as dark patterns frequently involve unlawful practices and relying on subjective judgment is inadequate for providing evidence-based assessments or informing enforcement and compliance efforts. Evaluation based on objective criteria is also important for ensuring con-

sistency between different use cases and to minimize the variability in expert judgments. In addition, objective metrics make decisions traceable, explainable, and transparent. To address this gap, the first research objective of this thesis is defined as:

Research Objective 1: *Defining measurable instruments for the objective assessment of dark patterns.*

While tools that enable objective evaluation are essential for accurate identification and false positive minimization, effectively combating dark patterns also requires additional components to build scalable, efficient, and automated solutions, given that dark patterns can manifest themselves in highly variable contexts within the vast and continuously growing ecosystem of digital platforms, such as websites, mobile applications, IoT devices, and other emerging technologies. In this context, a critical component that supports automated solutions is the availability of dark pattern datasets. Such datasets are particularly important for primarily two reasons: first, they enable the evaluation of the performance of dark pattern detection tools; and second, they provide training data for artificial intelligence model-based solutions. To be effective, these datasets must be structured and labeled, serving as reference data that allow reliable benchmarking and comparison of different detection approaches. Although there are various dark pattern data sources, there remains a significant gap: comprehensive, structured datasets with a sufficient number of instances and consistent, reliable annotations are still lacking. To address this gap, the second research objective is defined as:

Research Objective 2: *Creating a structured and reliable labeled dark pattern dataset.*

The third research gap identified in this thesis is the lack of an automated dark pattern detection tool that supports transparency, is efficient in addressing the multimodality of dark patterns, and is validated by experts in the dark pattern domain. Several detection tools exist, but they face significant reliability and applicability issues due to quality problems in the datasets and dark pattern identification criteria. For example, when one of the dark pattern datasets [35] labels interfaces with small close buttons as dark patterns even though experts and up-to-date and established ontology definitions contradict [29]. The important problem is that, while this leads to misclassifications of interfaces and increases false positives, it is not seen in the detection tools' evaluations because the reference is already incorrect. Another major limitation relates to transparency. Although some tools partially address this by spotting the

area of dark patterns, they fail to provide clear, complete, and verifiable explanations for why a specific instance is recognized as a dark pattern. These explanations are essential for both validation and explainability. Considering the potential users of this tool, we position it as a decision-support tool, so the explainability and transparency of the tool is critical. The final decision will be made by human experts; therefore, supporting them with interpretable evidence and explanations is important. In this direction, third and final research objective defined as:

Research Objective 3: *Developing an expert-validated, automated and AI-based dark pattern detection tool which supports transparency.*

All of these research objectives are related to and feed each other. For example, objective dark pattern instruments can be used while labeling the dark pattern dataset. And the dataset can be used to train AI-based automation tool. While measurable dark pattern assessment features can be used in the automated detection tool to enhance explainability of the tool, the tool can help to augment the dataset.

1.3 Research Questions

The studies conducted throughout this thesis were structured around the three main research objectives defined in Section 1.2. For each objective, the corresponding research questions were formulated as follows:

RQ1: What are the measurable functional properties of dark patterns and the assessment instruments to define them?

Dark patterns, while varying depending on their use cases and the environments in which they appear, share recurring and measurable characteristics. For example, *False Hierarchy*, a widespread dark pattern type can be encountered in a website's cookie consent banner as well as in a mobile application's permission request pop-up screen. Although the context, purpose, and content differ, the common feature is that the option unfavorable to the user is designed to be far less visible or selectable compared to the alternatives. And the question (**RQ1**) seeks to explore whether it is possible to formalize this feature in a measurable way that minimizes ambiguity.

RQ2: What considerations and methods should be taken into account when building labeled and structured dark pattern datasets?

Since the dark pattern dataset is intended to be used for training and/or testing automated AI-based dark pattern detection tools, it must be structured, consistent, and machine-readable.

That is the same for the tool developed under the third research question (**RQ3**). A collection of website screenshots or social media posts that contain dark pattern complaining need to be processed and transformed into a form that can be used by AI models. This research question investigates what attributes a dark pattern should have. At the same time, the evaluation of the developed detection tools also requires a labeled dataset. Another part of this question concerns the context and criteria for such labeling. Furthermore, this question requires an investigation of methods for creating datasets to contribute to the data scarcity problem.

RQ3: Which AI-based approaches can be developed for the automatic detection of dark patterns while providing explanations supporting informed decision-making, and what challenges and potential solutions do they involve?

Developing an automated detection tool in digital deception detection domain or in cybersecurity involves several common challenges such as limited access to data for decision-making and the continuous emergence of novel deceptive strategies. In this thesis, the focus is not on general detection issues but on those specific to dark patterns and on identifying effective strategies to address them. For example, rule-based detection systems, whether designed for identifying spam emails, phishing websites, or dark patterns, tend to be vulnerable to new or previously unseen cases (e.g., zero-day attacks or novel interface designs). Moreover, while verifying and labeling a phishing website is typically straightforward even for non-experts, recognizing a dark pattern often requires nuanced interpretation of context and intent, making annotation and validation more complex. Additionally, unlike URL-based phishing detection, which relies only on textual or structural data, dark pattern detection requires multimodal analysis that incorporates visual and linguistic cues. Considering these characteristics, this thesis addresses the use of AI-based approaches.

Table 1.2: Chapters and the corresponding research questions.

Chapter #	RQ1	RQ2	RQ3
Chapter 3	✓		
Chapter 4	✓		
Chapter 5		✓	
Chapter 6		✓	
Chapter 7	✓	✓	✓
Chapter 8		✓	✓

1.4 Thesis Overview

This thesis is structured in ten chapters. Chapter 2 presents the literature review on dark patterns. Chapters 3 to 9 address the research questions given in the previous section. Finally, Chapter 10 presents the discussions, limitations, and closing remarks with open problems for future work that could extend the contributions of this thesis. Table ?? portrays the correspondence between the chapters and the research questions they address.

Chapter 3: Towards Assessing Features of Deceptive Designs: Cookie Consent Processes inspects a common dark pattern use case, i.e., cookie consent process, and assesses the measurable features that help to identify dark patterns. Dark patterns are common in cookie banners where they are used to influence users to accept being tracked for more purposes than a data protection by default principle would dictate, thereby making them illegal. After discussing that objective, transparent, and verifiable assessment of dark patterns' qualities is missing, this chapter contributes to bridging this gap by studying several cookie processes, in particular their multi-layered information flow, and by identifying a list of observable and measurable features that can help describing the presence of dark patterns in digital consent flow use cases. It proposes thirty one of such properties that can be operationalised into metrics and therefore into objective procedures for the detection of dark patterns.

Chapter 4: A Systematic Approach for A Reliable Detection of Deceptive Design Patterns Through Measurable HCI Features extends the previous work. Detecting dark patterns among the various existing design patterns and discerning what an illegitimate design practice may depend on the subjective interpretation of expert users (such as regulators, civil society organizations, and academic researchers) who may not fully agree. It proposes a systematic approach to describe the characteristics of deceptive design patterns that are intended to reconcile the interpretations of expert users. In particular, the contributions are: (i) identifying use case-specific dark pattern types using the ontology drafted by Gray et al. [29]; (ii) clarifying the relationships between these types and the dark patterns' attributes proposed by Mathur et al. [18]; (iii) propose a list of observable and measurable user-interaction features of dark patterns that cover visual, process, and language design aspects, and (iv) describe the attributes based on our measurable features to lower the subjectivity of user' interpretation. Finally, this chapter discusses the proposal's cross-domain applicability and the potential for future work, including how to improve the descriptions of the attributes via semiformal languages, to generate an objective and usable framework to assess the presence of deceptive design patterns in digital interfaces.

Chapter 5: Dark Pattern Datasets presents an overview of the existing dark pattern

datasets. It discusses the challenges in the field regarding datasets and identifies quality issues, empirically demonstrating how label inconsistencies and varying dark pattern descriptions can affect the evaluation of dark pattern detection tools. After illustrating problematic data instances in existing dark pattern datasets, this chapter introduces a new dataset that is curated within the scope of this thesis. Each instance in the dataset is validated by domain experts who have at least one peer-reviewed publication on dark patterns, using majority voting, so every instance is annotated by three different experts. Additionally, a recent and comprehensive ontology was used as a reference to ensure consistency with established descriptions of dark patterns.

Chapter 6: Dark Pattern Data Augmentation and Fine-tuning discusses that the variety and quantity of structured and labeled dark pattern datasets are limited. This chapter focuses on leveraging Large Language Models' (LLM) sophisticated text data generation ability and proposes a dark pattern text data augmentation method by utilizing a state of art open-source language model and multi-agents framework, which has generator and controller models. Evaluation of the augmentation demonstrates that while increasing the data size, the proposal-based augmented dark pattern data preserves the same dark pattern characteristics of the source data and maintains its diversity. This study also demonstrates that dark pattern text data can be generated even on the basis of a few examples via prompt engineering techniques on the LLMs. Moreover, it shows that the augmented data can be used to fine-tune pre-trained language models using Low-Rank Adaptation to enhance their performance in detecting dark patterns.

Chapter 7: An Approach supporting Transparency in Dark Pattern Detection based on a Multimodal Large Language Model first discusses the need for multimodality in dark pattern detection and argues that the capabilities of Multimodal Large Language Model (MM-LLM) can achieve wider coverage with superior performance in the detection, while providing reasoning behind each decision. It proposes and implements an MM-LLM-based approach (DeceptiLens) that analyzes graphical user interfaces and assesses the presence of dark patterns. This study utilizes the Retrieval Augmented Generation (RAG) process in tasking the model with capturing the dark patterns, classifying their category, e.g., *False hierarchy, confirmshaming, etc.*, and explaining the reasoning behind the classifications by employing recent prompt engineering techniques, such as Chain-of-Thought (CoT). After creating a dataset by collecting UI screenshots from the literature and web sources, the agreement between the model's output and a few experts' opinions on dark pattern instances is quantified. Additionally dark pattern experts are asked to gauge the transparency of the proposed system's explanations for its classifications in terms of recognized metrics of clarity, correctness, completeness, and verifiability.

The results indicate that the proposed approach is capable of capturing the deceptive patterns in UIs with high accuracy while providing clear, correct, complete, and verifiable justifications for its decisions. One of the outputs of this study is the release of two curated datasets, one with expert-labeled UIs with dark patterns and one with AI-based generated explanations.

Chapter 8: Evaluating the Automated Detection of Dark Patterns in GUI Datasets with Open-Source Multimodal Large Language Models questions the performance of the most widely used open-source Multimodal Large Language Models (MM-LLMs) in the task of dark pattern detection considering the call of the academic research community for fostering open-source technologies, and the growing need for effective automated detection of dark patterns to mitigate the risks in digital platforms and information systems. First, this study inspects the existing dark pattern datasets with a focus on their use in model evaluation, and identifies several quality issues, such as inconsistency, mislabeling, too broad labeling, etc. The analysis is extended with the annotations of the DP experts and shows how they could cause a misleading interpretation of the performance of the models. Second, a dataset, which is evaluated by 14 dark pattern researchers and is labeled using a recent and comprehensive taxonomy, is selected to reduce the identified issues. The dataset, which contains web/mobile screenshots of DPs and non-DPs, is expanded by increasing the number of categories and instances, and its suitability for benchmarking is maintained by ensuring consistent inter-rater agreement. Then, this work investigates the performance of open source LLMs, i.e., LLaVA, Janus-Pro, Mistral, in the task of dark pattern recognition using the proposed expanded benchmark dataset. Performance comparison is conducted between the open-source models and the state-of-the-art proprietary model, GPT-4o. Then, common prompt engineering techniques, such as Chain-of-Thought, and a category-specific prompting strategy, which embeds expert generated rules, are applied to improve the DP recognition capabilities of the open-source models. This study presents the first empirical evaluation of open-source MM-LLMs in the task of DP detection and demonstrates that prompt engineering alone is not sufficient to close the gap between the open-source models and GPT-4o in this task, although minor performance improvements are observed in most models. The findings provide crucial insights for practitioners, such as researchers, designers, and regulators, who are working on DP detection, by presenting current limitations of open-source MM-LLMs and dataset quality issues in the context of dark pattern detection.

Chapter 9: Discussion, Limitations, and Closing Remarks with Future Work presents the main conclusions of this project. Through revisiting the research objectives and questions, we discuss how this thesis addresses them and enhances the knowledge of the dark pat-

tern research field. In addition, it discusses the limitations and open problems. After presenting potential directions for future research, this chapter concludes the thesis with final remarks.

1.5 Scientific Contributions

The following scientific papers are produced as part of this thesis:

1. Kocyigit, E., Rossi, A., & Lenzini, G. (2022). ***Towards assessing features of dark patterns in cookie consent processes.*** In *IFIP International Summer School on Privacy and Identity Management* (pp. 165-183). Cham: Springer Nature Switzerland.

- This publication forms the basis of **Chapter 3**. As a contribution to **Research Objective 1**, it examines the information flow in cookie consent processes and identifies measurable features that can be operationalized into objective metrics to assess the presence of dark patterns. This study also represents the first step in this thesis in defining measurable instruments for the objective evaluation of dark patterns.

2. Kocyigit, E., Rossi, A., & Lenzini, G. (2024, September). ***A systematic approach for a reliable detection of deceptive design patterns through measurable HCI features.*** In *Proceedings of the 2024 European Symposium on Usable Security* (pp. 290-308).

- This publication forms the basis of **Chapter 4**. It also contributes to **Research Objective 1** and complements the previous publication by developing a systematic approach for reliable dark pattern detection. In particular, it establishes relationships between the attributes of dark patterns and categories, integrates visual and textual features, and formulates dark patterns to provide objective evaluation criteria.

3. Kocyigit, E., Liga, D., & Lenzini, G. (2024, November). ***Augmenting Dark Patterns Text Data by Leveraging Large Language Models: A Multi-agent Framework and Parameter-Efficient Fine-Tuning.*** In *International Workshop on Causality, Agents and Large Models* (pp. 65-87). Cham: Springer Nature Switzerland.

- This publication forms the basis of **Chapter 6** and contributes to **Research Objective 2** by developing a framework for dark pattern text data augmentation. The study leverages the advanced language understanding and generating capabilities of Large Language Models within a multi-agent architecture and demonstrates the impact of the augmented data through comparative evaluations of fine-tuned open-source language models.

4. Kocyigit, E., Rossi, A., Sergeeva, A., Negri Ribalta, C., Farjami, A., & Lenzini, G. (2025, June). *DeceptiLens: an Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model*. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1942-1959).

- This publication contributes to **Chapter 5** and forms the basis of **Chapter 7**. First, it develops a Multimodal Large Language Model-based approach for dark pattern detection, employing prompt engineering techniques such as Chain-of-Thought and advanced methods for minimizing hallucinations, including Retrieval Augmented Generation. In addition to achieving an accurate and sensitive classification, the approach provides clear, correct, complete, and verifiable explanations that provide transparency. Furthermore, the outputs were validated by 13 dark pattern experts, making this work both novel and significant within the field. While this study primarily contributes to **Research Objective 3**, one of the outputs, an expert-annotated dataset, also contributes to **Research Objective 2**.

5. Kocyigit, E., Liga, D., Rossi, A., Sergeeva, A., & Lenzini, G. *Evaluating the automated detection of dark patterns in GUI datasets with open-source Multimodal Large Language Models*. (Manuscript submitted to *PeerJ Computer Science*).

- This work contributes to **Chapter 5** and serves as the basis of **Chapter 9**. It extends the previously developed dataset by incorporating new dark pattern categories while following the same expert validation methodology. In addition, it identifies quality issues in existing dark pattern datasets, contributing to **Research Objective 2**. Furthermore, the study evaluates the performance of open-source multimodal Large Language Models on dark pattern detection and compares with a leading proprietary model, thereby contributing to **Research Objective 3**.

For all of these works, I served as the primary and corresponding author, providing intellectual leadership in conceptualization, methodology development, coding, data curation, experimentation, and manuscript drafting. In keeping with established academic convention, the plural pronoun "we" is used throughout this thesis as the narrative voice of the research. This reflects both instances of joint action, acknowledging interdisciplinary collaboration in several studies, and the general presentation of the methodology and findings within this single-author document.

Other contributions conducted during the course of this thesis:

1. Rossi, A., Arenas, M. P., Kocyigit, E., & Hani, M. (2022, June). **Challenges of protecting confidentiality in social media data and their ethical import.** In *2022 IEEE European symposium on security and privacy workshops (EuroS&PW)* (pp. 554-561). IEEE.

- This publication represents my first collaboration at the beginning of this thesis. My contribution involved the development of Python-based pseudonymization solution to protect the confidentiality in social media data containing posts related to dark patterns. Although this work is not included in the thesis, examining real user complaints through the collected data provided valuable insights into the dark pattern phenomenon and users' perspectives. Moreover, the ethical considerations surrounding the collected data informed my subsequent studies, particularly in the creation of a datasets.

2. Maria W. Botes, Emre Kocyigit, Arianna Rossi, Lorena Sanchez-Chamorro, Anastasia Sergeeva, Kerstin Bongard-Blanchy, Philippe Valoggia, Cristiana Santos, Rachele Carli (2022). **Feedback to the Guidelines 3/2022 on “Dark patterns in social media platform interfaces: How to recognize and avoid them”.** Available at: <https://orbilu.uni.lu/bitstream/10993/52741/1>

- This collaborative work aimed to provide feedback on the European Data Protection Board's Guidelines 3/2022 and took place at the beginning of my PhD. Following internal discussions, the collective public comments prepared by members of the DECEPTICON (Deceptive Patterns Online) project at the University of Luxembourg and its collaborators were published.

The following research artifacts were produced as part of this thesis and made publicly available to contribute to the research community:

- **Python code:**

1. Fine-tuned Large Language Model for text-based dark pattern detection
2. Implementation for LLM and Multimodal LLM-based dark pattern detection, including Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) techniques

- **Dataset:**

1. Augmented dark pattern text data
2. Expert-annotated dark pattern image data

3. LLM-generated annotations for dark pattern image data

All code and datasets are publicly available at the project repository.⁶

The work presented in this thesis was conducted within the scope of the DECEPTICON project, which is funded by the Luxembourg National Research Fund (FNR, Grant No. IS/14717072). The project involved researchers from diverse fields, including user experience, HCI, law, ethics, and computer science. Its interdisciplinary nature significantly contributed to both the development and the novel, interdisciplinary contributions of this thesis.

⁶https://gitlab.com/uniluxembourg/snt/irisc/irisc-software/mitigating_deceptive_design

Chapter 2

Background and Related Work

Deceptive designs, i. e., dark patterns, are manipulative design techniques that aim to favor certain purposes of digital services at the price of user’s autonomy, such as collecting as many personal data as possible, and are very often illegal. Discussions about what exactly dark patterns are have arisen from different perspectives, for instance user experience, data protection law, interaction design. Each view has added useful insights on their determining characteristics that can help digital ecosystem stakeholders who are working on dark patterns, such as researchers, legislators, regulators, product developers, and designers, recognize them, avoid their use, or mitigate their risks. In this chapter, we summarize these views and related work, clustering the literature under four main areas: dark pattern definitions and taxonomies chronologically, dark pattern attributes, application domains and use cases, and automated detection of dark patterns. While presenting and reflecting on each aspect, we also highlight research gaps and briefly indicate how this thesis contributes to addressing them, referring to the relevant research objectives and chapters.

Terminology. Throughout this thesis, the following terminology is used to discuss dark patterns. **Type** refers to a specific category of dark pattern within a classification system (e. g. “confirmshaming”, “forced action”). **Attribute** refers to general characteristics of dark patterns that are not type-specific (e. g. *Asymmetric*, *Covert*, *Deceptive*). **Category** and **label** are used interchangeably with **type**, though we prefer **type** for low-level classifications and **category** for higher-level groupings where taxonomies distinguish between multiple levels. Detailed definitions are provided Chapter 4, Section 4.2.

2.1 Taxonomies and Descriptions of Dark Patterns

The history of deceptive designs influencing people’s decisions dates back to the days even when digital services did not exist. For instance, fake advertisements of store closings existed even before the internet with the goal of attracting customers and boosting purchases [36]. If we consider the formalized bargaining example by Ettinger and Jehiel, as described in *A Theory of Deception*[37], a seller may claim that there is another potential buyer to the client to avoid offering a discount, i.e., a scenario that favors the client. Similarly, online shopping platforms can display that a product has been viewed by certain number of people to give the impression of *high demand* [19]. In both cases, if the seller or designer is not truthful, they are misleading the client or user. The main difference is that the latter, as a dark pattern, occurs in a digital environment. Today’s online deceptive designs can be more complicated and hard to respond because digital services affect thousands, even millions of people, unlike a traditional store. Moreover, these designs are omnipresent: for example, 97% of the most common websites and applications exploit at least one deceptive design in the European Union [38]. Within the scope of this thesis, the literature on **online** deceptive designs that we may encounter in our increasingly digital world is addressed.

After Harry Brignull proposed the term of ”dark patterns” for online deceptive designs in a user experience design oriented blog post [39] ¹, he introduced different categories in *darkpatterns.org* [33]. For example, *Roach Motel*, refers that subscription is easy but canceling is difficult for users, and *Bait and Switch*, describes the pattern in which a user assumes performing one action but instead unintended one occurs. That attracted attention and people started to react by sharing their own experiences about ”potential” dark patterns, which led to first conceptual taxonomy of dark patterns. It has served as a foundation for subsequent academic taxonomies.²

In 2013, Zagal et al. questioned the unethical design in games with an inspiration from the concept of dark patterns [2]. After inspecting popular games, such as social media games and multiplayer online video games, they presented three high-level dark pattern categories in game designs: time-based, financial-based, and social capital-based. For example, time-based dark patterns such as *Grinding* and *Pay to Skip* place players in situations where they must either invest significant time or spend money to progress; otherwise, they remain stuck at a certain level. Another well-known dark pattern type identified in their taxonomy is *Social Pyramid*

¹Harry Brignull, *Dark Patterns: Dirty Tricks Designers Use to Make People Do Stuff*, 2010, web archive, Accessed: October 24, 2025.

²The initial **Dark Pattern categories** introduced by ”darkpatterns.org” included: *Roach Motel*, *Bait & Switch*, *Voice Theft*, *Sneaky Credit Card Roll-over*, *Price Comparison Prevention*, *Disguised Ads*, *Price Sensitivity Targeting*, and *Forced Information Disclosure* [40].

Schemes, which refers to a design that encourages players to share the games with their friends or directly invite them in exchange for in-game rewards. Although these are now well-established dark patterns, earlier academic studies, primarily aimed to raising awareness and define dark patterns from an academic perspective using systematic methods. We can consider this objective achieved, as one year later, a book [41] addressing game design and player engagement dedicated three chapters for the aforementioned three high-level dark pattern categories, and it attracted the attention of researchers working on ethics in game designs [42].

Not every non-user friendly design or poor design solution that causes inconveniences for the users is considered as a dark pattern. For example, practices that should be avoided to solve continuous design problems are referred to as "anti-patterns" [43]. In contrast to them, dark patterns describe actions taken in the service of malicious or manipulative design. One of the earliest academic effort on dark patterns unsurprisingly came from human computer interaction scholars and they identified dark patterns by examining interaction between device (system) and human (system users) [44]. Another aspect in which dark patterns negatively impact users is privacy. In 2016, Bösch et al. systematically investigated malicious strategies against privacy principles like "data minimization", "informing data subjects", "being in control of personal information processing", "demonstrating compliance with privacy-related legal requirements". They identified corresponding dark strategies like "maximizing the collected data", "obscuring personal data processing information", "deprivation of control over personal data", "fake compliancy" (respectively) and finally proposed a taxonomy of seven dark pattern types [4]. For example, *Bad Default* dark pattern presents a pre-selected configuration that favors sharing usually more than necessary personal data, making it difficult for users to go through all settings, as a implementation of "obscuring personal data processing information" strategy.

Following ethical game design and privacy-oriented dark pattern taxonomies, Gray et al. developed a user experience-driven taxonomy and proposed five high-level dark pattern strategies, which have since been widely used by stakeholders in the dark pattern domain [5]. After collecting dark pattern examples from various sources, such as user experience expert websites or social media platforms, they conducted a gradual analysis and consequently classified motivational-based dark pattern strategies as: *Nagging*, *Obstruction*, *Sneaking*, *Interface Interference*, *Forced Action*. For example, they classified *Social Pyramid* [2] and *Privacy Zuckering* [33] under the *Forced Action* strategy. Eventually, their taxonomy presented 15 low-level dark pattern types and helped to better understand the dark pattern phenomenon.

Dark patterns appear on online shopping platforms, as well as on games and social media platforms [23]. The first big scale study on electronic commerce-related dark patterns has been

conducted by crawling 11 thousands online retail websites [19]. Then, they apply hierarchical clustering-basis data analysis to the crawled data and their research team labeled the dataset with substantial agreement. They used previously proposed dark pattern classifications [33, 5] and presented seven high-level categories like *Sneaking* (as in the taxonomy of Gray et al. [5]) and *Urgency*, which refers driving users toward faster decision-making and buying behavior. Under these categories, they listed 14 different dark pattern types. For example, *Countdown Timer* and *Limited Time Message* as a low-level types of *Urgency*. While these efforts significantly contribute to the systemization of dark pattern research, they were not sufficient to capture all possible dark pattern classes. For example, Mathur et al. considered only text-based dark patterns.

In addition to academic taxonomy studies, significant efforts have also begun on the regulatory front. The enforcement of regulations such as the GDPR has prompted growing interest in dark patterns not only among scholars in user experience design, computer science, privacy, and gaming, but also among legal researchers and policymakers. For example, the Federal Trade Commission (FTC) prepared a comprehensive report dedicated to dark patterns in September 2022 [6] and developed a business practices-oriented taxonomy. While most of the categories were taken from the previously developed taxonomies, such as *Sneaking* [5], *Urgency* [19], they adapted some of the existing classifications with their own terminology. For example, they prefer *Endorsements* category for *Social Proof* [19]. OECD published a report one month after FTC's report as **Dark Commercial Patterns** and merged the existing taxonomies under seven high-level categories and listed both descriptions and examples of 24 dark pattern types [7]. In the same year, another regulatory association European Data Protection Board (EDPB) published a guidelines for identification and avoiding dark patterns [9]. Next year, they updated and released a new version with their own taxonomy [8]. In the recent version, they presented seven high level categories like *Overloading*, *Skipping*, *Stirring*. For example, the first category corresponds to designs where users encounter large options or information leading them to share more data. The report concentrated on data protection use cases and presented real examples or scenarios for each category. Furthermore, each dark pattern type was mapped with the corresponding GDPR provisions.

All of these taxonomies, specific dark pattern descriptions, and frameworks covering domains such as privacy, data protection, e-commerce, etc., helped us to understand both existing and potential implementations of dark patterns. They have raised awareness of these manipulative practices and contributed to the conceptualization of this wide-spread, deceptive, and harmful design phenomenon. However, two important limitations remain. First, while these taxonomies

provide valuable insights, they do not eliminate the need for **objective and operational criteria** for the detection of dark patterns. Second, the proliferation of taxonomies introduces inconsistencies in the label nomenclature. For example, the categories *Forced Action* and *Nagging* [5] correspond to *Coerced Action* and *Continuous Prompting* in the EDPB taxonomy [8].

Recent work by Gray et al. [29] has addressed the main issue of lack of a standard ontology by consolidating multiple taxonomies and creating an ontology that integrates academic and regulatory perspectives, providing a more structured mapping of dark pattern categories. They presented **60 different dark pattern types** organized under five high-level categories: *Nagging*, *Forced Action*, *Interface Interference*, *Obstruction*, and *Social Engineering*. During the mapping process, some dark patterns were directly adopted from prior taxonomies, while others were inferred. For example, *Hidden Information* was taken directly from the FTC [6] and Gray et al. [5], but also inferred from Bösch et al.'s *Hidden Legalese Stipulations* [4] and the EDPB's *Left in the Dark* [8]. In this thesis, dark pattern descriptions and classifications from this ontology were used as a reference.

Nevertheless, the first-mentioned problem, i.e., a lack of objective and operational criteria for dark pattern identification, is not fully solved. Many dark pattern definitions allow for subjective interpretation, leading to variability in labeling even among experts [34]. For instance, the definition of *Disguised Ads* specifies that advertisement content must be clearly marked, which raises practical questions such as: how can "clearly" be quantified, and what constitutes being "marked" in a measurable way? This thesis contributes to this gap by proposing a **systematic approach** and **operational and measurable features** associated with dark patterns (see **Research Question 1**). The contributions are detailed in Chapter 3 and Chapter 4.

2.2 Attributes of Dark Patterns

The previous section discussed numerous studies on the definition and classification of dark patterns. On the other hand, research focusing specifically on recurring characteristics that are fully or partially shared across different dark pattern types has been more limited. In 2019, Mathur et al. identified five fundamental dark pattern attributes in their comprehensive study oriented to crawled e-commerce data [19]. They grounded these attributes on previous work on online manipulation and associated each attribute with dark pattern types. The first attribute is **asymmetry**, refers to a design in which users must expend unequal effort across different choices. Although different types of asymmetries, such as "information asymmetry" [45, 46], can be involved when defining dark patterns, the asymmetry referred to here specifically relates to the effort required by users for the choices they encounter in the user interface. An example

illustration of asymmetric choice architecture, which is a common dark pattern in cookie consent banner [47], is shown in Figure 2.1.

The second common dark pattern attribute is **restriction**, which refers to designs that limit options that should be available to users [19]. A good example of restrictive pattern design is when users are required to accept marketing notifications in order to agree the terms of use [18]. The third attribute, **information hiding**, refers to postponing or concealing the presentation of information to the user [19]. It is similar to one of the deception techniques described by Kollmer et al. [1], namely the "omission technique". One striking example of dark pattern exhibiting the **information hiding** attribute is the case where Google was fined €50 million for hiding privacy-related information [36].

Another dark pattern attribute is **deception**, which refers to a design that create false impressions through deceptive wording [19]. *Countdown Timer* dark pattern type is a typical example of deception, as it misleads users into believing that a discount is available for a limited time in order to create a sense of urgency, whereas in reality, the timer simply refreshes itself. At first glance, one might assume that all dark patterns displays the deception attribute, partly because they have recently been referred to as "deceptive designs". However, certain dark pattern types, such as *Confirmshaming* or *Hard to Cancel* [19], and *Roach Motel* or *Pay to Skip* [18], were identified as not exhibiting deception at all .

Mathur et al. added **disparate treatment** as a new dark pattern attribute in [18], which was not present in prior work [19]. This attribute refers to designs that place certain users at a disadvantage by treating them differently from others. Unlike the other attributes, this one considers the differences among users who are exposed to dark patterns. Users' awareness of dark patterns does not necessarily prevent them from being manipulated [48]. Moreover, factors such as education level and age can influence their ability to recognize dark patterns. A design that may not demonstrate dark pattern characteristics for a typical user might have a different effect on older adults. For example, in a study focusing on interactions between elderly users and dark patterns, it was found that the older participants reported encountering *Hard to Cancel* and *Forced Enrolment* dark pattern types in online shopping context more frequently than other types [49] despite the fact that these were identified less in the e-commerce oriented dark pattern dataset created by [19].

Defining the attributes of dark patterns alongside the types of dark patterns is important; however, there is still a lack of **objective and operational criteria** for evaluating dark patterns. Mathur et al. outlined that **specificity is required for actionable definitions** [18]. This thesis addressed this problem by formalizing the aforementioned attributes with measur-

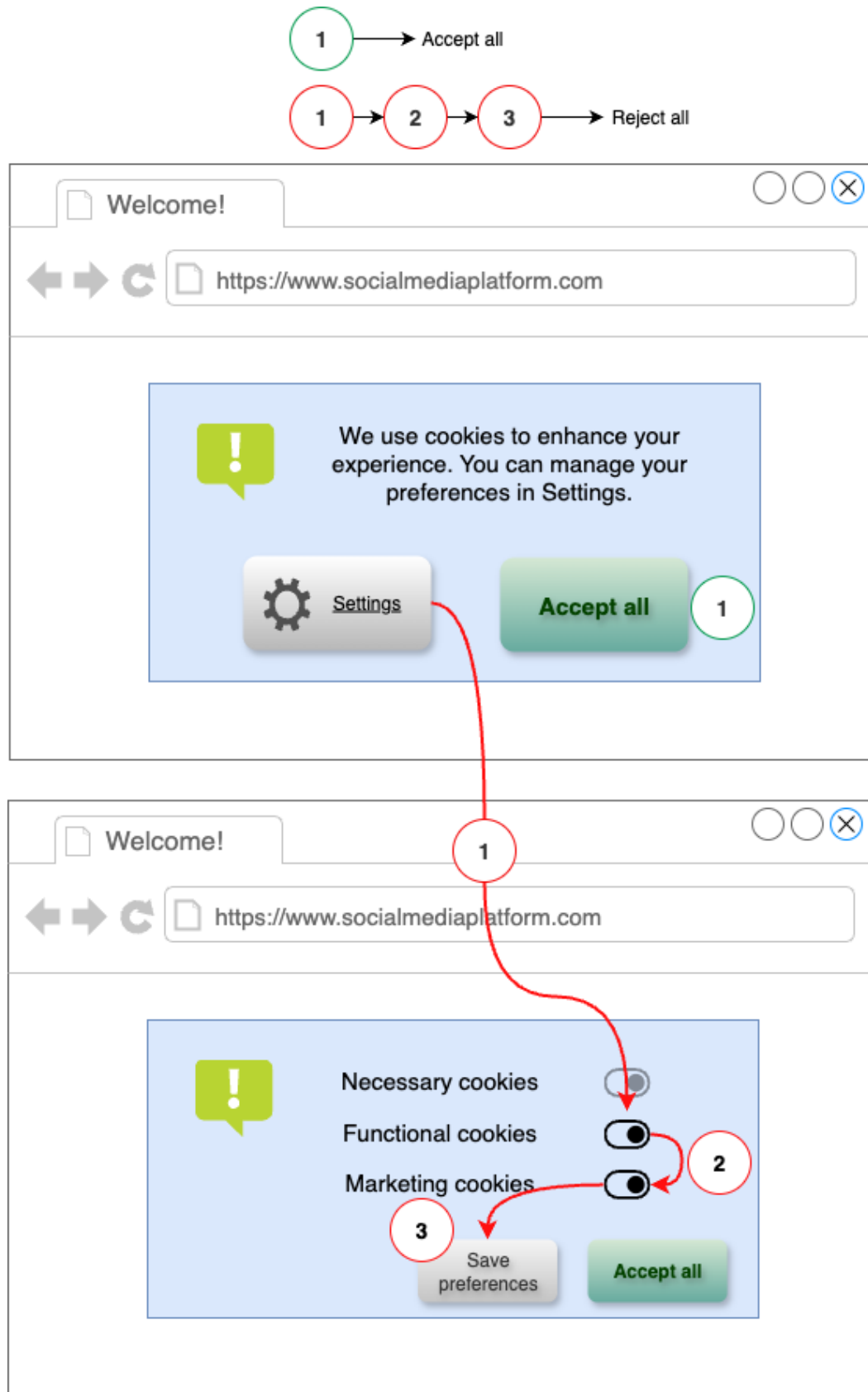


Figure 2.1: An example of cookie consent process illustrating asymmetric design with click counts for “accept all” and “reject all.”

able features. As detailed in Chapter 4, for example, an asymmetric pattern encountered in a cookie consent banner is expressed in terms of measurable features such as "the number of clicks". Precisely, if number of clicks for accepting all cookies differs from the number required to reject all cookies, then the design exhibits asymmetry. This asymmetry can also be formalized through visual and linguistic features, as illustrated in Figure 4.2, in Chapter 4.

2.3 Applications and Use Cases

Dark patterns are present in various digital applications and services such as online video games [2], mobile applications [50], social networks [51], finance applications [52], home robots [30], IoT devices [52], web pages of travel agency websites [53], Virtual Reality [54], and more. Although the use cases in which dark patterns appear are highly diverse, their prevalence is much more common in certain use cases than in others. For example, cookie consent banners, where manipulative designs nudge users toward privacy-unfriendly options, frequently presents dark patterns [55]. Studies showed that scenarios like cookie consent processes violate the requirements of GDPR and ePrivacy Directive [56, 57, 58, 16, 17, 59, 60, 61, 62]. These manipulative practices include pre-checked options, asymmetric choices, and misleading language. Since cookie consent banners frequently contain dark patterns and draw the attention of various stakeholders due to their implications for privacy violations, the **cookie consent process** was selected as the use case to test the approach developed in Chapters 3 and 4, which address **Research Objective 1**.

When examining dark patterns in terms of their domains of application, another important observation emerges: although the ontology study identified more than 60 dark pattern types, many of them are highly context-dependent. Some dark patterns are prevalent in e-commerce environments, while others commonly appear in scenarios that involve consent or choice mechanisms. For example, in the context of cookie consent, many of the existing dark pattern types have been observed to be not applicable or rarely implemented [63].

Dark patterns can appear across different modalities such as web and mobile platforms. Gunawan et al. demonstrated that while some dark patterns are present in both modalities in a similar way, others exhibit notable differences between web and mobile environments [64]. Games are one of the primary areas where dark patterns are commonly identified in mobile applications. Niknejad et al. crawled the website "www.darkpattern.games" and found that even the games that are considered as "healthy" according to the website rules contain dark patterns [25]. In particular, monetary dark patterns have been observed to be more prevalent in "dark" games. Recent studies show that dark patterns are continuously present in in-app

purchases, and previously unreported dark pattern types have also been identified in these contexts [65]. Dark patterns not only affect players but also game developers, which has led to ongoing discussions about the need to increase mutual awareness and promote more ethical design practices.[24].

Along with emerging technological developments, the environments of human-technology interaction are expanding and continually producing potential new dark patterns. From chatbots, which users frequently engage with [66], to home Internet-of-Things (IoT) devices [67, 68], dark patterns continue to deceive and manipulate users. The wide range of application contexts makes it difficult to develop an universal or standardized method for effectively detecting dark patterns. As proposed in Chapters 3 and 4, the specific use case and the related context must be considered when designing a detection approach. For example, in both chapters, we examined **cookie consent processes** and indicated that only 19 dark patterns are not present or potentially applicable in this use case, highlighting the need for use case-specific measurable features. Dark pattern attributes, such as "asymmetry", can be described differently in different use cases. For example, it is represented by the imbalance between "number of clicks to accept all cookies" and those required to reject all [69]. On the other hand, monetary-related features that are often seen in mobile applications are not relevant or applicable to this use case. In the next section, we will present dark pattern detection-focused literature.

2.4 Automated Detection of Dark Patterns

In order to detect manipulative designs, several automated detection solutions have been developed [70, 71, 35, 72, 73, 74]. These solutions employ diverse methods mostly relying on features extracted from text on UI, HTML, CSS, images, or user interaction-based metadata. Machine learning and deep learning-based techniques have been utilized during both input data processing and identification decisions. Some studies combined computer vision and natural language processing techniques to detect DP in screenshots or interfaces [72, 75], while others integrate back-end tracking information to verify consent-related manipulations [27, 76].

Although AI-based implementations were successful in recognizing certain dark patterns, challenges remain due to variation in implementation, multimodality, and dataset biases [77, 71]. As the number of studies focusing on automated dark pattern detection has increased, the limitations of current systems indicate that further work is required before they can be applied effectively in real-world scenarios. Nie et al. [78] reviewed five dark pattern detection solutions developed up to 2024 and found that, among 64 defined dark pattern types in taxonomies, only 32 were covered. Their analysis addressed three main issues observed in existing approaches: (i)

inconsistencies in dark pattern classifications, (ii) restrictions in the detection tools and their capabilities, and (iii) limited access to data needed to evaluate the presence of dark patterns. The absence of standardized dark pattern definitions in the literature is also reflected in dark pattern detection systems, which use inconsistent labeling and terminology. For instance, the *Countdown Timer* dark pattern type is categorized under the "Aesthetic Manipulation" and labeled as *Toying with Emotion* by describing it as "countdown offer/reward" in one system [35], whereas another system [72], it is categorized under "Urgency" with the label *Countdown Timer*. To mitigate such inconsistencies, this thesis adopts the labels and definitions proposed in the most recent and comprehensive ontology, as discussed in Section 2.1.

The second challenge identified in dark pattern detection systems, namely the **restrictions of detection tools and their capabilities**, underscores the necessity for detection approaches that are not limited to a single modality. One motivation for employing multimodal LLMs in the detection solutions presented in Chapters 7 and 8 was to overcome this singularity limitation. MM-LLMs integrate numerous data modalities, such as text, images, and voice data, overcoming the constraints of text-only LLMs [79, 80]. They have been applied to automated dark pattern detection, enabling reasoning across visual and textual features, and generating interpretable explanations [81, 82]. LLMs and MM-LLMs may generate incorrect, biased, or infeasible outputs [83, 84, 85, 86]. Techniques like Retrieval-Augmented Generation (RAG) improve accuracy and explainability [87, 88]. Human-in-the-loop approach and reinforcement learning with human feedback (RLHF) further enhance the performance by adding additional verification and alignment processes [89, 90]. Sensitivity to visual noise, dataset biases, and differences in modalities (text vs image) are major challenges. Automated detection is not yet fully reliable, and evaluation metrics include clarity, usefulness, and verifiability of explanations [91, 92, 93].

Another significant limitation involves the lack of transparency and interpretability of detection systems for end users. To address this issue, Chapter 7 introduces a transparency-supporting detection model, contributing to the mitigation of this research gap. Nevertheless, there are also several external limitations, which are independent factors of the detection systems, still influence the detection process. The most prominent among these is the limited access to data.

The issue of data accessibility for dark pattern detection can be examined from two perspectives: (i) the datasets required for training and testing during system development, and (ii) the data needed at the time of detection to make an informed decision. First, regarding dataset creation, several studies have been conducted and developed dark pattern datasets. The size, quality, and diversity of these datasets are crucial for ensuring reliable evaluation of detection

systems. For instance, Geronimo et al. [50] created a video dataset of 240 mobile applications annotated with dark pattern categories. Mathur et al. [19] crawled 11K e-commerce websites, labeling instances and providing structured datasets, later extended by Yada et al. [94]. Both UI-Guard [75] and AidUI [72] constructed UI-based corpus for dark pattern detection. The former utilized the Rico dataset to build a large-scale mobile UI collection, whereas AidUI included 301 instances in mobile and web applications. By this thesis, a dark pattern dataset that supports transparency was built that contains explanations about each detection process [34]. Chapter 5 of this dissertation also provides a detailed analysis of the existing dark pattern datasets including samples and identifies critical quality issues. We constructed a benchmark dataset by avoiding these issues and considering **Research Objective 2**. Furthermore, to address the data scarcity problem in the dark pattern domain, we develop a data augmentation framework and demonstrate it in Chapter 6. Specifically, we propose an approach that uses advanced language understanding and generation capabilities of LLMs. The augmented dataset successfully preserved the characteristics of dark patterns exist in the original dataset and contributed to the performance of detection models.

Second, from the perspective of data required at the moment of detection, it is often not feasible to reach a definitive conclusion without accessing the service provider’s server for certain dark patterns. For example, *High Demand* or *Low Stock Message* dark pattern types depend on dynamic inventory information that cannot be verified from the user interface alone. Therefore, this thesis argues that detection systems should instead issue potential dark pattern warnings, allowing the final decision to be made by stakeholders who have access to the necessary contextual and back-end data.

In situations where data collection is limited or prohibitively expensive, data augmentation techniques can help expand dataset coverage. Data augmentation constructs synthetic data from existing datasets to increase their quantity and diversity while maintaining representativeness [95]. In NLP, augmentation methods operate at different levels, such as character-level, word-level, sentence-level, and document-level. The methods include synonym replacement, paraphrasing, noising, sampling, and fine-tuning large language models [96, 97, 98]. More computationally efficient methods like prompt tuning with LLMs are also applied [99]. In this regard, Chapter 6 contributes by proposing a textual dark pattern data augmentation framework that considers both semantic, and structural similarity with the original data. The results show that the augmented dataset improves the accuracy of language models in detecting dark patterns [63].

Chapter 3

Towards Assessing Features of Deceptive Designs: Cookie Consent Processes

We first investigate the measurable and objective criteria of dark pattern properties by examining back-end of one of the most common dark pattern use cases: cookie consent process. In particular, we extract information flow of cookie consent process from typing URL of website to final cookie consent-relevant action of user and identify measurable and operational features based on interactions between user and browser, and browser and web server to recognize dark patterns.

*This chapter answers **Research Question 1**: "What are the measurable functional properties of dark patterns and the assessment instruments to define them?" and its content is primarily based on the following publication:*

- *Kocyigit, E., Rossi, A., & Lenzini, G. (2022). **Towards assessing features of dark patterns in cookie consent processes**. In *IFIP International Summer School on Privacy and Identity Management* (pp. 165-183). Cham: Springer Nature Switzerland.*

3.1 Introduction

Dark patterns are manipulative design techniques that aim to favour certain purposes of digital services at the price of user's autonomy, such as collecting as many personal data as possible,

and are very often illegal. Discussions about what exactly dark patterns are have arisen from different perspectives, for instance user experience, data protection law, interaction design. Each view has added useful insights on their determining characteristics that can help researchers, legislators, regulators, product developers and designers alike to recognize them and avoid their use.

Dark patterns in cookie banners have been particularly studied [56, 17, 71, 57] due to their ubiquitous presence, which has increased dramatically since the GDPR strengthened the consent and transparency requirements for personal data processing, therefore new ways of sidestepping the rules and extorting consent have been experimented. Still, even within this very specific domain, there is a lack of criteria capable of capturing the essence of dark patterns and that can be operationalised into measurable variables. Now, even if not all the characteristics of dark patterns can be measured, like the intentionality of the deception, this approach would offer a framework that can be leveraged to reliably discuss, argue, verify claims on the supposed presence of dark patterns on a digital service and, eventually, detect them.

Moreover, dark patterns are not only related with Human-Computer Interaction (HCI) layers, like the graphical user interface, but also with the back-end of applications. Even though inspecting the user interface level can help to determine some dark pattern characteristics such as unbalanced weight of options or hard-to-notice buttons, there are certain manipulative strategies that are hidden in the back-end and are therefore invisible without inspecting the elements of a web page such as the cookie content, the cookie size, etc. For example, some websites do not respect the user's consent refusal decision and continue to collect their data nevertheless, like Matte et al. indicated in [27]. Therefore, we consider both HCI and Machine-to-Machine Interaction (MMI) layers, like the Web Server, in this study.

Our contribution. For the purpose of providing objective descriptions on whether there are dark patterns in cookie consent flows, we study different websites' cookie consent processes and extract their activity diagrams through HCI and MMI layers, i.e. User Action, User Interface, Browser and Web Server. Consequently, we propose thirty-one features of cookie consent processes that could be useful to recognize dark patterns via the objective assessment of their characteristics. We define such features based on the information flow layers like User Interface, Web Server etc. and label them according to their values, which can be quantitative or binary. By leveraging these objective descriptions which can be operationalised into metrics, automated detection and analysing tools can be developed to help different stakeholders to detect, prevent and avoid dark patterns.

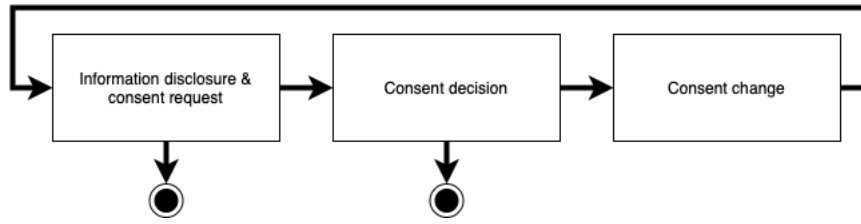


Figure 3.1: Cookie consent flow phases

3.2 Cookie Consent Process Background

In order to obtain measurable criteria of the cookie consent process that provide an objective description of the dark patterns presence in this process, we firstly studied the cookie consent flow and defined its phases as shown in Fig.3.1. While the flows have different interface design elements such as buttons, banner size, text etc. or provide various user actions such as “Decline all”, “Refuse unnecessary cookies” etc., they show a regular/common structure, probably also as a result of regulations like the GDPR.

When users visit a website for the first time, they get informed about the presence of cookies and are asked to consent, following Article 7.3 of the GDPR and Article 6.3 and 6.4 of the ePrivacy Directive. This is why, we defined the first phase of the cookie consent flow as “Information disclosure and consent request”. At this phase, cookie banners usually present links or buttons that redirect users towards a complete privacy policy, if they initiate that further action. Often there is also the possibility to customize consent options through user configuration.

After users are informed about the cookie policy, they usually have two options: leave the website without any further action or decide whether to grant or refuse consent. For this reason, the cookie consent process may end after the first phase or continue with the second phase, namely “Consent decision”. At this stage, users can grant their consent for all data processing purposes, for some (with customization of consent option), or refuse consent. After the consent decision, user can surf the website without any further consent action. According to the user’s decision, cookie-related requests and responses run between browser and web server. Thus, cookies which will be loaded and saved to the user’s device are known when user revisit the website after the consent decision.

Finally, users may change their previous consent decision (i.e., right to consent withdrawal, Article 7 GDPR). Hence, we defined the third phase as “Consent decision change”. At this stage, users may withdraw their consent, provide consent for previously refused processing purposes, or give consent for all or some of the previously refused purposes. This change reshapes cookie content which will be handled between the browser and web server for the user. Also, the third phase triggers the first phase as shown in Fig. 3.1.

We created a matrix to extract all possible cookie consent scenarios as shown in Figure B.1, in appendix. Because there is a loop from the third phase to the first phase, infinite scenarios are possible, but we ignored the last phase because it only determines if the loop is over or not. Eventually, we listed 14 possible scenarios in the consent management flow in terms of user actions and extracted the features by considering these scenarios.

3.2.1 Human-Computer and Machine-to-Machine Interaction Layers of the Cookie Consent Process

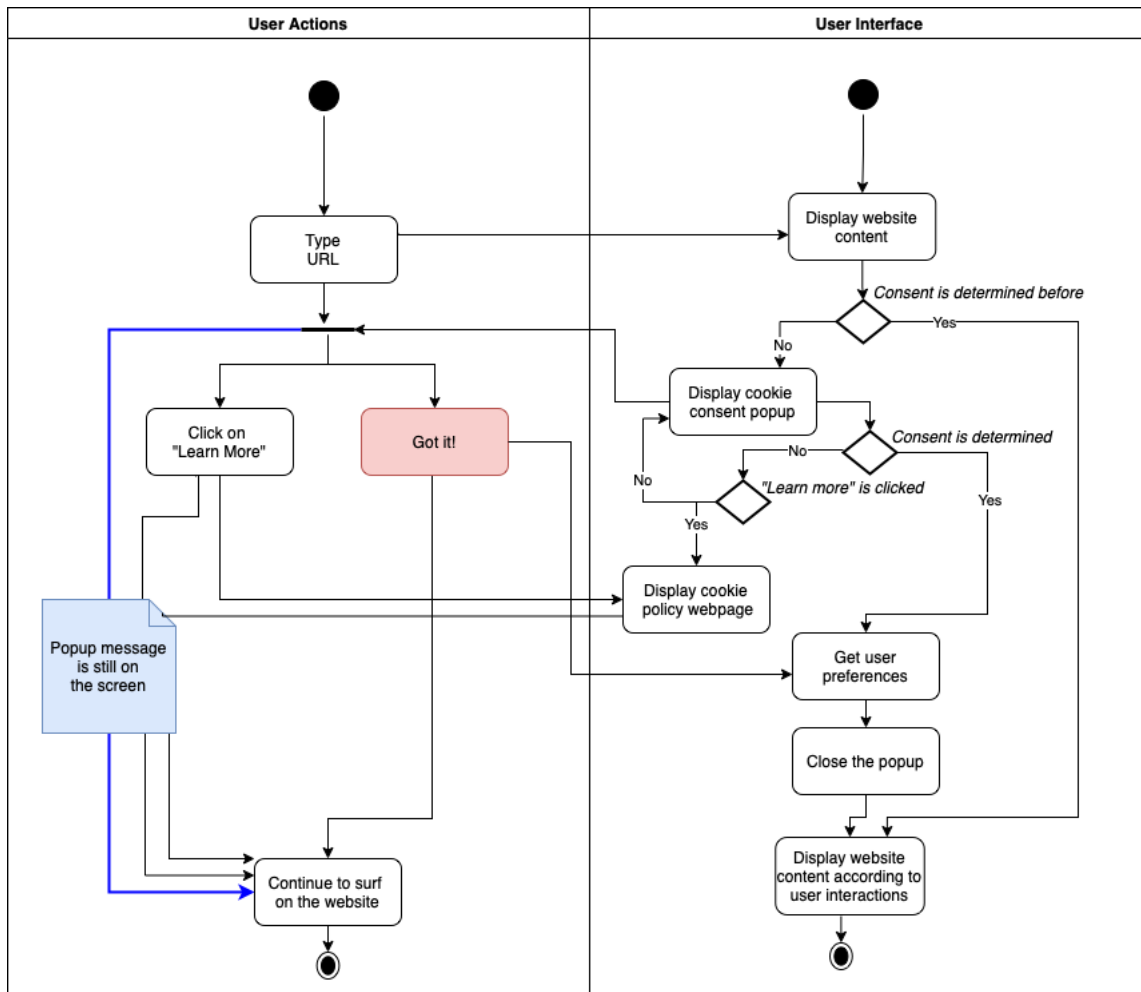


Figure 3.2: User Action and User Interface layers of cookie consent flow of website 2

A web flow is performed via a harmony of different components by synchronous and asynchronous requests and responses online. These requests and responses occur both between user and browser interface, i.e., within front-end of websites, and between the browser engine and the web server, i.e., within back-end. We create a Unified Modeling Language (UML) activity diagram which has different layers such as “User Action”, “Browser - User Interface”, “Browser - Engine” and “Server” as shown in Figure 3.4 to inspect human-computer interaction and machine-to-

machine interaction of consent management flows. Vertical flows demonstrate possible cookie consent scenarios within each layer, whereas horizontal flows show interactions between layers. For example, following the "Type URL" user action, there are three vertical arrows which represents possible user actions: accepting cookies by clicking "Got it!", displaying the cookie policy page by selecting "Learn More", or continuing to surf on the website while keeping the cookie banner visible on the screen, as shown in Figure 3.2. On the other hand, a horizontal arrow, from "Type URL" under "User Actions" layer to "Display website content" in "Browser - User Interface" layer represents a request, i.e., an interaction, between layers.

Various consent management flows are available on websites, thus we selected 10 websites (see Figure B.2, in appendix) that provide different consent management flows to examine and drew a diagram for each to make a comparative study. This comparison contributed to define measurable features in the cookie consent flow which may be descriptive for Dark Patterns. For example, while the "User Action" layer of a website's cookie consent flow has 4 actions as shown in Fig. 3.2, another website's cookie consent flow is more complex because it has more than 20 possible actions as shown in Fig. 3.3. We focused on this type of differences to obtain the measurable features. The feature extraction steps and the methodology are detailed in the next section.

3.3 Methodology

In this chapter, we aim to define the functional characteristics of dark patterns. As a first step towards this goal, we choose the cookie consent process as a use case to elaborate the objective assessment of dark patterns' characteristics. Thereby, our first research question is: "*What are the features in cookie consent processes to objectively describe the presence of dark patterns?*". The answer to this question will be a starting point to formulate the metrics that can be helpful to define the functional characteristics of dark patterns. Secondly, most studies have only looked at the most visible side of the online world, like the user interface as we mentioned in Section 3.1, thus we have an imperfect understanding of dark patterns. Another motivation of this paper is to find out if there are features that are visible/measurable only by examining the HCI and MMI layers of the cookie consent flow. A holistic approach is essential to define a comprehensive feature list. Accordingly, another question is: "*Are there any dark pattern-related features of cookie consent processes that can be identified only by inspecting multiple layers of HCI and MMI flows?*". We believe that answers to these questions will aid to formulate the characteristics of dark patterns and to extract suitable features for the automated detection models in future work.

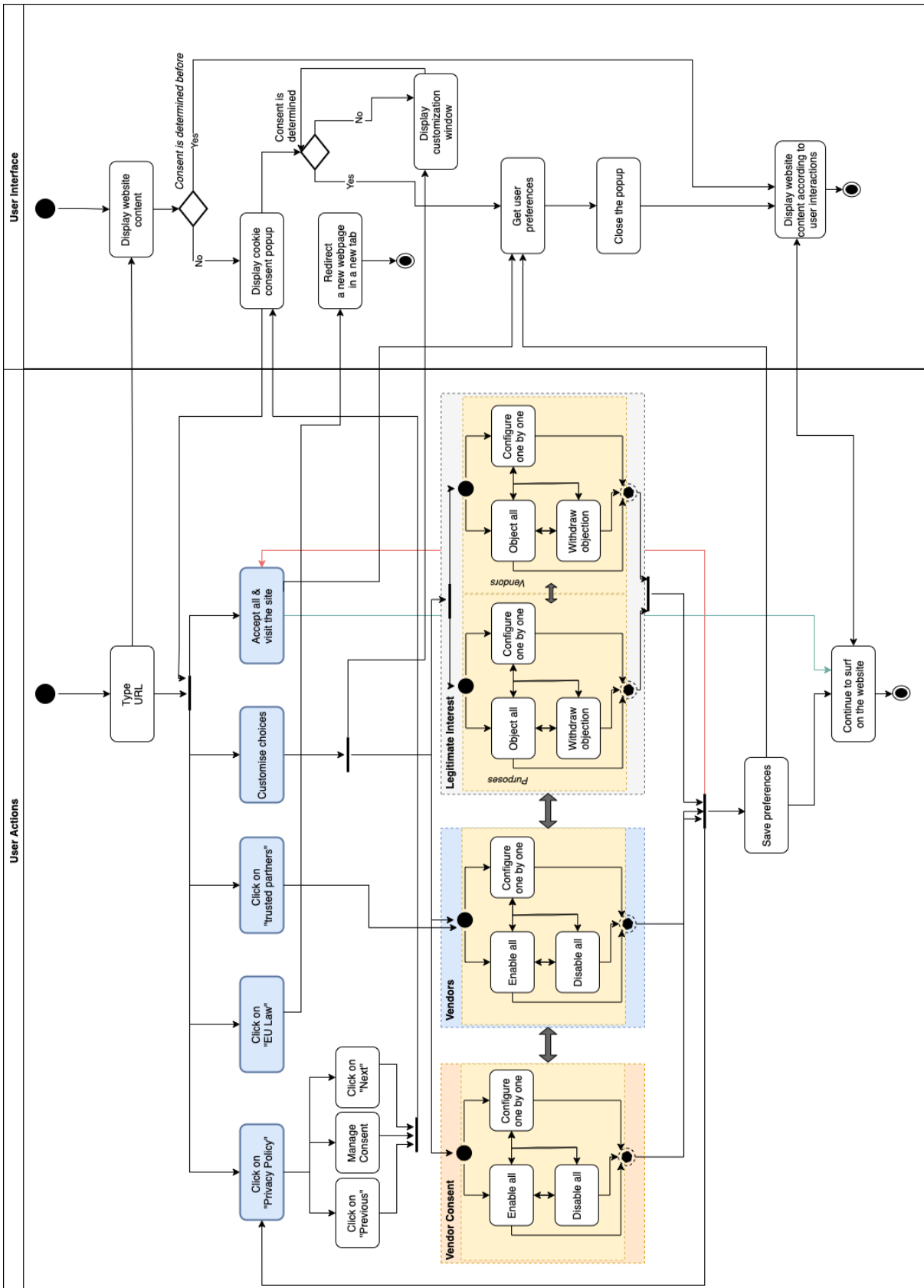
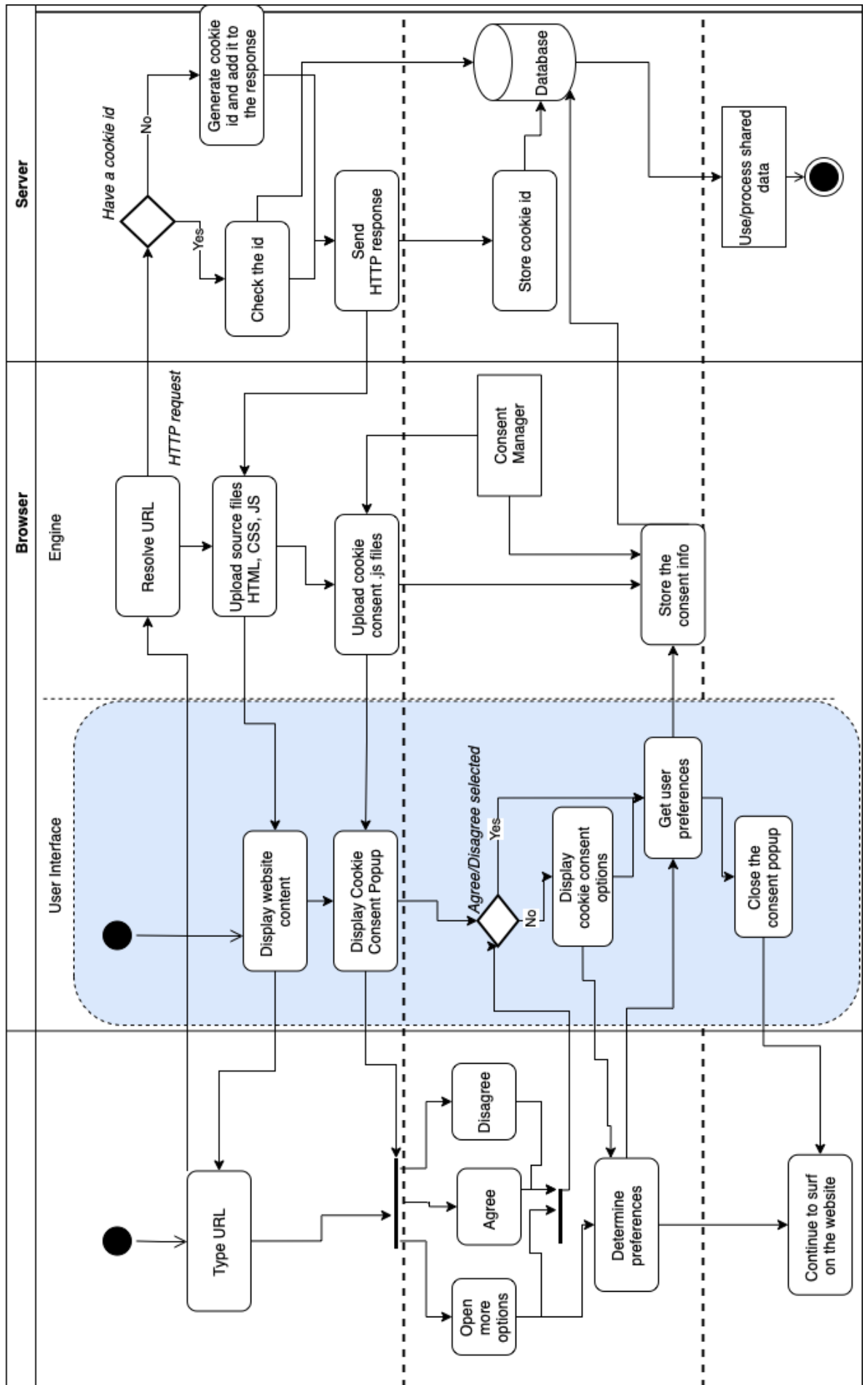


Figure 3.3: User Action and User Interface layers of cookie consent flow of website 3

This is an activity diagram of "researchgate.com" cookie consent flow



Information closure & consent request

consent decision

As a first step, we selected cookie consent flows of 10 different websites which are listed in Table B.1. For some even without deeper analysis, it is seen that they contain clear dark patterns such as "Forced Action" and "False Hierarchy" [5]. However, a more detailed examination is also needed because there could be more dark patterns like "Too Many Options" [9]. In this regard, the second step is creating UML activity diagrams for each cookie consent flow to represent information flows in multiple layers as: User Action - User Interface - Browser - Server. By these diagrams, we inspected horizontal and vertical flows between and within layers to understand which patterns can be associated with dark patterns. We organized the analysis of information flow into two subsections. First, we inspected the number of elements, the possible routes from the starting user action to the final one, the patterns possibly related to dark pattern types, and cookie related information such as cookie size and cookie number of each cookie consent flow separately. Then, we compared the same layers of different cookie consent flows with each other. For example, we extracted routes for same action from different cookie consent processes and compared them to find their differences, such as number of actions required to the user. After the analysis steps, we defined features according to the HCI and the MMO layers by using the analysis outputs. Finally, we discussed the relationship between existing dark pattern attributes and our proposed features to answer the research questions.

3.4 Features in Cookies Consent Processes Relevant for Dark Patterns Detection

After extracting the multiple layers of the cookie consent flows of ten different websites that are listed in Table B.2 and carrying out the extraction process that was detailed in Section 3.2, we listed thirty one features according to different layers of cookie consent information flow – i.e., User Action, User Interface, Browser Engine, Web Server – and types – i.e., Binary and Quantitative – as given in Table B.2, in appendix. As a result, we categorized the features into: Human-Computer(User Action and User Interface) Interaction-based in Section 3.4.1 and Machine-to-Machine(Browser Engine and Web Server) Interaction-based in Section 3.4.2. While extracting and identifying these features, the main objective was not to ensure that each one directly represents a specific dark pattern, but rather to define measurable and operational characteristics that could help detect potentially asymmetric, restrictive, or unnecessarily complex patterns within the cookie consent context. Some features are sufficient on their own to indicate the presence of a dark pattern, whereas others become meaningful when used in combination. The common property of these features is that they have potential to establish objective and

actionable criteria for dark pattern assessment.

3.4.1 Human-Computer Interaction-based Features

1. **Forced decision:** This feature describes if the cookie consent banner prevents website usage without a consent decision. In the UML activity diagram, if there is no direct link from the “Type URL” user action to the “Continue to surf on the website” user action, it means that the website forces the user to make a consent decision. The value of the feature would be “1” then. Some websites allow to navigate the website while keeping the consent banner on the screen without requiring a user action (see blue line illustrated in Figure 3.2). In this case, the value would be “0”.
2. **User options at first visit:** Consent banners vary in terms of the first possible actions available. While some of them provide three options such as “Agree”, “Disagree” and “Customise”, others can provide only one option like “Got it”, or two options “Accept all” and “Learn more”. This feature describes the number of user actions at the first connection.

Full consent

3. **Availability:** This feature describes if there is a full consent option that allows the user to accept all processing purposes at once, like an “Accept all” button. The value 1 is assigned when this option is available, while 0 is assigned when it is not. Generally, consent banners provide this option.
4. **Total number of routes:** This feature gives the total number of possible routes from the “Type URL” user action to the “Accept all/Agree all etc.” user action. Some websites provide more than one way to consent to all cookies, while providing one or no way to reach the option of refusal.
5. **Minimum length of routes:** This feature describes the shortest route from the “Type URL” user action to the “Accept all” action, and can be helpful to define if a flow provides asymmetric options.
6. **Maximum length of routes:** This feature describes the longest route from the “Type URL” user action to the “Accept all” action and can be helpful to define if a flow provides asymmetric options. If loops exist, then only the first loop will be considered for this feature.

Full consent refusal

7. **Availability:** This feature describes if there is a full consent refusal option like the “Decline all” button, where 1 means that that option is available. While websites generally have a full cookie acceptance option, it is not always the case that they have a full consent refusal option on the first layer. Since websites may use various phrases to express consent refusal (for example, while one shows “Refuse unnecessary cookies”, another one can show “Refuse all”), we accept all of these as full consent refusal indications.
8. **Total number of routes:** This feature gives the total number of all possible routes from the “Type URL” user action to “Decline all/Disagree all” user action. Therefore, this feature can be helpful to understand if both options are provided to user in a symmetric way.
9. **Minimum length of routes:** Users perform actions to complete a consent decision flow and this process contains a different number of steps. While some websites provide shorter or easier ways to get to consenting, they may make the consent refusal path harder. This feature describes the shortest route from the “Type URL” user action to the “Decline all” action and can be helpful to define asymmetry within the website or comparatively with other websites.
10. **Maximum length of routes:** This feature describes the longest route from “Type URL” user action to “Decline all” action, and can be helpful to define if a flow provides asymmetric options. If loops exist, then only the first loop will be considered for this feature.

Customised consent Some websites provide the customization of consent preferences to users according to the data processing purposes. This process usually takes more time and requires more user actions than granting full consent.

11. **Availability:** This feature describes if the cookie consent flow contains a configuration or customization option concerning processing purposes and/or vendors, where 1 means that the customization is available on the cookie banner, while 0 signifies that it is not available.
12. **Total number of routes:** This feature gives the total number of all possible routes from “Type URL” user action to “Customize/preferences/options etc.” user action. Therefore, this feature can be helpful to understand the complexity of the flow.

13. **Minimum length of routes:** This feature describes the shortest route from the “Type URL” user action to the “Partly accept/save and exit” action after configuring the consent settings. This feature can also be helpful to define if a flow provides asymmetric options.
14. **Maximum length of routes:** This feature describes the longest route from the “Type URL” user action to “Partly accept/save and exit” action. If loops exist, then only the first loop will be considered for this feature.
15. **Total user action options:** This feature describes the total number of actions on the User Actions layer of cookie consent flow. It can be helpful to express the complexity of consent management flow.
16. **Total consent flow routes:** This feature describes all possible routes from the first user action to the end on the User Actions layer. Website designs can provide different consent scenarios, where some of them offer “Accept all” scenarios more often than “Decline all” scenarios. This feature can be helpful to understand asymmetry through all possible scenarios.
17. **Total hyperlinks on user interface:** This feature concerns the Browser - User Interface layer, and describes the total number of hyperlinks (e.g. buttons) on it. It can be helpful to understand the complexity of interactions on the user interface.
18. **Consent decisions management availability:** This feature describes if there is a consent change button, link, dashboard etc. on the website to edit a previously given consent decision, e.g. withdrawing consent. If the value of this feature is 1, that means that the consent decision can be updated by the user.
19. **Dead end:** This feature describes if there is a link or button which redirects user from the cookie consent flow to a new web page outside of the flow without providing the possibility to come back into the cookie consent flow. For example, a cookie banner can open a new window when a user clicks on the ”privacy policy” link and the new page doesn’t provide any ”back” button.

3.4.2 Machine-to-Machine Interaction-based Quality Features

In the network section of browser inspection, requests between browser and web server can be observed. The number of requests, the size and number of cookies, and the size of transferred files vary according to whether the visit is the first one or not and whether consent decisions are taken or not.

20. **Number of cookies at first visit:** The cookies that are installed on a device and their details can be unveiled when the web browsers are inspected. This feature describes the total number of cookies installed when a user visits a website for the first time.
21. **Number of cookies in full consent:** This feature describes the total number of cookies installed when a user visits the website after full consent grant.
22. **Number of cookies in full consent refusal:** This feature describes the total number of cookies when a user visits the website after full consent refusal.
23. **Cookie size at first visit:** This feature describes the total cookie size when a user visits the website for the first time. Each cookie has its own size value, but we obtain the total size to simplify the feature and generate a quantitative one.
24. **Cookie size in full consent:** The cookie size value should be associated with consent grant decision. This feature describes the total cookie size when consent is fully granted.
25. **Cookie size in full consent refusal:** This feature is the opposite of cookie size in consent grant feature, and describes the total cookie size when consent is refused. When user does not allow all cookies, this feature should be lower from the above one in most cases.
26. **Total requests at first visit:** This feature describes the total number of requests between browser and web server at the first stage, before the consent decision. Some of the requests are not related with consent but with functionality of the website, e.g., source files of the web page, while others are directly related with the consent management process or the consent decision. Therefore, this feature can be helpful when it is used with the below two to check if the consent decision is correctly implemented.
27. **Total requests in full consent:** This feature describes the total number of requests between browser and web server when consent is granted.
28. **Total requests in full consent refusal:** This feature describes the total number of requests between browser and web server when consent is refused.
29. **Transferred file size at first visit:** Transferred files between browser and web server can vary based on consent decision. This feature describes the total transferred file size in the first interaction between the user and the website.
30. **Transferred file size in full consent:** This feature describes the total transferred file size when the user types a URL after consent is granted.

31. **Transferred file size in full consent refusal:** This feature describes the total transferred file size when the user types a URL after consent is refused.

3.5 Discussions

We have listed 31 features which can be operationalizable into metrics to assess dark patterns after analysing the cookie consent processes in the previous section. One of the purposes of this work is to promote objective grounds to discuss qualities about online patterns and eventually provide evidence to discuss, confirm or refute claims on whether a particular design pattern is “Dark”. To the best of our intention, the features that we have identified are all operationalizable. It is possible, in other words, to define measures for the features that can be evaluated on the layered message sequence charts that we have provided. Some of them can be already used to objectively capture certain qualities and characteristics that have been only informally described in the literature. For instance, in [18], Mathur et al. defined six attributes and two of them, i.e. asymmetric (unequal weights on the provided options to the user) and restrictive (discarding specific options that should be available to the user) can be described with the features we proposed. The asymmetric patterns can be evaluated by comparing these features: “Total number of routes in full consent grant (#4)”, “Total number of routes in full consent refusal (#8)”, “Total number of routes in customized consent grant (#12)”. If their values are not equal, we can conclude that the design is asymmetric in an objective way. The restrictive attribute can be described by the following ones: “Forced decision (#1)”, “Full consent availability (#3)”, “Full consent refusal availability (#7)”, and “Customized consent availability (#11)”. For instance, if the value of “Full consent refusal availability (#7)” is zero, then we can argue that users do not have the option to reject cookies in one action, i.e., the cookie consent flow contains a restrictive pattern.

The features can provide descriptions of dark patterns when they are used in different cookie consent flows, not only within one cookie consent flow. We believe that our proposals are not only providing direct description of dark patterns’ attributes such as “asymmetric” or “restrictive”, but also indirect metrics like “complexity”. For example, the complexity of a large amount of different flows can be measured with the help of “Total consent flow routes (#16)”, “Total user action options (#15)”, and “Total buttons on User Interface (#17)”. In this case, reasonable inferences can be made for individual samples after calculating the mean values. For example, “choice complexity”, which is defined as required effort for an action or selection in [38], can be measured with these features.

Dark patterns can also be hidden in the back-end of the cookie consent flows, and our MMI-

Website	Cookie number			Cookie size			Request number			Transferred file size		
	in first visit	full consent	consent refusal	in first visit	full consent	consent refusal	in first visit	full consent	consent refusal	in first visit	full consent	consent refusal
w3schools.com	3	5	5	66	519	401	46	39	37	1.32 MB	144.01 KB	47.13 KB
researchgate.net	4	10	6	387	946	663	41	35	31	1.18 MB	15.31 KB	14.40 KB
stackoverflow.com	5	6	6	360	379	403	44	36	33	609.51 KB	126.69 KB	104.17 KB
dropbox.com	10	65	46	677	2796	1757	108	405	90	1.80 MB	1.59 MB	375.26 KB

Figure 3.5: Extracted machine-to-machine interaction features from the cookie consent process.

based features can evaluate their presence since the features check the interaction within/between Browser and Web Server layers of the cookie consent flow. One of the measurable elements in the back-end is the cookies themselves. A basic cookie has information such as *name*, *value*, *host*, *path*, and *expires* [100] which are easily accessible and can serve as important cues when analyzing dark pattern attributes such as asymmetry. For example, when “Cookie size in full consent (#24)” and “Cookie size in full consent refusal (#25)” are assessed together, if they are equal we can deduce that the consent refusal option is not working or that the cookie consent flow is deceptive, as it shows to users the option of refusal but user’s choice is not correctly registered. We measured these features of different websites when users firstly arrived to the website before consent decision and after full consent grant or full consent refusal as shown in the Fig.3.5. As it can be observed, in most instances the values of the features are unsurprisingly lower, rather than equal, when a user refuses to consent comparing to the full consent grant.

As we highlighted in **Chapter 2**, previous categorization and definition studies offer a variety of perspectives to examine dark patterns, but the literature on the topic still lacks objective and measurable criteria to assess the presence of dark patterns on online services. We believe that measurable metrics are not only needed for an objective evaluation, but also to develop automated detection tools, which require a concrete and measurable feature set and can be helpful for developers and designers during testing processes, and to regulators and researchers during evaluation and inspection processes. Automated detection tools like machine learning-based systems require structured data in the training process. These metrics can be helpful to constitute feature sets and data sets thanks to their quantifiable nature.

3.6 Limitations and Future Research Directions

We examined the cookie consent flow as a use case of dark patterns and proposed thirty one features that can be operationalised into metrics based on our multi-layers interaction flow

analyses. We provided clear descriptions for them, but we have not yet defined procedures, quality measures, nor guidelines to interpret them for any of the features that we have described. We have informally given a few exemplifying arguments to show that this exercise is possible, but its full development is left for future work.

We selected ten different websites and extracted the interaction flows of their cookie consent processes. We believe this sample is sufficiently representative for this initial analysis, but of course additional different examples may help to explore other features. We excluded the "legitimate interest" paths in the cookie consent flow and the flows that repeat over time, e.g., nagging users to consent in their later visits. Moreover, the choices we made in selecting certain elements of the process in the message sequence chart affect what features we can define. For instance the edges in our charts are unweighted. One could, for instance, consider the cost of a certain choice which can provide new features. We quantify the asymmetric patterns based on the number of clicks, without accounting for the cognitive burden associated with each click. For instance, one option may be longer in terms of text length and contain more complex language, while another may be short and simple. Such differences are not captured through our information flow analysis, which limits the scope of the extracted features.

Furthermore, while studying multiple layers of interaction flows can contribute to obtain metrics describing dark pattern characteristics, we excluded the language. Therefore these features are not capable of describing examples like "confirmshaming" [101], which exploits human emotions like guilt to manipulate users. We also did not include visual elements such as button size, colour, contrast ratio etc. into this study but we intend to consider them in our future work.

The back-end of websites would need further investigations to cover other aspects of Dark Patterns and extend their features. For instance, we focused on the total cookie size in different consent conditions, but we did not extract the content of the cookies, and we observed the total number of requests between browser and web server, but we did not detect the ones that are directly related with cookies. Hence, the features can be elaborated into their breakdowns such as "Total cookie-based requests after full consent" or "Advertisement-based requests after full consent refusal". Since all cookies do not represent same mission, e.g., some can be designed the functionality of cookie consent mechanism. Of course, it should be noted that cookie values are often stored in encrypted form, and such detailed feature extraction is therefore subject to certain limitations. In this context, compared to external analyses that rely on restricted data access, internal efforts with full access to all relevant data can provide a more accurate and comprehensive evaluation of dark patterns.

3.7 Wrap-up and Perspectives

One of the harms associated with dark patterns is the manipulation, coercion, or misdirection of users during cookie consent processes, leading them to disclose their personal data even if they do not intend to do and infringes the legal requirements. Although regulations such as the GDPR prohibit the collection and sharing of user data by third parties without explicit consent, dark patterns are often employed to secure consent in a deceptive manner. Detecting dark patterns is therefore essential to mitigating the associated privacy risks. However, existing categorizations and definitions are not sufficient for objective evaluation, making systematic assessment challenging.

In this study, we analyze the information flows in different cookie consent processes—from the user, through the browser, to the web server—and derive measurable and operational features that can be used for the detection of dark patterns in consent interfaces. These features do not constitute a final or exhaustive list; however, they are significant in that they highlight an important research gap and provide concrete proposals for the objective evaluation of designs in order to detect dark patterns. Dark patterns cannot be captured solely through process-based metrics. Visual and language-based metrics must also be defined. The outputs of this chapter have provided an important contribution to ***Research Objective 1: Defining measurable instruments for the objective assessment of dark patterns.*** This work will be expanded from multiple perspectives in Chapter 4 and serves as a foundation for the systematic approach proposed in this dissertation, as well as for the transparent supporting detection framework presented in Chapter 7.

Chapter 4

A Systematic Approach for A Reliable Detection of Deceptive Design Patterns Through Measurable HCI Features

Following the analysis and proposed process-based measurable features of dark patterns in the previous chapter, this chapter further develops the concept of measurable features and proposes **a systematic approach** for a reliable detection of dark patterns. In particular, we first identify use case-specific dark pattern types and clarify the relationships between these types, such as Privacy Maze, False Hierarchy, etc., and the dark pattern attributes, such as asymmetry, restriction, etc.. Then, we describe these attributes based on our newly proposed set of measurable human-computer interaction features to lower the subjectivity of users' interpretation and to reconcile the interpretations of expert users on dark patterns.

This chapter answers **Research Question 1**: "What are the measurable functional properties of dark patterns and the assessment instruments to define them?" and its content is primarily based on the following publication:

- Kocyigit, E., Rossi, A., & Lenzini, G. (2024, September). **A systematic approach for a reliable detection of deceptive design patterns through measurable HCI features**. In *Proceedings of the 2024 European Symposium on Usable Security* (pp. 290-308).

4.1 Introduction

Privacy and data protection requirements of users are frequently violated in the user experience design of websites, social media platforms, mobile applications, and video games. Online deceptive designs, i. e., dark patterns, often circumvent privacy-protecting mechanisms by forcing, misleading or manipulating users to make them take potentially harmful decisions that lower, or even nullify, the protection of their personal data and weaken their rights. For example, a restrictive design pattern can be implemented in cookie consent processes to coerce users to disclose their personal data with third parties for profiling and advertisement purposes. Such a practice is explicitly against the international privacy engineering principle standardized in ISO/IEC 29100 [102] concerning "consent and choice" which clearly states that users should have the ability to control the collection, use and disclosure of their personal information. In the EU, this also contravenes the requirements about the freedom of consent enshrined by the General Data Protection Regulation (GDPR - Regulation 2016/679) and the ePrivacy Directive (Directive 2002/58).

Detecting and reporting the presence of dark patterns in digital services holds significant

importance for underpinning the scrutiny of potential violations of the law by independent authorities or administrative courts, for supporting organizations in their data protection compliance efforts, and for assisting privacy consultants in their advising activities. Nevertheless, several challenges hinder the identification of dark patterns. First, dark patterns are present across diverse environments, encompassing mobile applications, large online platforms, Virtual Reality environments, and all sorts of websites and digital services. This necessitates a large scale examination which is broader than an analysis directed at a specific environment, as well as a comprehensive understanding of the technologies used to design deceptive interactions in various environments. Second, dark patterns can be employed in various interfaces where users perform a variety of tasks such as registration pages, cookie consent banners, payment screens, content sharing on social media, as well as flight and hotel reservations. Each scenario differs in terms of design choices: while a deceptive design pattern such as a "bad default" [29] may steer users' decision toward a privacy-invasive option to obtain their personal information in a particular scenario, in a different scenario the same pattern may sneakily add an item to their shopping cart to increase the e-commerce vendors' profit . This situation calls for multiple strategies to detect all types of dark patterns, while developing a detection tool requires use case-specific proxies which help to recognize the dark patterns. Lastly, dark patterns can adopt distinct strategies of implementation characterized by different Human Computer Interaction design elements, such as reducing the visibility of an option, diverting user attention, imposing restrictions by eliminating the affordances for certain user actions, or manipulating language to exploit users' emotions. This emphasizes the necessity of considering multiple data types such as images, text, etc. when devising a detection solution.

Various studies have been instrumental in defining the phenomenon of deceptive design practices and in developing taxonomies to categorize and recognize them, such as [4, 5, 19, 29]. However, a precise and objective definition of dark patterns based on measurable elements that directly impact the identification process, remains elusive. To address this gap with concrete tools and quantifiable elements, we selected one of the most observed dark pattern use case (i. e., the cookie consent process) with the intention of producing objective descriptions of dark pattern characteristics. With our proposal, we aim to provide a process that relies on measurable features and that can be employed by anyone to reach reliable, reproducible conclusions on the detection of dark patterns.

While devising objective instruments to assess the presence of dark patterns, we focused on established attributes (such as asymmetric, restrictive, and deceptive) which have been defined by Mathur *et al.* [19]. We believe that the definitions of the attributes are high-level abstract

concepts that can be implemented in various manners. For instance, the definition of asymmetry concentrates on the unequal burdens placed on available choices [18], but what exactly is this unequal load and how it can be measured remains uncertain. Quantifiable features that are observable and measurable (e.g. the presence of buttons that provide the option to grant or deny consent to all available processing purposes) are critical for describing the dark pattern attributes. To prevent ambiguity and operationalize these abstract attributes, we investigated measurable features of cookie consent processes that can objectively describe the attributes in applied settings.

The deceptive design pattern types are mostly context-specific: for example, the dark pattern type "Hidden Costs" [40] can occur in check-out or payment web pages while it does not occur in other use cases. Moreover, each pattern type is not characterized by all attributes: for instance, Mathur *et al.* presented "Forced Action", i.e., a deceptive design pattern type, with the only required attribute "restriction" among the six existing ones [18]. This led us to identify the deceptive design pattern types that are present in the target use case of cookie consent banners and to propose the identification of the defining attributes of relevant types.

Research questions. We formulated the research questions below to meet the research objective of developing a measurable and objective process for the detection of dark patterns:

1. Which deceptive design pattern types are commonly employed in cookie consent processes?
2. Which deceptive design pattern attributes define the deceptive design types that are present in cookie consent processes?
3. What are the measurable human-computer interaction design features (i.e., visual, process-based and linguistic features) in cookie consent processes that can be operationalized to define the deceptive design pattern attributes?
4. How can the dark pattern attributes be objectively described through the measurable features?

Contributions. With respect to the research questions, our contributions are listed below in corresponding order:

- we identify the potential privacy-impacting dark pattern types in cookie consent processes.
- we elicit the relationships between the dark pattern attributes and the dark pattern types of cookie consent processes by presenting required, optional and non-required characteristics of each type.

- we propose measurable features that consider three main design angles (namely, visual, process and language) to recognize when a cookie consent process contains a dark pattern attribute, e. g. asymmetry.
- we describe the attributes based on the measurable features that we propose aiming to provide objective descriptions which can be utilized in the dark pattern detection.

4.2 Terminology

Before examining the methods and results of our study, it is necessary to clearly define some of the terms that we use frequently in the article, such as *type*, *attribute* and *feature* (composed of *entity* and *metric*), to avoid any confusion.

4.2.1 Type

Although dark patterns have the similar objective of influencing users' behaviors online, they are implemented in a variety of forms that can trick individuals, e. g. by hiding controls and information from them, forcing them to take predefined decisions or steering their actions towards desired outcomes in a predictable manner. The term "type" refers to a specific category of dark pattern within a system of classification that considers the tactics they employ and the scenarios where they can be found. For instance, there are 16 types in [40], such as "confirmshaming", "disguised ads", "forced action", "hidden costs", etc. However, there are many regulatory reports and academic papers that list and categorize dark pattern types. For instance, the European Data Protection Board's latest guidelines on dark patterns [8] listed 6 family types, e. g. "overloading", "skipping", etc., and 16 subcategories such as "continuous prompting", "privacy maze", "dead end", "emotional steering", etc. . In order to discuss the types in our study, we used a recent comprehensive taxonomy published by Gray *et al.* [29] as our main source (refer to Section 4.3.1 for the reasons behind this decision).

4.2.2 Attribute

This term refers to the characteristics of dark patterns that are not dark pattern type-specific, but rather represent general properties that characterize online deceptive designs. One dark pattern type can be described through one or more attributes. Mathur *et al.* first specified 5 attributes [19], and then added a sixth one in a following study [18]. We based our study on these attributes, namely *Asymmetric*, *Restrictive*, *Covert*, *Deceptive*, *Information hidden* and *Disparate treatment* (the definitions are given in Table 4.1). In the following sections, the

attributes will be written in *italics*.

Table 4.1: Dark pattern attributes with their descriptions from Mathur et al. [18]

Attribute	Description
Asymmetry	Unequal burdens on the choices available to the user
Restriction	Eliminate certain choices that should be available to users
Information hidden	Obscure or delay the presentation of necessary information to users
Covert	Hiding the influence mechanism from users
Deception	Induce false beliefs in users either through affirmative misstatements, misleading statements, or omissions
Disparate Treatment	Disadvantage and treat one group of users differently from another

4.2.3 Measurable HCI Feature

To define features, entity and metric need to be described first.

Entity: An entity refers to the specific element within the web design that is measurable. For example, ‘full consent grant button’ indicates whether there is a button that, once clicked, signifies user’s consent to all processing purposes. This is one of the measurable elements on the cookie consent banner and can be detected through CSS components of a web page as was done in [103]. Another example of an entity is ‘full consent grant process’ which identifies the path that the user needs to follow to provide consent to all processing purposes. This entity can be observed manually or via semi-full automated tools, crawlers, etc. as performed in the large scale analysis of cookie notices in [104].

Metric: A metric indicates and measures a quantitative aspect of the entity. For instance, the entity “full consent grant button” can be evaluated via different metrics such as ‘size’, ‘color contrast’, ‘position’ etc. The first one simply calculates the height and width of the entity in pixels. The second one focuses on the color values, e. g. RGB, of the entity and its surrounding, while the last one can measure the euclidean distance between the position of the entity and the center of the cookie consent banner. In other words, a metric can be conceptualized as a function where the entity serves as the input. The output of this function should be the same across different evaluators, thereby providing an objective assessment of the entity.

Feature: An entity measured on a predefined metric is a feature. Features are helpful to describe the dark pattern attributes. For example, ‘full consent grant button’ is an entity and ‘size’ is a metric. ‘The size of the full consent grant button’ is a feature which is quantifiable and measurable (and can be used to identify the presence of an attribute, as we will show later). Features refer to various aspects of the human-computer interaction HCI, including visual/graphical, process-based and linguistic/text-based aspects. For example, while the ‘color contrast of the full consent grant button’ represents a visual feature, the language used in the

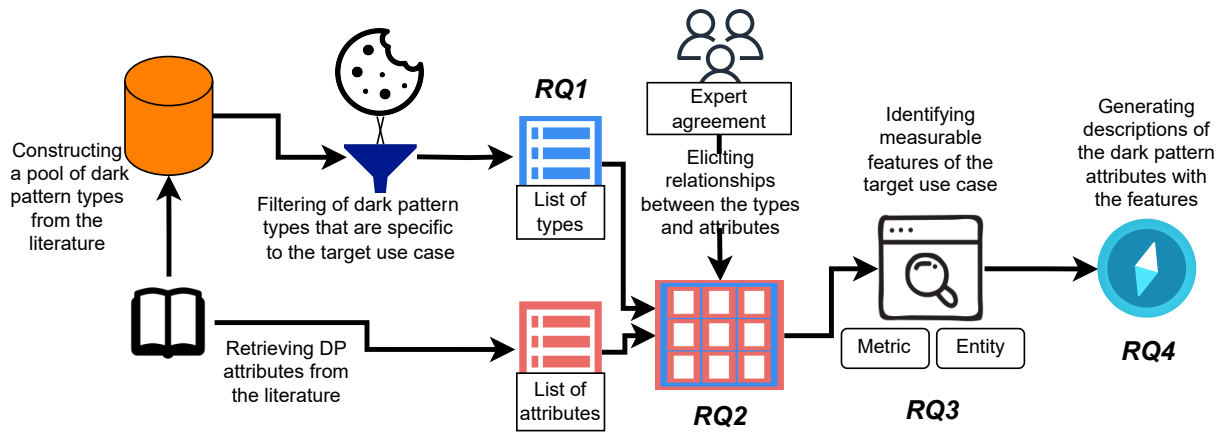


Figure 4.1: Methodology workflow exhibiting the steps and corresponding research questions.

button can be featured as ‘readability of the full consent grant language’ as a linguistic feature, and the user’s interaction with the system can be featured as ‘number of clicks for full consent grant process’ as a process-based feature. Methods and instruments to quantify the features differ for each of them. An example of a feature is shown in Figure 4.2 by annotating entities and utilized metrics.

4.3 Methodology

We follow a simple workflow composed of four steps as shown in Figure 4.1.

4.3.1 Identifying Dark Patterns in Cookie Consent Processes

Before understanding the characteristics of deceptive design patterns in the cookie consent process, the identification of the types that are specific to that use case is crucial because (i) the detection of deceptive design patterns requires identifying measurable elements that are related to the specific use case and (ii) each use case is characterized by its own design elements. This is why searching for a “Countdown Timer” in the cookie consent processes is pointless. Therefore, we identified the specific deceptive design pattern types that can be present in our target use case as a first step, to answer RQ1.

Construction of a Pool of Deceptive Design Pattern Types

The dark pattern literature is rife with taxonomies and categorizations. Such an abundance causes problems such as different labels associated with the same design pattern and varying levels of granularity in the categorization. For instance, the pattern based on constantly interrupting users’ tasks through prompts is called “Nagging” in the European Commission’s report

[105], while it is named "Continuous prompting" in the European Data Protection Board's guidelines [8].

Recently, Gray *et al.*, who are well-known scholars of the deceptive design research community, published an ontology that maps deceptive design pattern types contained in the academic literature and authoritative sources' reports into a hierarchical table [29]. We used this study as the main source for the construction of the deceptive design types pool because it is the most up-to-date and comprehensive on the matter (e.g. encompassing dark patterns relevant to EU data protection law [8] and US consumer protection law [6]), while the mapping table solves the conflict of different naming conventions. The table has 64 patterns in total.

It is paramount to consider granularity for assessing the presence of deceptive design patterns because the detection task requires descriptions that are as specific as possible, as opposed to abstract ones, as Mathur *et al.* highlighted in their cornerstone study [18]. Moreover, the attributes that characterize the low-level (i.e., the most granular) patterns do not necessarily correspond to the attributes that characterize the related higher level (i.e., the less granular) patterns. To enable the detection of specific instances, it is necessary to accurately characterize the attributes that help recognizing why a certain design pattern is deceptive - or not. In light of these considerations, we decided to employ the most granular available types to construct a pool of deceptive design pattern types.

Filtering of deceptive design pattern types that are specific to the target use case

Only a selection of dark patterns among the 43 types applies to the cookie consent process. Independently of each other, three experts, who have composite interdisciplinary backgrounds covering computer science (software development and artificial intelligence), data protection law, human aspects of privacy and security, and linguistics, and who have a research track on privacy-invasive design patterns, selected the dark patterns that are relevant for cookie consent processes and filtered out the irrelevant ones. Each expert was asked to perform a binary classification on the 43 types, where they selected the patterns they considered fitting and excluded the others. For each deceptive design pattern type, its definition from the main source [29], as well as an example from reports [8, 7, 106, 6, 105] or papers [4, 30] were given to the experts, together with criteria for exclusion and inclusion. To exclude a type, it had to be solely related to (a) purchasing decisions, (b) ways of spending time and attention, (c) implemented in social robots or (d) proper of the privacy domain but not applicable to a cookie consent decision. If the expert assessed that the type under consideration did not meet any of the exclusion criteria, the expert evaluated whether the relevant type was a plausible fit for the target use case, i.e.,

"cookie consent process." The voting results of the three experts were combined to compose the final list that consists of the types that achieved full consensus. Agreement through internal discussions was sought for the three cases where there was no full consensus.

4.3.2 Attributes of Deceptive Design Patterns in the Target Use Case

Contrary to the prosperous work concerning the establishment of types and taxonomies, there has been a limited focus on defining overarching deceptive design pattern attributes. In this two-round step, first, three experts independently evaluated whether each filtered dark pattern type could be described with one or more of the 6 selected attributes, to answer RQ2.

Two of the three experts also performed the previous filtering of types (see Section 4.3.1), with the third expert additionally providing a domain knowledge in formal methods and logic, in addition to human aspects of privacy and security. During this assessment, the relationship between the type and the attribute was expressed as `-Attribute is CODE in TYPE-`, where `CODE` could take one of three values: *"required," "optional,"* or *"not required"*. Definitions of the types and the attributes, as well as one example to validate their understanding, were shared with the experts. A total of 112 evaluations were conducted in the first round, covering 17 types and 6 attributes by each expert. When combined, each evaluation can result in a label representing one of four categories:

- "Consensus": all experts agreed on one of three `CODE`;
- "Majority decision without conflict": agreement only between two experts on *"required"* or *"not required"*, with the third expert indicating *"optional"*;
- "Majority decision with conflict": agreement only between two experts on a *"not required"* code, with the third indicating *"required"*, or viceversa;
- "High uncertainty" each expert provides a different evaluation (i.e., *"required," "not required," "optional"*).

The second round was aimed at solving the conflicts classified as "High uncertainty" and "Majority decision with conflict" and reach consensus. Three sessions for a total of 10 hours enabled the mutual sharing of ideas, the discussion of definitions, and the provision of arguments and counterarguments, after which the three experts reviewed their votes and made their final decisions to solve any conflict or high uncertainty cases. Given that each evaluation is categorized into three discrete values — *"required," "not required," "optional"* - and the participation of three experts, the permutations amount to 3 over 3, resulting in 27

possible combinations. Considering 112 evaluations, the total number of possible scenarios is 3024. Since consensus can be provided in 3 ways (i. e., "required-required-required", "optional-optional-optional", or "not required-not required-not required"), the probability of reaching consensus for one evaluation is 11,11%, while the probability for full consensus is $3/27^{112}$, which is close to zero. For this reason, experts aimed to solve conflicts during the second round.

4.3.3 Identifying Measurable Features of Cookie Consent Processes

This stage of the study answers RQ3. For doing so, we identified the measurable **feature** concept by clarifying its components, i. e., **entity** and **metric**, which were explained in Section 4.2.3. The necessity of this notation is better explained through an example. Habib *et al.* presented "readability of the notice" as a design parameter that can be built by different sub-elements such as "fonts", "colors", "contrast" etc. [47]. This parameter can be named "Readability of Privacy Notice Language" and classified as a **feature** according to our conceptualization. The privacy notice text, i. e., the language, is a measurable element that can be classified as an **entity**, and its readability is a (linguistic) **metric** which reflects whether the text is in plain language. "Readability of Privacy Notice Banner" can be classified as another **feature** that utilized the banner as an **entity**. Moreover, the readability here is a different **metric** and concerns the contrast, which is a visual design element. As explained in these examples, the notation we propose intends to reduce potential ambiguity and help the formalization of deceptive design patterns in the next step. As a consequence, the conceptual framework expressed by the equation below will be employed in defining measurable features of the cookie consent processes.

$$\text{Feature} = \text{Metric}(\text{Entity})$$

We extracted a set of features after reviewing [24, 45, 59, 54, 56, 9, 51], with specific attention to the measurable features of cookie banners provided in [29], that we validated and enhanced through the authors' collaborative analysis of some examples of cookie banners. We thus decided to group the entities (i. e., the measurable elements) into three categories: **visual features** (e. g. "full consent grant button"), **linguistic features** (e. g. "cookie policy text") and **process-based features** (e. g. "full consent grant user path"). An example of the method for extracting and defining the measurable elements of a cookie consent process is shown in Figure 4.2. As stated in Section 4.2, buttons are evident measurable design elements and the "agree and close" button on the cookie banner in Figure 4.2 can be identified as an **entity**. This **entity** can be measured via different **metrics**, e. g. "background color", "size", "position" etc., and this metric-entity pairs describe visual features such as "background color of full consent grant button".

“Readability of cookie consent banner language” is a linguistic feature that is composed of the text on the banner, i.e., an **entity**, and a **metric** such as “readability”. The “Learn more” button on the first layer of the cookie banner can direct the user to the second layer as shown in Figure 4.2, and the user path, i.e., the process, can be extracted as an **entity**. “Number of clicks”, “time” etc. can be **metric** for this **entity**, and together they identify a process-based feature. With this approach, we extracted a finite list of linguistic, visual, process-based features in Section 4.4 that can be expanded at will as new examples and new features are analyzed.

4.3.4 Describing Deceptive Design Pattern Attributes Based on the Features

In order to define the attributes of deceptive design patterns with observable and measurable features, they should be expressed as within a certain logic (e.g. as a rule), taking into account the relationship between the features. For instance, the features “size of the full consent grant button” and “size of the full consent refusal button” do not reveal the presence of dark patterns. However, a logical expression can elicit the relationship between them, such as “if the size of the full consent grant button is not equal to the size of the full consent refusal button”, indicates an asymmetric pattern, which is one of the attributes of dark patterns. In this regards, we defined rules using the features determined by the methods described in the previous section to minimize subjective interpretations. Eventually, the outputs of this step answer the last research question, i.e., RQ 4.

4.4 Identifying, Mapping and Describing Deceptive Design Patterns

Following the process described in Section 4.3.1, we identified 17 deceptive design pattern types that can be found in cookie consent processes out of the 43 types within the pool of deceptive design pattern types. The types and their definitions can be seen in Table 4.2, which is not a definitive list since other, or even novel, types of dark patterns can influence consent decision-making as new technologies and scenarios arise. While “Privacy Zuckering” was included with the majority voting, three experts unanimously selected the other 16 types. The “Forced Continuity” and “Personalization” types were selected by only one expert, and have been excluded after discussion. The remaining 24 types were not included in the list with full consensus. The filtering process’ inter-rater agreement score is nearly 0.926 according to Cohen’s Kappa metric. While there are 43 types in total in the pool, more than one third (39.5%) are relevant for cookie consent processes.

Figure 4.2: Examples of visual, linguistic, and process-based feature extraction from a cookie consent process.

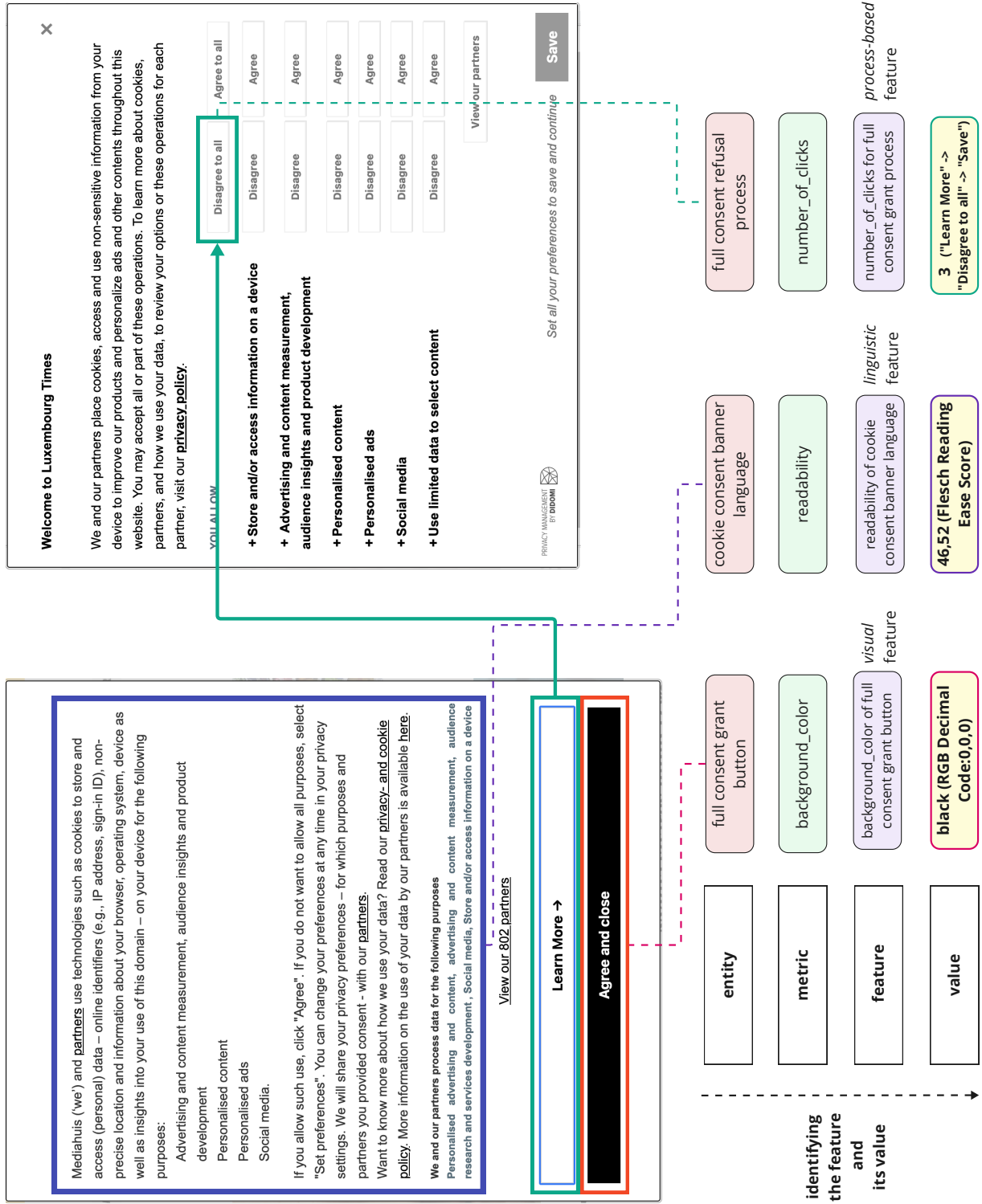


Table 4.2: Dark pattern types used in cookie consent processes (the definitions are given in Table C.1, in appendix)

Dead End	Privacy Maze
Nagging	Wrong Language
Complex Language	Confirmshaming
Information without Context	Hidden Information
Positive-Negative Framing	Conflicting Information
False Hierarchy	Bad Defaults
Visual Prominence	Feedforward Ambiguity
Trick Questions	Choice Overload
Privacy Zuckering	

4.4.1 Identifying the Attributes Characterizing Deceptive Design Patterns Types

We followed the methodology detailed in Section 4.3.2 to associate the attributes with the 17 types that were selected in the previous step. The detailed results of the mapping between the two is reported in Figure 4.3.

Our first finding was the difficulty of the process of eliciting the relationships between the types and the attributes that we will discuss in Section 4.5. This can be seen on the results of Round 1, which are shown in Figure C.1, in appendix. There was **Consensus** on 53 out of 119 evaluations (44.54%), 25 (21%) conflicts (encompassing **Majority decisions with conflict** and **High uncertainty**), and 41 (34.45%) **Majority decisions without conflicts**. Expert users pointed out that the definitions of the attributes are so broad that utilizing them for a concrete evaluation is hard. This challenge was also observed in the executive discussions held after the first round, as it took 10 hours of intense dialogical exchanges to eliminate all conflicts. Eventually, we concluded the experiment with **Consensus** on 89 evaluations (74,79%) and 30 (25.21%) **Majority decisions without conflicts**.

The overall findings of this expert analysis are:

1. No single attribute is consistently present in all dark pattern types, even though *Deceptive* is a frequently recurring attribute across the types.
2. No dark pattern type includes all attributes: each type has its own characteristics.
3. *Disparate treatment* is *required* only for "Wrong Language" and *optional* in "Nagging", i. e., it is not a common attribute in cookie consent processes.
4. Wherever there is the *Covert* attribute (always *optional*), the *Deceptive* attribute is also present.
5. "Conflicting Information" is not associated with any of the existing attributes. "Complex

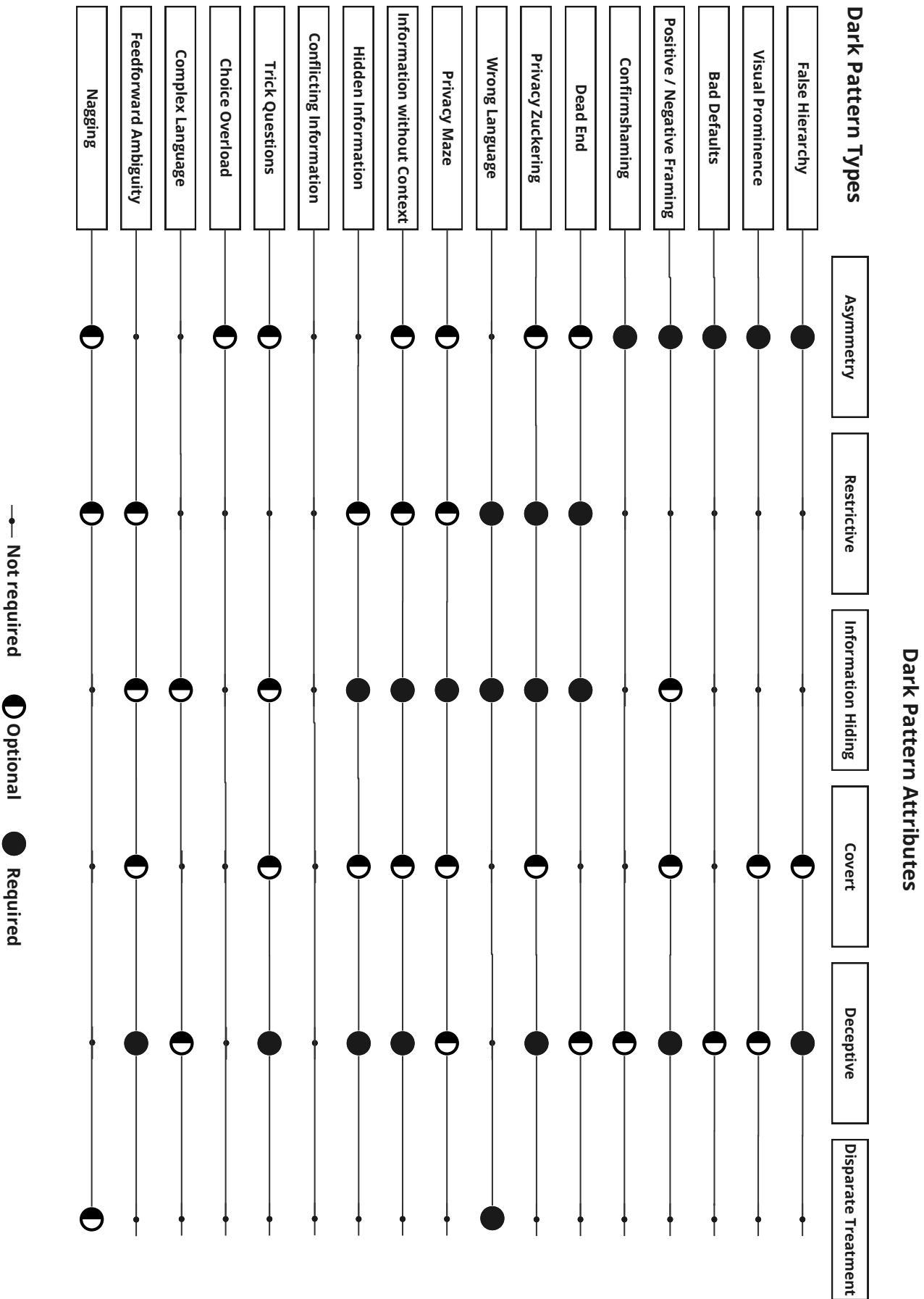


Figure 4.3: Association between dark pattern attributes and types in cookie consent processes, classified as required, optional, or not required

Language” and ”Choice Overload” also lack the required attributes. We believe this indicates that the current list of the attributes is not sufficient to represent the characteristics of all dark patterns (for instance, *Complexity* could be added, i. e., unnecessary cognitive load for users). We discussed this issue in detail in Section 4.5.

4.4.2 Defining Measurable Features of Dark Patterns

An attribute can appear in more than one form. For instance, asymmetric design can be based on graphical elements (e. g. full consent grant (FCG) and full consent refusal (FCR) buttons are not equal in terms of size), process-based elements (e. g. FCG and FCR processes are not equal in terms of number of clicks), or linguistic elements (e. g. FCG and FCR options’ language is not equal in terms of clarity). Following the methodology explained in Section 4.3.3, we defined measurable **features** for assessing whether a design pattern is potentially dark. Some examples of **features** such as ”size of the FCG button”, ”readability of the cookie consent info text”, and ”availability of the FCG process” are given with their **entities**, **metrics** and descriptions in Table C.1 in appendix, considering each category: graphical, linguistic and process-based, respectively.

These **features** can be used alone or in combination to define dark pattern attributes. For instance, a common dark pattern attribute in cookie consent processes, i. e., *Asymmetry*, which is a *required* attribute in five different types and *optional* in seven others, can be measured by employing the ”FCG button” and ”FCR button” **entities**. The comparison between two **entities** can be performed with different visual **metrics** such as ”size”, ”color”, ”contrast”. That is, ”size of the FCG button” and ”size of the FCR button” can be defined as **features** and if their values are not equal, there is an *Asymmetry* on the cookie consent banner. The *Asymmetry* can also be detected via **linguistic** or **process-based features**. For instance, for the process-based entities such as ”FCG process” and ”FCR process”, the **metrics** can be ”number of clicks”, ”time” etc., the values of which can be compared to detect the *Asymmetry*. The **features** can be quantitative (e. g. size, number of clicks, etc.) or binary (e. g. availability), according to their **metrics**. For instance, the *Restrictive* attribute can be evaluated by checking the availability of certain **entities** (e. g. choices). If the ”availability of FCR process” is false, this is a clear indication of a *Restrictive* pattern.

4.4.3 Measurable Feature-based Descriptions of the Dark Pattern Attributes

Without intending to be exhaustive, we enumerated a set of potential descriptions for each attribute with an easily applicable rule-based approach, as reported below. The features are

shown in italic.

Asymmetry can be described as:

A1 If *minimum number of clicks for the FCG process* and *minimum number of clicks for the FCR process* are not equal.

A2 If *availability of the FCG process* is True and *availability of the FCR process* is False.

A3 If *number of paths for the FCG process* and *number of paths for the FCR process* are not equal.

A4 If *size of the FCG button* and *size of the FCR button* are not equal.

A5 If *clarity of the FCG text* and *clarity of the FCR text* are not equal.

As seen in the Algorithm 1 that shows an example for A1, these descriptions are easily convertible to programmable rules, supporting automated detection and classification.

Algorithm 1 An Example for Asymmetry Detection with Measurable Features

Ensure: *asymmetry* (Boolean indicating presence of asymmetry)

```

1: Step 1: Detect all possible FCG and FCR processes
2: processesFCG ← detectProcesses(FCG)
3: processesFCR ← detectProcesses(FCR)
4: Step 2: Calculate minimum number of clicks for each process
5: minClicksFCG ← min(clicks required for each path in processesFCG)           ▷ Feature
6: minClicksFCR ← min(clicks required for each path in processesFCR)           ▷ Feature
7: Step 3: Initialize asymmetry variable
8: asymmetry ← False
9: Step 4: Check for asymmetry
10: if minClicksFCG ≠ minClicksFCR then
11:   asymmetry ← True
12: end if

```

Restriction can be described as:

R1 After *action status of the Consent Request (CR)* is FCR or FCG, if *availability of the CR process* is False.

R2 If *availability of the FCR process* is False.

R3 If *language of the CC information text* and *language of country the website serves* are not equal.

Information Hidden can be described as:

I1 If *completeness of the CC information text* is False.

I2 If *semantic of the cookie policy heading* and *semantic of the CC information text* are similar.

Covert can be described as:

C1 If *availability of the FCG icon* is True and *availability of the FCR icon* is False.

Deception can be described as:

D1 If the *action status of the CC process* is Refusal and *availability of unnecessary cookies (UN)* is True.

D2 After *expiration date of the cookies* exceeded, if *availability of the CR process* is False.

Disparate Treatment can be described as:

T1 After *action status of CC process* is FCG for user1 and FCR for user2, if *availability of the CRR process* is False for user1 and *availability of the CRR process* is True for user2.

T2 If *frequency of the CRR process* for user1 and *frequency of the CRR process* for user2 are not equal.

4.4.4 Applied Example of Feature-Based Dark Pattern Detection

The cookie consent process example in Figure 4.2 was assessed to check if it has dark patterns. Firstly, *Asymmetry* was detected by comparing the “number of clicks for the FCG process” and “number of clicks for the FCR process” **features**, i.e., A1 defined in the previous section. Users can perform the FCG with one click on the “Agree and close” button while they need to perform 3 clicks for the FCR through ‘Learn More’, ‘Disagree to all’ and ‘Save’ buttons. Therefore, the two patterns are asymmetric and the second one corresponds to the dark pattern known as “Privacy Maze”. Another assessment was performed by employing the visual features such as ‘size of the FCG button’ and ‘size of the FCR button’. Considering only the second layer, these **features** are equal, because the ‘Disagree to all’ and ‘Agree to all’ buttons’ size is equal. Therefore the design is symmetric on the second layer. *Asymmetry* was also detected through visual features related to the color of the available buttons, i.e., “background color of full consent grant buttons” is different than the “background color of the consent setting button”. Making a certain choice more visible than another may be labelled as “False Hierarchy”. Lastly, the linguistic feature “readability of cookie consent banner language” was used to compute the clarity of the text: the Flesch Reading Ease Score equals 46,52 which corresponds to the College grade educational level needed to understand the text, a symptom of *Complexity*. Therefore

it is a “Complex Language” dark pattern. The cookie consent management process in Figure 4.2 may also contain other dark patterns, but hereby, we merely aim at showing the concrete application of our approach to a real example.

4.5 Discussion

4.5.1 Identification of Dark Pattern Types in Cookie Consent Processes

Not all dark pattern types can be found in all use cases. Different types can be combined but they are not necessarily all present at the same moment. In the cookie consent process, we have identified 17 different types of deceptive design patterns that can potentially be present, presumably related to the fact that there are great economic interests in luring users to accept web tracking for profiling and advertising purposes. Almost perfect agreement was observed among experts regarding the identified 17 types. However, since the field and regulations continuously evolve and regulation-driven changes can occur in the cookie consent process, or new technologies may introduce new dark pattern types, the list is not definitive and should be updated in accordance with current taxonomies.

4.5.2 Attributes Characterizing Dark Patterns

In our study, certain attributes seem to constitute the basic characteristics of certain dark pattern types. For example, *Asymmetry* is required in five deceptive design types - therefore, the assessment of those dark patterns can be carried out by solely or predominantly focusing on the asymmetric nature of certain features. Yet, only a handful of types can be detected through a single attribute. Most times, dark patterns can (and should!) be analyzed under various perspectives that go beyond the superficial level offered by the graphical user interface (reflected in the various **features** that can be examined) and in different contexts they may expose different characteristics, as the need for *optional attributes* shows. For instance, *Restrictive* and *Information hiding* are essential attributes of the “Dead End”, while they are optional in “Feedforward Ambiguity”. From this follows that certain dark pattern types may be easier to detect through their essential attribute(s), whereas others have a multifaceted nature that requires more complex descriptions.

Disparate treatment is only required in “Wrong language” and occasionally found in “Nagging” in the case under examination. This may mean that *Disparate treatment* is not a frequent quality of the deceptive design patterns implemented on cookie banners, whereas it could be in other cases (e.g. online purchases). As a methodological choice, we decided to

interpret *Disparate treatment* not in terms of its outcomes (i. e., whether it has a disproportionate outcome on certain users or group of users) but in terms of its implementation (i. e., whether the dark pattern is implemented differently for different users or group of users). For instance, continuous prompts (i. e., nagging) may be used to exasperate those users who did not consent to advertising and make their navigation more cumbersome until they agree, as a recent sanction by the Italian DPA shows [107, 108]. This is a case of disparate treatment in the implementation. However, certain dark patterns may have more severe implications on certain users rather than others (i. e., disparate treatment in terms of outcomes). For instance, "Wrong language" may disproportionately impact the people that do not master that language, while "Privacy maze" may more severely affect users with poorer digital skills. We decided to focus on the first interpretation as it is more objective to establish, even though a growing body of literature seeks to prove how certain people or groups of people may suffer from deceptive design patterns' influence and harms more severely than others [109, 110].

Eliciting the relationships between the types and the attributes was a complex task, one reason being that the definitions of the attributes are broad and abstract, thus using them for a concrete evaluation is challenging. Furthermore, we reached different conclusions than Mathur *et al.* [19] for certain patterns. For instance, we deliberated that "Trick question" is always *Deceptive* due to its misleading nature, while it is not labelled as such in their study. Unlike the previous study, we also argue that it can optionally be *Asymmetric*, since it may place unequal burdens on two or more options, and *Hiding information*, since it may obscure necessary information, but not necessarily. This evidence suggests that defining measurable features is a necessary work and that future discussions among experts should elaborate more specific definitions of the attributes.

The *Covert* attribute needs to be discussed on its own, since we couldn't find that there is no case such that a pattern is *Covert* but not *Deceptive* in our analysis: since the influence mechanism is hidden from users, covert dark patterns are always misleading (i. e., deceptive). Therefore, it seems that *Covert* patterns can be categorized as a subset of *Deceptive* ones, even though further work should determine whether this conclusion applies to a broader range of use cases. The application of this attribute was lively discussed among the experts in this step as the consensus rate was generally low, likely reflecting the low usability of such an attribute for providing an objective criterion of detection of dark patterns. This mostly depends from the interpretation of "influence mechanism": does this notion refer to the influence strategy (e. g. for positive/negative framing: focusing on the positive consequences of a user decision, while purposefully withholding the negative ones) or does it rather refer to the effect (e. g. the swayed

decision)?

Complexity. Moreover, some dark pattern types cannot be characterized by any attribute that can be found in the literature. For instance, "Choice Overload" increases the number of options and thus the cognitive burden placed on users to complete the process. However, none of the attributes by [18] can account for this behavior. Therefore, we found it necessary to add *Complexity* as an attribute, which describes the additional unnecessary burden for users in terms of time, actions, cognitive effort, etc. Similarly, the European Commission proposed a behavioural taxonomy on design practices infringing consumer protection provisions [105] that considers complexity as the essential attribute of the dark patterns mentioned in the report. Cognitive load has been recognized as a harm [18] since it causes individuals to waste time, energy and attention, especially when one considers the cumulative effects of deceptive design practices. Another example that indicates the need for adding *Complexity* is that even though "Complex Language" can include *Information hiding* or *Deceptive* attributes, there can be scenarios where none of them are present. For example, all necessary information can be present in the cookie banner without any omission, but the language may be too complicated for users to enable an informed decision. Thus "information hidden" or "deception" are not constitutive elements of this deceptive design type, but complexity is. This insight leads to the hypothesis that the list of attributes may be expanded, also to account for other types of interfaces: for example, voice interfaces may embed dark patterns that play on linearity and volume of voice [111].

Context and expertise. Moreover, we have observed that including context and domain expertise is paramount for the assessment process. For example, a design pattern may be *Restrictive*, as it forces the user to undertake a certain course of action. However, sometimes restrictive designs are necessary, one reason being that legal obligations impose that user consent is asked before collecting their personal data for certain purposes. A design pattern may be *Restrictive* but mandated by law, thus not necessarily illegitimate, even though it is open to discussion determining whether a lawful design choice that does not serve the user interest (e. g. "Choice Overload" in specific consent regimes) amounts to a dark pattern. A too narrow focus on GUI elements should also be avoided, as many dark patterns are hidden in the interaction between user and system. All these considerations further reinforce the necessity of elaborating interdisciplinary approaches for evaluating whether a design pattern is dark or not in a reliable manner. A future development of this work should seek to map these attributes to legally relevant attributes, such as the definition of abusive commercial practices or the violation of equal treatment in specific jurisdictions. For example, it should be determined what kind of asymmetry in the visibility of certain options is illegal, since a simple difference cannot be

automatically labelled as dark.

4.5.3 Measurable Features

The measure of certain features can indicate whether a design pattern is potentially dark, especially when multiple features hint at the (co-)presence of one or more attributes. However, certain attributes are not black or white properties, and rather have degrees. For example, *Complexity* can be measured through the number of required clicks, but there is not necessarily a threshold for such a number. Moreover, even though measurable features are critical components of the dark patterns' detection process, they vary according to the use case. The list of process-based, visual-based and linguistic features we identified may be expanded when the number and typology of examined websites is broadened. Furthermore, these features are use case-specific, thus in other use cases (e.g. "e-shopping"), the features should be updated to detect the attributes in that use case.

Further, the **metrics** of the feature must be clearly defined before the feature is built to avoid measuring in different units. For example, for "readability of cookie consent (CC) info text", which is a linguistic feature, the "readability" **metrics** can be defined via different instruments such as the Flesch Reading Ease Score, the Flesch-Kincaid Grade Level, the Gunning Fog Index, etc. Another example is the "size" **metric**, which can commonly be measured in pixels with the order of width and height. Therefore, the description, scale and units of the **metrics** should be clearly defined. In addition, customized **metrics** can be defined for particular use cases. For instance, "completion status" can be a binary **metric** which can take the values of "True" and "False" for the "cookie consent process" **entity**. A crucial aspect of defining and using **metrics** is that they must generate consistent and objective results. This ensures that every implementation of the **metrics** equates to the same conclusion. Moreover, it should be considered that selection metric is crucial because different metrics can demonstrate varying levels of effectiveness depending on the context. For instance, when measuring the "clarity" metric of the text, various readability metrics can be employed, such as the Gunning Fog Index or the Flesch Reading Scale.

4.5.4 Relevance

While our focus has been on the cookie consent process, the method employed in this study can serve as a blueprint for investigating dark patterns in other use cases. For example it can be employed in any consent management UI (beyond cookies), in privacy settings of digital services as well in service registration processes or e-commerce check-outs (where asymmetries

and restrictions are common). It can also be applied to uncover information hidden dark patterns in e.g., legal documents such as privacy policies and consumer contracts. Even though this work has focused on GUIs, further work that defines the features of e.g., voice interactions could leverage our method to determine with higher certainty if those interactions are manipulative. The potential applications are many and go beyond these few examples.

Studies like ours will reinforce the world of bright design patterns in privacy [112], because they ultimately aim at establishing a set of requirements to avoid dark pattern design attributes. If, for instance, it is established that *Symmetry* is a required attribute for design patterns that counter *Asymmetric* dark patterns, it becomes possible to establish clear, tangible requirements for designers and developers, thereby going beyond predominant current research on dark patterns that exclusively "tell[s] a designer what not to do, when a designer is usually seeking advice on what to do" [113, p.1]. However, it is important to lay down requirements that offer lawful guidance but do not deviate from what empirical research demonstrates about the effect of certain design choices on users. The two do not always coincide, as the gaps identified by Bielova *et al.* [114] show.

We stress the necessity of developing an easy, reliable, (semi-)automatized detection of dark patterns which may enable fairer interactions in digital services. Our work, i.e., both the precise descriptions of the extracted features and the rule-based descriptions of the dark pattern attributes, can be utilized in automated dark pattern detection tools such as artificial intelligence-based and rule-based models. For instance, our feature concept can be employed in the feature engineering step of machine learning pipelines to automatically detect dark patterns in cookie consent processes.

Describing dark patterns in a feature-based manner can facilitate the work of various expert users: academic researchers that study the effects of design elements on user behaviour and that develop automated approaches to dark pattern detection; designers and developers that need tangible "how-to" guidance in their everyday work, rather than abstract principles and no-gos; supervisory authorities that inspect interfaces to investigate the compliance of businesses and provide guidance on technology design by interpreting the law; as well as civil society organizations that scrutinize and denounce problematic design practices. Observable HCI features and objective descriptions can represent a common language for enabling strategic collaboration across these sectors and across various domains (computer science, law, and UX/UI design).

4.5.5 Future Improvements

Natural languages, e.g. English, are powerful in terms of expressiveness, but their complexity and their inherent semantic ambiguity can generate unclear, ambivalent descriptions. Therefore, while describing the dark pattern attributes or defining rules that assess the presence of dark patterns, formal or semi-formal languages can be employed to bring precision and clarity. Considering the multidisciplinary nature of the dark pattern research field, a "Controlled Natural Language", which is placed somewhere between natural languages and formal languages (e.g. First-Order Logic) [115], could be adapted to further define the descriptions of the dark pattern attributes, because it is close to human natural language and allows everyone to reach the same conclusion on the presence of dark patterns by eliminating ambiguity, instead of a formal language that is technically difficult to understand and thus requires expertise. Such languages can be used to define requirements which instead of specifying bad practices (i.e., dark patterns), they specify good practices what should be implemented (i.e., fair patterns or bright patterns) and then detect the deviation from such ground truth to identify dark patterns. In both cases, the same features that we defined in this study can be used. However, first, best practices should be established and widely accepted by the community, which is a matter of current discussions.

4.5.6 Limitations

In this work, we identified a set of useful GUI features without striving for completeness. Beyond the GUI related features that we defined (following [29], the ones related to human-computer interaction), future work should include machine-to-machine interaction features (e.g., size, typography, expiration date of cookies) and maybe other features that are related to the effect of certain designs on users (cognitive features?) that could be useful for detecting dark patterns. As Mathur *et al.* stated in the "What makes a dark pattern dark pattern" study [18], actionable dark pattern definitions require as much specificity as possible, and actual use cases should be taken into consideration while increasing the specificity. In this study, we carried out feature extraction and dark pattern attribute description focusing on the "Cookie Consent Process" use case, which could limit the aspects of examining dark patterns. Moreover, as we have found necessary to add "complexity" to the set of attributes, it may be that other attributes should be identified, especially when it comes to non-GUI dark patterns, such as the linearity in voice interactions [111].

4.6 Wrap-up and Perspectives

In the previous chapter, we argued that objective assessment of dark patterns requires measurable and operational features and descriptions. As an initial step toward this goal, we introduced process-based features, such as "number of clicks per option 1". In this chapter, we extend that scope to include visual and linguistic features and propose a systematic approach that formalizes the dark pattern attributes, i.e., recurring characteristics across various dark pattern categories. Rather than relying on subjective judgments or non-actionable high-level descriptions, we demonstrate how these proposed features can be used to specify dark pattern design in a structured manner.

Using the cookie consent scenario as a focused use case, we illustrate the importance of user context and how it narrows the potential dark pattern types while aiming their detection. At the same time, our approach can be applicable to other use cases. Additionally, we elicit the relationship between potential dark pattern types and dark pattern attributes within the cookie consent context. This chapter addresses the lack of objective detection criteria that can be used to explain dark pattern identification decision process, thereby supporting both transparency and interpretability. The findings of this chapter have contributed to **Research Objective 1: Defining measurable instruments for the objective assessment of dark patterns**, and provided a comprehensive answer to the **Research Question 1: What are the measurable functional properties of dark patterns and the assessment instruments to define them?**

Having established a systematic approach for the reliable assessment of dark patterns (Chapters 3 and 4), this thesis pivots to address the next major challenge in achieving effective automated detection: the absence of a gold-standard dataset suitable for training and rigorously evaluating AI-based detection models. Consequently, the next chapter provides a comprehensive review of the current dark pattern dataset landscape, focusing on existing limitations, present quality issues, and the necessary attributes of a benchmark dataset.

Chapter 5

Deceptive Design Datasets

We examine the existing dark pattern datasets, which are critical for evaluating the performance of dark pattern detection solutions, and identify issues such as labeling inconsistencies and overly broad classification, which make it difficult to achieve a reliable evaluation of detection solutions. We then describe the process of creating our own dark pattern dataset, developed for benchmarking purposes, and compare it with existing ones through expert-based and model-based assessment.

*This chapter answers **Research Question 2**: "What considerations and methods should be taken into account when building labeled and structured dark pattern datasets?" and its content is primarily based on the following papers:*

- *"Kocyigit, E., Rossi, A., Sergeeva, A., Negri Ribalta, C., Farjami, A., & Lenzini, G. (2025, June). **DeceptiLens: an Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model.** In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1942-1959)."*
- *"Kocyigit, E., Liga, D., Rossi, A., Sergeeva, A., & Lenzini, G (Under review). **Evaluating the automated detection of dark patterns in GUI datasets with open-source Multimodal Large Language Models.** (Manuscript submitted to *PeerJ Computer Science*, October, 2025)"*

5.1 Introduction

Following the systematic approach for reliable dark pattern detection developed in the previous chapter (Chapter 4), this chapter now focuses on the dark pattern datasets because the choice of the evaluation dataset for dark pattern detection solutions is crucial for a reliable understanding of the capabilities of these detection methods. However, the datasets used for the training and/or testing of these tools have significant limitations. First, there is a lack of reliable and standardized benchmark datasets, as each tool [19, 50, 72, 75] was assessed using different datasets. Second, each dataset uses a different labeling convention, which exacerbates the problem of comparability across approaches. This is also due to the plethora of taxonomies of dark patterns that have been created. For example, the presence of a small close button was considered sufficient to classify a UI as a DP in a dataset while DP experts from various fields, such as computer science, human computer interaction, linguistic, UX, and law, disagree. Third, as indicated in a study that inspects existing detection tools and their datasets [78], almost half of the DP categories are underrepresented or absent in existing datasets, which makes them skewed towards certain types of dark patterns.

While no single dataset can fully resolve these challenges, we decided to create a dataset constructed with transparent multi-expert validation to provide a more reliable foundation for evaluating detection tools. To construct this dataset, we adopted the ontology of Gray et al. [29] as a foundational taxonomy. Starting from 43 categories, we filtered and merged them into 9 labels representing 13 dark pattern types. The dataset consists of real user interface screenshots from web and mobile sources, including true and deliberately selected false-positive examples. The experts independently annotated the instances and the final label was determined on the basis of majority voting (each instance was evaluated by three experts). While not exhaustive, this expert-validated dataset offers transparent, ontology-aligned labels to support more reliable benchmarking of detection tools.

This chapter makes the following contributions:

1. Developing an open, expert annotated benchmark dataset of graphical user interfaces for dark pattern detection.
2. Identifying the quality problems of existing dark pattern datasets.

After providing background on dark pattern data and datasets in Section 5.2, we identify the dataset-related research gap in Section 5.3. We then describe our methodology in Section 5.4 and present the empirical results in Sections 5.5. Finally, we discuss the limitations in Section 5.6 and conclude the chapter with closing remarks on the dark pattern dataset and

future research directions in Section 5.7.

5.2 Background

The primary properties that make a design a dark pattern may stem solely from textual elements, from visual components, or from the interplay of both. For example, a large scale dark pattern dataset [19] constructed by crawling online shopping websites consists mainly of text data. In this dataset, data curators specified whether an instance constitutes a dark pattern, assigned high-level and low-level categories, and documented additional metadata such as its placement within the interface. A set of samples from the dataset is given in Table 5.1.

ID	Text	Category	Type	Where	URL
1	9 people are viewing this.	Social Proof	Activity Notification	Product Page	https://...
2	No thanks I definitely measured correctly	Misdirection	Confirmshaming	Cart page	https://...
3	Your cart is reserved for 09:35 minutes.	Urgency	Countdown Timer	Checkout Process	https://fuzz.online...
4	ONLY 7 LEFT!	Scarcity	Low Stock Message	Product Page	https://jskis...
5	No I'd rather not hear about exclusive offers..	Misdirection	Visual Interference	Checkout process	https://www.greenf...

Table 5.1: Samples from dark pattern text dataset [19]

We analyzed this dataset and observed empty and duplicate rows in the "Text" column, which stores the primary dark pattern text data. After removing all entries with empty or duplicate text, the number of instances decreased from 2356 to 1178. The final dark pattern category distribution is shown in Figure 5.1. Another dataset [94] builds upon this corpus by adding non-dark pattern examples from the same crawled e-commerce websites. However, since both datasets are text-only, they rely on additional metadata to encode visual dark pattern types. For example, instance 5 in Table 5.1 includes an annotator note stating "negative option .. is grayed out". Without this metadata, the *Text* field alone would be insufficient to identify a *Visual Interference* dark pattern. In such cases, raw image data are needed for comprehensive assessment and quantification of the degree of gray-out.

Beyond text-only datasets, other datasets include user interface representations of dark patterns in image and video formats. For example, Geronimo et al. [116] created a collection of video recordings of mobile applications in categories family, music, news, etc. and analyzed the presence of dark patterns while interacting and using the applications. As shown in Table 5.2, the study documented timestamps of dark pattern appearances and include descriptive notes ex-

plaining each classification. While this manual annotation by users provide qualitative insights, processing long video segments is not feasible for most AI models.

More recent work aimed at automated detection of dark patterns focused on user interface-centric approaches, collecting structured datasets of screenshots in 2023 [35, 72]. Mansur et al. [72] created a dataset to support the automated identification of dark patterns in web and mobile application interfaces. Their dataset comprises 501 screenshots, of which 301 instances (55% of 544 annotated segments) were classified as dark patterns. To categorize these patterns, they compiled three existing taxonomies [5, 33, 19] and obtained seven high-level categories. However, the dataset includes examples across five categories, such as *Nagging*, *Misdirection*, *Forced Action* etc., and it also incorporates visual cues and localization analysis, capturing the

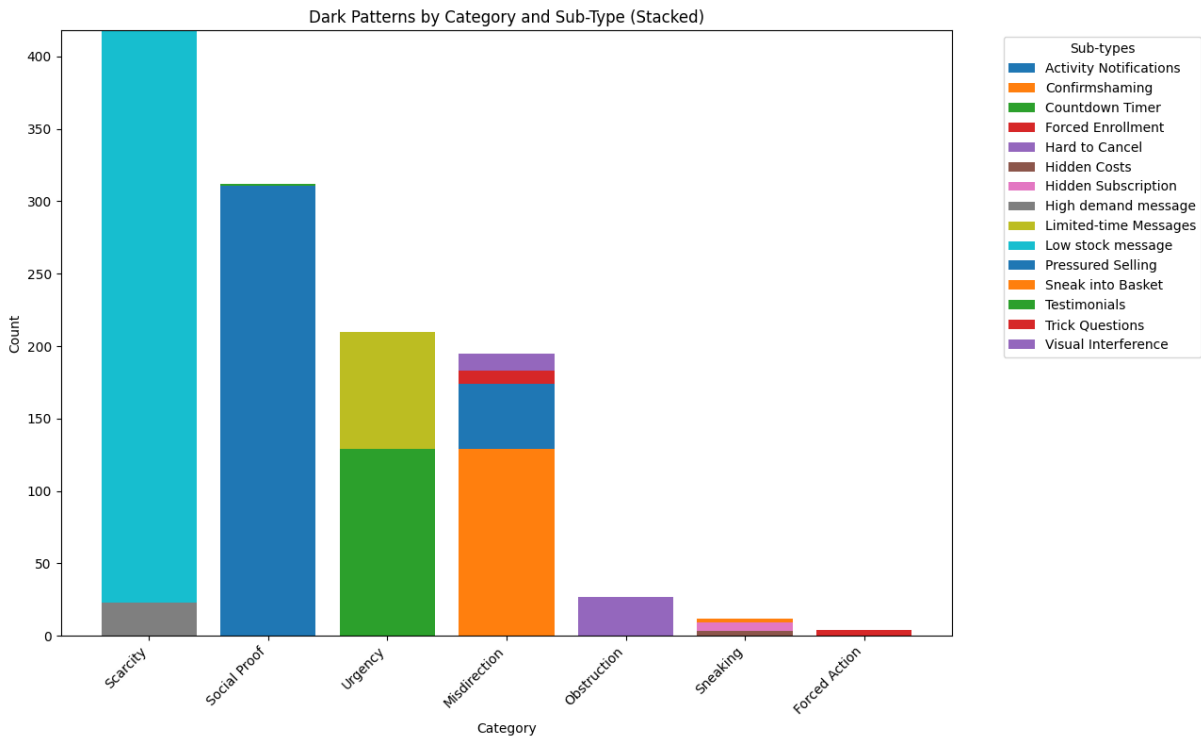


Figure 5.1: Distribution of dark pattern categories after removing empty and duplicate entries in [19]

Table 5.2: Example annotations from a family-category game app video. FA: Forced Action, DA: Deceptive Advertising, NG: Nagging

Timestamp	Class	Notes
00:16	FORCED ACTION: Watching ad to unlock feature	watch an add to get some weird coins
01:18	DISGUISED ADS: Icons/buttons are ads, but it's not clear	a star appears as normal content instead it brings you on the google play
09:56	NAGGING: A popup appears and interrupts the user in their task	when finishing a level they ask you to rate the app, "never" "later" "ok"

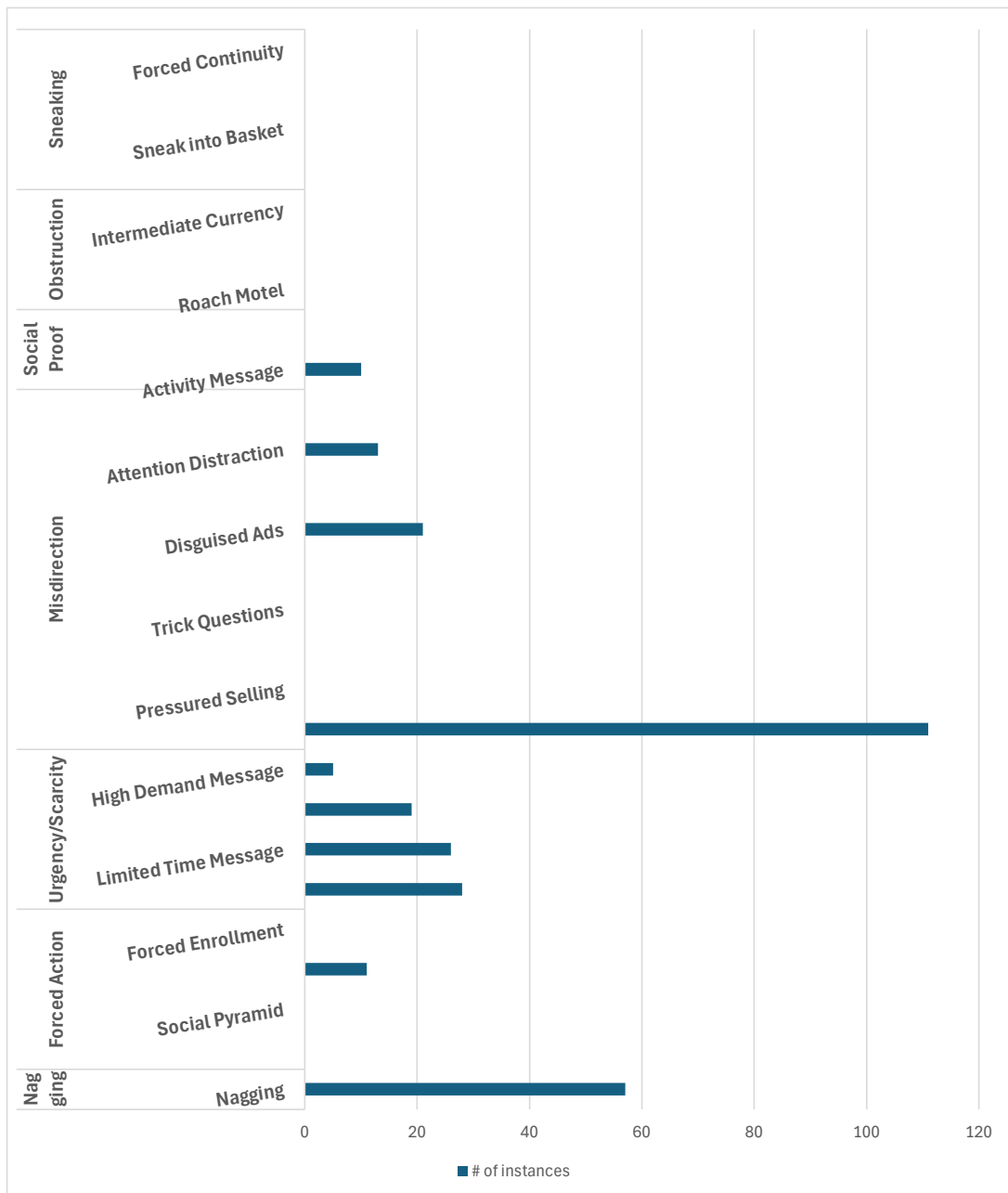


Figure 5.2: Dark pattern category - type distribution of UI-based dataset by Mansur et al. [72]

graphical elements that influence dark pattern classification (see Figure 5.2 for the distribution of dark pattern types).

In another study focusing on the user interface [35], the researchers constructed a dataset consisting of 6,352 screenshots of mobile applications. 1,660 instances were identified as dark patterns. The screenshots were sourced from a large mobile user interface dataset [117], and each dark pattern instance was annotated with metadata specifying the exact interface region in which the pattern appeared. The dataset curators synthesized three existing dark pattern taxonomies and introduced additional labels, e.g., *General Type*. For example, within the *Inter-*

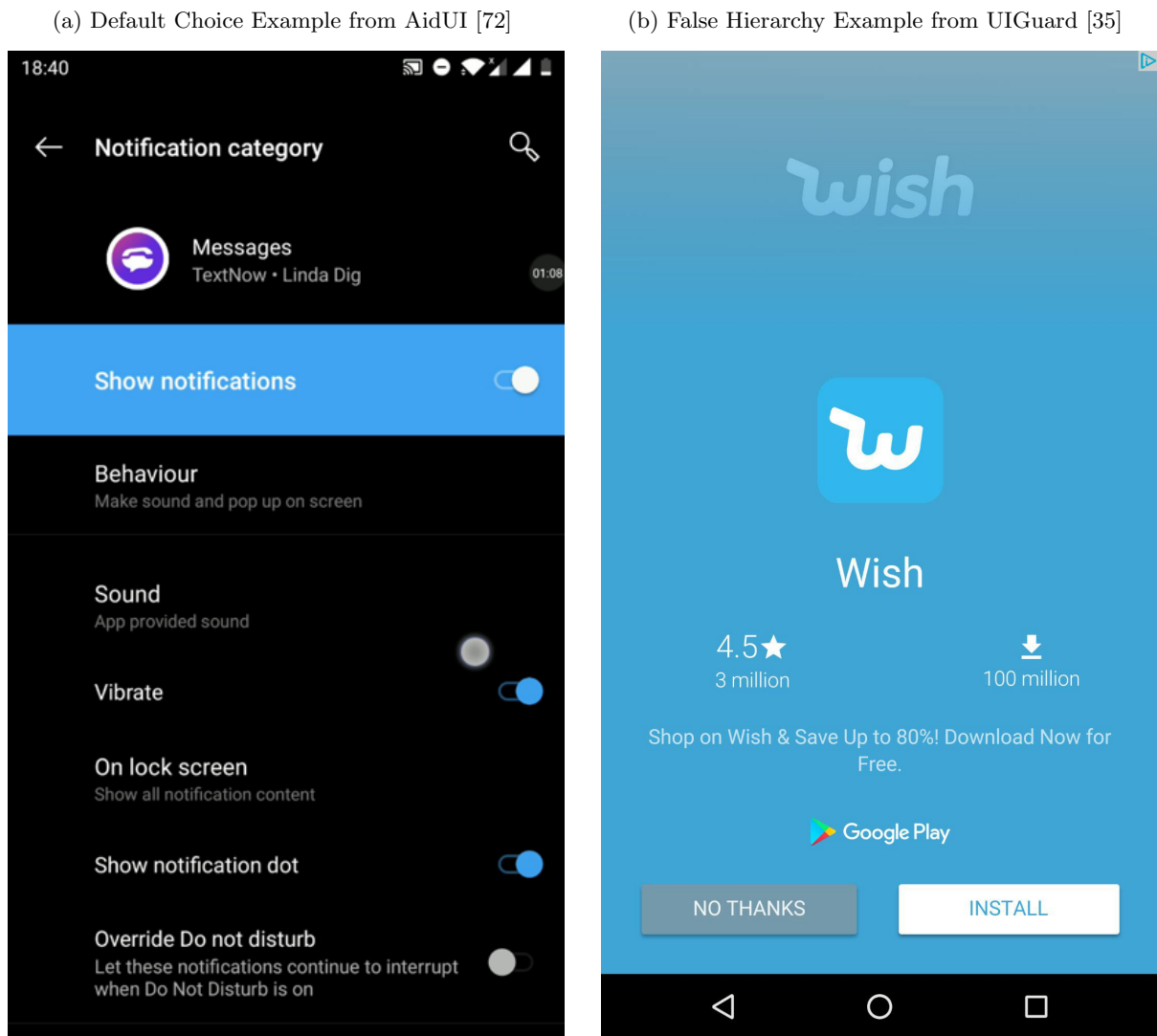
face Interference category, interfaces containing a small close button were assigned to *General Type* label.

5.3 Research Gap

Based on a review of the literature, we selected the existing available DP datasets [19, 50, 75, 72, 63, 34] and investigated whether they are usable for the reliable evaluation of the performances of multimodal AI-based DP detection models. Given the multimodal nature of DPs, we excluded datasets that only included text data like [19, 63]. Even though we first intended to merge the selected image-based datasets to have a comprehensive DP dataset, we identified issues that could lead to misleading evaluations, i.e. risks associated with dataset quality [118].

First, some datasets label screenshots as DP based solely on visual clues, without considering the context in which they are used. For example, a mobile UI screenshot (see Figure 5.3a) that displays notification toggles that are turned on by default is labeled as a "Default choice" DP (which may correspond to "Bad default" in [29]). While it could be argued that notifications turned on by default are detrimental to users because they compete for their attention, this may need to be determined contextually as notifications can also be useful. Moreover, in the literature, many examples of bad defaults refer to privacy-invasive defaults, e.g., in [8]. Similarly, design patterns that play on scarcity and urgency should be considered deceptive only when they mislead users into believing false information. For example, countdown timers are deceptive only when they reset automatically after a preset period, thereby indicating that the limited offer is not limited after all. On a similar vein, low-stock messages or high-demand messages are deceptive only when they refer to goods that are not scarce or are not in demand. When they refer to actual facts, these design patterns may genuinely inform consumers of limited-time offers or goods in short supply, and they may thus help them make decisions in their best interest. Only contextual information can disambiguate whether certain design patterns are deceptive or not.

Figure 5.3: Examples labeled as dark pattern in existing datasets



Second, some definitions of DPs are interpreted in a way that risks being overly comprehensive and thereby include UIs that are not necessarily deceptive. For instance, UIGuard dataset [75] applies the label "False Hierarchy" to UIs where "one option is more salient than other equal options" (see Figure 5.3b). However, visual hierarchy is a fundamental principle of UI design that helps the user to navigate various options without being overwhelmed. Making all options appear the same is not necessarily the solution, as visual hierarchy can be used for noteworthy goals, such as security-critical cases where it is meant to add friction to the user journey and thereby discourage users from taking mindless decisions. In a similar deterministic way, the label "Hidden information" is applied to any "grayed small text" that refers to relevant information such as terms of service. Whereas graying out text is a common deceptive strategy that disguises information that should be visible, sometimes it is simply a UI design strategy that avoids cluttering (mobile) interfaces with too much information with equal visual salience,

which would cause information overload.

5.4 Methodology

First, we choose the taxonomy of Gray et al. [29] as a starting point for the creation of our dataset. We apply the granular specification as we demonstrated in our systematic approach proposal presented in Chapter 4 and list 43 categories of dark patterns. We selected 28 types among them by applying filtering criteria, which is described in Section 5.4. Afterwards, we collected screenshots of dark patterns in web or mobile user interfaces from several sources as given in Table D.1. We also include false positives in the dataset. We prefer both user interfaces without dark patterns and those with dark patterns but labeled differently. For example, Figure 5.4 demonstrates an example of *Countdown Timer* dark pattern that was reported in [19]. On the other hand, Figure 5.5, which has a similar design considering having timer, illustrates a false positive example of same category. We then exclude categories for which we do not find at least 5 UI examples reported as dark patterns. Moreover, we combine similar and overlapping categories, such as "High Demand and Low Stock, or *Countdown Timer* and *Limited Time Messages*, and obtain 9 labels covering 13 dark pattern categories as given in Table D.2.

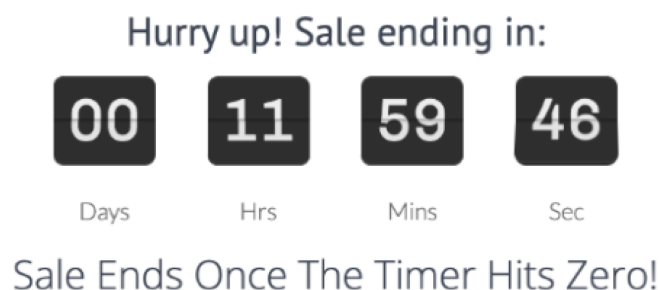


Figure 5.4: A true positive example of *Countdown Timer* dark pattern

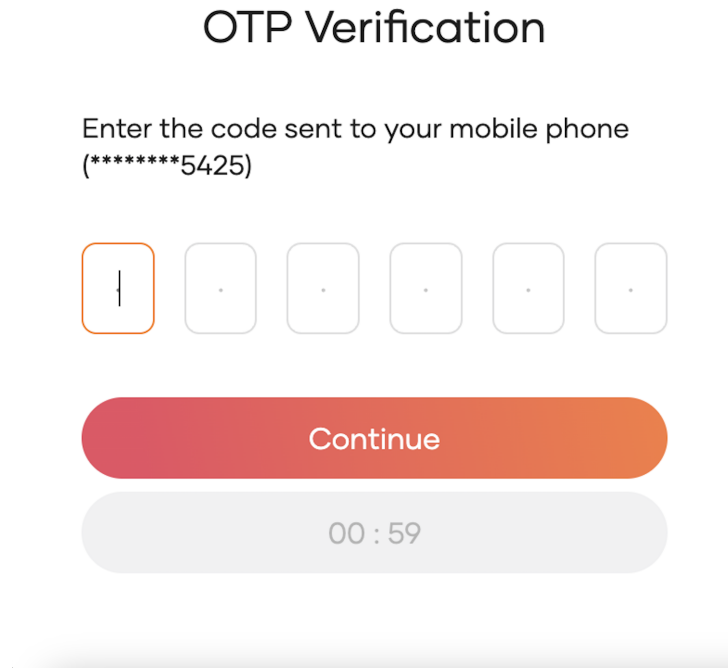


Figure 5.5: A false positive example of *Countdown Timer* dark pattern

We further expanded this dataset by adding new DP categories, and labeled each instance by three different DP researchers with peer-reviewed publications specifically on DPs. The annotation process is executed via the majority voting approach, as did in the first stage. This level of expert agreement, which is also grounded in a recent and established ontology [29] that merges and maps different DP taxonomies, and methodological transparency help mitigating the aforementioned issues. Although we do not claim to entirely address the plurality of taxonomies issue or thoroughly capture all types of DPs, our proposal offers a concrete step toward improving label reliability, comparability, and consequently, a more accurate evaluation of the tools.

To address **RQ1a**, we begin by identifying existing DP datasets and exclude those that are not publicly available or that do not include graphical UIs. We then inspect the labeling process of each dataset, the reference taxonomy used for the labeling, the label descriptions, and individual instances. Our analysis reveals several critical issues, such as unclear label descriptions, overly broad definitions of the DPs, that could negatively affect the reliability of model evaluation. Moreover, to better understand the limitations of the existing datasets, we perform both human- and machine-based analyses on them. Three DP experts assess a random set of DP instances pulled from the existing datasets in the former, and a full dataset is given to open-source MM-LLMs to observe their behavior and performance.

To avoid the issues that are also detailed in Section 5.3, we decided to proceed with the

DeceptiLens dataset [34], which is supposed to be less affected by them, as it used well-established, recent and comprehensive taxonomy [29], and was assessed by 14 different DP experts using a majority vote approach, which ensured moderate inter-annotator agreement, i.e., their Fleiss Kappa score is 0.5743. The dataset has 74 UI screenshots that are classified either as DP and non-DP. The instances that are tagged as DP are also categorized with specific DP types, such as *Confirmshaming*, *False Hierarchy*. We extended this dataset for our study by independently annotating it by three domain experts from our team, and with a 0.8672 Fleiss Kappa score. We then compared their annotations with those provided by 14 experts to evaluate labeling consistency between the two groups, i.e. inter-group agreement, and found the Cohen Kappa score of 0.6798, which means substantial agreement [119]. Then, we expanded this dataset adding new DP categories, such as *Positive or Negative Framing*, *Nagging* etc., thereby increasing the number of labels from 9 to 16, taking class distribution into account, with labeling conducted by the same three domain experts. The DP categories that are covered in our dataset are listed in Table 5.3.

Table 5.3: Dark Pattern (DP) categories in our expanded dataset. The newly added DP categories are shown as **bold**.

False Hierarchy	Disguised Ads
Bad Defaults	Trick Questions
Sneak into Basket	Hidden Costs
High Demand	Low Stock
Confirmshaming	Endorsements and Testimonials
Countdown Timer	Limited Time Messages
Immortal Accounts	Activity Messages
Pay-to-Play	Grinding
Price Comparison Prevention	Choice Overload
Ambiguous or Complex Language	Nagging
Positive or Negative Framing	

The workflow for creating the dataset is shown in Figure 5.6.

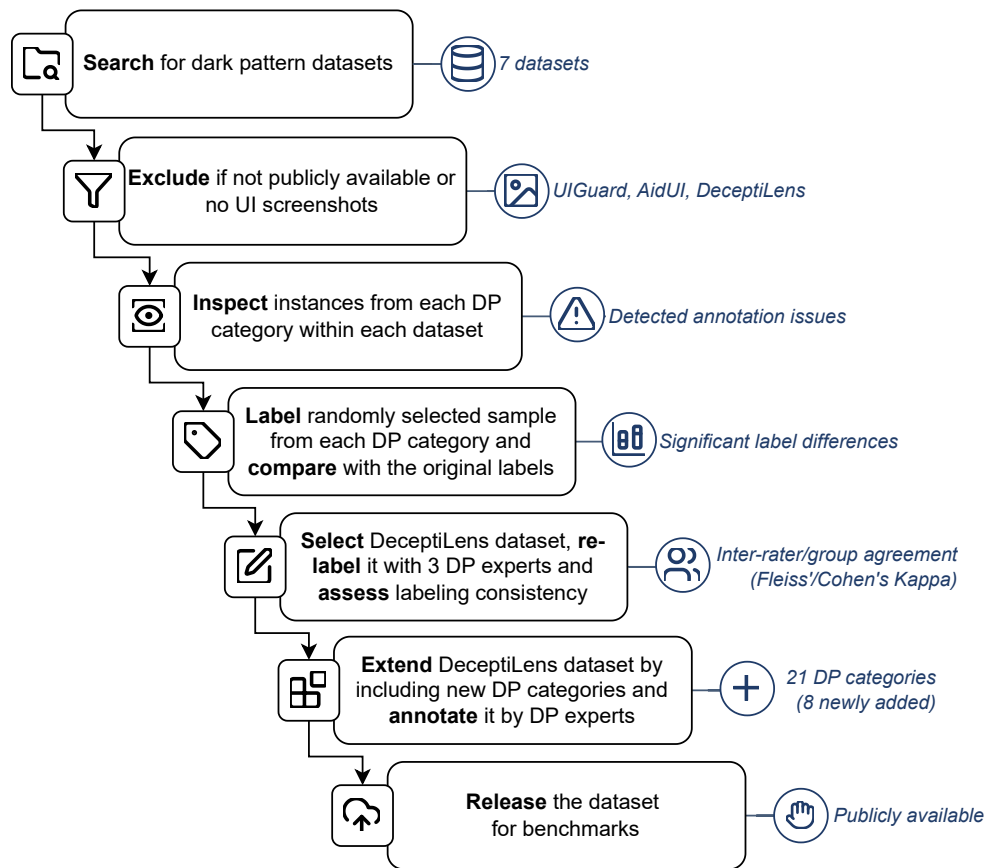


Figure 5.6: Overview of dataset creation process

In creating the data set, we take the descriptions of the ontology created by [29] as the main reference, as did [34]. Then we construct criteria on top of the descriptions listed in Table 5.4 during labeling because the descriptions are usually high-level and contain ambiguity. For example, in the description¹ of the *Immortal Account*, "difficult" and "impossible" are the keywords. The latter refers to the website or mobile application holding some data even after users have deleted their accounts [120]. However, the first adjective is ambiguous since it is not possible to assess the difficulty without having information about the entire user journey. Therefore, we decided to label the screenshots if they clearly state that they will keep the data even if the user deletes their account. We also combine some DP types into one category, such as *Grinding or Pay-to-Play* as done by [6], even though they are present separately in the ontology. The reason for this combination is similar to the *Countdown Timer or Limited Time Message* merging that was performed by [34]. Both *Grinding* and *Pay-to-Play* carry the characteristics of the "Forced Action" and "Gamification" higher-level dark pattern categories. The ontology

¹Immortal Account is defined as "they create a Roach Motel and use Obstruction to make it difficult or impossible to delete a user account once it has been created. [29]"

descriptions and our criteria for each category are given in Table 5.4.

5.4.1 Evaluation

We assess the quality of dataset annotation process by calculating inter-rater agreement in two ways: (i) using Fleiss’ Kappa score to measure the agreement among our DP experts, and (ii) Cohen’s Kappa to measure agreement between our experts’ majority voted labels and the labels of DeceptiLens experts [34], by following the inter-rater reliability framework in [119]. Moreover, we conduct an evaluation process with DP experts and open-source AI models for empirical assessment of the consequences of the issues that we outlined in Section 5.3.

Expert-based evaluation

We randomly picked a set of samples from the largest existing dataset [75], stratified by its class distribution, hereafter referred to as **UDS**, and our DP experts annotated it taking into account the recent unifying taxonomy [5] and their own expertise on the topic. The experts annotated the samples answering the following questions: i) Is there a DP in the given UI? ii) If yes, which DP type?

Model-based evaluation

We choose three open-source Multimodal Large Language Models (MM-LLM) which are **Janus Pro 7B**, **LLaVA 1.5 B**, and **LLaVA-v1.6-Mistral-7B** for the model-based evaluation stages because they are considered to be among the best-performing open-source multimodal models currently available. All open-source models were executed on an NVIDIA A100 GPU using a *temperature* to 0.7 and the *top-p* value to 0.9 with a generation cap of 512 tokens (*max_new_tokens* = 512). On the model-based evaluation side, we provided the entire **UDS** dataset to the open source MM-LLMs, but in this case, we instructed the models to use the same descriptions and labels of the dataset, not our reference taxonomy. Then we compared the results with those we obtained by using our benchmark dataset. By this experiment, we aim to investigate how differences and the issues in the DP datasets that we elaborate in Section 5.3 affect the performance and evaluation of the DP detection models.

5.5 Results

The dataset that we proposed as a benchmark for the evaluation of DP detection tools, see Section 5.4, was annotated with high agreement, i.e., 0.8672 Fleiss Kappa score, in the first round. After the first round, the three DP experts examined each case where consensus could

DP Class	Description	Criteria
Immortal Accounts	Create a Roach Motel and use Obstruction to make it difficult or impossible to delete a user account once it has been created.	Check if the screenshot indicates that the account data will be kept even after deleting the account. <i>Close</i> is accepted as the synonym of <i>Delete</i> .
Positive or Negative Framing	Uses Emotional or Sensory Manipulation and Interface Interference to visually obscure, distract, or persuade a user from important information they need to achieve their goal.	Check if the system provides only one aspect, such as positive effects, to lead the user.
Grinding or Pay-to-Play	Uses Gamification as a type of Forced Action to require repeated, often cumbersome and labor-intensive actions over time in order to obtain certain relevant functionality, OR to initially claim that aspects of a service or product are available via purchase or download, but then later charging users to actually obtain that functionality.	Check if there is an option to skip repeating the same action multiple times AND check if the option is paying money or upgrading.
Choice Overload	Subverts the user’s expectation that the choices they make should be understandable and comparable, instead providing too many options to compare or encouraging users to overlook relevant information due to the volume of choices provided.	Count the number of choices. If there are fewer than 4, do NOT label it as “Choice Overload”. If there are more than 4 options and no master control is present, label it as “Choice Overload”.
Price Comparison Prevention	Creates Barriers and uses Obstruction by excluding relevant information, limiting the ability of a user to copy/paste, or otherwise inhibiting a user from comparing prices across two or more vendors.	Check if user must switch screens,tabs or views to see other options and compare prices and features.
Nagging	Subverts the user’s expectation that they have rational control over the interaction they make with a system, instead distracting the user from a desired task the user is focusing on to induce an action or make a decision the user does not want to make by repeatedly interrupting the user during normal interaction.	Check if the interruption is unrelated to the main activity the user is engaged in. OR Check if the interruption blocks progress or covers important content AND the topic is not essential for the user’s main activity.
Ambiguous or Complex Language	Leverages Language Accessibility, using Interface Interference to make information difficult to understand by using obscure word choices and/or sentence structure, or using ambiguous and vague terms when giving information to users.	Check if the text uses complex vocabulary, legal jargon, unclear sentence structure or ambiguous wording allowing multiple interpretations.

Table 5.4: Descriptions of the newly added DP categories [29] and our criteria determined for their labeling.

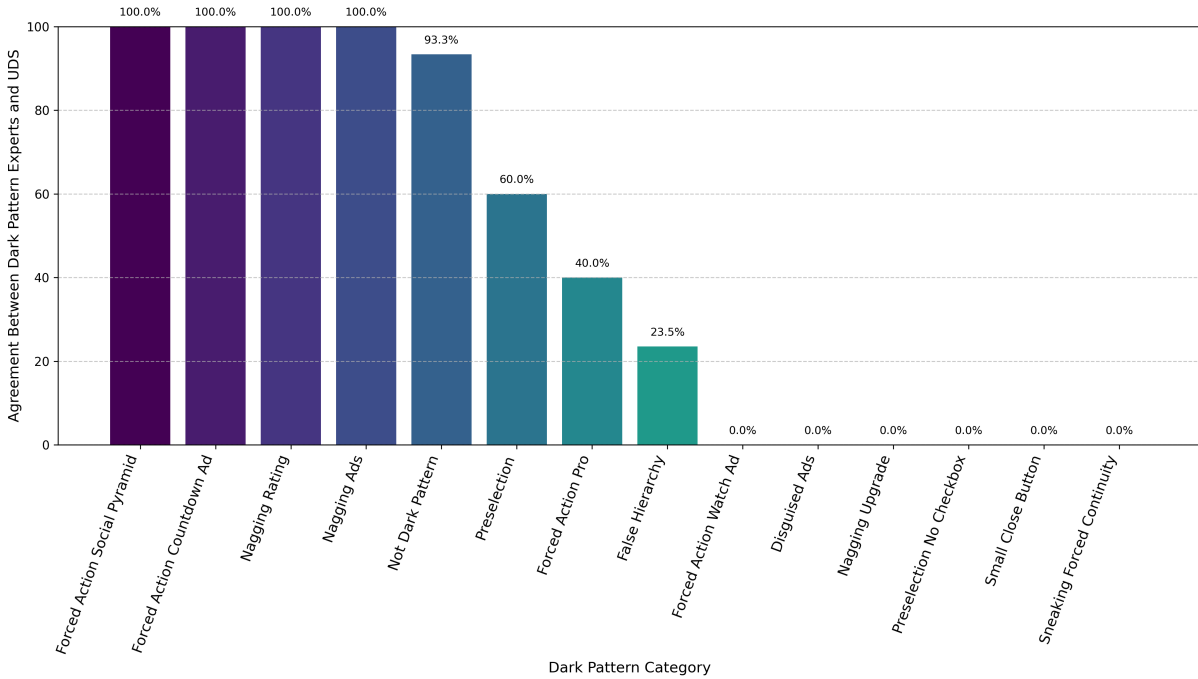


Figure 5.7: Case study: dark pattern expert opinion regarding UDS labeling.

not be reached and commonly defined a set of rules (which is available in the dataset repository) for each DP category. The practical rules were meant to increase agreement of interpretation since the DP definitions are high-level. These rules are also used in the "expert-chain-of-thought" prompting strategy, see Section 8.3.3 - Listing 1 for the implementation of the prompt. The proposed dataset is annotated with 16 dark pattern labels, which cover the categories listed in Table 5.3, and the raw screenshots are publicly available along with their annotations in the following repository: [LINK]

We also examine UDS and found that the agreement between our DP experts and the actual labels of UDS is low, particularly for the instances labeled as dark patterns, as seen in Figure 5.7. Although our experts made similar decisions on certain classes, such as "Not Dark Pattern", "Nagging" and "Social Pyramid", they evaluated that some screenshots labeled as DP, such as "Privacy Zuckering", "Small Close Button", are not DP at all. Another important observation is that while there is a 93.33% agreement between our experts and the labels of UDS for the instances labeled "Not Dark Pattern" (i.e., the negatives), there is a dramatic decrease in agreement for the instances labeled DP, namely 30.88% (i.e., the positives). When we assume the expert opinion as actual labels, overall accuracy, precision and recall scores amount to 50.00%, 30.88%, 91.30% respectively: this shows that UDS tends to over-label instances as DP.

We also demonstrated the consequences of dataset-related issues that we detailed in Section 5.3 by running the open-source MM-LLMs on the full UDS dataset, 6581 instances, with its label descriptions. The details of the agreement between the models and UDS are given in Table

5.5. While this setup resulted in relatively high accuracy and recall scores, the performance of the models decline dramatically when they are evaluated with our benchmark dataset as seen in Figure 5.8. This difference shows how the label descriptions of DP datasets can lead to divergent conclusions when analyzing the effectiveness of DP detection models.

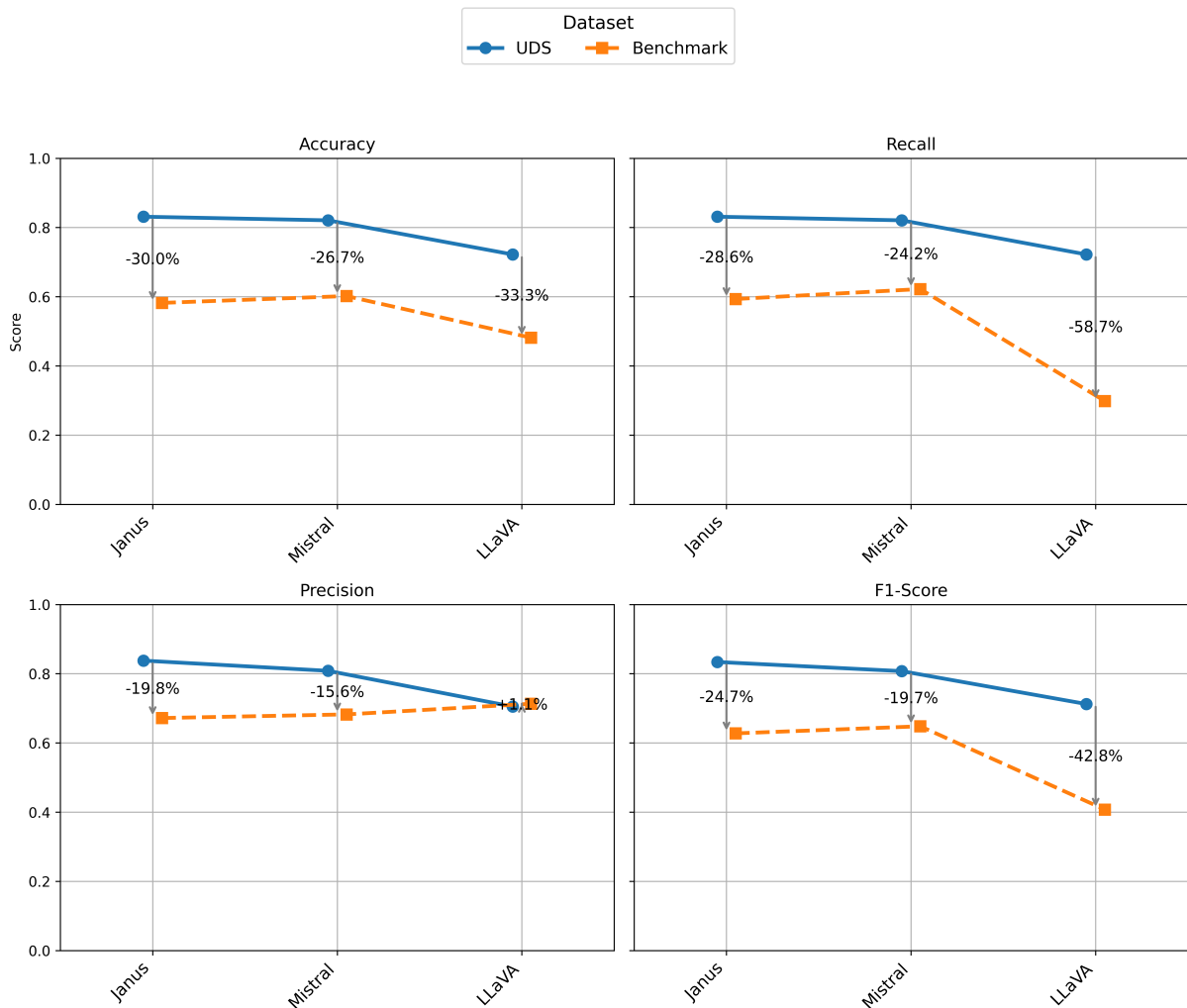


Figure 5.8: Performance of open-source models across datasets (UDS vs our benchmark dataset)

Model	Overall Agreement (%)	Non-DP Agreement (%)	DP Agreement (%)
Janus Pro 7B	83.12	87.25	70.04
LLaVa v1.6 - Mistral 7B	82.07	93.43	46.14
LlaVA v1.5 7B	72.21	84.53	33.25

Table 5.5: Agreement (%) between UI Guard labeling and open-source MM-LLMs on overall, non-dark-pattern (Non-DP), and dark-pattern (DP) samples.

5.6 Limitations

The reliance on static screenshots constitutes the primary limitation of the proposed dataset, as a single image can be insufficient for comprehensive assessment, particularly when evaluating dynamic elements or specific interaction patterns. Certain dark patterns, such as *Immortal Account*, require a dynamic understanding through website interaction, necessitating temporal data beyond a static image. A second limitation is the restricted availability of data instances for certain dark patterns, such as Intermediate Currency, since we aim the publicly reported dark pattern instances. To increase the dataset’s comprehensiveness, dark pattern category-specific collection methods can be developed for these underrepresented dark pattern types. Finally, while the annotation process successfully incorporated dark pattern experts, a limitation remains regarding the annotation count. Although each instance was evaluated by three different experts, which inherently provides a level of confidence, increasing the number of annotators per instance would further enhance the reliability and trustworthiness of the labels.

5.7 Closing Remarks

In dark pattern detection research, datasets that are sufficient in terms of size and quality are required to reliably evaluate proposed solutions. However, progress in this area is challenged by several factors, including the diversity of platforms in which dark patterns occur (e.g., web and mobile), variation in use cases (e.g., cookie consent banners, gaming applications, and e-commerce systems), and differences in data formats (e.g., textual, visual, or dynamic interaction data that may require video). Furthermore, the absence of standardized taxonomies and labeling procedures complicates cross-dataset comparison and benchmarking. In this chapter, we identify the issues of existing datasets that hinder the reliable evaluation of dark-pattern detection systems. We then introduce a new dataset annotated through a majority-voting process by domain experts and structured according to a contemporary ontology, thereby addressing current dataset gaps. We also empirically showed how the division in datasets affects dark pattern detection models’ performance results. While not intended as a final or exhaustive resource, this dataset contributes to the field by offering a more robust foundation for benchmarking and further research.

Chapter 6

Dark Pattern Data Augmentation and Fine-tuning

We contribute to the research on structured and labeled dark pattern datasets by developing a framework for dark pattern text data augmentation. The framework utilizes state-of-the-art open-source language models within a multi-agent architecture composed of generator and controller models, which are built to maintain semantic and structural similarity with the original text data during the augmentation process. After building the augmented dataset, we fine-tune language models and evaluate the dataset through expert annotations and by comparing the performance of fine-tuned models.

*This chapter contributes **Research Question 2**: "What considerations and methods should be taken into account when building labeled and structured dark pattern datasets?" and its content is primarily based on the following paper:*

- *Kocyigit, E., Liga, D., & Lenzini, G. (2024, November). **Augmenting Dark Patterns Text Data by Leveraging Large Language Models: A Multi-agent Framework and Parameter-Efficient Fine-Tuning**. In *International Workshop on Causality, Agents and Large Models* (pp. 65-87). Cham: Springer Nature Switzerland.*

6.1 Introduction

Dark patterns are online design strategies that are manipulating, deceiving, or coercing internet users to make decisions against their interest, e. g. options, which contain emotional manipula-

tion i. e., *confirmshaming*, can direct users into the expensive plan in purchasing decisions [121], and some of them can be issued illegal [108]¹. Even though taxonomy, categorizations, and descriptive analysis, which are crucial to identifying dark patterns, are well studied by scholars [29, 18, 124, 3] and associations [105, 6, 7, 8], existing structured and labeled dark pattern datasets are limited.

Although existing data sets are useful for automatic dark pattern detection or a further analysis in the literature, they do not fully meet the need for processed and labeled text datasets because most of them consist of video recordings or screenshots [50, 125, 57, 35, 72] or simpler dark patterns with string matching, not more complex dark patterns that require understanding the semantics of the text, such as “*confirmshaming*” [72]. The dark pattern dataset obtained by Mathur *et al.*’s landmark study after crawling of 11K e-commerce websites [19], and the dataset obtained by Yada *et al.* by adding non-dark pattern text samples to the aforementioned study, are structured and labeled text dataset that involves “*confirmshaming, trick questions*” etc. However, particularly considering that artificial intelligence-based detection models’ need for training, they need to be improved in terms of quantity and diversity [126]. Collecting more data is one of the ways of data expansion, but it is laborious due to the fact that each example should be assessed by experts and there is not established databases such as “phishtank” which is an publicly available phishing database. On the other hand, data augmentation techniques offer scalable solutions that enrich existing datasets by generating synthetic variations of existing examples.

Language models have made significant progress, especially with the introduction of pre-trained models, which consist of a very large number of parameters, namely Large Language Models (LLMs) or general-purpose task solvers, after statistical or neural network-based models [127]. There are many closed-source LLMs, e. g. GPT-4, and publicly available LLMs of different sizes, e. g. T5, Mistral etc. [128]. Scaling up the language models towards LLMs brings abilities that are not present in small models, i. e., emergent abilities [129], and causes successful performance of these models in generating text, language understanding, logical and arithmetic reasoning skills, contextual comprehension and common natural language processing tasks, e. g. translation, question-answering, etc. [130]. Llama-3 was introduced as the most capable openly available LLM at 8B and 70B scale [131]. To solve the problem of scarcity of structured and labeled dark pattern text data, we propose a multiple agent-based text data augmentation method, which employs open-source LLMs, specifically Llama-3, and assess the augmentation

¹It should be noted that while some LLMs are trained or fine-tuned specifically on legal data [122, 123], this does not eliminate the need to create ad hoc classifiers specifically “designed to categorize data that is (or might be) legally relevant. Dark patterns, for instance, may fall into this category of legally relevant data.

performance by state-of-the-art language models, e.g. GPT-4, and human evaluators. In this regard, we formulate our **research questions** as below:

1. How can state-of-the-art open source² LLM be utilized to augment dark pattern text data?
 - (a) How can multiple agents, LLMs, preserve the intent of original dark pattern text data during data augmentation?
 - (b) Which strategy can be applied to preserve the diversity of original dark pattern text data in LLM-augmented data?
 - (c) How can the augmented data increase the performance of the dark pattern detection model?

Our **contributions** are listed as below:

1. proposing a LLM-based dark pattern text data augmentation method.
2. creating a new dark pattern text dataset by augmenting the original dark pattern text data.
3. increasing the detection capability of a language model by fine-tuning it with the augmented data.

6.2 Background

6.2.1 Dark Patterns

Deceptive design patterns, i. e., dark patterns, can be crafted in various forms by utilizing different data types, e.g. visual, linguistic and process-based. For instance, cookie consent process, which is one of the most common dark pattern use cases [69], may have an asymmetric button design leading users to a specific choice, i. e., full consent grant, may have deceptive and confusing text data on the banner, or may restrict users to force them to take a certain action, e.g. accepting all cookies. Both industry and academia have been working on dark patterns since 2010 [40], and one of the most up-to-date ontology of dark patterns study presenting 64 categories, such as “obstruction”, “sneaking”, “interface interference”, “forced action”, and “social engineering”, by mapping different published taxonomies [29]. Some dark pattern categories are certain domain specific, e.g. “*Pay-to-Play*” in online video games, or particular data specific, e.g. “*Disguised Ad*” by utilizing graphics.

²We acknowledge that many prefer to describe Meta’s models as open-weight rather than open-source, as only the final weights are made available, not the complete training process or data. For simplicity, we will use the terms open-source and open-weight interchangeably here.

6.2.2 Source Dataset

In our study, we focus on text data, i. e., linguistic dark patterns such as “confirmshaming”, “trick questions” etc., and used Mathur et al.’s dataset [19], henceforth referred to as *DS1*, and Yada et al.’s non-dark pattern examples [94], henceforth referred to as *DS2*, as source data of the data augmentation since they contain text data that have the characteristics of the dark patterns and they are labelled with some of the well-known dark pattern categories, such as “confirmshaming”, “obstruction”, “trick questions”, etc.

Table 6.1: Dark pattern categories and subcategories in *DS1* [19]

Category	Subcategory
Forced Action	Forced Enrolment
Misdirection	Confirmshaming Pressured Selling Trick Questions Visual Interference
Obstruction	Hard to Cancel
Scarcity	High-demand Message Low-stock Message
Sneaking	Hidden Costs Hidden Subscription Sneak into Basket
Social Proof	Activity Notification Testimonials of Uncertain Origin
Urgency	Countdown Timer Limited-time Message

Dataset characteristics

DS1 has 1818 instances labeled with 7 dark pattern categories and 15 subcategories which are listed in Table 6.1. The distribution of the data set categories is not balanced as seen in Figure 6.1. For example, while “scarcity” is dominant with 679 instances, “forced action” involves only 6 instances. The dark pattern instances of *DS2* were sourced from the *DS1*; therefore, we only include the 1178 negative samples, i. e., non-dark pattern examples, from the *DS2*.

6.3 Methodology

6.3.1 Pre-processing

We perform the pre-processing steps, see Algorithm 2, to clean the dataset and to have representative data before applying the data augmentation. This preprocessing is critical because the success of the data augmentation is directly linked to the quality of the original data. In partic-

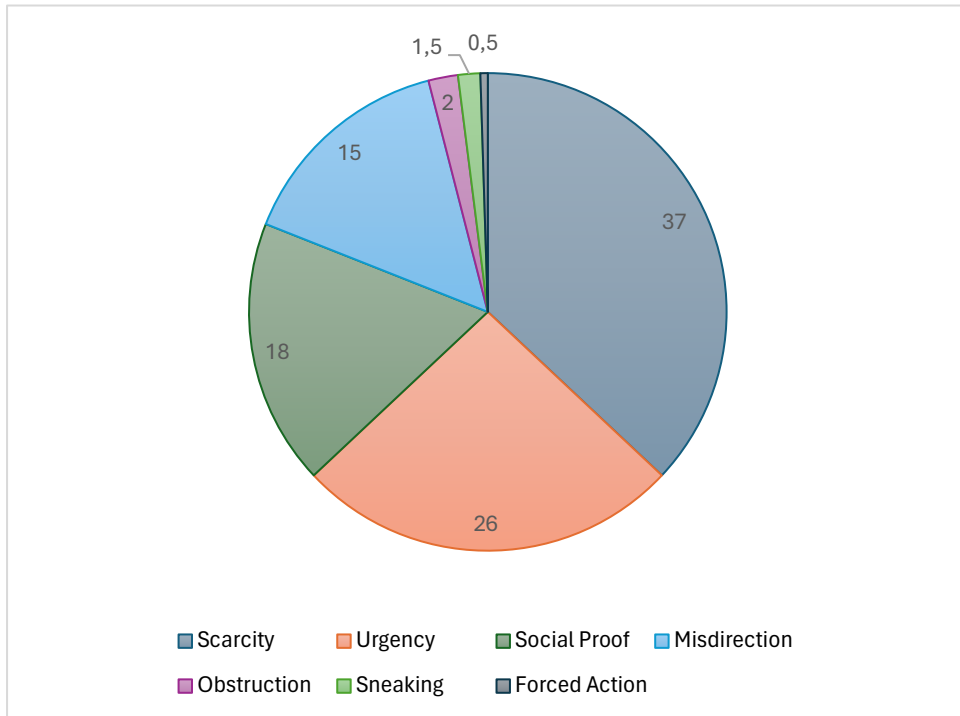


Figure 6.1: Distribution of dark pattern categories of DS1 (%)

ular, *DS1* contains blank text, duplicate entries, or text that does not reflect dark patterns in the absence of visual information. Furthermore, some text instances have different dark pattern characteristics, even though their label is the same. For example, while complexity, confusing language and the use of double negatives are properties of the *trick questions*, these features are not present in *confirmshaming* even though both belong to the same high-level label, *misdirection*. After applying the preprocessing, the total number of instances in the source data is reduced from 2996 to 2287, with dark patterns and non-dark patterns comprising %48.49 and %51.51 of the distribution, respectively.

Algorithm 2 Data Pre-processing before applying data augmentation

Require: *DS1*, *DS2*

Ensure: Cleaned and representative datasets

- 1: Load datasets *DS1* and *DS2*
 - 2: Remove missing values and duplicates from datasets
 - 3: Remove entries where text is insufficient or additional visual data is needed
 - 4: Check categories for necessity of using lower level categories
 - 5: **for** each category in dataset **do**
 - 6: **if** category is high level, e.g. *misdirection*, **then**
 - 7: Utilize its subcategory, e.g. *confirmshaming* or *trick questions*,
 - 8: **end if**
 - 9: **end for**
-

6.3.2 Augmentation

Let D be the source dataset, where each data instance d_i in D consists of text t and a category c . G represent the generator model and C the controller model, and both of which are distinctively instructed LLMs, specifically Llama-3. The data augmentation process is controlled by a parameter k , which determines the maximum possible number of augmented instances for each original text entry. The augmented dataset D' can then be defined as:

$$t'_{i,j} = G(t_i) \quad (\text{augmentation of instance } i \text{ by } G) \text{ for } j \in \{1, 2, \dots, k\}$$

$$t''_{i,j} = C(t'_{i,j}) \quad (\text{validation of augmented instance by Controller})$$

$$D' = \{(t''_{i,j}, c_i) \mid (t_i, c_i) \in D, j \in \{1, 2, \dots, k\}\}$$

Here, $G(t)$ generates a new text based on t , and C validates t' to align with the category c , ensuring that the augmented text preserves the same dark pattern category and its characteristics. The size of D' may vary as not all augmentations produced by the generator pass the validation step.

Generation

We use the Llama-3 8B version as a generator model which is one of the leader open source LLMs [132]. While the original dark pattern text, i.e., t and its category, i.e., c , are given to the generator model, data augmentation task description, dark pattern category descriptions and examples are also added to the prompt design. Prompting techniques, such as using clear and precise language, providing examples, framing the prompt as task, context, query, and syntactic indicators, are applied to increase the quality of augmented dark patterns. Moreover, we also dictate that the generator model should take into account the length of t into consideration to maintain not only semantic but also visual similarity. Finally, some strategies that the model should follow within the task are specified in order to have more diversity in the generated texts and to avoid potential dark pattern category shift. For instance, the following rule is added to the prompt to avoid the potential scarcity to urgency shift during the augmentation: *Augmented version of 'Scarcity' should not have time information or temporal expressions such as "hours", "day", "late", "soon" etc..*

Controlling

We also use Llama-3 8B as a controller model to validate the generated data. The controller model is designed to validate whether the augmented data preserve its original dark pattern category after the augmentation. If the controller model agrees that the dark pattern category of the augmented text is the same as the source data, then it will be included in the final list of augmented texts.

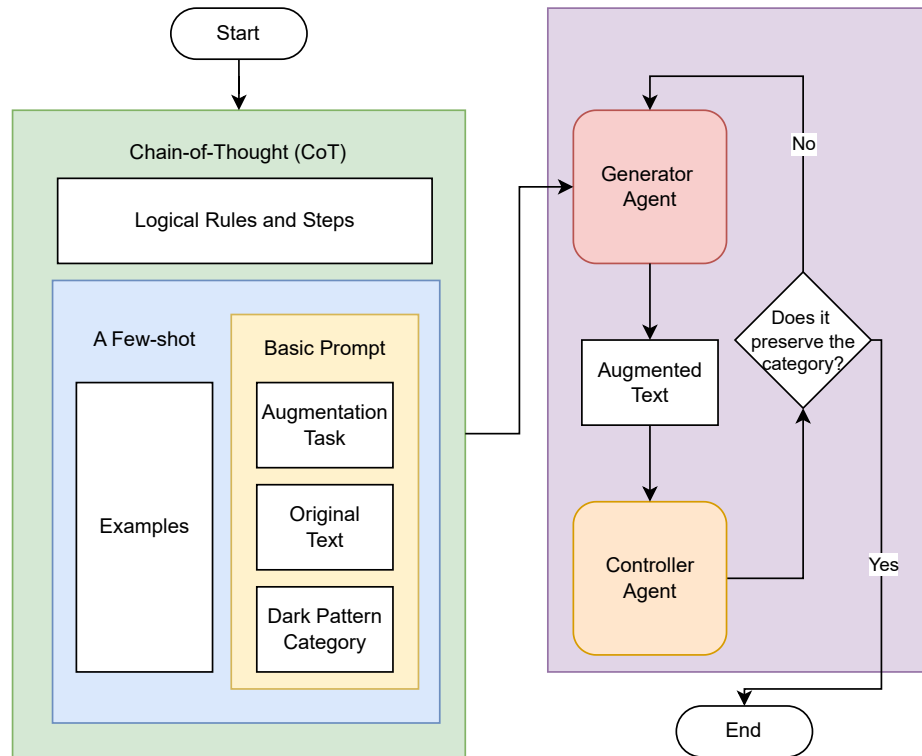


Figure 6.2: Flow of the dark pattern text data augmentation by multiple agents

Prompting

We apply the following prompting techniques during both the generation and controlling stage:

1. Zero-Shot Prompting: Crafting a prompt without any prior examples.
2. Few-Shot Prompting: Adding a small number of examples to the prompt to exhibit the relationship between the category and the text.
3. Chain-of-Thought (CoT) Prompting: Breaking down the task into steps or reasoning about the process in the prompt.
 - (a) determine the context

- (b) find the reason that the text is classified with a dark pattern category
- (c) generate a new sentence keeping the same category

In addition to the aforementioned prompting techniques, we follow several strategies, such as *describing precise instructions, stating specific constraints, providing context and controlling output verbosity* [133] to increase the diversity of the augmented text, to prevent a potential dark pattern category shift, and to restrict the frame of the augmentation. An example prompt that employs CoT principles is shown in Figure 6.3. Finally, the CoT prompt of the controller is reported in Figure 6.4.

```
prompt_cot = dedent("""
I will show you a random text from a dataset containing potential dark patterns divided into categories.
The available categories are: Social Proof, Confirmshaming, Trick Questions, Urgency, Scarcity, Obstruction, Forced Action.
The rest is not a dark pattern (i.e. the category "Not Dark Pattern").
Your goal is to augment this dataset, generating {number_of_augmentations_per_instance} sentences which are very similar to
the one I will show you, or at least reflect the given category.

To generate these {number_of_augmentations_per_instance} augmentations, you can add/remove/substitute words, or you can just
create a new sentence similar to the original.
The more augmentation techniques you employ, the better.
Try to consider the original length of the text when you augment it (whenever possible, DO NOT make the length of the
augmented text too distant from the original length).
But the most important thing is that these {number_of_augmentations_per_instance} augmented text MUST PRESERVE the original
category from the original sentence.
THIS IS SUPER IMPORTANT! THE AUGMENTED SENTENCES MUST REFLECT THE ORIGINAL CATEGORY!
Try to reproduce the communicative intentions of the original text whenever you can, but in case the text is too limited or
difficult to augment, you can just invent a totally different text, as long as it fits well the original category!

For example, if the category is 'Not Dark Pattern', then make sure that the augmented text preserve this without falling into
any other dark patterns category.

Step-by-Step Thought Process:

1. First of all, determine the context of the text I will give you
2. Secondly, find out the reason why the given text is classified in its given category
3. Finally, generate {number_of_augmentations_per_instance} new texts which preserves
the same dark pattern category.
Please, return the new augmented texts as a list, so one can digest it in python as
JSON (each element of the list should be a single augmented text).
While preserving the original manipulative purpose, remember that manipulations are often "hidden" in the language and not
necessarily explicit.
Therefore, for example: the obstruction category do not say "To cancel the subscription you have to fill very annoying and
lengthy forms and call us only in the weekends", instead it says "To cancel the subscription, please fill the dedicated forms
and call our dedicated phone number".

Your goal is to apply the following 3-step process to the text I will show you and generate
{number_of_augmentations_per_instance} augmented sentences.
Remember to follow the following step-by-step process to ensure clarity and accuracy in your responses.
{cot_process}
""")
```

Figure 6.3: The prompt to generate augmented dark pattern text data (CoT part zoomed).

6.3.3 Evaluation

The main objective of data augmentation is to increase the number of dark pattern examples while preserving their dark pattern category. We perform two evaluation strategies, which are language model-based and human-based.

```

cot_instructions = dedent(f"""
    1. First of all, describe the dark pattern category "{the_pattern_name}"
    2. Secondly, find out whether the {number_of_augmentations_per_instance}
    texts in the list correctly belong to the category "{the_pattern_name}"
    3. Finally, write "OUTCOME:True" if ALL the
    {number_of_augmentations_per_instance} sentences belong to the category "
    {the_pattern_name}", otherwise write "OUTCOME:False" if one or more
    sentences do not clearly fall into the category "{the_pattern_name}"
    4. If you see less than {number_of_augmentations_per_instance} sentences,
    then return "OUTCOME:False" too.
""")

```

Figure 6.4: The CoT prompt for the controller model.

Language Model Evaluation

E represents the evaluator. It can be one LLM or multiple LLMs that decide together. We consider accuracy, precision, recall and F-1 score, harmonic combination of precision and recall, metrics, which are detailed below:

$$\text{Accuracy}_{\text{category}} = \frac{|\{d_i \in D' \mid C(d_i) = E(t_i)\}|}{|D'|} \quad (6.1)$$

$$\text{Precision}_{\text{category}} = \frac{|\{d_i \in D' \mid C(d_i) = c \wedge E(t_i) = c\}|}{|\{d_i \in D' \mid E(t_i) = c\}|} \quad (6.2)$$

$$\text{Recall}_{\text{category}} = \frac{|\{d_i \in D' \mid C(d_i) = c \wedge E(t_i) = c\}|}{|\{d_i \in D' \mid C(d_i) = c\}|} \quad (6.3)$$

$$\text{F1-score}_{\text{category}} = 2 \cdot \frac{\text{Precision}_{\text{category}} \cdot \text{Recall}_{\text{category}}}{\text{Precision}_{\text{category}} + \text{Recall}_{\text{category}}} \quad (6.4)$$

Diversity

The augmented examples can be utilized to fine-tune the language models aiming to increase the detection capability of them. In this regard, the diversity of the dark pattern samples is crucial to prevent overfitting of the fine-tuned model into a small set of frequent phrases that may be the output of text data augmentation. We measure lexical diversity of the augmented data in addition to the above metrics and control if the diversity of the data significantly decreased.

$$\text{Diversity}_{\text{lexical}} = \frac{|\text{unique words}(D)|}{|\text{total words}(D)|} \quad (6.5)$$

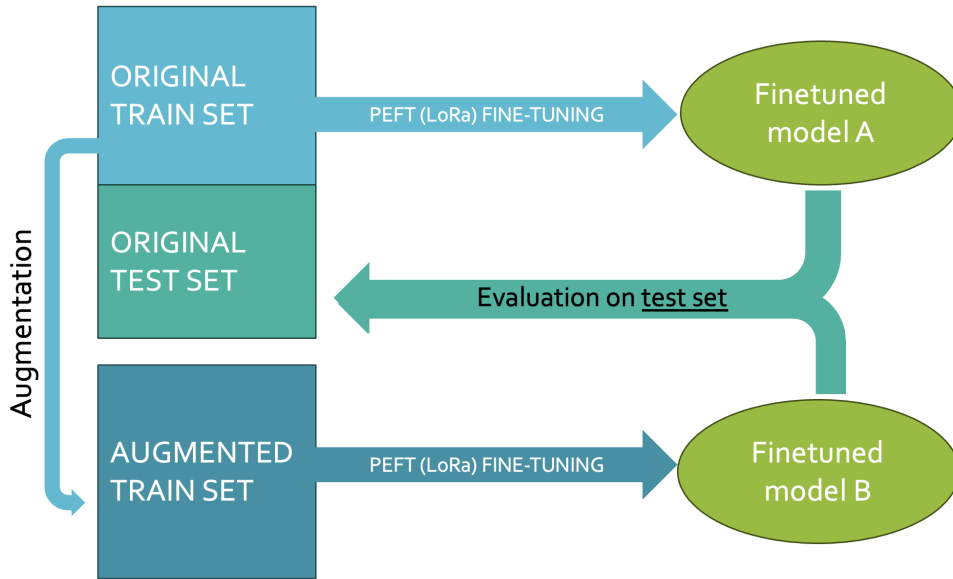


Figure 6.5: Pipeline of the PEFT approach. Model A is the model fine-tuned on the original training set, while Model B is one of the 3 models fine-tuned on the augmented datasets.

Human Evaluation

The amount of augmented data will be too time consuming for human evaluators. Therefore, we randomly select a sample considering the class distribution of the entire dataset. Then, human evaluators independently assess the samples and annotate each of the dark pattern categories or as “not dark pattern”. Then, we calculate the accuracy, precision, recall and f-1 score according to common decision of human evaluators.

Parameter-Efficient Fine-tuning

To demonstrate the usefulness of the data augmentation, we fine-tuned an open-weight LLM, namely Llama3.2-3B (a recent version of Llama with 3 billion parameters). We fine-tuned the same LLM four times to validate our approach. First, we fine-tuned Llama3.2-3B on the original training set (considered the baseline to be surpassed). Then, we fine-tuned the model on three augmented datasets generated through the aforementioned augmentation scenarios: zero-shot, few-shot, and CoT. For all four cases, the test set remained the original test set, consisting of 25% of the dataset we used (DS1 + DS2). The overall process is illustrated in Figure 6.5.

Regarding the fine-tuning methodology, we employed the LoRa approach [134], which allows fine-tuning a minimal subset of trainable parameters while keeping the original weights frozen. The LoRa hyperparameters were set as follows: the r value was fixed at 8, and the α value was kept at the standard value of 16. In addition, a dropout rate of 0.1 was applied. The fine-tuning process was conducted on a consumer-grade M1 GPU.

6.4 Results

We first checked the detection performance of two LLMs, GPT-4 and Llama3, with the source data before employing them to evaluate the performance of augmentation, which is detailed in Section 6.4.1. Then, we carried out dark pattern text data augmentation by multiple-agents, i. e., generator and controller LLMs. We applied three prompting techniques while augmenting the data: **zero-shot**, *few-shot* and *chain-of-thought*. Their results are presented, respectively, in Sections 6.4.2, 6.4.3, and 6.4.4. The diversity of the texts augmented with all the techniques, i. e., the average lexical diversity score is 0.3366, is close to the diversity of the source data, i. e., 0.3491. We also assessed the agreement among the GPT-4 and Llama-3 models and present it in Section 6.4.5. Finally, we randomly select a sample from the augmented dataset considering their class distribution and annotated them by human evaluators as part of the qualitative analysis of augmentation. The human evaluation results are shown in Section 6.4.6.

6.4.1 Detection Performance of GPT-4 and Llama3 on Source Data

We first used the GPT-4 and Llama3-8B models to classify the source data, aiming to assess their reliability for evaluating the augmented data. The corresponding classification prompts are provided in Figure 6.6.

```

prompt_binary = dedent(f"""
I will show you a random text from a dataset containing potential
dark patterns.
There are different types of dark patterns divided in 7
sub-categories: Social Proof, Confirmshaming, Trick Questions,
Urgency, Scarcity, Obstruction, Forced Action.
The rest is not a dark pattern (i.e. the category "Not Dark
Pattern").
Your goal is to classify the text that I will show you as either
"Dark Pattern" (if you see one of the above-mentioned 7
categories) or "Not Dark Pattern".
{categories_addition_binary}
{giving_some_shots_binary}

After I show you the text, you should answer only "Dark Pattern"
or "Not Dark Pattern". Don't say anything else, just one of these
two strings. You are performing a binary classification. DO NOT
say "Social Proof", "Confirmshaming", "Trick Questions",
"Urgency", "Scarcity", "Obstruction", or "Forced Action". These 7
categories are all "Dark Pattern".

""")

prompt_multi = dedent(f"""
I will show you a random text from a dataset containing potential
dark patterns. There are different types of dark patterns divided
into 7 sub-categories: Social Proof, Confirmshaming, Trick
Questions, Urgency, Scarcity, Obstruction, and Forced Action.

Your goal is to classify the text I show you into one of these 7
categories.
{categories_addition_multi}
{reduce_confusion_multi}
{giving_some_shots_multi}

After I show you the text answer by stating one of the
above-mentioned categories. Don't say anything else, just the
category name.

""")

```

Figure 6.6: Classification prompts for the binary and multiclass classifications.

As shown in the classification results of GPT-4 and Llama3-8B, in Tables 6.2 and 6.3, respectively, in the appendix, there is a significant difference between the models. GPT-4's performance is remarkably good while Llama3 is not able to detect some dark pattern categories, such as *confirmshaming*, *obstruction* and *trick questions*. We also checked the models' agreed decisions to calculate the same metrics. This approach indicated more reliable results, as shown in Table 6.4.

Table 6.2: Classification report of the GPT-4 for the source data

Class	Precision	Recall	F1-Score	Support
Not Dark Pattern	0.99	1.00	0.99	1153
Scarcity	1.00	0.86	0.92	418
Social Proof	1.00	0.98	0.99	312
Urgency	0.77	0.99	0.87	210
Confirmshaming	1.00	0.98	0.99	129
Obstruction	1.00	0.89	0.94	27
Trick Questions	0.88	0.78	0.82	9
Forced Action	0.80	1.00	0.89	4
Accuracy			0.97	2262
Weighted Avg	0.97	0.97	0.97	2262

Table 6.3: Classification report of the Llama-3 for the source data

Class	Precision	Recall	F1-Score	Support
Not Dark Pattern	0.76	1.00	0.86	1153
Scarcity	0.97	0.86	0.91	418
Social Proof	1.00	0.60	0.75	312
Urgency	0.87	0.62	0.72	210
Confirmshaming	0.94	0.13	0.23	129
Obstruction	1.00	0.26	0.41	27
Trick Questions	1.00	0.11	0.20	9
Forced Action	0.50	0.75	0.60	4
Accuracy			0.82	2262
Weighted Avg	0.85	0.82	0.80	2262

Table 6.4: Classification report of agreed decisions of the both models for the source data

Class	Precision	Recall	F1-Score	Support
Not Dark Pattern	0.99	1.00	0.99	1153
Scarcity	1.00	0.95	0.97	339
Social Proof	1.00	0.99	0.99	188
Urgency	0.90	0.98	0.94	131
Confirmshaming	1.00	0.89	0.94	19
Obstruction	1.00	0.70	0.82	10
Trick Questions	1.00	0.33	0.50	3
Forced Action	0.75	1.00	0.86	3
Accuracy			0.98	1846
Weighted Avg	0.99	0.98	0.98	1846

6.4.2 Zero-shot Prompting

Classification of dark pattern and non-dark pattern examples is quite successful, as shown in Table E.1, in Appendix. However, according to the multi-class classification of dark pattern categories, see Table E.2 in the Appendix, even though zero-shot prompting technique-based data augmentation presents representative examples for certain categories, e.g. *social proof*, overall accuracy score is significantly lower than binary classification.

6.4.3 Few-shot Prompting

The evaluation of the generator model that employs *few-shot prompting* strategy indicates slightly better performance scores as shown in Table E.3 and Table E.4, in the appendix. The diversity of the augmented data and the original data is consistent, the lexical diversity scores of them are 0.43 and 0.44, respectively.

6.4.4 Chain of Thought (CoT) Prompting

We also observed successful binary classification of dark patterns and non-dark patterns in CoT-based data augmentation, as shown in Table E.5, in the Appendix. The performance of the dark pattern category preservation is similar to the previous techniques, as shown in Table E.6 in the Appendix, in terms of precision, recall and accuracy. However, the total number of augmented dark pattern data, i.e., , the text data generated by validated controller model, is more than previous ones via the CoT technique.

6.4.5 Agreement Among GPT-4 and Llama-3 While Classifying the Augmented Data

We also compare the agreement of GPT-4 and Llama-3 on the augmented data for each prompting technique. The results, as shown in Figures 6.7 and 6.8, indicate that the number of successfully augmented data points increases in the order of zero-shot, few-shot, and CoT. This proves that the best version is CoT-based data augmentation considering all metrics, e.g. quantity, diversity, and representativeness.

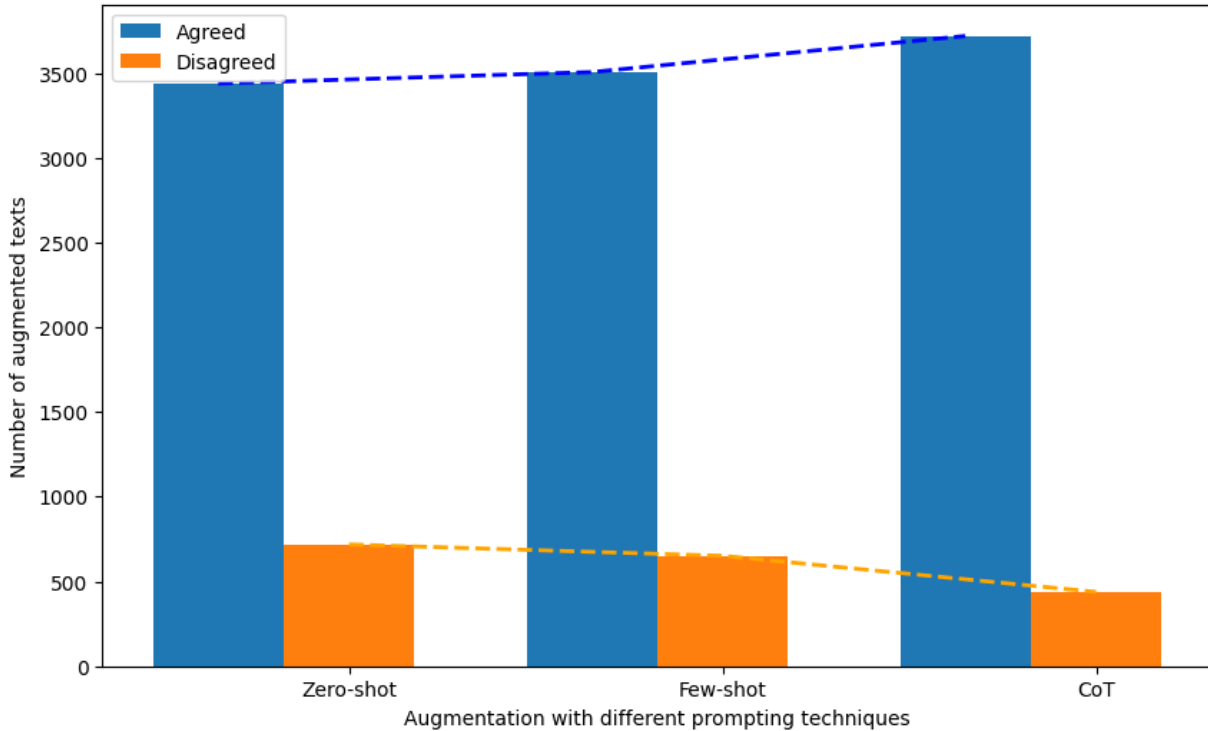


Figure 6.7: Number of agreed and disagreed augmented text data according to the binary classification of the evaluator models for each prompting technique

6.4.6 Human Evaluation of Augmented Dark Pattern Data

We performed a human evaluation by two experts. 300 augmented texts were randomly sampled from the augmented dataset, which was built by the CoT technique. Human evaluators had the same opinion on 272 examples. We calculated the classification metrics considering these examples, as we did in the previous sections, and we obtained %98 accuracy, precision, recall and f-1 score. On the other hand, we observed that while some of the dark pattern categories were successfully identified by the evaluators, incorrect classifications tended to cluster in certain categories. For instance, *social proof* and *obstruction* samples were classified with %100 accuracy while half of the *trick questions* samples were missed by the human evaluators. Moreover, as shown in Figure 6.10, %28 of the *scarcity* samples were identified as *urgency*, which indicates

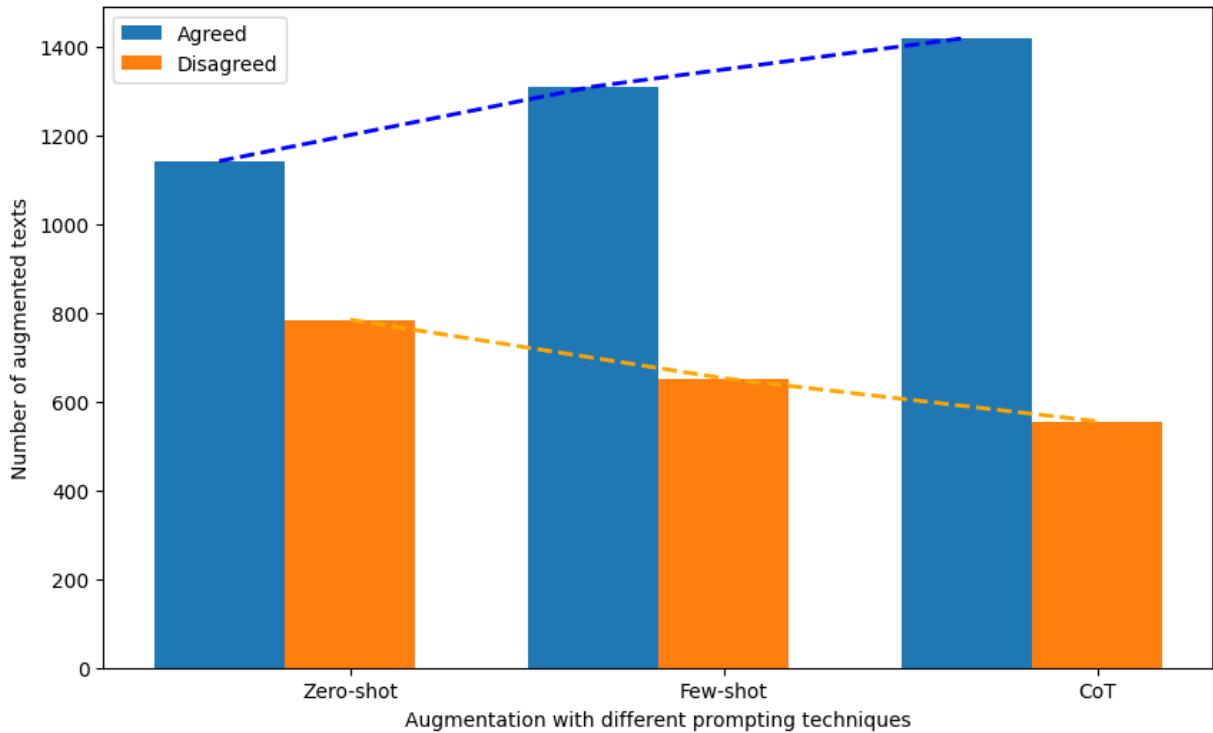


Figure 6.8: Number of agreed and disagreed augmented text data according to the multiple classification of the evaluator models for each prompting technique

that it is difficult to distinguish these two classes even for humans. When we inspected more, we observed that some of the examples can have multiple labels. For instance, an augmented text, “*Fewer options left, act now!*” sentence’s first three words reflect *scarcity* while the last part *urgency* with temporal expression.

6.4.7 Parameter-Efficient Fine-Tuning Low-Rank Adaptation Over the Augmented Data

The results of the fine-tuning are presented in Figure 6.9. Overall, the first column (mean F1-score of the model fine-tuned on the original training set) is consistently outperformed by the other three columns, which represent the models fine-tuned on the three distinct augmented training sets. This trend is observed in both the binary and the multiclass scenarios, although the binary classification tasks exhibit relatively smaller improvements, likely because binary classification is inherently simpler than multiclass classification. Unexpectedly, we observed a peak in performance with the few-shot augmentation, which outperforms the CoT-based augmentation. This counterintuitive result is currently under investigation and warrants further exploration.

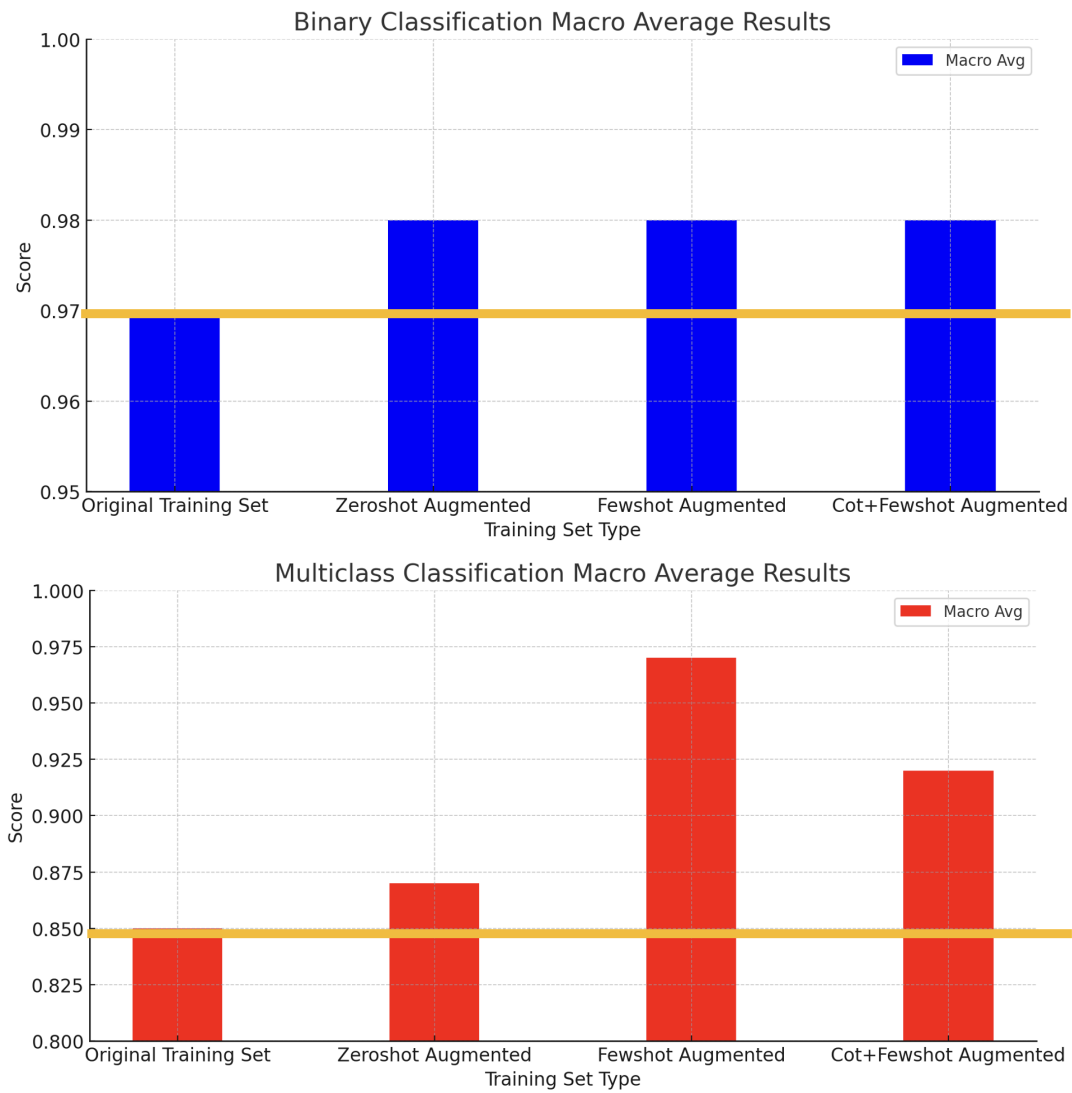


Figure 6.9: Mean F1-score results of the PEFT using the various augmentations, considering both the binary (top) and the multiclass (bottom) scenarios.

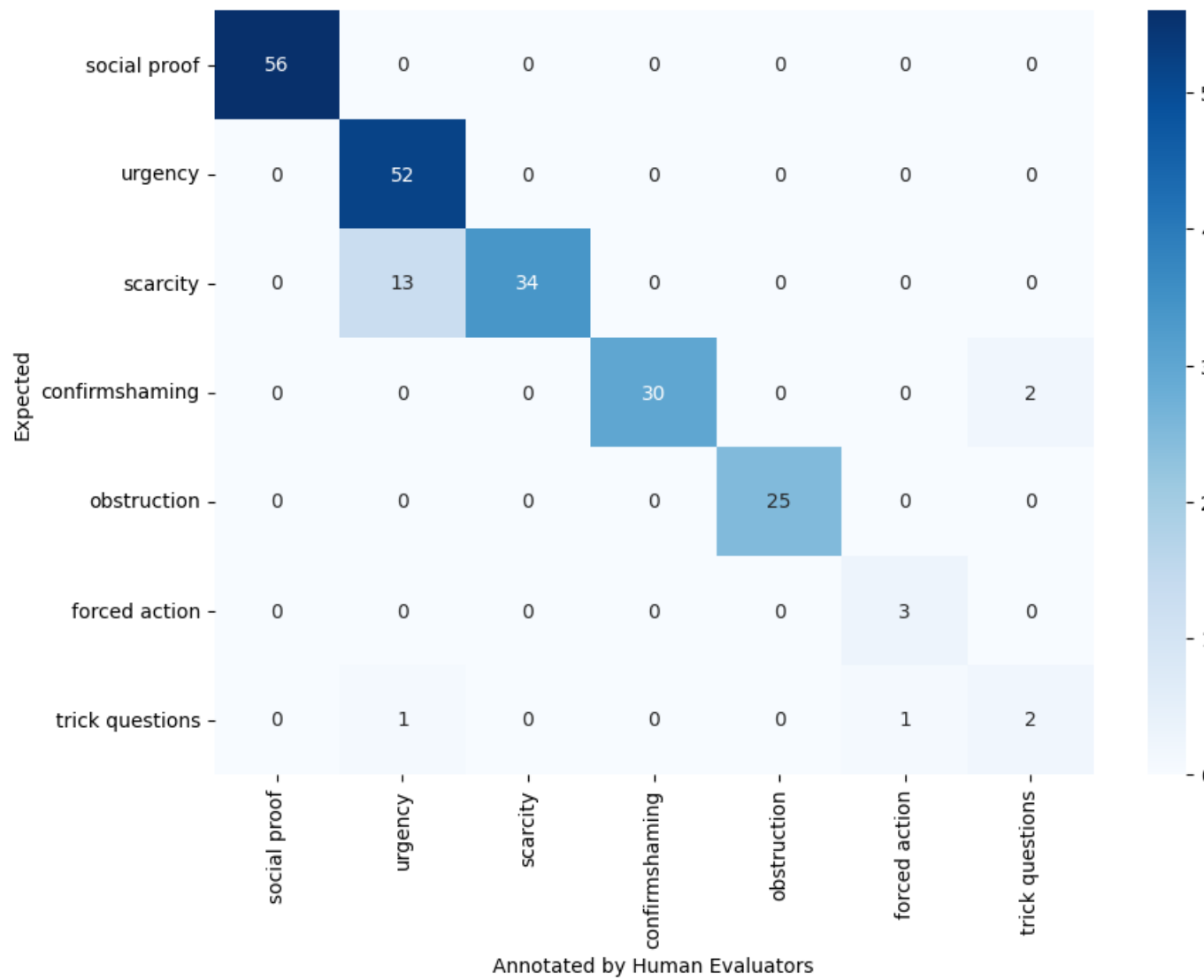


Figure 6.10: Confusion matrix for human evaluators' annotations on augmented data

6.5 Discussion

Applying traditional data augmentation techniques to dark pattern data alone is not sufficient because the text may lose its dark pattern feature as a result of applying methods such as replacing synonyms or deleting random words. Dark pattern data augmentation process should consider the context of the text and characteristics of the dark patterns because while the context is e-commerce and the category is *trick questions*, other one's context can be newsletter pop-up message and the category can be *confirmshaming*. Therefore, we prefer state-of-art advanced language models which are able to understand sophisticated semantics, e. g. irony, manipulation, etc., and the context of the text successfully. We prefer to utilize open-source model, Llama-3, to generate dark pattern text data. However, other leading models, such as GPT-4o, Gemini, Claude etc., can be utilized. Furthermore, other methods, such as fine-tuning, can be performed and data augmentation performance scores can be enhanced as in [135].

“Garbage in garbage out” is popular quote in machine learning to highlight the importance of the training data on the model performance. It can be also said for the data augmentation. The limitations of the source data will reflect on the augmentation. The source dataset of this study has also a few issues. Firstly, it has several dark pattern categories that require additional information to be perceived as dark patterns. For instance, *visual interference* needs to be assessed with graphical elements on the web page, i. e., text data is not sufficient to be annotated as *visual interference* without considering color or form of the option. Secondly, negative samples of the dataset, i. e., *DS2* [94], involve such text with only one character as “x” which is not representative or descriptive. Quality of the source data will affect the quality of the data augmentation. Moreover, number of dark pattern categories are limited in the dataset. With more categories, the prompting strategies need to be updated. For instance, we added a specific statement to prevent possible confusion between urgency and scarcity. Other categories may require similar strategies.

6.6 Wrap-up and Perspectives

The development and evaluation of dark pattern detection models require structured and labeled datasets. Although this phenomenon is widely common, gained attention with the launch of “darkpatterns.org” in 2010 and has since growing in prominence, available datasets remain insufficient. In response to the data scarcity problem, there are two primary strategies: collecting new data or augmenting existing data. In this chapter, we introduced an effective augmentation method developed specifically for text-based dark pattern data. We leverage the advanced

understanding and generation capabilities of open-source large language models to construct a multi-agent framework composed of generator and controller models. Data quality is critical for augmentation; as commonly expressed in machine learning, "garbage in, garbage out" also applies here. Moreover, when producing synthetic dark pattern text data, it is essential that the new data preserve the deceptive characteristics of the originals while maintaining comparable structural patterns. We evaluate the effectiveness of our approach through expert annotation and model-based comparison, which includes fine-tuning models with the original data and the augmented data separately. The results show that the augmented data improve the detection accuracy of the models while maintaining the dark pattern attributes.

Chapter 7

An Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model

We propose and implement an Multimodal Large Language Model-based approach (DeceptiLens) that analyzes graphical user interfaces and assesses the presence of dark patterns. We utilize Retrieval Augmented Generation (RAG) process in our design and task the model with capturing the dark patterns, classifying its category, e.g., False Hierarchy, Confirmshaming, etc., and explaining the reasoning behind the classifications by employing recent prompt engineering techniques, such as Chain-of-Thought (CoT). We quantify the agreement between the model's outputs and experts' opinions. We additionally ask experts to gauge the transparency of the system's explanations for its classifications in terms of recognized metrics of clarity, correctness, completeness, and verifiability. The results indicate that our approach is capable of capturing the dark patterns in graphical user interfaces with high accuracy while providing clear, correct, complete, and verifiable justifications for its decisions.

*This chapter answers **Research Question 3**: "Which approaches should be implemented to detect dark patterns automatically? How can AI-based approach be used to detect them? What are the challenges and possible solutions for automated assessment tools?" and contributes to **Research Question 2**: "What considerations and methods*

should be taken into account when building labeled and structured dark pattern datasets?”.

The content of this chapter is primarily based on the following paper:

- *Kocyigit, E., Rossi, A., Sergeeva, A., Negri Ribalta, C., Farjami, A., & Lenzini, G. (2025, June). **DeceptiLens: an Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model.** In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1942-1959).*

7.1 Introduction

Deceptive (design) patterns (or Dark Patterns), which we abbreviate as DPs, are digital practices that manipulate, trick, or coerce users in ways that may harm them but serve the interests of the companies that implement them [18]. DPs are increasingly under scrutiny by the academic community that, among the others, also contributes to determine their harm [111]. Most DPs are recognized as mechanisms to circumvent the law by data protection authorities [8], consumer protection agencies [136, 6, 137] and their use is explicitly prohibited in recent regulatory efforts [138, 139, 140, 141]. In light of this, the use of some of these deceptive practices has been sanctioned by relevant watchdogs [142, 108, 143].

Since 2010, when Harry Brignull first named these practices as "dark patterns" and introduced categories such as sneaking, trick questions, and nagging [40], the list of DPs have expanded to more than 70 categories [29, 144], demonstrating that these patterns can manifest in various ways, contexts, and use cases. The usage of this practice is widespread, with the European Commission (EC) reporting in 2022 that 90% of the most popular apps in the EU contain at least one DP [11]. To curb the proliferation of these practices and equip stakeholders with tools that can contrast the existing automation asymmetry (where slow-paced manual detection methods compete with large-scale implementations of DPs), efforts have led to the development of various automated detection tools employing techniques such as machine learning (ML), natural language processing, and computer vision [72, 75, 145]. However, these tools have limited identification coverage and robustness [78], and we claim that this is because they lack multimodality, which makes them type-specific, i.e., they are working on specific aspects, such as text, images.

This study focuses on pre-trained Multimodal Large Language Models (MM-LLMs), which are built using Large Language Models (LLMs) and can simultaneously handle multiple data types, including text and image, performing at the same levels as humans in specific contexts

[146]. In principle, MM-LLMs can address the limitations and robustness issues of the existing detection tools, outperform them while being more efficient to use, i.e. less labor intensive.

It is not hard-coded like a rule-based model, which could enhance its robustness when addressing previously unseen deceptive design examples. Instead, these advanced language models learn from the data via the neural network components in their architecture [147]. Furthermore, they can provide reasoning behind their decisions [148], which is particularly critical for enhancing the transparency of a model that can be used as a decision-support system. Transparency is a prerequisite for the accountability of the decisions taken by its users. We propose our pre-trained MM-LLM-based approach (which we call DeceptiLens) to detect deceptive designs on the web and mobile user interface (UI), employing one of the state-of-the-art models, GPT-4o [149], along with an advanced prompt engineering technique, the Chain-of-Thought (CoT[150]), and a framework that minimizes errors by retrieving relevant information (Retrieval Augmented Generation - RAG) [87].

With this research study, our aim is not only to propose an implementation that identifies a potential DP on a UI, but also to provide relevant information that can enhance user trust in the functioning and the results of the tool. Thus, we guide our model to present explanations of the reasoning behind each decision. Explanations should be clear and objective to prevent misinterpretations and enable users to make informed decisions. Thus, we encourage the model to use the concept of measurable features proposed by [63] and include the information of the reference document that is used to reduce the uncertainty within explanations [151]. We specifically aim to answer the following research questions:

- **RQ1: Is the accuracy of the MM-LLM-based approach in tasks of automated detection of UI-based deceptive design patterns equivalent to that of human experts?**
- **RQ2: How clear, correct, complete, and verifiable is the reasoning generated by our deceptive pattern detection tool for experts?**

To address these questions, we first narrow our focus to 12 DP categories and create a dataset by gathering labeled or reported GUIs with DPs. We then test whether an MM-LLM (GPT-4o) empowered with the CoT technique and the RAG framework, is capable of recognizing these patterns and providing explanations for its decisions. Later, we ask DP experts to assess if the same GUIs contain DPs and we calculate the agreement between the experts' opinion and the model's output. Beyond calculating the accuracy of the performance of the system, we also ask the same experts to evaluate its transparency which we operationalize through

clarity, correctness, completeness, and verifiability of the system’s explanations. In this light, our contributions are:

- The first approach that proposes, designs and implements an automated deceptive design detection based on Multimodal Large Language Model and RAG process.
- A thorough DP expert evaluation of the accuracy and transparency of the approach based on recognized metrics of clarity, completeness, correctness and verifiability.
- The creation and publication of two datasets: the first one includes screenshots of UIs and deceptive design category annotations, and the second one includes AI-based generated explanations for each label’s reasoning.

7.2 Methods

7.2.1 Research Gaps & Research Questions

Our research aims to empirically test our proposal, which aims to support stakeholders detecting DP in a reliable and scalable manner. Our approach can be useful for identification of problematic practices in digital services [27, 4, 2]. Our tool should also be useful for regulatory bodies, seeking to investigate deceptive practices (see sec. 7.4.4).

Modern AI models face significant challenges, such as the need for huge amounts of structured and well-labelled training datasets and data-type-specific processing techniques. They are also restricted by their singular design, which does not accommodate the multimodal nature (text, graphics, *etc.*) of DPs. This makes ML-based systems insufficient for comprehensive DP detection. We check whether MM-LLMs can assess the presence of DP as well as human experts (**RQ1**) by bypassing the limitations of the aforementioned ML models via their multi-modal design and ”emergent abilities” [152].

On the other hand, since transparency helps the auditability of systems [153], we aim for our MM-LLM-based DP detection system to provide meaningful explanations for each detection reasoning to users. Thus, we investigate whether these explanations are transparent, i.e., whether they are clear, correct, complete, and verifiable by human experts (**RQ2**). These concepts are well-defined criteria in the literature. Moreover, clarity, correctness and completeness are commonly considered fundamental pillars of transparency (e.g., the definition of transparency in Art. 13 GDPR or Art. 13 AI Act). Verifiability of the explanations is also key, as it is linked to their usefulness: ”explanations are only useful to the extent that they allow a human decision maker to verify the correctness of the AI’s prediction” [154]. None of the existing approaches have ever been evaluated by DP experts in this regard.

Table 7.1: Target deceptive design pattern categories

1	False Hierarchy	8	Bad Defaults
2	Disguised Ads	9	Countdown Timer
3	Sneak into Basket	10	Limited Time Message
4	Hidden Costs	11	Confirmshaming
5	High Demand	12	Trick Questions
6	Low Stock	13	Activity Messages
7	Endorsement and Testimonials		

7.2.2 DP Dataset

We consider the 2024 taxonomy of [29] as a starting point for the creation of our dataset. We apply the granular specification as demonstrated in [63] (i.e., the low-level patterns that exhibit specific, identifiable characteristics) and list 43 categories of DPs. We selected 28 DP categories that are present in more than one taxonomy considered by [29]. Afterwards, we collected screenshots of DPs in web or mobile UIs, from sources given in Table D.1. The authors included false positives in the data set. We prefer both UIs without DPs and those with DPs but labeled differently. For example, Figure 7.1 demonstrates an example of "Countdown Timer" that was presented in [19]. On the other hand, we included false positives, such as Figure 7.2, which has a similar design. We then exclude categories for which we do not find at least 5 UI examples reported as DPs. Finally, we combine similar and overlapping categories, such as "high demand and low stock" or "countdown timer and limited time messages", and obtain 9 labels covering 13 DPs categories as given in Table 7.1. We use this dataset as a test set to compare the experts' opinion and our model's evaluation. We present data and code¹.

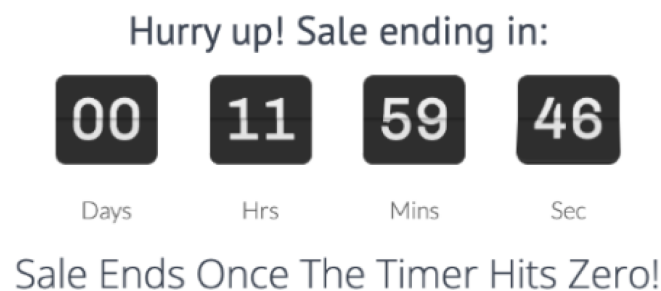


Figure 7.1: A true positive example of *Countdown Timer* dark pattern

¹<https://github.com/kocyigitmre/deceptilens.git>

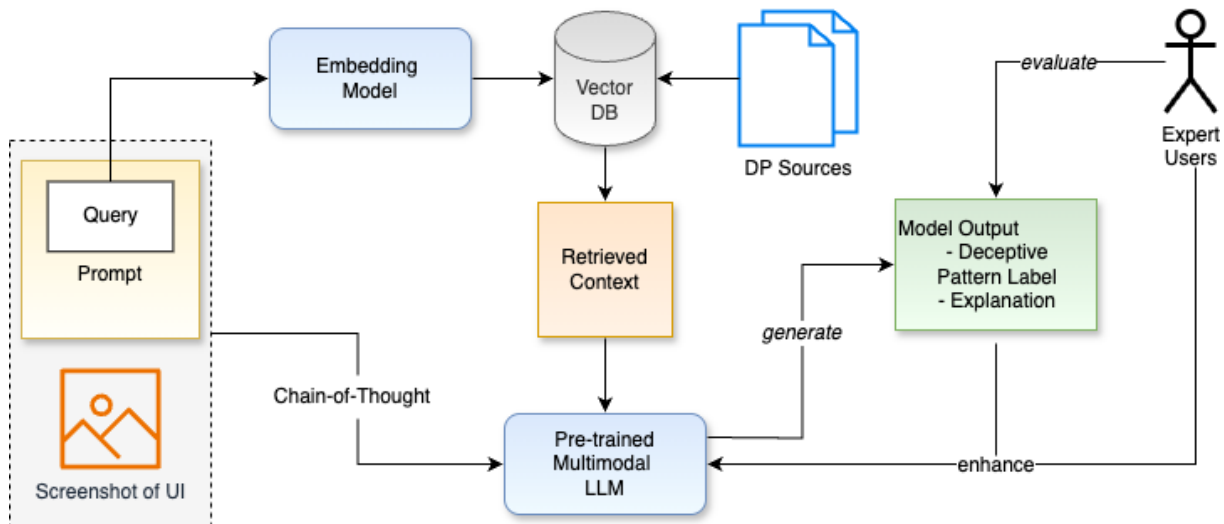


Figure 7.3: Overview of the deceptive pattern detection and reporting approach design.

OTP Verification

Enter the code sent to your mobile phone
 (*****5425)

Continue

00 : 59

Figure 7.2: A false positive example of *Countdown Timer* dark pattern

7.2.3 Design

In order to answer the **RQ1** and **RQ2**, we design a frame shown in Figure 7.3 to create a DP detection tool that is capable of explaining the reasoning behind its decisions.

Conceptual motivation

We choose GPT-4o (<https://openai.com/index/hello-gpt-4o/>) as our MM-LLM due to its state-of-the-art performance in linguistic and vision-based tasks [155] and the ease of use due to its API support. It is one of the leading models in the LLM landscape [156]. In addition, it is multilingual and supports 20 languages, on top of English. Although advanced pre-trained MM-LLMs indicate significant success in various tasks, the well-known issue of generation of factually incorrect responses is still present. To address this problem, we employ Retrieval Augmented Generation (RAG) and Chain-of-Thought (CoT) techniques in our design, which are presented in Section 7.2.3. Furthermore, we force our tool to use a certain structure in its reasoning and in the output explanation, which contains (1) "DP measurable features" [63], (2) a step-by-step analysis, and (3) a reference to the documents used for reaching the conclusion.

Model, Prompt Engineering, RAG

We employ prompt engineering techniques, such as instructive and contextual prompts [157], and apply strategies, such as using clear language and dividing the tasks into subtasks. We construct a prompt given in Figure 7.4 based on the selected DP target category and retrieve a relevant text, such as a description or an example from a vector database that stores the full text of the ontology of Gray et al. [29] and the seminal paper on DP attributed by Mathur et al. [18], as part of the RAG process. These were selected because they contain DP definitions and examples and systematize existing knowledge. Moreover, they are highly cited, peer-reviewed publications. We did not strive to find the perfect database, but rather to create a knowledge base that would have been familiar to the study participants who are part of the academic community. The model uses this retrieved text while assessing the presence of DP in the given screenshot. An example input-output pair of the model is shown in Table 7.4.

7.2.4 Expert Evaluation

We gathered a diverse set of experts composed of 14 academic researchers with multi or interdisciplinary expertise on DPs (i.e., HCI, UI/UX design, consumer protection, data protection, behavioral economics). They all had at least one peer-reviewed publication on the topic of DP. The study was organized in three stages. First, they were asked to assess the presence of a specific DP category on given UIs, identically to the task provided to the system. Secondly, they evaluated the explanations about the DP assessment that were generated by DeceptiLens. Finally, we asked them to comment on their evaluations in an optional interview and asked for their feedback to improve our approach. The details of each stage are given in the following

```

prompt = (
  # context
  "You are a helpful assistant that detects and reports dark patterns, i.e., deceptive design patterns"
  "by giving the reasoning behind the decision! Inputs are screenshots of web pages."
  # task description
  f"Decide if there is a {target_dp} which is described as: {target_dp_description},"
  # chain-of-thought
  "First of all, analyse the input and don't decide until the analysis is completed."
  "Analysis should follow these steps:"
  "1 - Detect measurable features which are observable, quantifiable elements."
  "For instance, \"accept all\" button is a visual entity while \"accept all\" text is a linguistic entity."
  "Metrics are the functions take the entity as an input and perform measuring."
  "For instance, 'size' is a metric which calculates weight and height in pixels."
  "Therefore, 'size of the accept all button' is a visual and measurable feature."
  "Another example is 'clarity of the accept all text'. In this case, 'clarity' is the metric."
  "Very important issue is that the metric's formula should be clear."
  "If you list clarity, clarity formula should be clear and objective."
  "Now, list all measurable features that can be relevant with dark patterns."
  f"2 - Analysis should be step by step. After determining the measurable features, check if there is {target_dp}."
  "3 - At the end provide your decision with an explanation of the reasoning behind the previous output."
  # output format
  "The output should have two components and be formed as below:"
  "Output 1: Label. It can be True or False."
  "Output 2: Explanation. It should contain the steps of your reasoning behind the previous output. "
  "Use the following example structure for the explanation."
  "1. Measurable Features:"
  "2. Assessment:"
  "3. Conclusion:"
  # include quote from cited document
  f"Add quote from: {retrieved_document} and give the title and year info if available in the Output 2 by quoting. "
)

```

Figure 7.4: A prompt for the MM-LLM and `target_dp` refers to the targeted DP in the UI, and its description is stored in `target_dp-description` variable.

sections.

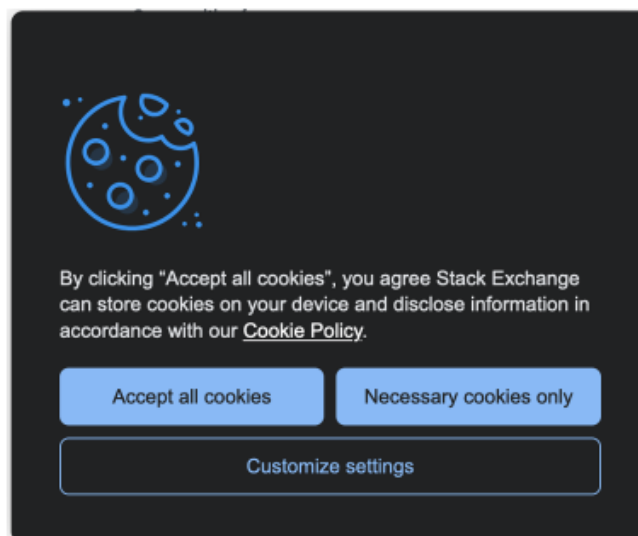
Task 1: Classification accuracy

Each UI of the dataset was assessed by three different experts who performed a binary classification considering the presence of a specific DP category on randomly distributed instances. An example question is given in Figure 7.5. Then, we compared our model classification with the classification of experts by taking the majority opinion into account to understand the agreement among them. We calculated the accuracy, precision, and recall scores of the model. We conducted these evaluations for both the majority opinion and the consensus, referring to cases where experts are in complete agreement. Furthermore, we measured the inter-rater agreement among the experts and examined how the examples included in our dataset and reported as DP are evaluated by the experts. We use absolute agreement scores and weighted measures of agreement: Cohen's Kappa [158] and Fleiss' Kappa [159].

Task 2: Explanation Evaluation

Whereas the previous task was geared towards computing the accuracy of the model compared to expert evaluations, the second task was mainly dedicated to assessing the transparency of the explanations. In this task, each expert evaluated five randomly assigned explanations that are generated by DeceptiLens for each DP assessment reasoning in terms of *clarity*, *correctness*,

Question 7



Definition	<i>Trick Questions subvert the user's expectation that prompts will be written in a straightforward and intelligible manner, instead using confusing wording, double negatives, or otherwise leading language or interface cues to manipulate a user's choice.</i>
1	Is there a "Trick Questions" dark pattern on the above UI?
	<input type="radio"/> YES <input type="radio"/> NO

Figure 7.5: An example question for the experts that was asked in the Task 1. In this task (more details are given in Section 7.2.4), the experts were asked to evaluate whether there is a specific deceptive pattern in the given UI. For each question, we also provide the related definition of the DPs that were taken from the recent ontology study [29] below the image.



Read carefully this AI-based automatically generated explanation(text in blue frame) about the above UI.

Note: The term "measurable features" refers to observable UI characteristics that can be helpful to identify the dark patterns.

1. Measurable Features:

- Visual Entity: Countdown timer displaying "00 Days 11 Hrs 59 Mins 46 Sec".
- Linguistic Entity: Text stating "Hurry up! Sale ending in:" and "Sale Ends Once The Timer Hits Zero!".

2. Assessment:

- The presence of a countdown timer suggests urgency, pressuring users to make a purchase decision quickly.
- The text reinforces this urgency by implying that the sale will end when the timer reaches zero.
- This can be classified as a "Countdown Timer" dark pattern, which uses urgency as a type of social engineering to indicate that a deal or discount will expire, potentially leading to uninformed purchases.

3. Conclusion:

- The use of a countdown timer and urgent language creates a sense of urgency, which is a characteristic of dark patterns. This aligns with the description from CHI '24, May 11–16, 2024, Honolulu, HI, USA: "Countdown Timers use Urgency as a type of Social Engineering to indicate that a deal or discount will expire by displaying a countdown clock or timer... As a result, the user may feel undue urgency and purchasing pressure, leading to their uninformed purchase of a product or service.

How strongly do you agree or disagree with the following statements? Please choose one for each statement.

		Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree	Reasoning (Optional)
1	The explanation is clear (i.e., the text is easily understood and unambiguous).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
2	The explanation is correct (i.e., the explanation accurately reflects the truth).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
3	The explanation is complete (i.e., the explanation contains the entire needed information).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4	The explanation is verifiable (i.e., the explanation provides information about the process of determining correctness through concrete and observable elements).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Figure 7.6: An example question for the experts that was asked in the Task 2. In this task(more details are given in Section 7.2.4), the experts were asked to evaluate *clarity*, *correctness*, *completeness* and *verifiability* of the explanations that were generated by our system.

completeness, and *verifiability* of the explanations. Questions were asked on a 1-5 Likert scale (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree). We provided the UI and the AI explanation, asking the experts to select "strongly agree" if they found the explanation to be clear (i.e., the text is easily understood and unambiguous), correct (i.e., the text accurately reflects the truth), complete (i.e., the text contains all needed information), and verifiable (i.e., the text provides information about the process of determining correctness through concrete and observable elements). An example question of this task is given in Figure 7.6.

Task 3: Interviews

While inviting the experts to participate in Tasks 1 and 2, we also asked them to join an optional (compensated) supplementary semi-structured interview session. Five participants joined the interviews (see demographics reported in Table 7.2) that lasted around 45-60 minutes. The objective of these sessions was to use an in-depth qualitative approach to find the features that contributed to the experts' perception of the explanation quality by the parameters described in the definition of Task 2. We applied a deductive-inductive approach to analyze the interviews [160]. To do that, we asked experts to reflect on the ratings they provided and compare the instances with high/low ratings to determine the main reasoning behind the ratings. We also asked experts to provide their general opinion about the characteristics of the system explanations and their perception of the system's utility. The full interview guide can be found in Appendix F. Based on that, Author 2 and Author 3 coded one interview of the dataset and then compared and discussed the coding categories. After the comparison, Authors 2 and 3 coded each 50% of the dataset; after full coding, Authors 2 and 3 discussed the final results. Similarly, Authors 2 and 3 coded the comments to Task 2 using the same code book.

Table 7.2: Demographics of the interviewed participants

No.	Professional domain	Years of exp. in the domain	Years of exp. on DPs
1	UX design	20	4
5	Behav. economics	3	2
6	HCI + law	5-6	5-6
12	Online privacy	7	5
15	Design, Data Prot., AI	10	6

7.2.5 Ethical and Legal Aspects

The participants were appropriately informed of the objectives and modalities of the research study and have freely consented to participate. This study has been carried out in accordance with the best practices for research with human subjects [161], the principles established in the EU Code of conduct for research integrity [162] and the recommendations provided by the EC’s guidelines regarding the implementation of generative AI in research [163]. The study has received ethical approval by the Ethical Review Board of the University of Luxembourg, under authorization 22-005 DPVADM. All participants were compensated for taking part in Task 1 and 2; those that agreed to be interviewed received an extra 15 EUR voucher. Any personal data collection has been carried out in compliance with the GDPR provisions concerning lawfulness, transparency and fairness, including data minimization, purpose limitation, and data protection by design and by default. The data processing activities have been documented in the dedicated institution’s register.

7.3 Results

7.3.1 Performance of DeceptiLens Compared to Experts

We first checked the consensus rate of the experts in the binary classification of the UIs. %71.62 of the instances were classified in complete agreement. In other words, %28.38 of the instances were classified by majority vote (i.e., two out of three experts agree). We also measured the agreement between the experts by calculating the Fleiss’ Kappa, which is 0.5743. This demonstrates that even DP experts had moderate agreement while evaluating whether there is a DP on the UIs. Then, we separately calculated the accuracy, precision, recall and f-1 (harmonic mean of precision and recall) scores by comparing our model’s predictions with the expert opinions. This evaluation was carried out for the consensus opinion of the experts and including the majority votes as shown in Table 7.3. In addition to achieving promising overall scores, our model reached a recall score of %100 based on the consensus of expert opinion. This indicates that our model did not miss any UI classified as DP by the full consensus of the experts. We also calculated the Cohen Kappa score as 68.26% (*consensus*) and 57.14% (*majority*) based on the true positive, true negative, false positive, and false negative numbers for this case. It shows that there is substantial agreement between the model and expert opinion when there is consensus. Therefore, answering our RQ1, the accuracy of the approach can be considered comparable to that of human experts.

Table 7.4: Example input and output of the DeceptiLens

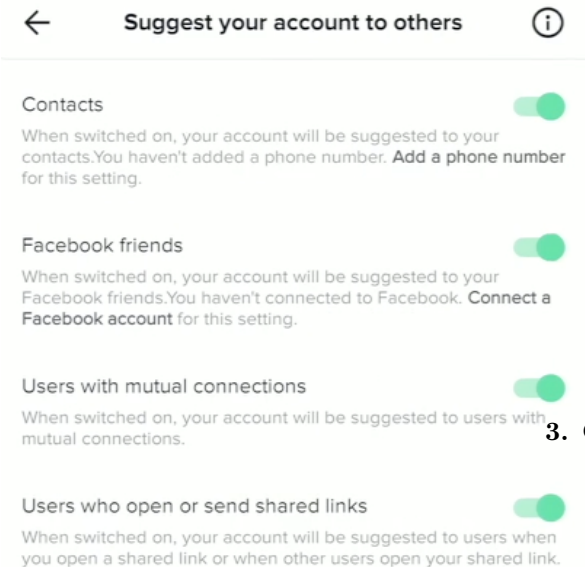




Input	Output
<ul style="list-style-type: none"> • {<i>Query</i>}: Is there a specific DP (e.g., Bad Default) in the given UI? • Context • Tasks • Chain-of-Thought-based assessment • {<i>Retrieved Text</i>} • Screenshot: 	<p>1. Measurable Features:</p> <ul style="list-style-type: none"> • Default Settings: All options are toggled on by default. • Visibility: The toggles are clearly visible, but the implications of being switched on are not immediately clear. • Language: The language used is neutral, but it requires users to actively change settings to protect their privacy. <p>2. Assessment:</p> <ul style="list-style-type: none"> • The default settings are privacy-invasive, as they automatically suggest the user’s account to various groups without explicit consent. • Users must take action to disable these settings, which can be considered a ”Bad Defaults” dark pattern. This subverts the expectation that default settings are in the user’s best interest. • The interface does not provide clear, immediate information about the privacy implications of these settings, which can lead to unintentional disclosure of information. <p>3. Conclusion:</p> <ul style="list-style-type: none"> • The design employs a dark pattern by using privacy-invasive defaults that require users to actively opt out to protect their privacy. This aligns with the concept of ”Bad Defaults” as described in the literature on dark patterns. • <i>Quote:</i> ”Subverts the user’s expectation that default settings are in their best interest by requiring them to actively change settings that could cause harm or lead to unintentional disclosure of information.” (CHI ’21, May 8–13, 2021, Yokohama, Japan, Mathur et al.)
 <p>← Suggest your account to others ⓘ</p> <hr/> <p>Contacts  When switched on, your account will be suggested to your contacts. You haven't added a phone number. Add a phone number for this setting.</p> <p>Facebook friends  When switched on, your account will be suggested to your Facebook friends. You haven't connected to Facebook. Connect a Facebook account for this setting.</p> <p>Users with mutual connections  When switched on, your account will be suggested to users with mutual connections.</p> <p>Users who open or send shared links  When switched on, your account will be suggested to users when you open a shared link or when other users open your shared link.</p>	

Table 7.3: DeceptiLens’s performance considering expert opinions. *Consensus* refers to cases where all experts share the same opinion on the evaluation of DP. *Majority* refers to cases where there is contradictory opinion among experts. *Overall* includes both *consensus* and *majority*.

	Consensus (%)	Majority (%)	Overall (%)
Accuracy	88.67	57.14	90.54
Precision	86.36	61.11	89.09
Recall	100.00	84.61	84.48
F_1 Score	92.68	70.97	86.72

7.3.2 Clarity, Correctness, Completeness and Verifiability of the DeceptiLens’ Explanations

Each expert evaluated five different (randomly pulled from the full set) explanations that our model generated about the reasoning behind each DP classification decision. Median values for each criteria, i.e., *clarity*, *correctness*, *completeness*, and *verifiability*, are given in Table 7.5. The results suggest that experts strongly agree that the overall explanations are clear and correct, while they reflect slightly weaker agreement for the completeness and verifiability of the explanations. Although the experts thought that the explanations were correct in general, they expressed lower ratings for the correctness of ”Sneak into Basket or Hidden Costs” category assessment and for the verifiability of the ”Countdown Timer or Limited Time Messages”. However, the sample size is so small that any conclusion about the reasons for these differences would be unreliable, thus we only discuss limitations of specific explanations as raised by the participants. Answering our RQ2, we can conclude that while there is no baseline evaluation of the transparency parameters from previous works, in general, all our results are at least on the ”Agree” side of the scale. In the following, we additionally provide the qualitative evaluation of these parameters via semi-structured interviews to grasp the nuances of the experts’ perception of the quality of the AI-based explanations.

Table 7.5: Evaluation results of the explanations by experts on a 1-5 Scale. (1: strongly disagree, 5: strongly agree)

Category	Clarity	Correctness	Completeness	Verifiability
Overall	5	5	4	4
Bad Defaults	5	4	4	4
Confirmshaming	4.5	4	3.5	4
Countdown Timer or Limited Time Messages	5	4.5	4.5	3
Disguised Ads	5	4.5	5	4
Trick Questions	5	5	5	4
Sneak into Basket or Hidden Costs	4	3	4	4
High Demand or Low Stock	4	5	5	4
Testimonials and Endorsements	5	5	4.5	5
False Hierarchy	5	5	4	5

7.3.3 Qualitative Feedback on the Explanations

Perceived quality of the system

Explanations were perceived as *”short, concise and overall well-written”* (P5) and useful as they provide a simple binary answer to the presence of DPs. The structured way of providing a step-by-step explanation was mentioned as one of the elements that supports understandability (*”it’s clearly understandable what’s the different parts are for. So the first one, it’s clearly about what has been detected by the system. Then the second is how the system is kind of analyzing those elements and then [there’s] the conclusion”* (P12)). It also enables the comparison between the models’ explanation and the given screenshot, which together with the granular assessment of UI elements in turns facilitates the human checking of the correctness and completeness of the answer. Providing the definitions from scientific publications was also positively valued, which contributes to *”a sort of authenticity”* (P6) and helps with the verifiability of the assessment, even though looking up the actual content of the paper *”is a lot of work”* (P2).

There were also some issues in the explanations that were raised by the study participants. Some underscored an occasional lack of coherence between the three parts of the explanations ("*conflicting information between the measured features, the assessment and the conclusion*" (P15)) or the mention of elements that are not related to the identified DP, sometimes raising the fear that the model has "*cherry-picked*" from the DP definitions (P13) or adopted "*a very specific definition*" (P15) that may not be generalizable or accurate.

Participants often lamented that the model would skip certain elements or DPs, without providing a complete assessment of the UI. Others commented that it was difficult to gauge what the model assesses and what it ignores. Some deplored incomplete citations of the references, which negatively impacts the verifiability ("*I am not sure if the description in the cited paper is correct and if this paper exists*" P2). Sometimes the tool is vague about which content it considers of the mentioned publications, using expressions such as "*aligning with Mathur et al.*" or making claims of similarity without explaining the reasons.

The experts underlined a set of issues with the correctness and completeness of the assessment. In a couple of cases, the participants noticed that the elements mentioned in the assessment did not correspond to the screenshots (e.g., "*the checkbox is not pre-ticked*" P1;). Some definitions of DPs were not considered accurate by the participants ("*The explanation uses a definition that is not necessarily agreed upon*" P2) or they used inaccurate terminology. More importantly, participants observed that the model occasionally applied the DP definitions in an overly comprehensive manner ("*may make users feel*" ... *is so carefully phrased that it might always be true*" P2) and that conclusions were seldomly based on an inherent assumption of "darkness" (P6) ("*the fact that the information is false is taken for granted*" P3). Certain instances contained what was perceived as mistaken assessments in the perspective of the experts ("*I don't think that the red text is a tactic to draw attention away*" P13). The dependence of a correct interpretation on contextual elements that were not provided to the LLM, was also raised, such as the completeness of the UI and the "*user's expertise and interpretation*" (P6).

The issue of detecting DPs on the UI

Many comments maintained that it is difficult, or even impossible, to assess the presence of DPs on a screenshot alone. Sometimes DPs are visible as parts of processes ("*I think my problem here was really about having just this piece of the interface and not having the whole user experience journey*" P15). In other cases, the UI is insufficient ("*It depends on the design's back-end to determine the truth*" P6) and context (or a comparison of the same UI at different times) is needed to avoid making a guess, for example when it comes to countdown timers. As a

consequence, P1: *"the conclusion is an overstretch, the AI cannot know if the claim is false or truthful"*.

This also raised the question of the objectivity of expert detection (*"I don't think we all agree on what is really a dark pattern [...] there is really quite a huge difference in interpretation."* P15) and of academic scholarship (*"just because you find a paper that says A, [it] doesn't mean it's the holy truth"* P1).

Trust in the AI model

The issue of trusting the LLM was raised by a few AI-savvy participants in terms of interpretation of the DP publications and fabrication of content (*"you cannot really trust AI to have properly interpreted the paper and is the citation coming from it or is it made-up?"* P1), even though the doubts can be counterbalanced by the confidence in the system designers (i.e., the authors) (*"But because you and your team and you are doing this, I believe in you [...] I trust the provenance of the researcher, not of any researcher"* P12). Even though the impossibility of research objectivity was underscored earlier, trust in research publications seems to be blind for some participants: *"It looks verifiable, and since I know the paper I trust"* P5; *"for me as a researcher, I look at citations and I'm like, yeah, great [...] Trust the definitions!"* P6).

Target users and the overall system usefulness

DeceptiLens was perceived as useful for different users, for example for researchers to support the analysis of digital services, to build evidence for the civil rights association doing advocacy work (including supporting the general public who wants to contribute) and to educate students and younger generations. It was not considered helpful for regulators, though, as it uses a research-oriented vocabulary and does not provide legal arguments about the lawfulness of the design practices.

Suggestions for the system's improvement

The experts proposed several recommendations to improve the quality of the system. First, for enhancing the comprehensibility for users and filling the lacuna in the explanations, the experts proposed to reorganise and enhance the content by (a) rearranging the order of information by priority, for example by including a concise conclusion first instead of starting from the measurable elements so the user first understands the system verdict and then moves towards the explanation (which could also be given as a second layer of information on demand that could address the informational needs of different users) or order statements according to what

is a more dangerous influence for decision-making; (b) avoiding repetitions since "in general the last assessment is already a conclusion" (P5); (c) separating better the design elements from the judgment on its potential outcomes on the user; (d) standardizing the citations of the references and including a link to support checking the information; (e) referring to only one source, i.e., the ontology, to provide more consistent definitions across DP and resolve the confusions of different interpretations; (f) adding the definition of the DP to the explanation; (g) improving the graphical presentation by e.g., colorcoding the result of the assessment so that the results of the assessment are clear at a glance even without reading the whole text.

In response to the question of verifiability of the AI assessment, experts proposed to show the users a step-by-step analysis or a decision-tree interpretation of analysis, combined with confidence rates of the element being a DPs. Finally, one of the experts proposed that the system should be able to interfere with the proposed interface, showing how the design without the DP could look like.

7.4 Discussions

7.4.1 Overall System Performance

The overall results about accuracy, clarity, completeness, correctness and verifiability are very remarkable. Completeness ratings were not always very high because the participants expected the assessment to be about all the DPs in the interface, whereas the LLM was instructed to look for the presence of specific DPs. When asked whether any other criteria would have been useful to assess the performance of the system, the experts found that these metrics were sufficient.

7.4.2 How Might the Presence of Dark Patterns Reliably Assessed by Human Beings and by MM-LLMs?

We observed that even experts do not reach full consensus on detection tasks: only 71.62% of the UI examples were classified in complete agreement. When we investigated the reasons through the qualitative analysis, we observed two main causes. First, UI screenshots alone are often insufficient to assess the presence of particular types of DPs: contextual information, user journeys and code is also needed. Other types of data, such as web source files of the related page, or a video record that shows the interactions are needed to be sure about the functionality of the buttons. Second, the lack of objective descriptions of the DPs types causes different interpretations. For example, a clear indication of an advertising objective must be present on the UI, otherwise it can be considered as a "Disguised Ad" according to the definition contained

in [29]. However, one of the experts remarked , even when the advertising mark is not clearly visible, contrast is sufficient to distinguish the advertisement from the main content of the web page. As a result, two out of three experts classified an instance that was reported ² as an example of DP as non-DP. Moreover, some participants explicitly stated that even the expert classification is not objective and pointed to individual characteristics of users that may also influence such classification (e.g., "difficult language" is not an absolute value and rather depends on the actual user).

These results point to structural problems of the current state-of-art DP definitions and the methods for analysis. First, in certain cases, the mechanics of DP necessitates the user-computer interaction beyond static UI representations. Therefore, including static examples in the datasets of DP creates uncertainty both for humans' and for the model's interpretations. Second, in many cases, the answer about DP presence can be represented as a continuum of options rather than a binary yes/no classification. Generally speaking, there are certain patterns that may be manipulative in one context but not in another (e.g., adding steps as a friction design pattern may be a legitimate strategy to slow down automated behavior and bolster reflective thinking to strengthen security [164], but it is demonstrated that it increases cookie consent acceptance rate when it is associated with consent refusal [114]), while certain DPs (e.g., DPs that omit relevant information) may intrinsically be more difficult to spot than others [48]. The ease of "visual detection" for human beings has also been pointed to as one of the reasons why many legal proceedings against illegal deceptive design practices only address issues at the interface level and disregard the underlying code [142].

An important measurement was that the recall score of our approach achieved 100% when experts unanimously agreed on the presence of DPs without any objection. This indicates that DeceptiLens did not miss any DP compared to the experts. However, when there are contradictory opinions among the experts, it is plausible to observe this situation reflected in the DeceptiLens' behavior. This issue also raises another discussion point: "Can we use the dataset to train an AI-based DP detection model even when the experts do not agree on it?". One solution to this would be to only include data where the label has full consensus.

Another observation related to this was the classification of a screenshot, which was presented as an example of DP in a well-known academic paper [19] published in 2019. The experts labeled it as non-DP while the model identified it as an example of "Hidden Costs" consistently with the aforementioned paper. This situation brings up the following question: "Should the previously collected data still preserve its label and can it be reliably used to train or instruct

²<https://darkpatterns.uxp2.com/pattern/relay-for-reddit-ads/>

the AI models?”. This should be critically considered, otherwise, no matter how reliable the data source, as observed in this case, it may no longer be a DP and mislead the model. A difficulty in this respect is that, at date, there are almost no curated, publicly available datasets of DP instances that have been carefully assessed by DP experts based on the current available knowledge produced by DP scholarship. With our dataset contribution, we seek to bridge this gap in current research, even though we acknowledge that this is just a first step in this direction.

7.4.3 How Might We Make the Explanations More Reliable While Avoiding Overreliance?

Experts generally agreed that the system provides clear, transparent, and verifiable explanations about its reasoning and results. This is a crucial system design consideration that the other AI-based approaches mentioned in Chapter 5 did not engage with. Yet, study participants still acknowledged some limitations to verifiability. We implemented verifiability through reference to the scientific sources and reference to the measurable elements in the interface, which could be the basis for DP implementations. While the experts mentioned that the latter mostly worked and helped to catch the system’s mistakes, the former occasionally became a source of confusion. The inclusion of bibliographic citations could even be perceived as an “anti-transparency” and authority cue [165] that may cause automation bias - an overreliance on decision-support systems’ output without looking for or simply ignoring counterfactual information [166]. Experts expressed doubts that DeceptiLens users would check the lengthy, complex publications themselves and would rather blindly trust the “scientific sources” mentioned by the AI system. We add RAG process by including another fully independent component of the DeceptiLens’ architecture, responsible for retrieving most relevant scientific sources from our database to minimize risks of made-up scientific references in the explanation of the core MM-LLM.

Since adding explanations seems to increase human overreliance on the system [167], an approach that has been proposed to counter automation bias is the implementation of cognitive interventions [168]. Namely interventions that break automatic, heuristic-based reasoning and engage the human being in reflective thinking [169], based on the dual-process theory [170]. For example, asking the user to make a decision before being shown the AI suggestion, delaying the AI answer so that the user can form their own opinion first and allowing the user decide whether they want to look at the AI recommendation, may reduce overreliance at the expense of acceptability [169]. As noted by the authors, beyond improving the intrinsic quality of explanations, efforts should be devoted to designing an interaction between AI systems and humans, to ensure that they make meaningful use of the explanations and make informed decisions. The

future versions of the systems should not only perform on a considerable level, but it should also have mechanisms for facilitating user reflection. For example, the system can ask the user about their perception of the interface and facilitate a step-by-step reasoning [171].

7.4.4 How Might We Enhance the Approach?

Enhancing the Approach for Researchers' use

We believe there are several ways to increase the transparency of the assessment results presentation. One of them includes showing the systems' (un)certainly about the results. Recent studies in this area showed that acknowledging the degree of the system's certainty can raise the trust in the system and help users make better decisions [172, 167, 151]. Furthermore, sharing the test dataset, limitations, and risks while publishing the evaluation results of the LLM-based DP detection system enhances transparency. [93].

Another idea is to implement the ability to switch between different guiding methodologies used for the system's decisions and to show the user the spectrum of outputs proposed based on the difference. It will help to acknowledge methodological differences in the DP scholarship and provide a nuanced understanding of it, leaving to the user the decision.

A more thorough analysis of the various uses-case contexts should also be carried out, in order to address the different informational needs of stakeholders. For instance, some participants suggested to provide a clear, concise, colorcoded conclusion upfront and the details as a follow-up that can be explored on demand. This could be useful especially when researchers intend to inspect digital services at scale: they could first check if the AI system answer corresponds to their expectation and only if it doesn't examine the explanation. In the end, to quote one of the interviewees, "making a better explanation would really depend of the context in which the explanation would be given" (P15).

Enhancing the Approach for Independent Authorities and the General Public

The experts mentioned that the perceived quality of the system's output can be very context-dependent, as the end-users can have different professional backgrounds and knowledge about DPs. Therefore, the following iteration of the tool should be specifically tailored to the context of use. Based on the current evaluation, we identify two additional use cases: a regulatory authority tool for checking compliance; and an educational tool to raise awareness.

Even though we developed DeceptiLens to specifically aid researchers (a decision that is reflected in the design of the system), since DPs often constitute unlawful practices (such as consumer protection), we plan that future versions may be used by regulators to support their

enforcement activities related to monitoring and detecting wrongdoing in digital products and services (i.e., EnfTech or Enforcement Technology [173]). In order to do so, there are challenges that need to be addressed. First, the definitions that were selected for the study [29] do not provide legal arguments. Thus there would be a further step that needs to be made: for example, the "Bad default" DP should be mapped to privacy-invasive defaults that are relevant under the GDPR and its national implementations. This information can be enriched with official interpretation guidelines such as the EDPB's on DPs that state that if one option among many is highlighted, then it should be "the most restrictive one regarding personal data" [8, p. 23].

Another database for the RAG should then be implemented, but it would not be a trivial task. The first question would concern selecting relevant documents, spanning across regulations, case law and official guidelines. One of the challenges may be the representativeness of this dataset: for example, [142] show a skew in enforcement decisions towards "visible" UI elements, which neglects other more covert deceptive techniques. Moreover, sometimes guidance on the same design differs across jurisdictions [114], which makes it hard to provide a uniform source of knowledge. Thus a second key aspect would be how to integrate divergent information sources. An idea would be to include multiple vector databases in the RAG process and to utilize LLM-based multi-agent systems, which can provide individual specialized agents that can collaborate effectively [174]. One of the participants with previous experience in a Data Protection Authority (DPA) suggested focusing on specific cases, such as the assessment of the legality of cookie consent banner design [175].

Then, it should be carefully analyzed if enforcement agencies would use such an AI-based tool and if so, according to which safeguards (e.g., only with certain demonstrated performances levels?). Most importantly, it would need to be carefully defined who would use the tool and at which stage of the enforcement process and for which objective: this would also inform the implementation of strategies that lower overreliance on the system, as mentioned earlier.

In contrast, for the general public use, it may be necessary to streamline the explanation and present the decision, even starting from the system verdict and showing the explanations supporting it. Additionally, it can be useful if the system also proposes several ways which can help eliminate the DP from the interface and replace them with a "fair" [124, 176, 177] or "bright" design pattern [46], in line with the interventions proposed by Lu et al.[178]. However, at date, whereas there is an increasing agreement on what a DP is, there is still no common understanding of what a "fair" or "bright" design pattern is. Finally, since our experts have already reported that the system can have an educational effect in helping users learn more about DP, it would be useful if the system could implement more explicit educational strategies

and leverage different types of motivation to facilitate learning about DP [179].

7.5 Compliance With the AI Act

Would our detection tool be subject to the requirements imposed by the AI Act (Regulation 2024/1689)? If this tool was intended for research purposes, it would fall under the research exemption (Art. 2(6)), but would need to respect the key principles for developing trustworthy AI and of research ethics (see Sec. 7.2.5). However, if the intended use was for enforcement agencies to support their decision-making and investigative process in supervisory tasks, then we would need to determine if the system would be considered a high-risk system under Art. 6(2) (Annex III(6)). This applies only if there is a "significant degree of power imbalance" in the actions of enforcement agencies that have "adverse impacts on fundamental rights" (Rec. 59). Since consumer protection and data protection authorities only investigate administrative offenses by organizations, our analysis suggests that our tool would not be classified as a high-risk system.

Nevertheless, enforcement agencies that adopt the detection tool would need to ensure that their staff reaches "a sufficient level of AI literacy[...]" (Art. 4) based on their technical knowledge, experience, education, training and the context of use. For example, the training should encompass the limitations and the risks of overreliance on the tool's outputs. If, instead, the tool was made available to the public to strengthen their ability to recognize DPs and it was not evident that it is an AI, the transparency obligations under Article 50(1) would apply.

7.6 Limitations

We inherit the cons due to using MM-LLMs, and we relied exclusively on the UIs to detect DPs. As we discussed in Section 7.4, the evaluation of certain DPs requires additional information, e.g., html of the web page or user journey, which we did not include. We agree that inputs should go beyond the UI. We took the recent ontology [29] as a reference and did not consider the types that do not exist in it. Since we created our test dataset by collecting reported DP examples, we could not include the categories that do not have sufficient reported examples. In parallel with this and considering the human resources, the size of the dataset is kept restricted. Another important limitation is that we did not include all the DP experts. Although the experience and focus area of our experts are diverse, they are researchers or have studied DP problems from a researcher perspective. Moreover, we asked experts whether there is a specific DP instead of asking any DP in the given UI. We are aware that the latter is valid in a real scenario. However,

we prefer the former for the simplification of tasks for human experts and measurement of the comparison between the model and experts.

7.7 Future Work

Beyond the recommendations for various users illustrated in 7.4.4, there are several manners in which we believe the present approach can be enhanced. A future development concerns the implementation of instruction fine-tuning of the core model based on the experts' feedback that we gathered. First, the examples annotated with the consensus of experts could be given to the model as references. Second, rules that are directly relevant to the quality of the explanations could be added within the prompt design. For example, the presentation of the bibliographic citation can be easily standardized, while various layers of explanations could be presented (e.g., a first concise informational layer about the conclusion and other layers that provide details about the reasoning, the definitions and the sources used). Third, the dataset should be enriched with other types of data (e.g., the user journey, metadata, HTML code), as UIs alone in many cases are insufficient to reliably establish the presence of a DP. Fourth, the dataset should be expanded with other types of DPs that are not currently included, in order to enhance the ability of recognizing DPs and the usefulness of the system. Then, it would then be meaningful to carry out a benchmark test that compares the "enhanced" explanations with the explanations produced in the present study with an expert assessment.

7.8 Wrap-up and Perspectives

This study that focused on achieving efficient and automated dark pattern detection ultimately presented that several crucial considerations extend far beyond the singular goal of accurate classification. The first critical challenge centers on the lack of objective and standardized definitions of dark patterns. This foundational problem and insufficient contextual information on assessed data instances lead to subjective interpretation and even domain experts fail to reach a definitive, unanimous consensus. The observed disagreements among the experts also cause the absence of gold-standard datasets currently impeding the field. This fundamental issue must not be disregarded in future detection tool development. A second significant empirical finding is related to the performance of Multimodal Large Language Models (MM-LLMs). MM-LLMs demonstrated high success (detection of dark patterns with perfect recall score) in cases where experts were in full agreement. However, the performance rate declined (when measured against the majority-vote reference) for several examples where expert opinions diverged. This confirms

that the primary constraint on reliable AI-based detection is not merely an algorithmic limitation, but rather the ambiguity embedded in the underlying human-defined ground truth. A third key insight is the necessity of viewing the detection method not as a fully autonomous system, but as a decision support tool. Consequently, the system must transparently articulate the reasoning behind its assessment to the human decision-maker. We proved that the advanced language generation capabilities of LLMs can be effectively harnessed for this purpose. We moved beyond raw text generation by utilizing the "measurable features" concept (proposed in **Chapters 3 and 4**) to instruct the LLM to generate explanations in a structured format. The experts also evaluated these explanations and rated them as clear, correct, complete, and verifiable. We achieved this verifiable output by deliberately enhancing our framework with the Retrieval-Augmented Generation (RAG) process. This integration was essential not only to successfully minimize hallucination but also to address the crucial need for factual verifiability by citing reliable sources in the explanations, which seen as a important capability. This architectural decision confers significant versatility upon the system. While our RAG implementation used a database populated with reliable research sources, the framework is readily adaptable to other specialized contexts. For instance, we anticipate successful application in a purely legal compliance scenario through the straightforward replacement of the source documents with relevant legal texts and minor adjustments to the prompt design. In summary, this work emphasizes that while MM-LLMs offer a promising and effective method for dark pattern detection, the overall success of it is contingent upon critical factors external to the utilized technology. These determining factors include the standardization of expert consensus, the quality of the foundational dataset, and the specificity of the usage scenario. The core component driving the framework presented in this study was the proprietary GPT-4o, which is selected specifically for its state-of-the-art performance in multimodality needed tasks. Given this reliance on a closed model, the following chapter focuses on a critical, and novel, comparative analysis: As the first such investigation, what specific performance outcomes and resource trade-offs are realized when utilizing **open-source multimodal LLMs** for the automated dark pattern detection task?

Chapter 8

Evaluating the Automated Detection of Dark Patterns in GUI Datasets with Open-Source Multimodal Large Language Models

We evaluate the performance of open-source Multimodal Large Language Models (MM-LLM), such as LLaVA, Janus-Pro, in the task of dark pattern detection on visual user interfaces. Moreover, we compare the open-source models' performance with one of the leading proprietry model, GPT-4o. While evaluating the performance of the models, we use different prompt engineering techniques including our own dark pattern category-specific expert Chain of Thought (CoT).

*By presenting the first empirical evaluation of open-source MM-LLM in detecting dark patterns, this chapter contributes to **Research Question 3**: "Which approaches should be implemented to detect dark patterns automatically? How can AI-based approach be used to detect them? What are the challenges and possible solutions for automated assessment tools?". The content of this chapter is primarily based on the following paper:*

- *Kocyigit, E., Liga, D., Rossi, A., Sergeeva, A., & Lenzini, G (2025, August). **Evaluating the automated detection of dark patterns in GUI datasets with open-source Multimodal Large Language Models.** (Manuscript submitted to PeerJ Computer Science).*

8.1 Introduction

Dark Patterns (DPs) are digital deceptive UI design practices that coerce, manipulate, or mislead users to take actions that often work against their own interests [18]. For example, often users are nudged to accept privacy-invasive marketing practices on the interface of cookie consent notices that implement *False Hierarchy*¹, which is a way of visually presenting an option with greater prominence than others [29]. DPs threaten individual and social welfare by variously impacting individual autonomy [180], and many regulatory bodies [8, 137] denounce that they circumscribe the rules on data protection, consumer protection and competition. This is why these digital practices are sanctioned, sometimes with hefty fines [181]. However, given the proliferation of digital services that sometimes incorporate dark patterns by default, while watchdog resources to identify them are finite, detecting these manipulative patterns is challenging.

User behavior may be influenced through the use of misleading language on digital interfaces, such as with overly *Complex Language* [8], or *Confirmshaming* [19] that plays with their emotions. In other cases, manipulation occurs through the design of graphical elements, such as when *Visual Interference* [19] or *Bad Defaults* [4] give prominence to certain options over others, or by making user journeys overly complex or even meaningless, e.g., *Dead End* or *Privacy Maze* [8]. Two widespread dark pattern examples are simulated in Figure 8.1 and Figure 8.2. The first one uses visual elements, such as more prominent and bigger buttons, to lead users to the "Accept All Cookies" option, while the second one exploits emotionally manipulative language on the second option to nudge users to subscribe. Given all the different modalities through which DPs can be implemented, a robust and efficient approach to DP detection requires multimodal analysis abilities.

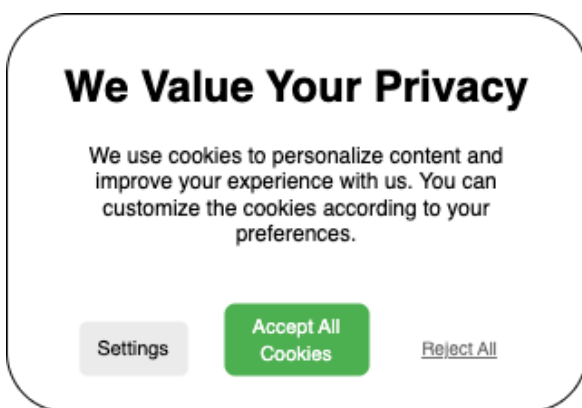


Figure 8.1: A simulated example of *False Hierarchy* dark pattern.



Figure 8.2: A simulated example of *Confirmshaming* dark pattern.

Recently, academia has developed approaches for the automated detection of DPs, as a

¹Dark Pattern categories are indicated in italics.

recent survey shows [78]. Even assuming the availability of a gold-standard reference dataset that addresses the aforementioned issues, certain challenges persist in the automated detection of DPs. Automated DP detection tools have the potential to be used as decision support systems for supervisory authorities that need to streamline their auditing tasks (through Enforcement Technologies or EnfTech [173]), companies that carry out internal compliance checks, consultants that support them in such tasks, and scientists that variously advance the research on the identification of manipulative design patterns. However, tools that rely on traditional rule-based text matching may struggle to adapt to ever-evolving manipulative design patterns. Moreover, tools that are built on top of conventional deep learning models require vast amounts of labeled data. We hypothesize that MM-LLMs can address these issues thanks to their emerging skills, such as customization of the model without the need to collect big data, and the capacity to understand both visual and textual inputs, and their adaptive nature [182].

In the European view of fostering open science and thereby lowering entry barriers of developers and researchers [183], the academic research community is calling for further developing open-source LLMs and preferring them to proprietary models. Moreover, since public funding should benefit society at large and re-use should be ensured to foster innovation even further, funding agencies often require the application of copyleft licenses to the release of computational models and tools, if these have been developed or trained through financial support coming from public sources. The reasons for encouraging open source are many: i) reproducibility of scientific results, as the parameters and training data of proprietary models can be modified unilaterally [184] and older models are deprecated [185]; ii) transparency and oversight, as proprietary models are not open to independent scrutiny and verification [184] also because their architectures, pre-training corpora, dataset composition, parameter count and fine-tuning methodologies are not publicly disclosed [186]; iii) control, as open source LLMs enable researchers to inspect why they work as they do, customize them and improve them [183]; iv) availability, as the companies developing proprietary models could raise prices substantively or even fail [184], thereby nullifying or destroying research endeavors that are dependent on the access to their models. Proprietary models, in other words, do not provide the necessary assurances of trust for their adoption in sensitive contexts, such as those supporting tasks with significant legal implications, as described in this paper. This is why we expand the work [34] carried out with GPT-4o to the use of open-source models in this paper.

In this regard, we evaluate the performance of open-source multimodal LLMs in the detection of dark patterns. We carry out experiments on the models through different prompt engineering strategies, such as Chain-of-Thought and "expert-crafted DP category-specific rules". Finally,

we compare the results of open-source models with GPT-4o, one of the leading proprietary MM-LLMs that the research community has widely used to this end. This chapter makes the following contributions:

1. Presenting first empirical evaluation of open-source multimodal large language models in the task of dark pattern detection, including a performance comparison with GPT-4o and practical investigation of prompt engineering strategies.
2. Developing expert-guided and category-specific Chain-of-Thought prompting strategies to improve the accuracy and recall of DP detection MM-LLMs.

8.2 Research Gap

Researchers have explored various approaches for effective detection of DP, including machine learning algorithms [71], deep learning techniques [187], hybrid approaches that combine deep learning with knowledge-based processes such as text pattern matching [72], or computer vision and natural language processing techniques [75]. Moreover, language model-based approaches, which employ an earlier approach such as BERT and its derivatives [94] or recent versions such as LLama [63], have been tested on text data. Although these approaches have shown promise, several issues hamper the effective detection of DP.

First, the traditional rule-based approach, e.g., defining hard-coded text matching conditions, is not an efficient way of detecting DPs because deception can be implemented via a slight modification of the text. For example, AidUI [72] defined three hard-coded textual patterns, such as "only + [number] + left OR available in stock", "limited stock OR supply OR quantity OR availability", "left OR available in stock" for *Low Stock* DP identification. However, instances that have alternative wording such as "almost gone", "few remain", or "run out" would not be recognized as *Low Stock*, despite sharing the same manipulative characteristic.

Second, although machine learning and deep learning detection models can deal with the aforementioned issue of having hard-coded rules, i.e., they are more robust against zero-day attacks [188], they have other challenges or limitations. For example, traditional machine learning algorithms, such as Random Forest or Support Vector Machine, require exhaustive feature engineering prior to training [189] and do not perform as well as deep learning models on image recognition problems [190]. Although deep learning models overcome these specific limitations, their performance is highly dependent on the availability of large and well-annotated representative training datasets [191], particularly for classification tasks such as dark pattern identification. Another crucial issue is the reliance on a single modality in dark pattern detection

tools, which narrows their dark pattern identification range. For example, a text-data-specific dark pattern detection model such as that proposed by [63] can achieve accuracy up to 97%. However, it remains ineffective in dark patterns that can be recognized by visual cues, e.g., *False Hierarchy*. This highlights the need for multimodal approaches in dark pattern detection.

To address the issues outlined in Section 8.2, the need for flexibility, adaptiveness, and capability of operating without large-scale labeled data and multiple modalities has led us to explore one of the state-of-the-art advances in AI, namely multimodal LLMs. A recent study [34] on the matter indicated that dark pattern instances, which are present in UIs and unanimously labeled as dark patterns by experts, were identified with 100% recall. However, the performance of open-source versions of these multimodal-LLMs on the same task has not been explored to our knowledge, despite the fact that the research community encourages and advocates open-source AI for the reasons outlined in introduction of this chapter. Consequently, we formulate the following research question:

RQ1 How well do open-source multimodal Large Language Models perform in the detection of dark patterns?

This main research question raises several sub-questions.

Recent benchmark results show that proprietary models, such as "claude-3-5-sonnet", "gpt-4o", and their various iterations across different periods, occupy the top 6 rank [192]. Furthermore, recent findings highlighted that the existing multimodal LLMs have not yet reached their maximum capacity [193]. Although fine-tuning is one of the common methods of improving the performance of an LLM in specific task, it is impractical in our case because our data source for fine-tuning is extremely limited and even low-source fine-tuning needs nearly 1K [194] instances for only training. These lead us to ask two sub-questions related to the main research question:

RQ1b. How do open-source multimodal LLMs compare to proprietary models like GPT-4o in the task of dark pattern detection?

RQ1c. To what extent can prompt engineering strategies bridge the performance gap between open-source MM-LLMs and the proprietary models?

8.3 Methodology

8.3.1 The Task of Dark Pattern Detection

To address **RQ1b**, the primary task assigned to AI models is to determine the presence of a specific DP in a given UI, which involves applying a binary classification. For each instance,

the definition of a DP type from [29] is provided to the models along with the UI assessed. We perform five runs for each model repeating the same prompting strategy to account for stochastic variation of LLMs [195], and the mean was reported. This experimental design aims at reaching a good balance between cost-effectiveness and evaluation robustness, and reflects established practices in previous works [196].

8.3.2 Model Selection

For the investigation of open source MM-LLMs’s performance in the DP detection task, in continuation of our work on **RQ1b**, we choose 3 models, which are **Janus Pro 7B**, **LLaVA 1.5 B**, and **LLaVA-v1.6-Mistral-7B** because they are well-known and are considered to be among the best-performing open-source models currently available. We select **GPT-4o** as our proprietary base model to evaluate the performance of open source models since it is listed as the leader MM-LLM on different benchmarks [197], it is still considered flagship model and had achieved promising results in DP detection [34].

8.3.3 Prompt Engineering Techniques

To address **RQ1c**, our prompt engineering strategy is to use common prompt engineering techniques, such as zero-shot learning, contrastive learning, and Chain-of-Thought (CoT), individually and in combination, in addition to utilizing recommended prompting tactics, e.g., using plain and clear language, structuring the prompt by explicitly stating the input and expected output format, and providing context [198]. After preliminary prompting practices, we initially formulate four prompting scenarios: (i) basic, (ii) basic + contrastive learning, (iii) basic + Chain-of-Thought, and (iv) basic + contrastive learning + Chain-of-Thought. In Chain-of-Thought, we instructed the models for extracting measurable features from the given image first, then make the assessment to identify dark patterns by checking the characteristics, such as asymmetry, restriction, etc. Finally, we develop an expert-guided prompting strategy based on strategy (iv). This approach includes DP category-specific instructions and reasoning from our DP experts instead of generic CoT for all categories. , and named this strategy (v) as ”expert-CoT”. A template of prompts for this strategy is shown in Listing 1. We defined the model as a dark pattern detection assistant in all scenarios.

Listing 1. Expert Chain-of-Thought Prompt Template for Dark Pattern Detection

```
You are a helpful assistant trained to detect and report dark patterns—i.e., deceptive design patterns—based on the analysis of web page screenshots.
```

Your goal: Determine whether a <SPECIFIC DARK PATTERN CATEGORY> exists in the provided image, answering **Yes** or **No**. Use the description of this type of dark pattern and the steps provided in the assessment section.

Definition of the dark pattern: <DEFINITION OF DARK PATTERN CATEGORY IN QUESTION>

Important Instructions:

- Provide a detailed explanation for your decision.
- Provide your answer after the explanation.
- Follow the steps strictly and only conclude after completing all.
- Avoid these common mistakes:
 - Do **not** rely on superficial appearances.
 - Do **not** overlook subtle asymmetries or minor restrictions.
 - Do **not** confuse partial visibility with absence.

Analysis Steps:

1. **Detect Measurable Features:** Identify all relevant visual or textual elements that can be objectively measured.
2. **Assessment:** Apply the following labeling steps:

<EXPERT CHAIN-OF-THOUGHT GOES HERE>

3. **Output Format:**

- Explanation: [...]
- Label: **Yes** | **No**

8.3.4 Evaluation

Model Evaluation

We evaluate the performance of the models by comparing their classification with the annotations of human experts. For the comparison, we used four main performance metrics: accuracy, precision, recall, and F-1 score since they are well-established metrics for evaluating classification performance [199]. Since "accuracy" is not sufficient alone in terms of informativeness and

distinctiveness, including "recall" and "precision", and "F-1 score", which is the harmonic mean of these two metrics, is crucial [200]. While accurately detecting DPs is important, minimizing false negatives is more critical in our work. False alarms are less harmful than missing the target class or misclassifying the data (i.e., classifying a negative class as positive). Therefore, we prioritize recall (i.e., sensitivity) over precision while keeping the balance between the two, including the F-1 score in the evaluation.

Ethical and legal aspects

This study has been carried out in accordance with the principles established in the EU Code of conduct for research integrity [162] and the recommendations provided by the EC's guidelines on the use of generative AI in research [163].

8.4 Results

We conducted the experiments in two main stages: dataset-related results and multimodal LLM-related results. The results of the experiments are presented in the order of these stages.

8.4.1 Open Source Models

Experimental setup

For all open-source models, we run the inference on a CUDA GPU A100, setting the *temperature* to 0.7 and the *top-p* value to 0.9. At each generation, we asked the model to generate a maximum of 512 new tokens (*max_new_tokens* = 512). For evaluation, we separately assessed the performance of the models with the original version of the DeceptiLens dataset[34], hereafter called **Dataset 1**, and our additions, hereafter called **Dataset 2**. We also calculated overall scores.

Baseline Results Across Datasets

We first tested open-source models with a basic prompt design, strategy (i). The mean scores calculated after five runs of each model are presented in Figure 8.3. While LLaVa v1.6 - Mistral 7B slightly outperformed Janus Pro 7B, the performance of LLaVA v1.5 7B is much lower in terms of accuracy and recall. Given that recall, i.e. sensitivity against dark patterns, is the most critical metric alongside accuracy, we focus on these two during the analysis of the results. Having less than 70% accuracy and recall indicates the limited reliability of the models in the task of detecting dark patterns.

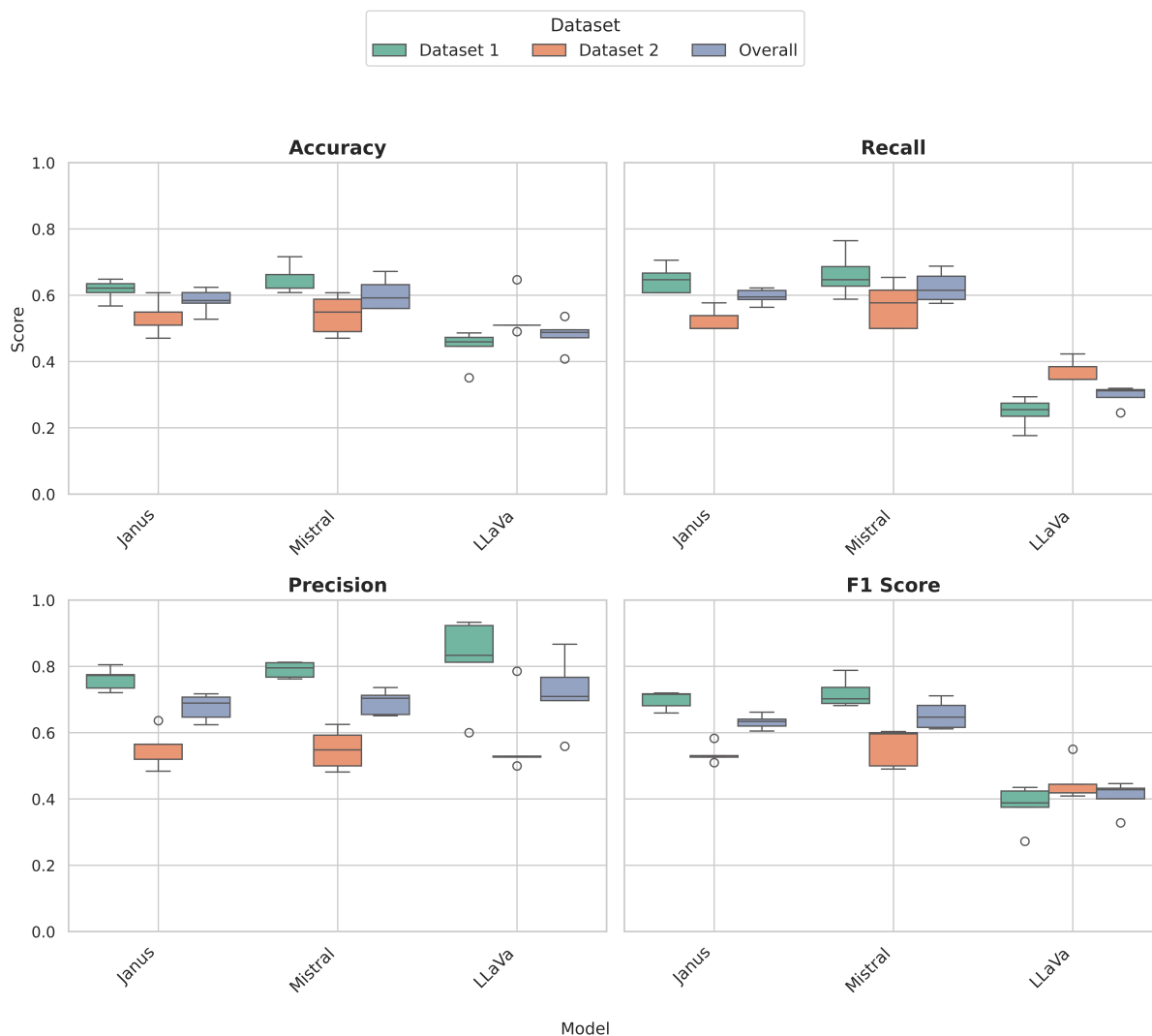


Figure 8.3: Baseline performance scores of open-source models with prompt strategy (i)

Implementation of Prompt Engineering Techniques

First, we embedded elements of **”contrastive learning”** into the prompt, as a method (ii) given in Section 8.3.3, and obtain the results shown in Figure 8.4. While there is a slight increase in the performance of Janus Pro 7B for both dataset 1 and dataset 2, the opposite is valid for LLaVa v1.6 - Mistral 7B. In general, the inclusion of **”contrastive learning”** elements alone did not cause significant improvement.



Figure 8.4: Performance scores of open-source models with prompt strategy (ii): basic + contrastive learning

As the prompting strategy (iii) given in Section 8.3.3, we implemented another important and widely used prompt engineering technique that works on the principle of step-by-step reasoning, “**chain-of-thought**”. The given results in Figure 8.5 show that the application of chain-of-thought without “contrastive learning” elements decreases the performance of the models. In other words, expanding the prompt with chain-of-thought compliant inputs introduced more noise than benefit to the models.

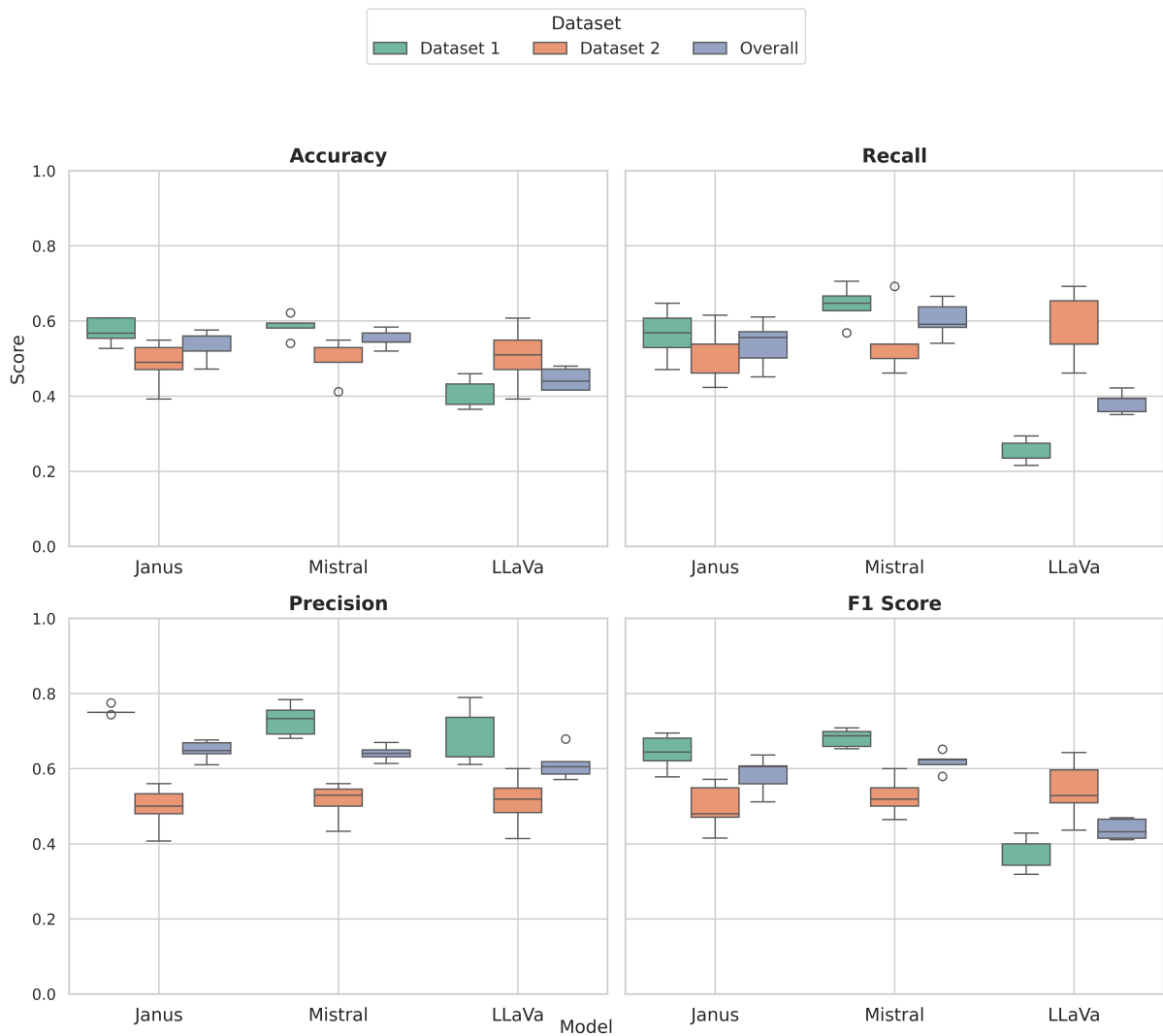


Figure 8.5: Performance scores of open-source models with prompt strategy (iii): basic + chain-of-thought

When we used the “contrastive learning” and “chain-of-thought” prompt engineering techniques together, which were previously evaluated independently, as prompting strategy (iv) given in Section 8.3.3, to see their combined impact on the models’ performances, we did not observe a significant enhancement as seen in Figure 8.6. On the contrary, a negative impact can be observed when compared to the “contrastive learning only” scenario, strategy (ii).

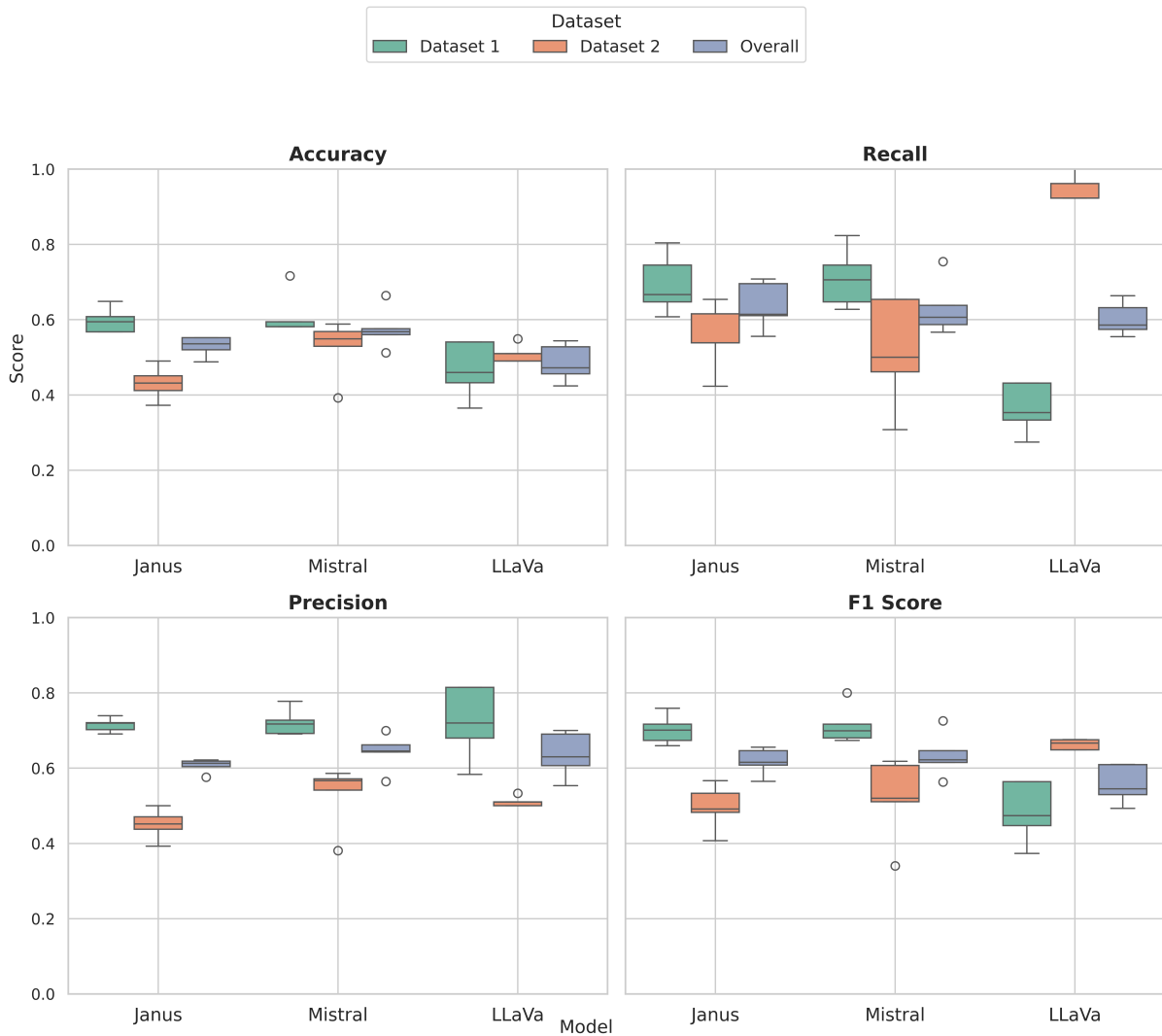


Figure 8.6: Performance scores of open-source models with prompt strategy (iv): basic + contrastive + chain-of-thought

Since we did not observe a significant improvement in the performance of the models with “contrastive learning” and “chain-of-thought”, separately or together, we tried another prompting strategy based on rules that our experts defined for each category, as the prompting strategy (v) given in Section 8.3.3. During the definition of the rules, the objective is to provide more concrete and observable features-based step-by-step category-specific reasoning. Although Janus Pro 7B slightly increased its performance with these expert chain-of-thought prompts, overall performances are more or less the same as those of the others, as seen in Figure 8.7.

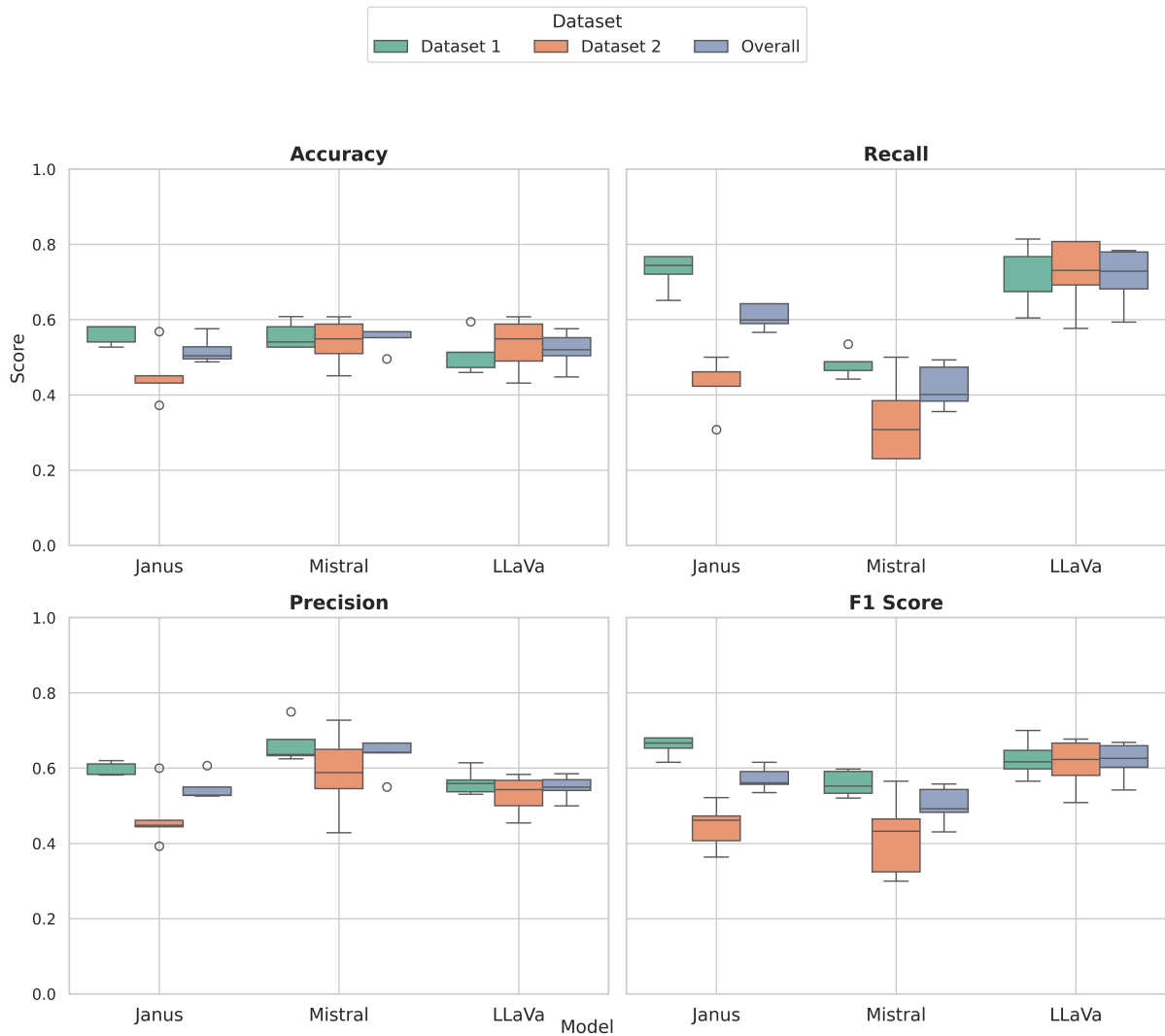


Figure 8.7: Performance scores of open-source models with prompt strategy (v): category-specific expert chain-of-thought

8.4.2 Comparison

We performed the same task with one of the best-performing proprietary LLMs, GPT-4o, for comparison purposes. The DP detection performance of this model is clearly better than all open-source models that we have evaluated, according to the results given in Figure 8.8. Furthermore, when the details of the performance of GPT-4o are inspected, in Figure 8.9, it is seen that prompt engineering techniques improved accurate and sensitive DP detection for GPT-4o compared to the baseline scores, i.e., basic prompting strategy.

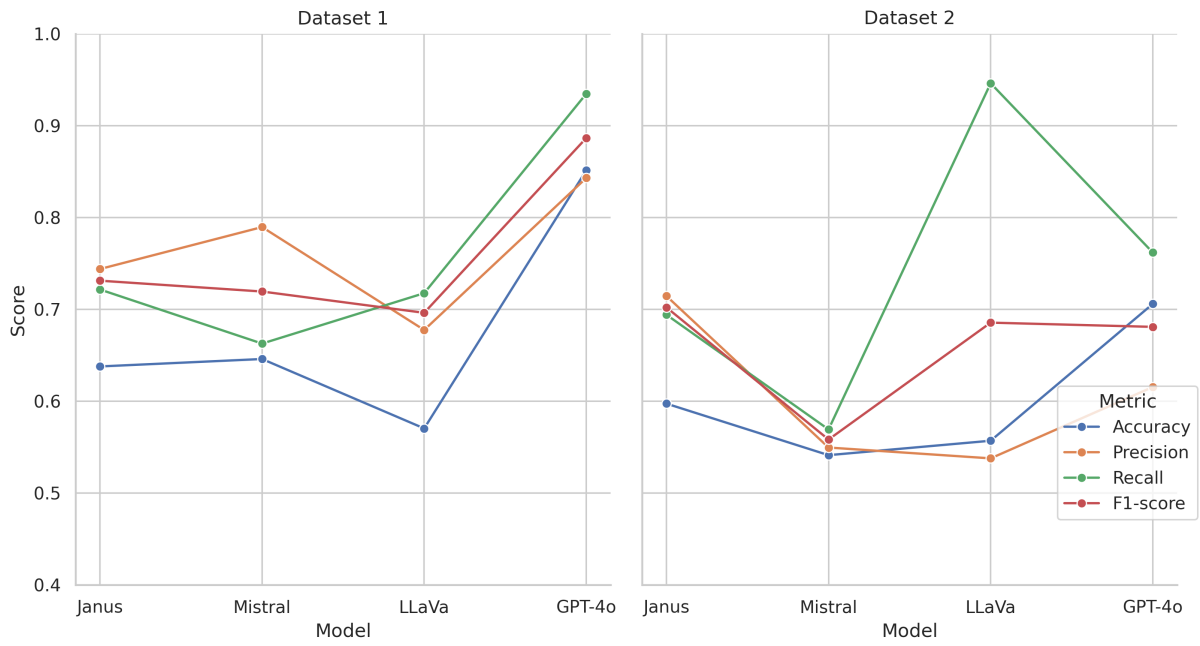


Figure 8.8: Comparison of the models' best performing scenarios

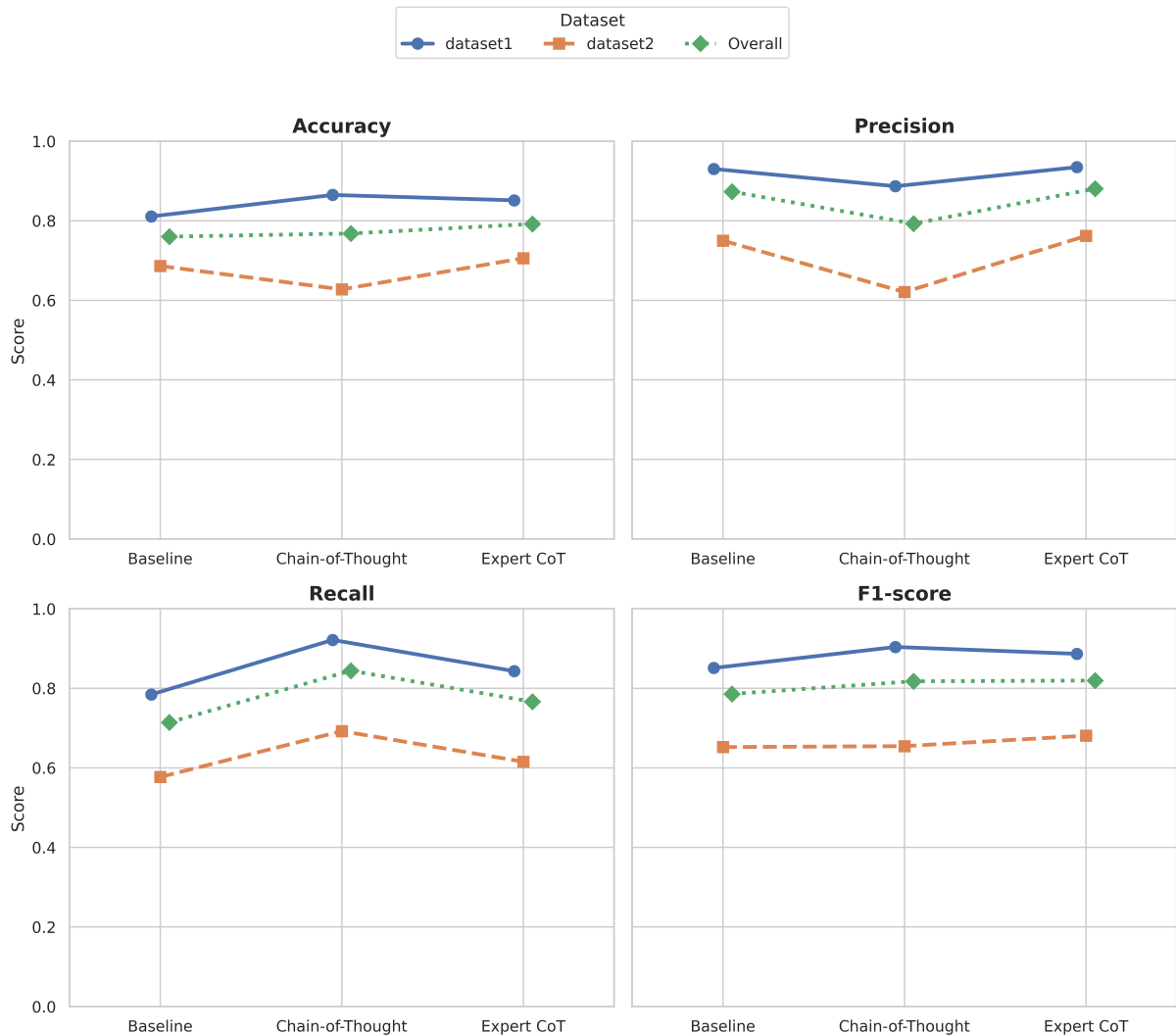


Figure 8.9: GPT-4o performance on both datasets

We also inspected the sensitivity of the best-performing model, that is, GPT-4o, for each DP category. We observed that, although the model did not miss certain DP categories, identification of some others was difficult, as seen in Figure 8.10.

8.5 Discussion

Detecting Dark Patterns (DP) is an activity that does not enjoy any crisp set of criteria even to verify whether a spotted potential instance is or is not part of the category. This is not only because gold-standard datasets are not yet available, a task which has its own challenge, but also because the decision on whether a design is a DP also depends on the context or additional information (see Section 8.2) that is not always available, or can be available, to the decision maker, and that can hardly be found in any dataset. This differs from other cases of deception detection, like phishing, where any user can verify whether an instance is a false positive or a

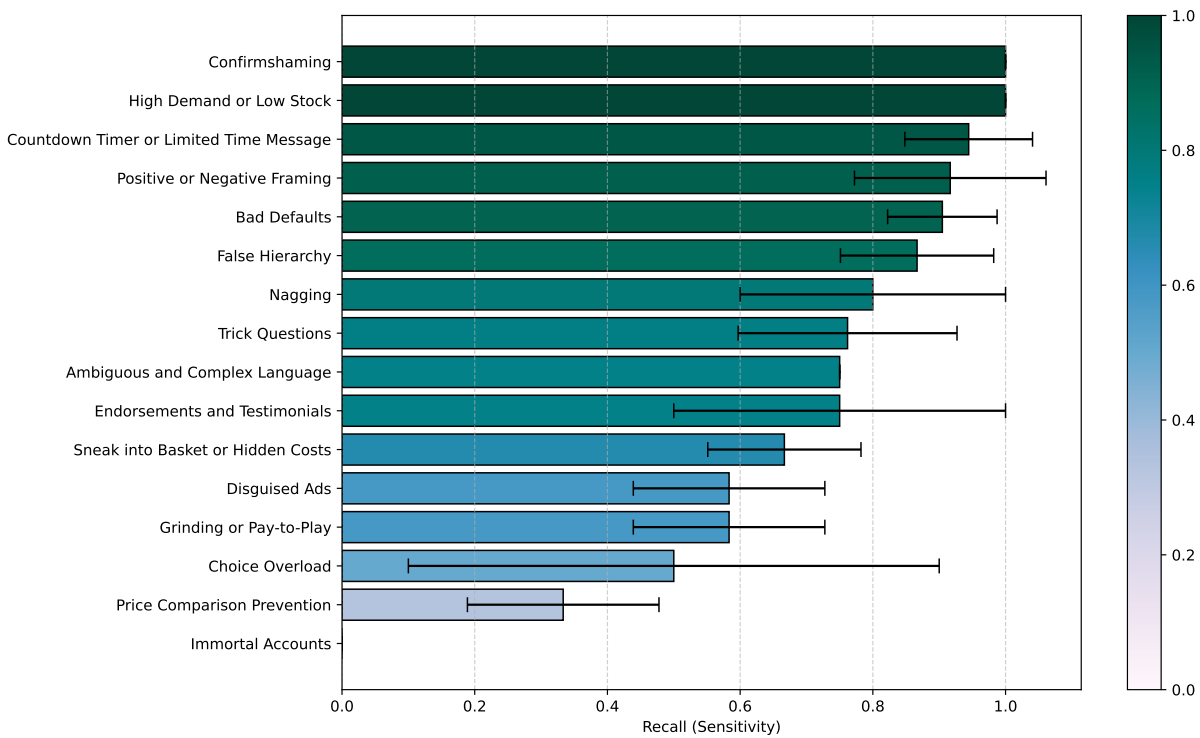


Figure 8.10: Dark pattern category-based sensitivity status of best performing model - GPT-4o

true positive with minimal effort and a bit of expertise.

For this reason, we believe that dark pattern detection tools should be featured as decision-support components, rather than independent decision-making instruments. For example, they could help researchers identify "suspicious" design patterns in large datasets to expedite their research tasks. These tools are also meant to aid independent authorities that perform large-scale sweeps of vast quantities of websites and digital services [201] in search of practices that circumvent the law. Among the so-called Enforcement Technologies that optimize such tasks, some models are purposely developed or fine-tuned to identify deceptive design patterns in websites and legal documents [173]. A few businesses are already developing solutions that audit organizations' interfaces in search of legal violations concerning their UI design choices². Once suspicious DPs are flagged, it is then the task of the human decision-maker to assess whether the flagged UI element is an actual deceptive design pattern, usually based on additional information (e.g., by screening the same website at different points in time) and knowledge (e.g., by comparing the actual data practices of an organization with its declarative practices on the UI). This is why we assume that even though the evidence is not supporting the detection decision with full confidence, raising a "potential dark pattern" flag (i.e., obtaining lower precision results) should be preferred to missing a positive class, which would signal an incapacity of identifying a dark pattern. Furthermore, adopting this decision-support role for DP detection

²See e.g., <https://www.fairpatterns.com/solutions/fairaudit-ai>

can help mitigate the risks associated with overly proactive AI-systems, which have been shown to introduce various challenges in human-AI collaboration [202].

That said, our attention was on understanding what current LLMs, and their multimodal versions, can offer in terms of decision support, and in particular by assessing open-source models *versus* the proprietary GPT-4o. Thus, we measured the effectiveness of *open-source* MM-LLMs in the task of detecting DPs from screenshots. With the prompt engineering strategies applied, e.g., Chain-of-Thought, GPT-4o achieved 86% accuracy and 93% recall, which we identify as the two primary metrics in DP detection (see Section 8.3.4 for details). On the other hand, none of the *open-source* MM-LLMs demonstrated performance over 65% accuracy and 75% recall, highlighting that they are not yet reliable decision-support tools in the same task.

We also noted that the same prompting strategies do not ensure the same quality of results when applied to different models. This increases the complexity of the discussion. This emerges clearly with the LLaVa v1.6 - Mistral 7B where our efforts to expand prompt with Chain-of-Thought introduced more noise than benefit, in neat contrast to other open-source models and GPT-4o. While LLaVa v1.6 - Mistral 7B uses the same MLP-based vision-language connector as LLaVA v1.5 7B, the increase in input resolution might have introduced representational shifts that could affect the connector's ability to retain class-relevant features, especially in longer context windows, possibly contributing to the observed performance difference in our classification settings.

Our findings are consistent with previous research that indicates that (i) while MM-LLMs achieve high accuracy in structured tasks, such as code generation, the models are still suffering in tasks requiring complex reasoning and do not even surpass 60% accuracy, (ii) a single prompting strategy does not consistently enhance the performance of all models across all tasks [203]. Thus, the lack of reason capacity of a model, also across modalities, cannot be overcome by changing the prompting technique. In our experiments, we also clearly noted that detection performance differs significantly across different DP categories (see Figure 8.10). Given this variegated nature of DPs, some improvement can be achieved by splitting the general task into sub-tasks, and by recombining the overall solutions that dedicated LLMs have found by working specifically on the sub-tasks (i. e., by following a Multi-Agent System approach).

8.6 Limitations

Assessing the performance of the open source model in DP detecting task by a screenshot of a web or mobile UI is a limitation of our approach because it is not reflecting full user journey and does not provide some meta data, such as backend, html, css. Although the complete user

journey is not needed for all DP types, such as *Confirmshaming*, certain DP types require this information, such as *Countdown Timer* which necessitates a couple of page refreshes to check if the timer resets itself each time. Another limitation of this study is the size and diversity of the DP datasets. A dataset enriched with alternative dark pattern cases will help measure the performance of the models more realistically. Moreover, data types other than text and images, such as audio and video, are not included.

8.7 Wrap-up and Perspectives

Detecting dark patterns, effectively and accurately, has been investigated by researchers, regulators, and UX designers. In particular, advances in technology, such as developments of LLMs, and legal arrangements, such as GDPR, have augmented these studies. However, we identified the need for a standardized benchmark dataset by demonstrating the issues present in previous work to enable a reliable evaluation of DP detection models. To address this need, we expanded a dataset that was originally annotated by 14 DP researchers and used a recent and unifying taxonomy [29] by including new dark pattern categories and published it according to the principles of accessible and findable data. In the second stage of our study, we examined the performance of multimodal LLMs in the task of detecting dark patterns using our benchmark dataset. Motivated by the growing and encouraging interest in open source AI, we conducted, to our knowledge, the first evaluation of the sensitivity of open source multimodal LLMs against dark patterns. It is seen that open source models perform poorly compared to latest proprietary models, such as GPT-4o, and implementation of various prompt engineering strategies, such as Chain-of-Thought, does not contribute significantly.

Our findings clearly show the need for an increase in data set size and diversity, a customization effort through fine-tuning the models rather than prompt engineering alone, and the employment of stronger models with a capacity for robust reasoning and visual alignment.

In general, the remaining open problem in the field of LLMs, and in particular, in the automatic detection of DPs, is to ensure not only the availability of open-source models but also the reproducibility, transparency, accountability, and reliability of the results. These are crucial interests that need to be safeguarded, especially in the European view of fostering open science, which in turn supports technological innovation. Pursuing such goals may also help the training of reliable and transparent DP detection tools that could be employed in various contexts, including business contexts. Furthermore, we encourage researchers and other stakeholders in the DP community to fine-tune the model(s) by expanding our dataset, which is accessible in

the repository of this work³.

³We will make the link publicly available here once the anonymity requirement has been lifted.

Chapter 9

Discussions and Closing Remarks with Future Work

This final chapter presents the principal conclusions of the PhD research, provides comprehensive discussions and includes a retrospective reflection on the core insights and challenges encountered throughout the study.

9.1 Why Detecting Dark Patterns Matter?

Detecting dark patterns matters because they are not merely superficial design issues, but socio-technical problems with legal and ethical consequences. This recognition forms the core foundation of this thesis. Online deceptive and manipulative designs have gradually increased their popularity since they were named dark patterns in 2010 [33]. While users share the problems they experience with dark patterns on social media, scholars in various domains such as user experience design and privacy have also begun to examine these patterns [204]. The academic perspectives contributed to better understanding and defining the characteristics of these problematic practices, which are recurring design patterns that include asymmetry, restriction, deception, or information hiding [18]. Moreover, dark patterns have attracted the attention of regulatory bodies [7, 8, 6] because they can lead to unlawful situations such as violations of personal data protection provisions, e.g. the GDPR consent, due to dark patterns' use in privacy-oriented decision processes [20]. The heterogeneity of approaches across different fields, domains, and stakeholders with distinct objectives has resulted in ontological confusion, i.e., not only in terminology but also regarding the definition and classification of dark patterns. As a result of this substantially evolving and fragmented landscape of the dark pattern field during the course of this work, **a primary objective of this thesis was to precisely define dark**

patterns and clarify their functional properties.

As our work into dark patterns deepened, we observed that the negative consequences directly or indirectly caused by dark patterns are diverse. These harms range from privacy violations and non-compliance with the GDPR requirements [16] to facilitating anticompetitive practices [205]. For example, they can manipulate users to make unintended purchases, leading to financial loss. While such practices may provide short-term benefits for designers or businesses, their long-term effects include a collective loss of trust in the market and reputational damage [206]. The complexity and multifaceted nature of these impacts is illustrated by categorizations across two dimensions: (i) individual versus collective, and (ii) material versus non-material [15]. These impacts are not entirely independent; rather they can interact with and reinforce one another. For example, compromised autonomy can be tied to addiction and privacy violations. Collectively, these demonstrate that dark patterns generate interconnected harms, such as individual and collective, material and non-material. These impacts can also be indirect and may not be immediately visible. For example, certain dark patterns can cause excessive and undesirable consumption and, indirectly, increase ecological deterioration [207]. Given the extensive and interconnected problems caused by dark patterns, **the ultimate goal of this thesis was to correctly detect these patterns** as a way of minimization of associated harm and risk.

The pervasive nature of these harms is exacerbated by several crucial factors that warrant careful consideration. First, dark patterns exploit a universal user vulnerability: virtually anyone who interacts with these designs can be affected regardless of their experience or knowledge [48]. Prior work revealed that reducing user effort, such as through single-click transactions, significantly increases users' susceptibility and strengthens the efficacy of the deception and manipulation, resulting in harmful outcomes like financial losses [208]. This demonstrates that deceptive design does not require high complexity; even simple implementations can be highly potent. Consequently, dark patterns can be deployed widely across diverse use cases with minimal investment, thereby amplifying their prevalence and aggregate impact. Despite the wide prevalence of such designs and their associated harms, these practices are often normalized due to the persistence of the status quo, which often translates into tacit acceptance by platforms and users [209]. Together with the extensive and interconnected harms outlined above, these factors (universal vulnerability, low deployment cost, and normalization) turns dark patterns into a problem whose magnitude exceeds the limited sources of researchers and regulatory bodies to manually investigate, identify, and analyze. As a result, there is a critical need for **effec-**

tive, automated, and scalable dark pattern detection methods. Accordingly, the main objective of this thesis is to develop reliable and scalable methods for this detection challenge. Furthermore, ongoing regulatory efforts, such as the proposed Digital Fairness Act, which explicitly identifies and targets dark patterns as unfair digital practices, demonstrate that dark patterns continue to be regarded as a significant and unresolved problem, and thus the critical need for effective detection methods remains an active concern.

Effective dark pattern detection is a critical mechanism for harm mitigation that serves the interests of multiple stakeholders. First, it directly aids individual users. Research shows that users who are aware of deceptive design are less likely to be manipulated by them [121]. Thus, one way of combating against dark patterns is to proactively alert or inform users to enhance their protection. However, relying solely on user-level awareness is insufficient as it does not guarantee immunity from manipulation [48] and is not scalable in the wide range of digital environments and use cases in which dark patterns may appear. This necessitates solutions that transcend individual vigilance and operate at a scalable, systemic level. Second, detection offers crucial support to researchers and regulatory bodies who monitor unfair and unethical digital practices, particularly through the use of Enforcement Technology [173, 210]. Scalable detection methods enable systemic examination and support creating a more equitable and dark pattern-free digital environment. For example, automated detection tools can be employed during compliance audits of a mobile application to identify manipulative and potentially illegal patterns, particularly those that violate legal requirements. Third, detection provides a mechanism for industry self-regulation. Since not all companies implement dark patterns intentionally, often following common, yet harmful, design practices, detection tools can assist companies that aim to build fair and ethical user experiences and preserve long-term user trust. Emerging services and consultancies also provide assistance to organizations transitioning from dark patterns to fair patterns in their design.¹ Ultimately, this multi-stakeholder relevance confirms that dark pattern detection is not just a technical endeavor but a complex socio-technical challenge demanding solutions that are scalable, transparent, and context-aware. These requirements fundamentally shaped the methodological and interdisciplinary approach taken in this thesis.

¹For example, these companies offer fair pattern services: <https://www.rsystems.com/dark-pattern-detection/> and <https://www.fairpatterns.com/solutions/fairaudit-ai>

9.2 Challenges of Detecting Dark Patterns

Developing reliable and scalable methods for dark pattern detection is nontrivial and involves several crucial challenges. First, dark patterns are not limited to a single digital environment; they can appear on websites, mobile applications, and platforms like voice assistants [111]. Because dark patterns can manifest in any user-system interface, their prevalence is projected to grow with the proliferation and diversification of these interaction points. Although traditional environments like online shopping platforms [19] and video games [2] are the most cited examples, emerging technologies, such as smart home devices, robotics, and other IoT systems [68], also represent new and expanding venues where dark patterns can arise. This cross-platform expansion presents a significant generalization challenge. Even a highly effective detection tool optimized for web environments requires significant methodological adaptation to extend its efficacy to interfaces like IoT, which possess fundamentally different interaction modalities, data access patterns, and technical constraints.

Second, dark patterns are not confined to a single use case; they exhibit high contextual variability, appearing across diverse domains such as cookie consent processes, online shopping, social networks, and many more.² To illustrate this variability, consider the dark patterns *bad default* and *nagging* across two different scenarios: a cookie consent interface versus an in-app purchase screen in a mobile video game. In the cookie consent setting, a pre-selected checkbox (*bad default*) may violate data minimization principles by unnecessarily collecting personal data or sharing it with third parties, resulting in privacy loss. Conversely, this same *bad default* pattern can be combined with *sneak into basket* to nudge players into making small, hard-to-detect additional purchases, leading to financial loss. Similarly, *nagging* may be used to push users toward an "accept all" option to maximize data collection in the first scenario, but it may also be employed to repeatedly request a five-star rating in the application stores. Thus, while the form of the pattern remains consistent, the underlying intention, semantics, and consequent harm differ significantly across contexts. This contextual variability must be carefully considered in detection tool development, as it dramatically increases the complexity of achieving reliable dark pattern detection across distinct use cases.

The third challenge arises from the inherent diversity of data modalities involved in dark pattern

²**Sneak into Basket** refers to a dark pattern that leads to unintentional purchase by adding unwanted items without user consent or explicit action [29].

design.³⁴ Dark patterns can be formed through multimodal elements, such as text, visual components, audio signals, or machine-readable code (e.g., HTML, CSS and JavaScript). For example, the *confirmshaming* pattern is detectable primarily from textual elements, whereas *false hierarchy* dark pattern relies on visual elements. In contrast, patterns like *immortal account* cannot be assessed using text or images alone; they require extraction and structured representation (e.g., JSON) of user journey information across multiple steps. The inherently multimodal nature of these patterns renders traditional ML and DL models inadequate since they are typically single task and modal-specific. The effectiveness, data processing pipelines, and overall applicability of such models are rigidly bound to the specific data modality on which they are trained, severely limiting their capability to generalize across diverse dark pattern manifestations.

The fourth and final challenge is that existing definitions of dark patterns often remain at a conceptual level rather than providing actionable and objective criteria. Over the years, numerous taxonomic studies have successfully classified these patterns and demonstrated their broad scope and application diversity. These descriptive studies have significantly enhanced the understanding of this phenomenon, yet they remain insufficient for establishing the objective criteria and operational descriptions required for automatic identification of dark patterns. Accordingly, given that reliable detection of dark patterns is crucial for mitigating the risks and harms of dark patterns, **the foundational objective of this thesis** was defined as the investigation and **development of instruments for the objective identification of dark patterns**.

Addressing the foundational objective, our initial contribution focused on cookie consent processes, one of the most ubiquitous dark pattern use cases. Our approach involved the examination of interaction flows: Human-Computer Interaction (HCI) (user-to-browser interface) and machine-to-machine (M2M) (interface-to-server) information flows were extracted and analyzed from real cookie consent examples. This data extraction was crucial for the operationalization of dark pattern attributes. For instance, the attribute of *asymmetry*, defined as "unequal user burden" [19], was formulated by measurable features such as "number of clicks required to perform an action". Through this process-based approach, a total of 31 distinct, measurable features were defined (see **Chapter 3**). While these features were limited to process-based metrics and did not include visual or text-based modalities, this work constituted a crucial initial step toward

³**False Hierarchy** refers to a dark pattern where an option, which usually does not serve the user's interest, is made more visually prominent than the other option [29].

⁴**Confirmshaming** refers to a dark pattern that uses emotional manipulation by framing one option in a shaming or guilt-inducing way, thereby leading users toward the option that benefits the designer [29].

establishing the required objective and actionable criteria within the scope of the thesis. The integration of remaining modalities was addressed in the subsequent research phase.

While some dark pattern types are widely applicable across various use cases, others emerge only within particular scenarios. During the examination of real cookie consent processes, we observed that certain dark pattern types, such as *cuteness at home robots*, were irrelevant to this specific use case, highlighting the critical role of contextual relevance. This finding underscored the need for a systematic and holistic approach for developing actionable and objective criteria for dark patterns through measurable features, as the required operational attributes are inherently context-dependent. We developed a four-stage framework for this purpose.

In the first stage (detailed in **Chapter 4**), we leveraged multidisciplinary expert validation, comprising experts from computer science, artificial intelligence, UX design, and law, to determine the subset of dark pattern types applicable to our target use case (cookie consent), finding high agreement that 17 of the over 40 dark pattern types in Gray et al.'s unified taxonomy [29] were relevant. Subsequent stages focused on (i) establishing expert-validated mappings between the dark pattern types (e.g., *Privacy Maze* and *False Hierarchy*) and their attributes (e.g., asymmetry, restriction, and covertness); (ii) identifying corresponding measurable linguistic and visual features (as a complementary to process-based features identified in our previous work); and (iii) reformulating the dark pattern attributes using these features.

Collectively, these systematic efforts directly addressed the foundational objective by: (i) proposing a systematic approach to extract objective criteria, i.e., functional properties, for detection; (ii) establishing expert-validated mappings between dark pattern types and attributes; and (iii) defining measurable features that operationalize dark pattern attributes for objective evaluation. The culmination of the studies detailed in **Chapter 3** and **Chapter 4** provided the answer to **RQ 1: *What are the measurable functional properties of dark patterns and the assessment instruments to define them?*** Answer: We determined that the measurable functional properties of dark patterns are quantifiable proxy elements that may take textual, visual, or process-based forms depending on the design of the dark pattern, its context, e.g., cookie consent or online shopping, and its environment, e.g., web or mobile. For example, one common dark pattern attribute in cookie consent process scenario is asymmetric choice. This can be defined through multiple measurable properties: unequal numbers of clicks required to select different options (a process-based feature); unequal visual prominence of options, such as

differences in button size or contrast (a visual feature); and unequal sentiment framing, where the option favoring the service provider is positively framed while the privacy-protective option is negatively framed (a textual feature).

Importantly, the context in which these properties appear plays a critical role when formulating measurable features, as the same property may have different implications depending on the specific use case and environment. Ultimately, these features are intended to eliminate subjectivity and ambiguity in dark pattern definitions, thereby assessing. However, determining the correct formulation of a dark pattern remains a research challenge in its own right. For example, the *choice overload* dark pattern can be objectively represented using the measurable features, such as the maximum number of user actions (clicks) to perform all possible scenarios which are operational and actionable. Nevertheless, the threshold for what constitutes the dark pattern may vary: should it be defined as more than three options, or should it follow cognitive memory limits such as seven items? Such parameters are not fixed and may depend on contextual and empirical considerations [211]. The principal goal is to establish objective criteria that can serve both as assessment tools and as input to detection models, thereby supporting the explainability of overall detecting mechanisms.

9.3 Construction and Validation of Dark Pattern Datasets

In this thesis, whose ultimate goal is the effective detection of dark patterns, a second major challenge emerged: the absence of a gold-standard dataset suitable for training, testing and reliably evaluating the developed methods or AI-based detection models. Beyond this fundamental absence, the limited number of available datasets [78] suffer quality issues, including inconsistent labeling and overly broad category definitions, thereby hindering reliable benchmarking of existing detection solutions. Accordingly, the second research objective of this thesis was defined as **creating a structured, reliable and expert-labeled dark pattern dataset**. To mitigate this critical limitation, researchers typically employ two strategies: (i) the collection of new data and (ii) the augmentation of existing datasets. In this thesis, we performed both strategies.

In **Chapter 5**, we initiated our work by reviewing existing datasets and identifying several critical issues that impede their utility and reliability. These issues include taxonomy misalignment and the use of disparate dark pattern terminology and descriptions, which cause labeling inconsistencies across studies. Furthermore, we found instances of overly broad labeling, such as one dataset [35] where a small close button was deemed sufficient for a dark pattern label,

despite complete disagreement among dark pattern researchers who attended our experiments. Such mislabeling and reliability issues risk misleading detection models and invalidating realistic evaluations. Beyond reliability issues, the datasets also suffer from technical defects, including missing values, duplicate entries, and severe imbalances in dark pattern category distribution. This identification of flaws is particularly crucial as these datasets are still used by other researchers who assume they are representative and reference resources.

To mitigate this critical limitation, and as defined by **Research Objective 2**, we addressed this gap by curating a expert-validated, balanced, Graphical User Interface (GUI) dataset that consists of screenshots of web and mobile interfaces. The annotation process involved experts who had published a minimum of one peer-reviewed publication in the field and the majority vote approach, resulting in a robust dataset encompassing 21 categories of dark patterns listed in [29], including false positive examples. In parallel to the creation of the GUI dataset, we developed a text-based data augmentation framework. This framework was initiated starting with the merged form of two existing datasets [19] and [94]. Since standard augmentation techniques (e.g., synonym replacement or random deletion) risk altering the deceptive nuance inherent in dark pattern language, we leveraged the advanced natural language processing capabilities of LLMs. To mitigate the risk of losing the deceptive characteristic during augmentation, a multi-agent architecture consists of *Generator* and *Controller* agents was developed. The *Controller* agent served as a guardrail, verifying that the augmented data preserved semantic and structural fidelity to the original text. The resulting augmented corpus was evaluated both for diversity via expert annotation of random samples and through a comparative fine-tuning experiments. LLMs were fine-tuned using the augmented data and evaluated against models trained solely on the original data. The results (detailed in **Chapter 6**) indicated a significant improvement in detection performance following the fine-tuning on the augmented data.

In pursuit of the second research goal, the following contributions were delivered: (i) a novel data augmentation method for dark pattern text data; (ii) a reliable, expert-validated dark pattern dataset; and (iii) an empirical demonstration of critical quality issues that are previously under-explored in existing dark pattern datasets. These contributions collectively provided the answer to **RQ2: *What considerations and methods should be taken into account when building labeled and structured dark pattern datasets?*** We conclude that labeled and structured dark pattern datasets are essential prerequisites for the reliable evaluation of detection tools and the training AI-based detection models. While the quality of the labeling process is

fundamental to all supervised learning AI tasks, dark patterns introduce peculiar and acute challenges. For instance, labeling clear, visually stable concepts such as animal image classification, or established domains like phishing emails classification, are comparatively straightforward. In contrast, dark pattern labeling often allows ambiguity and subjective interpretation due to abstract, high-level descriptions, taxonomic heterogeneity, or the lack of data availability, context and meta data. Consequently, even researchers demonstrated 71.62% consensus on annotation, as we showed in Chapter 7. Therefore, one important consideration in the dark pattern dataset building process is the rigorous labeling process and the source reference descriptions used for each dark pattern category. Label descriptions must be up-to-date and grounded in established taxonomies to prevent inconsistencies or mislabeling.

A second aspect arises from the lack of consensus and standardized criteria in the field: the labeling process should be conducted by domain experts. While we accept that precisely defining experts is difficult, until a set of actionable and objective criteria is available, a few hours of training are insufficient to ensure a reliable and consistent labeling process. We defined experts as researchers with at least one peer-reviewed publication on dark patterns in **Chapter 7**, this definition must be adapted to the scope of the dataset. For example, if privacy-violating dark patterns and associated legal provisions are examined, the expert definition should be updated to include legal specialists. Finally, as we demonstrated by the findings in **Chapter 7**, expert annotators may have different opinions on the presence of dark patterns. Therefore, utilizing a mechanism that leverages multiple expert evaluations for each instance, e.g., using the majority voting approach in binary classification, throughout the labeling process is crucial for the reliability of dataset. This approach provides a distinction between consensus ground truth (high confidence) and majority-voted ground truth (standard confidence). This methodical differentiation between consensus ground truth and majority-voted ground truth represents a key novelty of our work and a transferable methodology that future researchers can adopt to improve data set reliability and guide systematic expansion. This distinction is also important when evaluating the dark pattern detection models. Our findings showed a significant difference in performance metrics, e.g., the recall score achieved by **DeceptiLens** was substantially higher when evaluated against the consensus data than when evaluated against the majority-voted data. This discrepancy highlights the critical need for transparency and rigor documentation in dataset construction, as it impacts the perceived efficacy of detection solutions.

9.4 Developing An Automated Dark Pattern Detection Tool

Finally, the third research objective focused on developing an expert-validated, automated, and AI-based dark pattern detection tool. This required the development of **multimodal AI-based automated detection methods for effective dark pattern identification**. Existing approaches include rule-based techniques, such as text matching, and AI-based methods, including machine learning and deep learning, are limited. Specifically, the multimodal nature of dark patterns requires detection systems to handle both textual and visual information; otherwise, models may detect only text-based dark patterns, while failing to capture even the simplest and most obvious visual patterns. To address these limitations, we used the advanced language understanding, generation, and multimodal capabilities of state-of-the-art LLMs. In this context, the work presented in **Chapter 8** developed **DeceptiLens**, a Retrieval Augmented Generation (RAG)-based framework centered on a multimodal LLM (MM-LLM).

The RAG process was adopted primarily to mitigate hallucinations, i.e., incorrect classifications or fabricated explanations generated by LLMs. Being trained with publicly available large data is not sufficient to deal with particularly domain-specific problems and one of the recent, less costly, and most successful technique is RAG. To this end, we first constructed a vector database composed of reliable sources that can be used in dark pattern evaluation, such as [18] and [29]. Next, the evaluated user interface was converted into vector form using an embedding model, and the most relevant pieces of text were retrieved from the trusted sources based on similarity. These sources can be adapted depending on context and user expectations. For example, if the focus is on dark patterns that violated the GDPR, the vector database can be reconfigured with legal sources, allowing the proposed framework to be applied to other problem domains as well.

The classifications of **DeceptiLens** were evaluated by 14 dark pattern experts who come from various backgrounds such as computer science, HCI, UX design, privacy, and law. Each instance was labeled by three different experts by majority voting. This process provided two important insights. First, the varying inter-rater agreement among experts revealed that expert opinion can differ significantly on the presence of dark patterns. If human experts cannot achieve full alignment, expecting an AI model's classification to align perfectly becomes unrealistic. This empirical variation strongly reinforces this thesis' first learning, which emphasizes the critical need for objective criteria of dark patterns and demonstrates how our contribution (see listed contributions given at the end of Section 9.2) of proposing operationalizable measurable features

is important.

The majority vote allows us to see confidence of experts in each instance and the second important finding related to this was the observed performance difference of the data annotated with full confidence versus the ones annotated with majority voting, i.e., one expert disagree with the majority, so less confident annotation. The proposed AI-based framework achieved 100% sensitivity (i.e. recall score) in instances where experts were fully agreeing. Conversely, the sensitivity dropped to 84.48% when measured against the majority-voted ground truth. This performance distinction indicates that our AI-based solution can work perfectly in cases where human interpretation is unanomous, but not surprisingly, their performance degrades in areas where human disagreement is high. This confirms that one of the major limitations of AI-based solutions is not algorithmic failure, but lack of standard in human-defined ground truth.

In addition to classifications, our proposal generated structured explanations detailing its reasoning, and the experts reviewed the quality of the explanations too. We demonstrated that our approach can detect certain dark patterns with high accuracy while providing transparent, explanatory outputs that support interpretability of the detection process. This transparency is critical because automated detection tools are not considered as fully autonomous. The final decision, particularly when the confidence score of AI is low, rests with a human verifier.⁵ The explanations provide necessary evidence and justification to user in order to enhance decision-making process.

The measurable features concept developed in the earlier stages of this thesis (see **Chapters 3 and 4**) was incorporated into the model explanations, enhancing their clarity, correctness, completeness, and verifiability. Furthermore, we embedded the source of the information in these explanations in order to increase verifiability and decrease uncertainty [151].

In addition to a quantitative evaluation of both classification and the quality of explanations by experts, we conducted semi-structured interviews to gather feedback on the overall assessment, usability of our proposed system and understand the reasoning of experts while evaluating user interfaces to check whether they have dark patterns or not. The experts found the explanations to be clear, complete, and mostly verifiable. They also provided specific recommendations that are crucial, as they define the necessary next steps for advancing the state-of-the-art and in-

⁵”A computer can never be held accountable, therefore a computer must never make a management decision.”
IBM Training Manual, 1979

creasing the usability of such tools from a researcher’s perspective. The interviews, specifically focusing on the perspective of dark pattern researchers, provided insights into the demands of the field. However, future work could explore the expectations and needs of a more diverse set of interviewees (e.g., industry professionals, regulators), which would likely present different priorities and potential future research directions for dark pattern detection technology. The recommendations included the adoption of a consistent language and reference format, and the presentation of content based on user roles. Furthermore, they explicitly stated that certain dark pattern assessments necessitate access to back-end or meta data, such as HTML or complete user journey, respectively, rather than relying solely on screenshots to render a confident decision. This work is novel due to its rigorous evaluation by 14 dark pattern experts, making it unique within the current literature.

We used GPT-4o, which is a proprietary pre-trained multimodal LLM, as the core model in our DeceptiLens framework because of its state-of-the-art image-to-text, language generation and understanding capabilities and was a leader in multimodal benchmarks [156]. The academic research community has increasingly called for the development of open-source LLMs and often favors them over proprietary models. However, the performance of open-source multimodal LLMs (MM-LLMs) is inferior compared to proprietary models [212, 213] and the performance of open-source models’ in dark pattern detection had not been examined. Within the scope of this thesis, open-source models were tested in dark pattern identification and empirically demonstrated the current status of them. Several prompt engineering strategies, such as Chain-of-Thought and contrastive learning, has been applied during the experiments. As detailed in **Chapter 9**, the results indicate that the performance of these open-source models is substantially lower compared to proprietary models in consistent with the benchmarks in other evaluations and highlight the need for further development in dark pattern detection.

The findings of this thesis answers ***RQ3: Which AI-based approaches can be developed for the automatic detection of dark patterns while providing explanations supporting informed decision-making, and what challenges and potential solutions do they involve?*** Multimodal AI models, characterized by advanced language understanding and generation, can address the limitations of traditional AI approaches, such as ML and DL models, which are typically single-task specific, lack multimodality, and possess limited capabilities for understanding the sophisticated semantics (i.e., complex meaning and hidden implications of the language) and nuance (i.e., subtle differences or context-dependent meaning) critical for accurate

dark pattern detection. Multimodal LLM-assisted decision-making processes for dark pattern assessment are capable of generating correct, complete, clear, and verifiable explanations, which thereby reinforces decision makers via evidence-based reports. However, given that access to complete data is a significant challenge, positioning automated AI-based detection solutions as the final decision-maker is neither practical nor advisable. Instead, these systems should be positioned within a human-in-the-loop concept to support human oversight.

All code and datasets developed in the scope of this thesis have been shared in accordance with the FAIR principles. The concrete measurable feature list, the expert-labeled dataset, the RAG-based LLM code, and the developed methodological approaches constitute significant contributions to researchers in the field of dark patterns. Furthermore, we emphasize that models developed for dark pattern detection should be positioned as decision-support tools, as humans ultimately make the final decisions; therefore, transparent and explainable tools are crucial.

9.5 Limitations, Open Problems and Future Work

9.5.1 Moving Beyond Static Visual Analysis: Developing Agentic AI-based Detection Systems

We demonstrate that certain dark pattern types, such as *confirmshaming*, *high demand*, and *low stock*, can be readily detected by LLMs, and other categories, e.g. *Immortal Account*, *Sneak into Basket*, require additional contextual information and difficult to identify with high confidence by static visual analysis. For instance, accurately detecting the *Sneak into Basket* dark pattern type necessitates contextual details such as information presented on previous pages and the country in which the purchase occurs, which are not available in a screenshot alone. In our research, we focused on the graphical interfaces of the mobile and web platforms, with screenshots used as the main basis for the analysis, and this is one of the main limitations. To deal with this limitation, an **agentic AI**⁶ approach should be explored particularly for detecting these types of dark patterns. Unlike standard automation, agentic AI can interact with each website or application considering its own specific conditions, enabling a more comprehensive assessment whether dark patterns are present by surpassing the limitations of single-screenshot analysis. In particular, domain specific agents, e.g., online shopping agent, or dark pattern type specific

⁶**Agentic AI**: represents a new paradigm in artificial intelligence, aimed at solving complex problems autonomously. Its primary purpose is to make decisions with minimal human input [214]. Rather than relying on a single LLM, each subtask can be assigned to specialized LLMs or other types of AI models, often coordinated by a central orchestrator LLM.

agents, e.g., *Immortal Account*⁷ controller agent, can be built under the orchestration of core AI model capable of assessing the data provided by these agents in order to decide whether dark patterns are present. Furthermore, the logs, recordings, and metadata gathered by these agents can contribute to improving the transparency and explainability of the dark pattern detection solutions.

9.5.2 Need for Objective Criteria: Towards Interdisciplinary Formalization

Although agentic AI approaches can contribute significantly to greater data collection and inclusion, establishing objective criteria to assess the presence of dark patterns remains essential in the decision-making process, regardless of whether it is conducted by human experts or advanced AI-based automated systems. A set of objective criteria for each dark pattern category is necessary, not only for detection methods but also for providing practical guidelines for designers. While this thesis contributes to the need for objective criteria through our measurable features and systematic approach (see **Chapters 3** and **Chapter 4**), establishing comprehensive and widely accepted objective criteria requires further research and broader collaboration. For instance, a semi-structured controlled language [115] could be developed for dark pattern definitions to eliminate ambiguity and subjective interpretations. Natural language expressions offer richness and creativity, but they often lack the precision needed for objective assessment. The measurable features that we identified can provide this precision and may be integrated into a controlled language specific to dark patterns. It is critical to consider all stakeholders in the dark pattern domain when developing such tools to ensure the new language is applicable and usable. Pure technical languages or formal logic can achieve high precision, however, their complexity and low readability for non-technical stakeholders can hinder usability. Given the interdisciplinary environment of the dark pattern field, a balance between understandability and precision is necessary. Controlled Natural Languages may provide a suitable solution, as they combine the clarity of natural language with a level of formal precision, making them both accessible and machine-readable.

9.5.3 Performance Enhancement Strategies for LLMs

LLM-based solutions naturally inherit the limitations of LLMs; consequently, these restrictions must be carefully considered and proactively mitigated during the development of an AI-based

⁷**Immortal Account**: refers to designs where user account deletion is impossible or too difficult [120].

dark pattern detection system that incorporates LLMs. A fundamental limitation of LLMs is that they generate output based on the likelihood of textual continuation, resulting in not being factual and the inability to represent logical connections, which requires a dedicated verification stage [215]. The limitations and risks associated with LLMs extend beyond factual unreliability to include biases inherited from the training data [216] and risks of privacy violation via prompting strategies [217]. Research on developing techniques to minimize these risks has introduced several methods such as advanced prompting techniques, RAG, and fine-tuning. No single technique offers an optimal solution for all aspects, as each has distinct advantages and limitations. In this thesis, we used the prompting and RAG because they do not require retraining and updating model parameters and are therefore substantially more cost-efficient. Moreover, given the scarcity of reliable labeled data in the dark pattern domain, prompting and RAG implementation are particularly suitable, since they do not necessitate large training datasets. RAG is additionally advantageous because it enables the system to store and use relevant long-form contextual information in pipeline without being constrained by model's context window. However, these techniques also have limitations. Prompting is often effective when underlying model already possesses sufficient domain-relevant knowledge; otherwise, its performance deteriorates especially in niche or specialized tasks [218]. On the RAG-based solutions, more substantial improvements can be achieved, but RAG also introduces its own limitations and inefficiencies. These include challenges related to similarity-scoring performance, vector-store scalability, search optimization, memory utilization etc. [219]. We believe that further research that focuses on advanced RAG methods can enhance the performance of DeceptiLens. The recent advanced RAG-based techniques are particularly crucial in the dark pattern detection to guarantee factual verifiability. Fine-tuning also remains an active further research area in dark pattern field, but it requires high-quality and sufficiently large datasets. Our expert-validated dataset can serve as a reference set for constructing future fine-tuning corpora. Finally, given the rapid evolution of AI, incorporating recent advancements (whether in model architectures, fine-tuning approaches or prompt engineering techniques) is an open research area for the improvement of existing solutions.

9.5.4 Open versus Proprietary LLMs for Dark Pattern Detection

Regarding AI models, specifically LLMs, another important consideration is that some models are "open" while others are not, i.e., proprietary. In the context of LLMs, i.e., foundation models, the term of open can have two different meanings: only model weights publicly available or

all other details such as model architecture, training data etc., are available [183]. As highlighted in **Chapter 9**, there is an encouragement to utilize and develop open models due to various reasons such as privacy [220] or research oriented concerns [221] and to support transparency, re-usability, and extensibility [222]. Importantly, open projects can also be more easily adopted by public administrations and regulatory bodies. Some open-source LLMs can provide results comparable to closed models in certain tasks [223]. However, as we illustrated in **Chapter 9**, the performance gap in the dark pattern detection task through graphical user interfaces remains significant. This underscores the need for further research efforts in the development of open-source AI models. In addition, open projects benefit from community support and contributions, enabling faster and more effective progress in improving model capabilities, transparency, and adaptability.

9.5.5 Dark Pattern Dataset: Establishing a Collaborative Platform for Dataset Expansion and Expert Validation

Another limitation of the thesis is the size of the dataset. Expanding the number of categories, as well as increasing the quantity of samples within existing categories, would enable more comprehensive and reliable evaluations. This limitation also points to a direction for future work. Specifically, **there is a need for a collaborative platform through which dark pattern instances can be collected, reported, and systematically evaluated**, and ultimately supporting the development of an up-to-date, accessible, and reliable database that has more coverage of dark pattern categories and more samples for each to increase sample diversity. Ideally, such data could be made publicly available with an open license so that activities that rely on high-quality datasets can advance. However, the labeling process for reported instances should be carefully designed and carried out by verified expert user profiles. As demonstrated in this thesis (see **Chapter 8**), even experts may hold different opinions on the same instance. Certain dark pattern categories also require assessment by trained experts, particularly considering the breadth and complexity of taxonomies, terminology, and required domain expertise. A majority voting approach could be adopted to enhance labeling reliability.

9.5.6 Where and How Dark Pattern Detection Tools Should Be Used

During the evaluation of dark patterns, high-confidence decisions cannot always be made due to various reasons. For example, a *low stock* dark pattern may be accurate, but understanding

it fully may require access to the service’s server or database. Although the available data does not allow for 100% certain evaluation, it has been deemed appropriate for dark pattern detection tools to operate within a human-in-the-loop decision support process, alerting users to potential dark patterns rather than making fully autonomous decisions. At this stage, user characteristics and expectations can also be considered to develop user-specific tools. For instance, a system could integrate GDPR-violating dark patterns or a Digital Fairness Act compliance checker, distributing detected violations across a report to provide actionable insights. We can envision two scenarios in which the detection tool is utilized. In the first scenario, an official responsible for GDPR compliance could perform more efficient operations by filtering content and enabling scalable monitoring. Beyond simple detection, a tool that supports transparency and explainability—providing a report detailing which GDPR provisions may be violated and why—would significantly facilitate the work of decision-makers. In the second scenario, the detection tool could serve as an internal testing resource for companies aiming to design ethically responsible systems, maintain long-term customer satisfaction, and build trust. In this context, evaluations would need to consider not only GDPR compliance but also user-friendliness, manageable user load, and overall usability. For example, developers or testers would understand that a notification related to security settings would not constitute a nagging dark pattern, whereas preselected notifications for advertising purposes would be considered an unethical practice. While the final decision remains with the user, the tool can highlight potential risks considering the user expectations and the related context.. Through collaborative efforts, such tools can contribute to creating fair digital experiences that minimize or eliminate dark patterns.

9.6 Closing Remarks

Dark patterns are socio-technical problems and their mitigation strategies require interdisciplinary collaboration. Because of information asymmetry between system designers and users, and the potential gains derived from guiding user decisions, dark patterns will likely persist as an enduring problem within digital interfaces. This could happen through conscious design choices or the adoption of established common industry practices. The emergence of new dark pattern types is also inevitable as technology evolves and interaction with digital interfaces diversifies. Collective efforts (spanning academic studies, regulatory guidance, seminars) to understand, define, and raise awareness against dark patterns have achieved notable progress, and they are crucial for mitigating these harmful patterns. When I began my research, the number of researchers familiar with dark patterns was quite limited; yet, the count of raised hands responding to the question (*"Has anyone heard of dark patterns before?"*) during my conference

presentations has increased dramatically over just a few years. However, the first core learning from this research is that **educational and awareness-based efforts constitute only one leg of the fight against dark patterns**. It is undeniable fact that even highly educated individuals can fall into traps set by dark patterns because they exploit human vulnerabilities and weakness. Therefore, effective mitigation requires solutions that transcend individual awareness.

The problem is fundamentally socio-technical, and it necessitates an interdisciplinary collaboration that integrates social and technical approaches. Given my background in computer science and artificial intelligence, I focused on developing technically grounded and scalable solutions, accepting that the whole solution landscape is much bigger. Ultimately, these **technical solutions must incorporate the perspectives of non-technical stakeholders in dark pattern field and avoid isolation from the contextual and social aspects of digital design** in order to be effective and widely adopted.

Expertise in the dark pattern domain is critical for developing effective and robust dark pattern-targeted technical solutions. An important insight from our work is that a lack of domain knowledge and expert validation cause major problems that lead to unreliable and not realistic solutions. As detailed in **Chapter 5**, pervasive quality issues such as overly broad labeling, mislabeling, and inconsistent label descriptions are indicators of the absence of rigorous interdisciplinary collaboration and nuanced understanding of deceptive designs. A notable example that is highlighted multiple times in this thesis, is the practice of labeling mobile user interfaces with small close buttons as dark patterns.

A uniqueness and novelty of this thesis lie in its interdisciplinary nature. Rather than approaching the problem solely from a computer science perspective, the work integrates insights from AI, HCI, UX design and law. Instead of uncritically adopting existing datasets or relying exclusively on LLM-based detection, which risks producing results misaligned with real world phenomena, the thesis grounds its methods in domain expertise and systematically examines the reliability of labels and models. Additionally, by empirically evaluating the performance of state-of-the-art LLMs, the research provides valuable insight into the current capabilities and limitations of open-source models for dark pattern detection. Beyond contributing to dark pattern research, the methods and insights presented in this thesis are applicable to adjacent domains, such as the forthcoming Digital Fairness Act (DFA) and similar regulatory frameworks. This work illustrates how objective, operational criteria and explainable detection mechanisms can be developed in those domains where ambiguity and subjective interpretation hinder evaluation.

In this thesis, three main research goals were established with the aim of mitigating dark patterns (online deceptive and manipulative designs), which cause both individual and collective harms as detailed in Section 9.1.

- **Research Objective 1:** *Defining measurable instruments for the objective assessment of dark patterns.*
- **Research Objective 2:** *Creating a structured and reliable labeled dark pattern dataset.*
- **Research Objective 3:** *Developing an expert-validated, automated and AI-based dark pattern detection tool which supports transparency.*

The work has produced several contributions through an interdisciplinary approach. First, the thesis identifies the pressing need for objective criteria for the assessment of dark patterns, and proposes operationalizable and measurable features to address this gap. Second, it identifies significant quality issues in existing datasets, such as labeling inconsistencies and inaccuracies that cause unreliable evaluations, and addresses this gap by creating a new expert-labeled dataset and developing dark pattern data augmentation method. Building on these foundations, the third research contribution focused on the development of an automated detection tool. By leveraging multimodal LLMs and recent techniques such as in-context learning and RAG, the thesis demonstrates that dark patterns in graphical user interfaces can be detected at an expert-level accuracy while generating correct, clear, complete, and verifiable explanations that informs users about dark pattern assessment process.

Bibliography

- [1] Tim Kollmer and Andreas Eckhardt. “Dark patterns”. In: *Business & information systems engineering* 65.2 (2023), pp. 201–208.
- [2] José P. Zagal, Staffan Björk, and Chris Lewis. “Dark Patterns in the Design of Games”. English. In: *Foundations of Digital Games 2013*. Conference paper, Refereed. Available from: 2016-10-31. RISE, Swedish ICT. RISE, Swedish ICT, Interactive Institute. GAME., 2013, p. 7. URL: <http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-24252> (visited on 12/01/2020).
- [3] Corina Cara et al. “Dark patterns in the media: A systematic review”. In: *Network Intelligence Studies* 7.14 (2019), pp. 105–113.
- [4] Christoph Bösch et al. “Tales from the dark side: privacy dark strategies and privacy dark patterns.” In: *Proc. Priv. Enhancing Technol.* 2016.4 (2016), pp. 237–254.
- [5] Colin M Gray et al. “The dark (patterns) side of UX design”. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–14.
- [6] Federal Trade Commission. *Bringing Dark Patterns to Light*. Sept. 2022. URL: https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report%209.14.2022%20-%20FINAL.pdf.
- [7] OECD. *Dark Commercial Patterns*. <https://www.oecd-ilibrary.org/docserver/44f5e846-en.pdf?expires=1707456299&id=id&accname=guest&checksum=063A07EB53611E9EB1941F5DECEF38C3>. No 336. Oct. 2022.
- [8] EDPB. *Guidelines 03/2022 on Deceptive design patterns in social media platform interfaces: how to recognise and avoid them - version 2.0*. https://edpb.europa.eu/system/files/2023-02/edpb_03-2022_guidelines_on_deceptive_design_patterns_in_social_media_platform_interfaces_v2_en_0.pdf. [Accessed 10-11-2023]. 2023.

- [9] EDPB. *Guidelines 03/2022 on Deceptive design patterns in social media platform interfaces: how to recognise and avoid them*. https://www.edpb.europa.eu/system/files/2022-03/edpb_03-2022_guidelines_on_dark_patterns_in_social_media_platform_interfaces_en.pdf. 2022.
- [10] European Parliament and Council of the European Union. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. Official Journal of the European Union, L 277, 27.10.2022, p. 1–102. Text with EEA relevance. 2022. URL: <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>.
- [11] Directorate-General for Justice et al. *Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation: final report*. eng. LU: Publications Office of the European Union, 2022. ISBN: 978-92-76-52316-1. URL: <https://data.europa.eu/doi/10.2838/859030>.
- [12] European Commission. *Review of EU Consumer Law*. https://commission.europa.eu/law/law-topic/consumer-protection-law/review-eu-consumer-law_en. Accessed: 2025-11-07. 2024.
- [13] The Economist. “Why bosses need to wake up to dark patterns”. In: *The Economist* (). Accessed: 2025-11-07. URL: <https://www.economist.com/business/2025/10/16/why-bosses-need-to-wake-up-to-dark%E2%80%91patterns>.
- [14] Euronews. “What do Europeans have to say about the Digital Fairness Act?” In: *Euronews Next* (). Accessed: 2025-11-07. URL: <https://www.euronews.com/next/2025/10/31/what-do-europeans-have-to-say-about-the-digital-fairness-act>.
- [15] Cristiana Santos, Viktorija Morozovaite, and Silvia De Conca. “No harm no foul: how harms caused by dark patterns are conceptualised and tackled under EU data protection, consumer and competition laws”. In: *Information & Communications Technology Law* (2025), pp. 1–47.
- [16] 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*. Washington DC USA: ACM, 2022, pp. 181–194. DOI: <https://doi.org/10.1145/3511265.3550448>.
- [17] Colin M Gray et al. “Dark patterns and the legal requirements of consent banners: An interaction criticism perspective”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–18.

- [18] 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*. 2021, pp. 1–18.
- [19] Arunesh Mathur et al. “Dark patterns at scale: Findings from a crawl of 11K shopping websites”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–32.
- [20] Agnieszka Kitkowska. “The hows and whys of dark patterns: Categorizations and privacy”. In: *Human factors in privacy research* (2023), pp. 173–198.
- [21] Colin M Gray et al. “Dark patterns and the emerging threats of deceptive design practices”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–4.
- [22] Vibhav Singh, Niraj Kumar Vishvakarma, and Vinod Kumar. “Unveiling digital manipulation and persuasion in e-commerce: a systematic literature review of dark patterns and digital nudging”. In: *Journal of Internet Commerce* 23.2 (2024), pp. 144–171.
- [23] Ilayda Karagoel and Dan Nathan-Roberts. “Dark patterns: Social media, gaming, and e-commerce”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 65. 1. SAGE Publications Sage CA: Los Angeles, CA. 2021, pp. 752–756.
- [24] Jacob Aagaard et al. “A game of dark patterns: Designing healthy, highly-engaging mobile games”. In: *CHI conference on human factors in computing systems extended abstracts*. 2022, pp. 1–8.
- [25] Sam Niknejad et al. “Level Up or Game Over: Exploring How Dark Patterns Shape Mobile Games”. In: *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. 2024, pp. 148–156.
- [26] Dan Fitton and Janet C. Read. “Creating a Framework to Support the Critical Consideration of Dark Design Aspects in Free-to-Play Apps”. In: *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. IDC ’19. Boise, ID, USA: Association for Computing Machinery, 2019, pp. 407–418. ISBN: 9781450366908. DOI: 10.1145/3311927.3323136. URL: <https://doi.org/10.1145/3311927.3323136>.
- [27] Célestin Matte, Nataliia Bielova, and Cristiana Santos. “Do cookie banners respect my choice?: Measuring legal compliance of banners from iab europe’s transparency and consent framework”. In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2020, pp. 791–809.

- [28] Shruthi Sai Chivukula et al. ““Nothing Comes Before Profit”: Asshole Design In the Wild”. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–6. ISBN: 9781450359719. DOI: 10.1145/3290607.3312863. URL: <https://doi.org/10.1145/3290607.3312863>.
- [29] Colin M Gray et al. “An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action”. en. In: *CHI 2024 Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2024, pp. 1–22. DOI: <https://doi.org/10.1145/3613904.3642436>.
- [30] Cherie Lacey and Catherine Caudwell. “Cuteness as a ‘dark pattern’ in home robots”. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. Daegu Republic of Korea: ACM, 2019, pp. 374–381.
- [31] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [32] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. “Phishing Detection: A Literature Survey”. In: *IEEE Communications Surveys & Tutorials* 15.4 (2013), pp. 2091–2121. DOI: 10.1109/SURV.2013.032213.00009.
- [33] Harry Brignull. “Dark Patterns: dirty tricks designers use to make people do stuff”. In: *Retrieved September 29 (2010)*, p. 2019.
- [34] Emre Kocyigit et al. “DeceptiLens: an Approach supporting Transparency in Deceptive Pattern Detection based on a Multimodal Large Language Model”. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '25. Association for Computing Machinery, 2025, pp. 1942–1959. ISBN: 9798400714825. DOI: 10.1145/3715275.3732129. URL: <https://doi.org/10.1145/3715275.3732129>.
- [35] Jieshan Chen et al. “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*. 2023, pp. 1–20.
- [36] Arvind Narayanan et al. “Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces”. In: *Queue* 18.2 (2020), pp. 67–92.
- [37] David Ettinger and Philippe Jehiel. “A theory of deception”. In: *American Economic Journal: Microeconomics* 2.1 (2010), pp. 1–20.

- [38] Francisco Lupiáñez-Villanueva et al. “Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation”. In: *European Commission, Directorate-General for Justice and Consumers, final report. May (2022)*.
- [39] Harry Brignull. *Dark Patterns: Dirty Tricks Designers Use to Make People Do Stuff*. <https://web.archive.org/web/20100722045159/http://www.90percentofeverything.com/2010/07/08/dark-patterns-dirty-tricks-designers-use-to-make-people-do-stuff/>. Accessed: October 24, 2025. 2010.
- [40] Harry Brignull. *Formerly darkpatterns.org*. <https://www.deceptive.design>. Accessed: 25 August 2024. 2010.
- [41] Chris Lewis. *Irresistible Apps: Motivational design patterns for apps, games, and web-based communities*. Apress, 2014.
- [42] Benedikte Mikkelsen, Christoffer Holmgård, and Julian Togelius. “Ethical Considerations for Player Modeling.” In: *AAAI Workshops*. 2017.
- [43] Fehmi Jaafar et al. “Evaluating the impact of design pattern and anti-pattern dependencies on changes and faults”. In: *Empirical Software Engineering* 21.3 (2016), pp. 896–931.
- [44] Saul Greenberg et al. “Dark patterns in proxemic interactions: a critical perspective”. In: *Proceedings of the 2014 conference on Designing interactive systems*. 2014, pp. 523–532.
- [45] Wen-Ting Yang and Mark Leiser. “Illuminating Manipulative Design: From” Dark Patterns” to Information Asymmetry and the Repression of Free Choice under the Unfair Commercial Practices Directive”. In: *Loy. Consumer L. Rev.* 34 (2022), p. 484.
- [46] Paul Graßl et al. “Dark and Bright Patterns in Cookie Consent Requests”. In: *Journal of Digital Social Research* 3.1 (2021), pp. 1–38.
- [47] Hana Habib et al. ““Okay, whatever”: An Evaluation of Cookie Consent Interfaces”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans LA USA: Association for Computing Machinery, Apr. 2022, pp. 1–27. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3501985. URL: <https://dl.acm.org/doi/10.1145/3491102.3501985>.
- [48] Kerstin Bongard-Blanchy et al. ““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. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 763–776. ISBN: 9781450384766. DOI: 10.1145/3461778.3462086. URL: <https://doi.org/10.1145/3461778.3462086>.

- [49] Kalya Win Aung, Ewan Soubutts, and Aneesha Singh. ““What a stupid way to do business”: Towards an Understanding of Older Adults’ Perceptions of Deceptive Patterns and Ways to Develop Resistance”. In: *Proc. ACM Hum.-Comput. Interact.* 8.CHI PLAY (Oct. 2024). DOI: 10.1145/3677113. URL: <https://doi.org/10.1145/3677113>.
- [50] Linda Di Geronimo et al. “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. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376600. URL: <https://doi.org/10.1145/3313831.3376600>.
- [51] Thomas Mildner and Gian-Luca Savino. “Ethical user interfaces: Exploring the effects of dark patterns on facebook”. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Yojohama Japan: ACM, 2021, pp. 1–7.
- [52] Ivana Rakovic and Yavuz Inal. “Dark finance: exploring deceptive design in investment apps”. In: *IFIP Conference on Human-Computer Interaction*. Springer. 2023, pp. 339–348.
- [53] Woo Gon Kim et al. “Dark patterns used by online travel agency websites”. In: *Annals of tourism research* 88 (2021), pp. 1–6.
- [54] Veronika Krauss et al. “What makes XR dark? Examining emerging dark patterns in augmented and virtual reality through expert co-design”. In: *ACM Transactions on Computer-Human Interaction* (2024).
- [55] Paul Graßl et al. “Dark and Bright Patterns in Cookie Consent Requests”. In: *Journal of Digital Social Research* 3.1 (2021), pp. 1–38.
- [56] Midas Nouwens et al. “Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence”. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–13.
- [57] Than Htut Soe et al. “Circumvention by design-dark patterns in cookie consent for online news outlets”. In: *Proceedings of the 11th nordic conference on human-computer interaction: Shaping experiences, shaping society*. 2020, pp. 1–12.
- [58] Chiara Krisam et al. “Dark Patterns in the Wild: Review of Cookie Disclaimer Designs on Top 500 German Websites”. In: *European Symposium on Usable Security 2021*. 2021, pp. 1–8.

- [59] Michael Toth, Nataliia Bielova, and Vincent Roca. “On dark patterns and manipulation of website publishers by CMPs”. In: *PETS 2022-22nd Privacy Enhancing Technologies Symposium*. 2022.
- [60] Carlos Bermejo Fernandez et al. “This website uses nudging: Mturk workers’ behaviour on cookie consent notices”. In: *Proceedings of the ACM on human-computer interaction* 5.CSCW2 (2021), pp. 1–22.
- [61] Lorenzo Porcelli, Massimo Ficco, and Francesco Palmieri. “Mitigating User Exposure to Dark Patterns in Cookie Banners Through Automated Consent”. In: *International Conference on Computational Science and Its Applications*. Springer. 2023, pp. 145–159.
- [62] Benjamin Maximilian Berens et al. “Cookie disclaimers: Dark patterns and lack of transparency”. In: *Computers & Security* 136 (2024), p. 103507.
- [63] Emre Kocyigit, Arianna Rossi, and Gabriele Lenzini. “A Systematic Approach for A Reliable Detection of Deceptive Design Patterns Through Measurable HCI Features”. In: *Proceedings of the 2024 European Symposium on Usable Security*. EuroUSEC ’24. Association for Computing Machinery, 2024, pp. 290–308. ISBN: 9798400717963. DOI: 10.1145/3688459.3688475. URL: <https://doi.org/10.1145/3688459.3688475>.
- [64] Johanna Gunawan et al. “A comparative study of dark patterns across web and mobile modalities”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–29.
- [65] Yuxuan Shang et al. “AdsDP: A Video Dataset for Recognizing and Examining Dark Patterns in iOS In-App Advertisements”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9.3 (Sept. 2025). DOI: 10.1145/3749515. URL: <https://doi.org/10.1145/3749515>.
- [66] Verena Traubinger et al. “In Search of Dark Patterns in Chatbots”. In: *International Workshop on Chatbot Research and Design*. Springer. 2023, pp. 117–132.
- [67] Monica Kowalczyk et al. “Understanding dark patterns in home IoT devices”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–27.
- [68] Diego Ramil-López, Paula Fraga-Lamas, and Tiago M. Fernández-Caramés. “A Practical Analysis of Persuasive and Dark Patterns for IIoT and Industrial Cyber-Physical Systems”. In: *2024 IEEE SENSORS*. 2024, pp. 1–4. DOI: 10.1109/SENSORS60989.2024.10784940.

- [69] Emre Kocyigit, Arianna Rossi, and Gabriele Lenzini. “Towards Assessing Features of Dark Patterns in Cookie Consent Processes”. In: *IFIP International Summer School on Privacy and Identity Management*. Springer, 2022, pp. 165–183.
- [70] Dino Bollinger et al. “Automating Cookie Consent and GDPR Violation Detection”. In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 2893–2910. ISBN: 978-1-939133-31-1. URL: <https://www.usenix.org/conference/usenixsecurity22/presentation/bollinger>.
- [71] Than Htut Soe, Cristiana Teixeira Santos, and Marija Slavkovic. “Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way”. In: *arXiv preprint arXiv:2204.11836* (2022).
- [72] SM Hasan Mansur et al. “Aidui: Toward automated recognition of dark patterns in user interfaces”. In: *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE. 2023, pp. 1958–1970.
- [73] Zhaoxin Cai et al. “DARPA: combating asymmetric dark UI patterns on android with run-time view decorator”. In: *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE. 2023, pp. 480–493.
- [74] Emre Kocyigit, Davide Liga, and Gabriele Lenzini. “Augmenting Dark Patterns Text Data by Leveraging Large Language Models: A Multi-agent Framework and Parameter-Efficient Fine-Tuning”. In: *International Workshop on Causality, Agents and Large Models*. Springer. 2024, pp. 65–87.
- [75] Jieshan Chen et al. “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. San Francisco, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606783. URL: <https://doi.org/10.1145/3586183.3606783>.
- [76] Cristiana Santos, Nataliia Bielova, and Célestin Matte. “Are cookie banners indeed compliant with the law?” In: *Technology and Regulation 2020* (2020), pp. 91–135.
- [77] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. ” O’Reilly Media, Inc.”, 2022.
- [78] Liming Nie et al. “Shadows in the Interface: A Comprehensive Study on Dark Patterns”. In: *Proc. ACM Softw. Eng.* 1.FSE (July 2024). DOI: 10.1145/3643736. URL: <https://doi.org/10.1145/3643736>.

- [79] Zijing Liang et al. “A Survey of Multimodal Large Language Models”. In: *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*. CAICE '24. Xi' an, China: Association for Computing Machinery, 2024, pp. 405–409. ISBN: 9798400716942. DOI: 10.1145/3672758.3672824. URL: <https://doi.org/10.1145/3672758.3672824>.
- [80] J. Wu et al. “Multimodal Large Language Models: A Survey”. In: *2023 IEEE International Conference on Big Data (BigData)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2023, pp. 2247–2256. DOI: 10.1109/BigData59044.2023.10386743. URL: <https://doi.ieeecomputersociety.org/10.1109/BigData59044.2023.10386743>.
- [81] Zhuosheng Zhang et al. “Multimodal Chain-of-Thought Reasoning in Language Models”. In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856. URL: <https://openreview.net/forum?id=y1pPWFVfvR>.
- [82] Yasin Sazid, Mridha Md. Nafis Fuad, and Kazi Sakib. “Automated Detection of Dark Patterns Using In-Context Learning Capabilities of GPT-3”. In: *2023 30th Asia-Pacific Software Engineering Conference (APSEC)*. Dec. 2023, pp. 569–573. DOI: 10.1109/APS EC60848.2023.00072. URL: <https://ieeexplore.ieee.org/abstract/document/10479438>.
- [83] Rajvardhan Patil and Venkat Gudivada. “A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)”. In: *Applied Sciences* 14.5 (2024). ISSN: 2076-3417. DOI: 10.3390/app14052074. URL: <https://www.mdpi.com/2076-3417/14/5/2074>.
- [84] Shuhan Qi et al. “What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing”. In: *Information Processing & Management* 60.6 (2023), p. 103510. ISSN: 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2023.103510>. URL: <https://www.sciencedirect.com/science/article/pii/S0306457323002479>.
- [85] Shengbang Tong et al. “Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 9568–9578.
- [86] Sadaf Ghaffari and Nikhil Krishnaswamy. “Exploring Failure Cases in Multimodal Reasoning About Physical Dynamics”. In: *Proceedings of the AAAI Symposium Series* 3.1 (May 2024), pp. 105–114. DOI: 10.1609/aaais.v3i1.31189. URL: <https://ojs.aaai.org/index.php/AAAI-SS/article/view/31189>.

- [87] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.
- [88] Gautier Izacard and Edouard Grave. “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 874–880.
- [89] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).
- [90] Paul F. Christiano et al. “Deep reinforcement learning from human preferences”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4302–4310. ISBN: 9781510860964.
- [91] European Data Protection Supervisor. *TechDispatch #2/2023: Explainable Artificial Intelligence*. Accessed: 2024-12-23. Nov. 2023. URL: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en.
- [92] Robert R. Hoffman et al. “Metrics for Explainable AI: Challenges and Prospects”. In: *ArXiv abs/1812.04608* (2018). URL: <https://api.semanticscholar.org/CorpusID:54577009>.
- [93] Q. Vera Liao and Jennifer Wortman Vaughan. “AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap”. In: *Harvard Data Science Review Special Issue 5*. Special Issue 5 (May 2024). <https://hdsr.mitpress.mit.edu/pub/aelql9qy>.
- [94] Yuki Yada et al. “Dark patterns in e-commerce: a dataset and its baseline evaluations”. In: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. 2022, pp. 3015–3022.
- [95] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. “Text data augmentation for deep learning”. In: *Journal of big Data* 8.1 (2021), p. 101.
- [96] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. “A survey on data augmentation for text classification”. In: *ACM Computing Surveys* 55.7 (2022), pp. 1–39.
- [97] Markus Bayer et al. “Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers”. In: *International journal of machine learning and cybernetics* 14.1 (2023), pp. 135–150.

- [98] Xiao Liang et al. “Task Oriented In-Domain Data Augmentation”. In: *arXiv preprint arXiv:2406.16694* (2024).
- [99] Haixing Dai et al. “Auggpt: Leveraging chatgpt for text data augmentation”. In: *arXiv preprint arXiv:2302.13007* (2023).
- [100] Aaron Cahn et al. “An Empirical Study of Web Cookies”. In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 891–901. ISBN: 9781450341431. DOI: 10.1145/2872427.2882991. URL: <https://doi.org/10.1145/2872427.2882991>.
- [101] Jamie Luguri and Lior Jacob Strahilevitz. “Shining a light on dark patterns”. In: *Journal of Legal Analysis* 13.1 (2021), pp. 43–109.
- [102] ISO. *Information technology - Security techniques - Privacy framework*. Standard. Geneva, CH: International Organization for Standardization, Dec. 2011.
- [103] Ali Rasaii et al. “Exploring the Cookieverse: A Multi-Perspective Analysis of Web Cookies”. In: *International Conference on Passive and Active Network Measurement*. Springer. 2023, pp. 623–651.
- [104] Ahmed Bouhoula et al. “Automated, Large-Scale Analysis of Cookie Notice Compliance”. In: *33rd USENIX Security Symposium (USENIX Security 24)*. 2023.
- [105] Directorate-General for Justice et al. *Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation: final report*. eng. LU: Publications Office of the European Union, 2022. ISBN: 978-92-76-52316-1. URL: <https://data.europa.eu/doi/10.2838/859030>.
- [106] Competition and Markets Authority. *Evidence review of Online Choice Architecture and consumer and competition harm*. <https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers/evidence-review-of-online-choice-architecture-and-consumer-and-competition-harm#taxonomy-of-oca>. Apr. 2022.
- [107] Garante per la Protezione dei Dati Personali. *Provvedimento prescrittivo e sanzionatorio nei confronti di Ediscom S.p.A. - 23 febbraio 2023 [9870014]*. it-IT. 2023. URL: <https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870014>.
- [108] Cristiana Santos and Arianna Rossi. “The emergence of dark patterns as a legal concept in case law”. In: *Internet Policy Review* (July 2023). URL: <https://policyreview.info/articles/news/emergence-of-dark-patterns-as-a-legal-concept>.

- [109] Arianna Rossi et al. “Who is vulnerable to deceptive design patterns? A transdisciplinary perspective on the multi-dimensional nature of digital vulnerability”. In: *Computer Law & Security Review* (In press).
- [110] OECD. *Consumer vulnerability in the digital age*. OECD Digital Economy Papers 355. Paris, 2023. URL: <https://www.oecd.org/publications/consumer-vulnerability-in-the-digital-age-4d013cc5-en.htm>.
- [111] Kentrell Owens et al. “Exploring Deceptive Design Patterns in Voice Interfaces”. en. In: *2022 European Symposium on Usable Security*. Karlsruhe Germany: ACM, Sept. 2022, pp. 64–78. ISBN: 978-1-4503-9700-1. DOI: 10.1145/3549015.3554213. URL: <https://dl.acm.org/doi/10.1145/3549015.3554213>.
- [112] Hellen Truong and Axel Dalbard. “Bright Patterns as an Ethical Approach to Counteract Dark Patterns. A Closer Investigation of The Ethics of Persuasive Design”. PhD thesis. Jonkoping: Jonkoping University, 2022. URL: <https://www.diva-portal.org/smash/get/diva2:1680425/FULLTEXT01.pdf>.
- [113] Evan Caragay et al. “Beyond Dark Patterns: A Concept-Based Framework for Ethical Software Design”. en. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, May 2024, pp. 1–16. ISBN: 9798400703300. DOI: 10.1145/3613904.3642781. URL: <https://dl.acm.org/doi/10.1145/3613904.3642781>.
- [114] Nataliia Bielova, Cristiana Santos, and Colin M. Gray. “Two worlds apart! Closing the gap between regulating EU consent and user studies”. In: *Harvard Journal of Law & Technology* 37 (2024), pp. 1295–1333. URL: <https://jolt.law.harvard.edu/assets/articlePDFs/v37/Symposium-12-Bielova-Santos-Gray-Two-Worlds-Apart-Closing-the-Gap-Between-Regulating-EU-Consent-and-User-Studies.pdf>.
- [115] Tobias Kuhn. “A survey and classification of controlled natural languages”. In: *Computational linguistics* 40.1 (2014), pp. 121–170.
- [116] Linda Di Geronimo et al. *UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception*. Zenodo, Jan. 2020. DOI: 10.5281/zenodo.3601501. URL: <https://doi.org/10.5281/zenodo.3601501>.
- [117] Biplab Deka et al. “Rico: A Mobile App Dataset for Building Data-Driven Design Applications”. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. UIST ’17. Québec City, QC, Canada: Association for Computing Ma-

- chinery, 2017, pp. 845–854. ISBN: 9781450349819. DOI: 10.1145/3126594.3126651. URL: <https://doi.org/10.1145/3126594.3126651>.
- [118] Youdi Gong et al. “A survey on dataset quality in machine learning”. In: *Information and Software Technology* 162 (2023), p. 107268. ISSN: 0950-5849. DOI: <https://doi.org/10.1016/j.infsof.2023.107268>. URL: <https://www.sciencedirect.com/science/article/pii/S0950584923001222>.
- [119] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3 (2012), pp. 276–282.
- [120] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. “Understanding Account Deletion and Relevant Dark Patterns on Social Media”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (Nov. 2022). DOI: 10.1145/3555142. URL: <https://doi.org/10.1145/3555142>.
- [121] Tasneem Naheyan and Kiemute Oyibo. “The Effect of Dark Patterns and User Knowledge on User Experience and Decision-Making”. In: *International Conference on Persuasive Technology*. Springer. 2024, pp. 190–206.
- [122] Ilias Chalkidis et al. “LEGAL-BERT: The muppets straight out of law school”. In: *arXiv preprint arXiv:2010.02559* (2020).
- [123] Davide Liga and Livio Robaldo. “Fine-tuning GPT-3 for legal rule classification”. In: *Computer Law & Security Review* 51 (2023), p. 105864.
- [124] Marie Potel-Saville and Mathilde Da Rocha. “From Dark Patterns to Fair Patterns? Usable Taxonomy to Contribute Solving the Issue with Countermeasures”. In: *Annual Privacy Forum*. Springer. 2023, pp. 145–165.
- [125] *Dark Pattern Visual Detection Dataset*. https://universe.roboflow.com/syd-help/dark-pattern-_visual-detection. Open Source Dataset. visited on 2024-06-07. Mar. 2023. URL: https://universe.roboflow.com/syd-help/dark-pattern-_visual-detection.
- [126] Ali Mohammad Saghiri et al. “A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions”. In: *Applied sciences* 12.8 (2022), p. 4054.
- [127] Wayne Xin Zhao et al. “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223* (2023).
- [128] *GitHub - eugeneyan/open-llms: A list of open LLMs available for commercial use.* — [github.com](https://github.com/eugeneyan/open-llms). <https://github.com/eugeneyan/open-llms>. [Accessed 10-06-2024].

- [129] Jason Wei et al. “Emergent abilities of large language models”. In: *arXiv preprint arXiv:2206.07682* (2022).
- [130] Yupeng Chang et al. “A survey on evaluation of large language models”. In: *ACM Transactions on Intelligent Systems and Technology* 15.3 (2024), pp. 1–45.
- [131] *Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com*. <https://ai.meta.com/blog/meta-llama-3/>. [Accessed 10-06-2024].
- [132] *Meta Llama 3 — llama.meta.com*. <https://llama.meta.com/llama3/>. [Accessed 24-06-2024].
- [133] Sabit Ekin. “Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices”. In: *Authorea Preprints* (2023).
- [134] Edward J Hu et al. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021).
- [135] Xinyu Lin et al. “Data-efficient Fine-tuning for LLM-based Recommendation”. In: *arXiv preprint arXiv:2401.17197* (2024).
- [136] NCC. “Deceived by design, how tech companies use dark patterns to discourage us from exercising our rights to privacy”. In: *Norwegian Consumer Council Report* (2018).
- [137] BEUC. “*Dark patterns*” and the EU consumer law acquisition. Last accessed 9 January 2023. 2022. URL: https://www.beuc.eu/sites/default/files/publications/beuc-x-2022-013_dark_patterns_paper.pdf.
- [138] European Parliament and Council of the European Union. *REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. Official Journal of the European Union, L 277, 27.10.2022, p. 1-102. Oct. 2022. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065>.
- [139] CPRA. *The California Privacy Rights Act of 2020*. Last accessed 9 January 2023. 2020. URL: <https://vig.cdn.sos.ca.gov/2020/general/pdf/top1-prop24.pdf>.
- [140] European Parliament and Council of the European Union. *Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)*. Official Journal of the European Union, L 2023/2854, 22.12.2023. Dec. 2023. URL: <https://eur-lex.europa.eu/eli/reg/2023/2854/oj/en.g>.

- [141] European Parliament and Council of the European Union. *REGULATION (EU) 2022/1925 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)*. Official Journal of the European Union, L 265/1, 12.10.2022. Sept. 2022. URL: <https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng>.
- [142] M. R. Leiser and Cristiana Santos. “Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface”. en. In: *European Journal of Law and Technology* 15.1 (2024). URL: <https://ejlt.org/index.php/ejlt/article/view/990>.
- [143] Harry Brignull et al. *Deceptive Patterns Database of Legal cases*. 2023. URL: <https://www.deceptive.design/cases>.
- [144] Thomas Mildner et al. “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. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580695. URL: <https://doi.org/10.1145/3544548.3580695>.
- [145] Dmitry Nazarov and Yerkebulan Baimukhambetov. “Clustering of Dark Patterns in the User Interfaces of Websites and Online Trading Portals (E-Commerce)”. In: *Mathematics* 10.18 (2022). ISSN: 2227-7390. DOI: 10.3390/math10183219. URL: <https://www.mdpi.com/2227-7390/10/18/3219>.
- [146] J. Wu et al. “Multimodal Large Language Models: A Survey”. In: *2023 IEEE International Conference on Big Data (BigData)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2023, pp. 2247–2256. DOI: 10.1109/BigData59044.2023.10386743. URL: <https://doi.ieeecomputersociety.org/10.1109/BigData59044.2023.10386743>.
- [147] Arun James Thirunavukarasu et al. “Large language models in medicine”. In: *Nature medicine* 29.8 (2023), pp. 1930–1940.
- [148] Jinqi Lai et al. “Large language models in law: A survey”. In: *AI Open* 5 (2024), pp. 181–196. ISSN: 2666-6510. DOI: <https://doi.org/10.1016/j.aiopen.2024.09.002>. URL: <https://www.sciencedirect.com/science/article/pii/S2666651024000172>.
- [149] Aaron Hurst et al. “Gpt-4o system card”. In: *arXiv preprint arXiv:2410.21276* (2024).

- [150] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- [151] Snehal Prabhudesai et al. “Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making”. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23. Sydney, NSW, Australia: Association for Computing Machinery, 2023, pp. 379–396. ISBN: 9798400701061. DOI: 10.1145/3581641.3584033. URL: <https://doi.org/10.1145/3581641.3584033>.
- [152] Jingfeng Yang et al. “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond”. In: *ACM Trans. Knowl. Discov. Data* 18.6 (Apr. 2024). ISSN: 1556-4681. DOI: 10.1145/3649506. URL: <https://doi.org/10.1145/3649506>.
- [153] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission, 2019. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
- [154] Raymond Fok and Daniel S. Weld. “In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making”. In: *AI Mag.* 45 (2023), pp. 317–332. URL: <https://api.semanticscholar.org/CorpusID:258686588>.
- [155] Dongping Chen et al. “MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark”. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org, 2025.
- [156] LMSYS. *LMSYS Org — lmsys.org*. <https://lmsys.org/blog/2024-06-27-multimodal1/>. [Accessed 09-01-2025]. 2025.
- [157] Louie Giray. “Prompt engineering with ChatGPT: a guide for academic writers”. In: *Annals of biomedical engineering* 51.12 (2023), pp. 2629–2633.
- [158] Matthijs J Warrens. “Five ways to look at Cohen’s kappa”. In: *Journal of Psychology & Psychotherapy* 5 (2015).
- [159] Joseph L Fleiss and Jacob Cohen. “The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability”. In: *Educational and psychological measurement* 33.3 (1973), pp. 613–619.
- [160] Udo Kuckartz and Stefan Rädiker. *Analyzing qualitative data with MAXQDA*. Springer, 2019.

- [161] British Psychological Society and thics Committee of the British Psychological Society. *Code of ethics and conduct*. en. Leicester: The British Psychological Society, 2018. ISBN: 978-1-85433-759-7.
- [162] ALLEA. *The European Code of Conduct for Research Integrity - Revised Edition 2023*. eng. DE: ALLEA - All European Academies, June 2023. ISBN: 978-3-9823562-3-5. URL: <https://doi.org/10.26356/ECOC>.
- [163] European Commission Directorate General for Research and Innovation. *Living guidelines on the responsible use of generative AI in research*. en. Brussels: European Commission, 2024. URL: https://research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en.
- [164] Verena Distler et al. “The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely”. en. In: *New Security Paradigms Workshop 2020*. Online USA: ACM, Oct. 2020, pp. 45–58. ISBN: 978-1-4503-8995-2. DOI: 10.1145/3442167.3442173. URL: <https://dl.acm.org/doi/10.1145/3442167.3442173>.
- [165] Xialing Lin, Patric R Spence, and Kenneth A Lachlan. “Social media and credibility indicators: The effect of influence cues”. In: *Computers in human behavior* 63 (2016), pp. 264–271.
- [166] Mary L Cummings. “Automation bias in intelligent time critical decision support systems”. In: *Decision making in aviation*. Abingdon, Oxfordshire, UK: Routledge, 2017, pp. 289–294.
- [167] Gagan Bansal et al. “Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445717. URL: <https://doi.org/10.1145/3411764.3445717>.
- [168] Kathryn Ann Lambe et al. “Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review”. In: *BMJ quality & safety* 25.10 (2016), pp. 808–820. URL: <https://qualitysafety.bmj.com/content/25/10/808.short>.
- [169] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making”. en. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (Apr. 2021), pp. 1–21. ISSN: 2573-0142. DOI: 10.1145/3449287. URL: <https://dl.acm.org/doi/10.1145/3449287>.

- [170] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
- [171] Valdemar Danry et al. “Don’t Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580672. URL: <https://doi.org/10.1145/3544548.3580672>.
- [172] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 295–305. ISBN: 9781450369367. DOI: 10.1145/3351095.3372852. URL: <https://doi.org/10.1145/3351095.3372852>.
- [173] Professor Christine Riefa and Liz Coll. *The transformative potential of Enforcement Technology (EnfTech) in Consumer Law*. en. 2024.
- [174] Xinyi Li et al. “A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges”. In: *Vicinagearth* 1.1 (2024), p. 9.
- [175] European Data Protection Board. *Report of the work undertaken by the Cookie Banner Taskforce*. Brussels: European Data Protection Board, Jan. 2023. URL: https://edpb.europa.eu/system/files/2023-01/edpb_20230118_report_cookie_banner_taskforce_en.pdf.
- [176] Hauke Sandhaus. “Promoting Bright Patterns”. en. In: *CHI ’23 Workshop: Designing Technology and Policy Simultaneously*. arXiv:2304.01157 [cs]. Hamburg, DE: arXiv, 2023. DOI: 10.48550/arXiv.2304.01157. URL: <http://arxiv.org/abs/2304.01157>.
- [177] René Schäfer et al. “Fighting Malicious Designs: Towards Visual Countermeasures Against Dark Patterns”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642661. URL: <https://doi.org/10.1145/3613904.3642661>.
- [178] Yuwen Lu et al. “From awareness to action: Exploring end-user empowerment interventions for dark patterns in ux”. In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1 (2024), pp. 1–41.

- [179] Pooria Babaei and Julita Vassileva. “Drivers and persuasive strategies to influence user intention to learn about manipulative design”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, pp. 2421–2431. ISBN: 9798400704505. DOI: 10.1145/3630106.3659046. URL: <https://doi.org/10.1145/3630106.3659046>.
- [180] Sanju Ahuja and Jyoti Kumar. “Conceptualizations of user autonomy within the normative evaluation of dark patterns”. In: *Ethics and Information Technology* 24.4 (2022), p. 52.
- [181] FCC. *Fortnite Video Game Maker Epic Games to Pay More Than Half a Billion Dollars over FTC Allegations of Privacy Violations and Unwanted Charges — Federal Trade Commission*. Dec. 2022. URL: <https://www.ftc.gov/news-events/news/press-releases/2022/12/fortnite-video-game-maker-epic-games-pay-more-half-billion-dollars-over-ftc-allegations>.
- [182] Jiayang Wu et al. “Multimodal large language models: A survey”. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023, pp. 2247–2256.
- [183] Luciano Floridi et al. “Open-Source AI made in the EU: Why it is a Good Idea”. en. In: *SSRN Electronic Journal* 5191372 (Mar. 2025). DOI: 10.2139/ssrn.5191372. URL: <https://papers.ssrn.com/abstract=5191372>.
- [184] Arthur Spirling. “Why open-source generative AI models are an ethical way forward for science”. en. In: *Nature* 616.7957 (Apr. 2023), pp. 413–413. DOI: 10.1038/d41586-023-01295-4. URL: <https://www.nature.com/articles/d41586-023-01295-4>.
- [185] Étienne Ollion et al. “The dangers of using proprietary LLMs for research”. en. In: *Nature Machine Intelligence* 6.1 (Jan. 2024), pp. 4–5. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00783-6. URL: <https://www.nature.com/articles/s42256-023-00783-6>.
- [186] Jiya Manchanda et al. “The Open Source Advantage in Large Language Models (LLMs)”. In: arXiv:2412.12004 (Feb. 2025). arXiv:2412.12004 [cs]. DOI: 10.48550/arXiv.2412.12004. URL: <http://arxiv.org/abs/2412.12004>.
- [187] S. Hrushikesava Raju et al. “Smart Dark Pattern Detection: Making Aware of Misleading Patterns Through the Intended App”. In: *Sentimental Analysis and Deep Learning*. Ed. by Subarna Shakya et al. Singapore: Springer Singapore, 2022, pp. 933–947. ISBN: 978-981-16-5157-1.
- [188] Mahdi Soltani et al. “An adaptable deep learning-based intrusion detection system to zero-day attacks”. In: *Journal of Information Security and Applications* 76 (2023), p. 103516.

- [189] Tim Verdonck et al. “Special issue on feature engineering editorial”. In: *Machine learning* 113.7 (2024), pp. 3917–3928.
- [190] Yunfei Lai. “A comparison of traditional machine learning and deep learning in image recognition”. In: *Journal of Physics: Conference Series*. Vol. 1314. 1. IOP Publishing, 2019, p. 012148.
- [191] Xue-Wen Chen and Xiaotong Lin. “Big Data Deep Learning: Challenges and Perspectives”. In: *IEEE Access* 2 (2014), pp. 514–525. DOI: 10.1109/ACCESS.2014.2325029.
- [192] Colin White et al. “Livebench: A challenging, contamination-free llm benchmark”. In: *arXiv preprint arXiv:2406.19314* 4 (2024).
- [193] Bohao Li et al. “SEED-Bench: Benchmarking Multimodal Large Language Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 13299–13308.
- [194] Xionghao Zhou et al. “An Empirical Study on Parameter-Efficient Fine-Tuning for Multimodal Large Language Models”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10057–10084. DOI: 10.18653/v1/2024.findings-acl.598. URL: <https://aclanthology.org/2024.findings-acl.598/>.
- [195] Berk Atıl et al. “LLM Stability: A detailed analysis with some surprises”. In: *arXiv e-prints* (2024), arXiv–2408.
- [196] Davide Liga, Réka Markovich, and Liuwen Yu. “Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI: Symbolic Constraints over Large Language Models via Prompt Engineering”. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAAIL)*. forthcoming. ACM, 2025.
- [197] Sakib Shahriar et al. “Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency”. In: *Applied Sciences* 14.17 (2024). ISSN: 2076-3417. DOI: 10.3390/app14177782. URL: <https://www.mdpi.com/2076-3417/14/17/7782>.
- [198] James Phoenix and Mike Taylor. *Prompt engineering for generative AI*. ” O’Reilly Media, Inc.”, 2024.
- [199] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information processing & management* 45.4 (2009), pp. 427–437.

- [200] Mohammad Hossin and Md Nasir Sulaiman. “A review on evaluation metrics for data classification evaluations”. In: *International journal of data mining & knowledge management process* 5.2 (2015), p. 1.
- [201] Global Privacy Enforcement Network. *GPEN Sweep 2024 - “Deceptive Design Patterns”*. 2024. URL: <https://www.privacyenforcement.net/system/files/2024-07/GPEN%20Sweep%202024%20-%20%27Deceptive%20Design%20Patterns%27.pdf>.
- [202] Christopher Diebel et al. “When AI-Based Agents Are Proactive: Implications for Competence and System Satisfaction in Human–AI Collaboration”. In: *Business & Information Systems Engineering* (Jan. 2025). ISSN: 1867-0202. DOI: 10.1007/s12599-024-00918-y. URL: <https://doi.org/10.1007/s12599-024-00918-y>.
- [203] Anwesha Mohanty, Venkatesh Balavadhani Parthasarathy, and Arsalan Shahid. “The Future of MLLM Prompting is Adaptive: A Comprehensive Experimental Evaluation of Prompt Engineering Methods for Robust Multimodal Performance”. In: *Submitted to Transactions on Machine Learning Research* (2025). Under review. URL: <https://openreview.net/forum?id=B1L8HrjoA1>.
- [204] Ike Obi et al. “Tracing the History and Evolution of Dark Patterns on Twitter from 2010-2021”. In: *Trans. Soc. Comput.* 8.3–4 (Mar. 2025). DOI: 10.1145/3716306. URL: <https://doi.org/10.1145/3716306>.
- [205] Gregory Day and Abbey Stemler. “Are dark patterns anticompetitive?” In: *Ala. L. Rev.* 72 (2020), p. 1.
- [206] Kawon (Kathy) Kim, Woo Gon Kim, and Minwoo Lee. “Impact of dark patterns on consumers’ perceived fairness and attitude: Moderating effects of types of dark patterns, social proof, and moral identity”. In: *Tourism Management* 98 (2023), p. 104763. ISSN: 0261-5177. DOI: <https://doi.org/10.1016/j.tourman.2023.104763>. URL: <https://www.sciencedirect.com/science/article/pii/S0261517723000456>.
- [207] Woon Chee Koh and Yuan Zhi Seah. “Unintended consumption: The effects of four e-commerce dark patterns”. In: *Cleaner and Responsible Consumption* 11 (2023), p. 100145. ISSN: 2666-7843. DOI: <https://doi.org/10.1016/j.clrc.2023.100145>. URL: <https://www.sciencedirect.com/science/article/pii/S2666784323000463>.
- [208] Amit Zac et al. “Dark patterns and consumer vulnerability”. In: *Behavioural Public Policy* (2023), pp. 1–50.

- [209] Leah Zhang-Kennedy, Maxwell Keleher, and Michaela Valiquette. “Navigating the Gray: Design Practitioners’ Perceptions Toward the Implementation of Privacy Dark Patterns”. In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1 (Apr. 2024). DOI: 10.1145/3637374. URL: <https://doi.org/10.1145/3637374>.
- [210] Giovanni Sartor and Thiago Raulino Dal Pont. *Artificial Intelligence for Monitoring the Application of EU Law*. Tech. rep. PE 777.915. Requested by the JURI Committee. European Parliament, Policy Department for Justice, Civil Liberties, Institutional Affairs, Directorate-General for Citizens’ Rights, Justice, and Institutional Affairs, Oct. 2025. URL: [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/777915/IUST_STU\(2025\)777915_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/777915/IUST_STU(2025)777915_EN.pdf).
- [211] Colin M. Gray, Thomas Mildner, and Ritika Gairola. “Getting Trapped in Amazon’s ”Iliad Flow”: A Foundation for the Temporal Analysis of Dark Patterns”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713828. URL: <https://doi.org/10.1145/3706598.3713828>.
- [212] Vellum AI. *Vellum LLM Leaderboard*. <https://www.vellum.ai/llm-leaderboard>. Accessed: 19 November 2025. 2025.
- [213] LMSys Organization. *LMSys Chatbot Arena Leaderboard - Vision*. <https://lmarena.ai/leaderboard/vision>. Accessed: 19 November 2025. 2025.
- [214] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. “Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey”. In: *IEEe Access* (2025).
- [215] Andrei Kucharavy. “Fundamental limitations of generative llms”. In: *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*. Springer Nature Switzerland Cham, 2024, pp. 55–64.
- [216] Muhammad Usman Hadi et al. “LLMs: A Comprehensive Survey of Applications, Challenges, Datasets, Models, Limitations, and Future Prospects”. In: (2024).
- [217] Bo Wang et al. “Unveiling privacy risks in llm agent memory”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 25241–25260.
- [218] Qingxiu Dong et al. “A survey on in-context learning”. In: *Proceedings of the 2024 conference on empirical methods in natural language processing*. 2024, pp. 1107–1128.
- [219] Tolga Şakar and Hakan Emekci. “Maximizing RAG efficiency: A comparative analysis of RAG methods”. In: *Natural Language Processing* 31.1 (2025), pp. 1–25.

-
- [220] Vincent Hanke et al. “Open llms are necessary for current private adaptations and outperform their closed alternatives”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 1220–1250.
- [221] Étienne Ollion et al. “The dangers of using proprietary LLMs for research”. In: *Nature Machine Intelligence* 6.1 (2024), pp. 4–5.
- [222] David Gray Widder, Meredith Whittaker, and Sarah Myers West. “Why ‘open’AI systems are actually closed, and why this matters”. In: *Nature* 635.8040 (2024), pp. 827–833.
- [223] David Li et al. “Comparative diagnostic accuracy of GPT-4o and LLaMA 3-70b: proprietary vs. open-source large language models in radiology”. In: *Clinical Imaging* 118 (2025), p. 110382.

Appendix A

Overview of Dark Pattern Taxonomies

Source	Year	High-level Categories	# Low-level Types
Brignull et al. [33]	2010	– First classification, non-academic	12
Zagal et al. [2]	2013	<ul style="list-style-type: none"> • Temporal • Monetary • Social Capital-based 	7
Bösch et al. [4]	2016	<ul style="list-style-type: none"> • Privacy-focused patterns (defaults, immortal accounts, legalese) 	6
Gray et al. [5]	2018	<ul style="list-style-type: none"> • Nagging • Obstruction • Sneaking • Interface Interference • Forced Action 	15
Mathur et al. [19]	2019	<ul style="list-style-type: none"> • Sneaking • Urgency • Misdirection • Social Proof • Scarcity • Obstruction • Forced Action 	14
FTC [6]	2022	<ul style="list-style-type: none"> • Endorsements • Scarcity • Urgency • Obstruction • Sneaking • Interface Interference • Coerced Action • Asymmetric Choice 	20
OECD [7]	2022	<ul style="list-style-type: none"> • Forced Action • Interface Interference • Nagging • Obstruction • Sneaking • Social Proof • Urgency 	18
EDPB [8]	2023	<ul style="list-style-type: none"> • Overloading • Skipping • Stirring • Obstructing • Fickle • Left in the Dark 	12

Table A.1: Chronological Overview of Foundational and Influential Dark Pattern Taxonomies, Illustrating the Evolving and Expanding Landscape

Appendix B

Cookie Consent Scenarios, Measurable Features, and Source Websites

Table B.1: Source websites that are selected to inspect cookies consent processes

No	URL	No	URL
1	https://www.researchgate.net	6	https://twitter.com
2	https://www.kaggle.com	7	https://www.dropbox.com
3	https://www.w3schools.com	8	https://wwwfr.uni.lu
4	https://stackoverflow.com	9	https://github.com
5	https://www.linkedin.com	10	https://medium.com

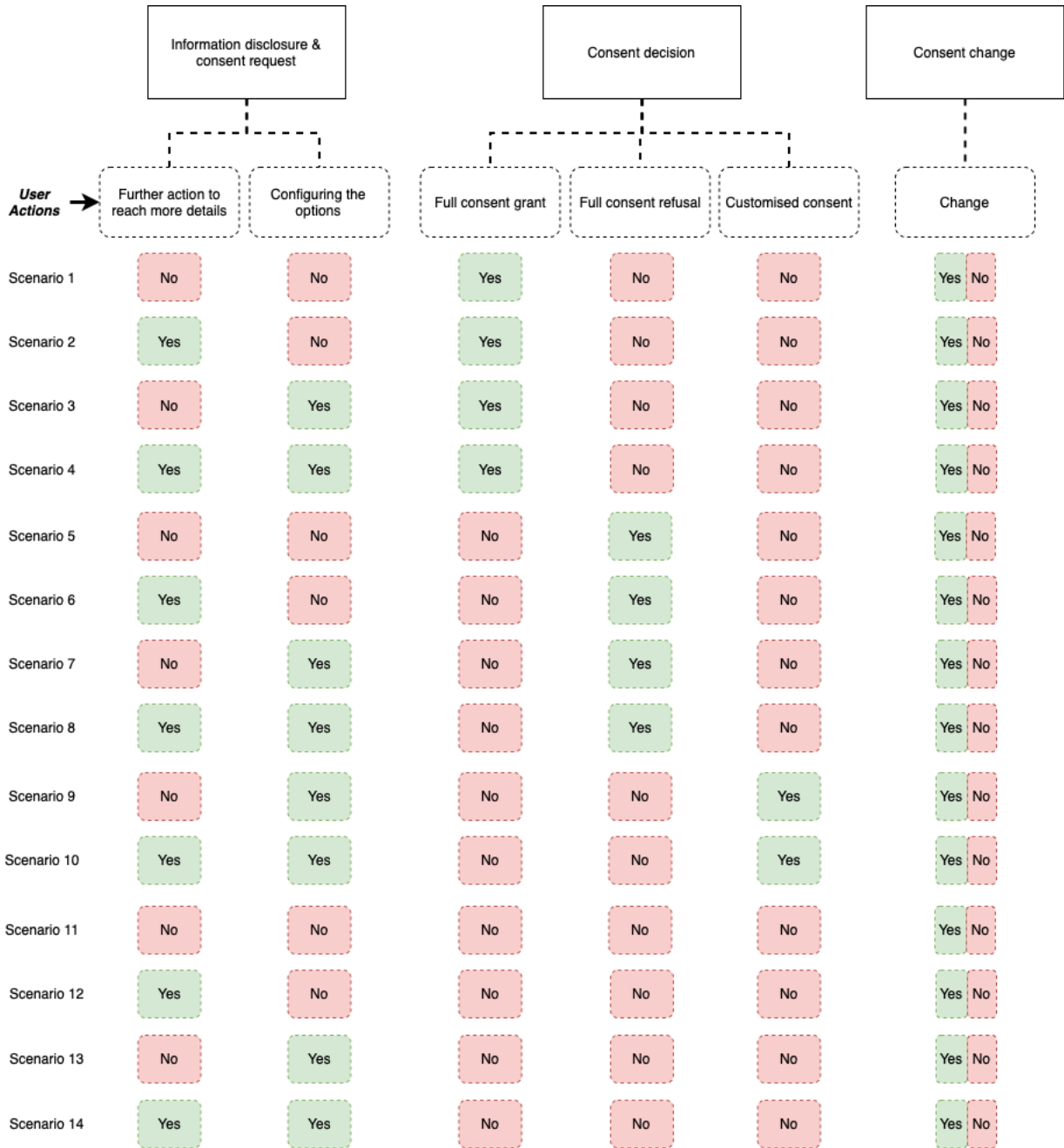


Figure B.1: Cookie consent scenarios based on user actions in the first two phases

Table B.2: Features of cookie consent process relevant for dark pattern detection

No	Feature	Layer	Value
1	Forced decision	User Action	Binary
2	User options at first visit	User Action	Quantitative
3	Full consent availability	User Action	Binary
4	Total number of routes in full consent grant	User Action	Quantitative
5	Minimum length of routes in full consent grant	User Action	Quantitative
6	Maximum length of routes in full consent grant	User Action	Quantitative
7	Full consent refusal availability	User Action	Binary
8	Total number of routes in full consent refusal	User Action	Quantitative
9	Minimum length of routes in full consent refusal	User Action	Quantitative
10	Maximum length of routes in full consent refusal	User Action	Quantitative
11	Customised consent availability	User Action	Quantitative
12	Total number of routes in customised consent	User Action	Quantitative
13	Minimum length of routes in customised consent	User Action	Quantitative
14	Maximum length of routes in customised consent	User Action	Quantitative
15	Total user action options	User Action	Quantitative
16	Total consent flow routes	User Action	Quantitative
17	Total hyperlinks on user interface	User Interface	Quantitative
18	Consent decisions management availability	User Action	Binary
19	Dead end	User Action	Binary
20	Number of cookies at first visit	Web Engine	Quantitative
21	Number of cookies after full consent	Web Engine	Quantitative
22	Number of cookies after full consent refusal	Web Engine	Quantitative
23	Cookie size at first visit	Web Engine	Quantitative
24	Cookie size after full consent	Web Engine	Quantitative
25	Cookie size after full consent refusal	Web Engine	Quantitative
26	Total requests at first visit	Web Engine-Web Server	Quantitative
27	Total requests after full consent	Web Engine-Web Server	Quantitative
28	Total requests after full consent refusal	Web Engine-Web Server	Quantitative
29	Transferred file size at first visit	Web Engine-Web Server	Quantitative
30	Transferred file size after full consent	Web Engine-Web Server	Quantitative
31	Transferred file size after full consent refusal	Web Engine-Web Server	Quantitative

Appendix C

Feature Association and Measurable Feature Taxonomy

	Expert 1	Expert 2	Expert 3	Result	Round 1		Round 2	
					Count	Overall	Count	Overall
Consensus	Required	Required	Required	Required	14	50	21	82
	Optional	Optional	Optional	Optional	6		21	
	Not required	Not required	Not required	Not required	30		40	
Majority voting without conflict	Required	Required	Optional		11	41	11	30
	Required	Optional	Required	Required				
	Optional	Required	Required					
	Optional	Optional	Required		6		8	
	Optional	Required	Optional	Optional				
	Required	Optional	Optional		9		7	
	Optional	Optional	Not required					
	Optional	Not required	Optional	Optional				
Not required	Optional	Optional		15	4			
Not required	Not required	Optional						
Not required	Optional	Not required	Not required					
Majority voting with conflict	Required	Required	Not required		4	8	0	0
	Required	Not required	Required	Required				
	Not required	Required	Required					
	Not required	Not required	Required		4		0	
	Not required	Required	Not required	Not required				
Required	Not required	Not required						
High uncertainty	Required	Optional	Not required	-	13	13	0	0
	Required	Not required	Optional	-				
	Optional	Required	Not required	-				
	Optional	Not required	Required	-				
	Not required	Required	Optional	-				
	Not required	Optional	Required	-				
Total					112			

Figure C.1: Comparison of first- and second-round expert mappings of attributes to dark pattern types, showing changes in required, optional, and not-required classifications.

Table C.1: Description of measurable features and their design categories

ID	Category	Entity	Metric	Feature	Description
1	visual	FCG button	size	size of the FCG button	Size of a button such as "accept all", e.g. in pixels.
2		FCR button		size of the FCR button	Size of a button such as "reject all", e.g. in pixels.
3		CC banner		size of the CC banner	Size of the cookie consent banner, e.g. in pixels.
4		FCG button	background color	background color of the FCG button	Color of a button such as "accept all", e.g. in RGB codes.
5		FCR button		background color of the FCR button	Color of a button such as "reject all", e.g. in RGB codes.
6		consent setting button		background color of the consent setting button	Color of a button such as "preferences", "options" etc., e.g. in RGB codes.
7	linguistic	FCG icon	availability	availability of the FCG icon	Binary feature. True if an icon is present on the FCG button. Otherwise, False.
8		CC image		availability of the CC image	Binary feature. True if an image is present on CC banner. Otherwise, False.
9		CC image	relevance	relevance of the CC image	Semantic relevance of an CC image with the use case.
10		CC information text	readability	readability of the CC information text	Readability score of the CC info text, e.g. Flesch Reading Ease score.
11		CC information text	comprehensibility	comprehensibility of the CC information text	Assessment of how easily text is understood by user, e.g. percentage of correctly answered questions related to the CC information text.
12		FCG text	clarity	clarity of the FCG text	Clarity of a text such as "accept all", e.g. based on a survey-based scoring from 1 to 5.
13		FCR text		clarity of the FCR text	Clarity of a text such as "reject all", e.g. based on a survey-based scoring from 1 to 5.
14		FCG text	sentiment	sentiment of the FCG text	Sentiment of a text such as "allow cookies", e.g. positive.
15	FCR text	sentiment of the FCR text		Sentiment of a text such as "disagree", "reject", e.g. negative.	
16	CC information text	consistency	consistency of the CC information text	Assessment of how consistent terminology is used through the CC text.	
17	CC information text	language	language of the CC information text	Language of the CC information text, e.g. English, Italian, Turkish etc.	
18	process	FCG process	availability	availability of the FCG process	Binary feature. True, if user has a process such as FCG process. Otherwise False.
19		FCR process		availability of the FCR process	Binary feature. True, if use has a process such as FCR process. Otherwise False.
20		FCG process	number of clicks	minimum number of clicks for the FCG process	Shortest user path to perform FCG process in number of clicks.
21		FCR process		minimum number of clicks for the FCR process	Shortest user path to perform FCR process in number of clicks.
22		FCG process	number of paths	number of paths for the FCG process	Total number of paths are available to the users for the FCG process.
23		FCR process		number of paths for the FCR process	Total number of paths are available to the users for the FCR process.
24		CC setting process	time	time for CC setting process	Duration of the CC setting configuration in seconds.
25		FCG process		time for the FCG process	Duration of the FCG process in seconds.
26		FCR process		time for the FCR process	Duration of the FCR process in seconds.
27		first interaction process		time for the first interaction process	Duration between the moment CC is presented to the user and user's first action in seconds.
28	CRR process	frequency	frequency of the CRR process	Frequency of the CRR after consent decision is given by user.	

Appendix D

Dark Pattern Categories and Sources of the Curated Dataset

Table D.1: Sources of the curated dataset

Source	URL
Web	https://www.deceptive.design/types/
Web	https://hallofshame.design/
Web	https://darkpatterns.uxp2.com/patterns/
Web	https://tuta.com/blog/outlook...
Publication	https://dl.acm.org/doi/10.1145/3359183
Report	https://www.edpb.europa.eu/system/files
Web	https://thomasmildner.me/darkpatterns.html
Web	https://givewp.com/addons/recurring-donations/
Web	https://paymentsplugin.com/blog/donation-page/
Web	https://uxplaybook.org/articles/ux-dark...
Report	https://www.ftc.gov/system/files...
Web	https://www.osano.com/articles/dark-pattern...
Web	https://think.design/blog/responsible-design...
Web	https://blog.crobox.com/article/dark-patterns
Web	https://things.qz.com/2019/dark-patterns...
Web	https://blog.mobiversal.com/dark-patterns...

Table D.2: Target deceptive design pattern categories

No.	Deceptive Design Pattern Type
1	False Hierarchy
2	Disguised Ads
3	Sneak into Basket
4	Hidden Costs
5	High Demand
6	Low Stock
7	Endorsement and Testimonials
8	Bad Defaults
9	Countdown Timer
10	Limited Time Message
11	Confirmshaming
12	Trick Questions
13	Activity Messages

Appendix E

Detailed Classification Reports: Comparative Analysis of Prompting Strategies

Table E.1: Binary classification report of the generator model that employs **zero-shot** prompting

Class	Precision	Recall	F1-Score	Support
Dark Pattern	0.96	1.00	0.98	2084
Not Dark Pattern	0.99	0.94	0.96	1354
Accuracy			0.97	3438
Weighted Avg	0.97	0.97	0.97	3438

Table E.2: Dark Pattern Category Classification Report of the generator model that employs **zero-shot** prompting

Class	Precision	Recall	F1-Score	Support
Scarcity	0.97	0.63	0.76	477
Social Proof	0.96	1.00	0.98	278
Urgency	0.65	0.98	0.78	326
Confirmshaming	1.00	0.83	0.91	42
Obstruction	1.00	1.00	1.00	13
Trick Questions	0.00	0.00	0.00	1
Forced Action	1.00	1.00	1.00	6
Accuracy			0.83	1143
Weighted Avg	0.88	0.83	0.83	1143

Table E.3: Binary classification report of the generator model that employs **few-shot** prompting

Class	Precision	Recall	F1-Score	Support
Dark Pattern	0.98	0.99	0.98	2021
Not Dark Pattern	0.98	0.97	0.97	1484
Accuracy			0.98	3505
Weighted Avg	0.98	0.98	0.98	3505

Table E.4: Dark pattern category classification report of the generator model that employs **few-shot** prompting

Class	Precision	Recall	F1-Score	Support
Scarcity	0.99	0.73	0.84	545
Social Proof	0.99	0.99	0.99	342
Urgency	0.71	1.00	0.83	358
Confirmshaming	1.00	0.87	0.93	55
Obstruction	1.00	0.80	0.89	5
Trick Questions	0.00	0.00	0.00	1
Forced Action	1.00	1.00	1.00	5
Accuracy			0.88	1311
Weighted Avg	0.91	0.88	0.88	1311

Table E.5: Binary classification report of the generator model that employs **Chain of Thought** prompting

Class	Precision	Recall	F1-Score	Support
Dark Pattern	0.98	1.00	0.99	2102
Not Dark Pattern	1.00	0.98	0.99	1616
Accuracy			0.99	3718
Weighted Avg	0.99	0.99	0.99	3718

Table E.6: Dark pattern category classification report of the generator model that employs **Chain of Thought** prompting

Class	Precision	Recall	F1-Score	Support
Scarcity	1.00	0.71	0.83	592
Social Proof	0.99	1.00	1.00	360
Urgency	0.69	1.00	0.81	377
Confirmshaming	0.99	0.97	0.98	72
Obstruction	1.00	1.00	1.00	16
Trick Questions	0.00	0.00	0.00	2
Forced Action	1.00	1.00	1.00	2
Accuracy			0.88	1421
Weighted Avg	0.91	0.88	0.88	1421

Appendix F

Semi-Structured Expert Interview Protocol

After a short introduction by the researchers on the research goal, the interviews were organized in three main parts.

Part 1: Demographics

1. What is your professional area of expertise and how long have you been working in the area?
2. How long have you been working on Dark Patterns problems?

Part 2: Explanation of assessment Now I would like you to reflect on the scores you provided to the explanations.

1. Can you tell us your opinion in general about the [clarity, completeness, correctness, verifiability] of the explanations you assessed?
2. (if there is one with extremely bad rating): You marked one of the explanations you received as particularly bad, comparing with the rest. Can you explain why? What was the difference between this and others explanations?
3. (if some of the ratings are much higher, compared to the rest): You marked one of the explanations you received as particularly good, comparing with the rest. Can you explain why? What was the difference between this and others explanations?
4. (if there are groups of different ratings, e.g 4-5 and 1-2, or a totally diverse sample): In your sample you marked some of the explanations as much better than others. Can you reflect on these two groups and explain the difference between them

5. Beyond these four criteria, would you like to add something more about your perception of quality of explanations?
6. Is there any other criteria that you believe would be useful to evaluate the quality of explanations?

Part 3: Usefulness of the AI-based explanations (Explanation about the tool) We have created a detection tool based on a multi-model large language model that, based on screenshots and prompts of the researcher, assesses the presence of dark patterns and provides explanations for the decision it takes. To increase its accuracy and reliability, instead of relying on publicly available data to answer the prompts, (the design contains Retrieval Augmented Generation techniques) our tool considers a few peer-reviewed academic articles on dark patterns, which contain definitions, attributes and examples.

1. In your domain of practice how useful a tool like this can be?
2. Can you envision possible applications?
3. Would you personally use this tool?
4. (if yes) For which purposes would it be suitable to you
5. (if no) Why? Is there any modification that it could make it useful to you?

Conclusions

1. Is there anything else that you would like to add?
2. Is there anything you would like to ask us?