Poster: 50 Shades of Deceptive Patterns: A Unified Taxonomy, Multimodal Detection, and Security Implications

Zewei Shi*†, Ruoxi Sun†, Jieshan Chen†, Jiamou Sun†, Minhui Xue†, Yansong Gao‡, Feng Liu*, Xingliang Yuan*
*University of Melbourne, Australia
†CSIRO's Data61, Australia
‡University of Western Australia, Australia

Abstract

Deceptive patterns (DPs) are user interface designs deliberately crafted to manipulate users into unintended decisions, often by exploiting cognitive biases for the benefit of companies or services. While numerous studies have explored ways to identify these deceptive patterns, many existing solutions require significant human intervention and struggle to keep pace with the evolving nature of deceptive designs. To address these challenges, we expanded the deceptive pattern taxonomy from security and privacy perspectives, refining its categories and scope. We created a comprehensive dataset of deceptive patterns by integrating existing small-scale datasets with new samples, resulting in 6,725 images and 10,421 DP instances from mobile apps and websites. We then developed DPGuard, a novel automatic tool leveraging commercial multimodal large language models (MLLMs) for deceptive pattern detection. Experimental results show that DPGuard outperforms state-of-the-art methods. An extensive empirical evaluation on 2,000 popular mobile apps and websites reveals that 25.7% of mobile apps and 49.0% websites feature at least one deceptive pattern instance. Through 4 unexplored case studies that inform security implications, we highlight the critical importance of the unified taxonomy in addressing the growing challenges of Internet deception.

I. Main Content

This work [1] was recently accepted to The 2025 ACM Web Conference (formerly known as the International World Wide Web Conference, abbreviated as WWW) and the assigned DOI is: https://doi.org/10.1145/3696410.3714593. The original abstract and author list are shown above. Since the work is not yet published, we are providing the paper link to the arXiv version[1] here.

References

[1] Z. Shi, R. Sun, J. Chen, J. Sun, M. Xue, Y. Gao, F. Liu, and X. Yuan, "50 shades of deceptive patterns: A unified taxonomy, multimodal detection, and security implications," in *Proceedings of the ACM Web Conference 2025 (WWW'25)*, Sydney, NSW, Australia, 2025.

[1] https://arxiv.org/abs/2501.13351

50 Shades of Deceptive Patterns: A Unified Taxonomy, Multimodal Detection, and Security Implications

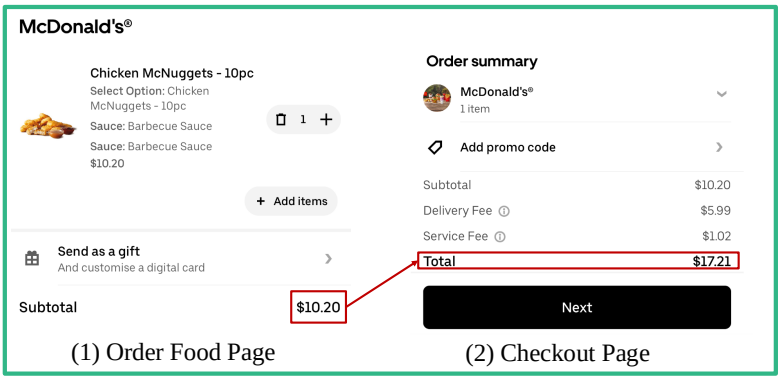
Zewei Shi[*†], Ruoxi Sun[†], Jieshan Chen[†], Jiamou Sun[†], Minhui Xue[†], Yansong Gao[‡], Feng Liu[*], Xingliang Yuan[*]

[*]University of Melbourne, [†]CSIRO's Data61, [‡]University of Western Australia

Introduction

Deceptive patterns (DPs) are user interface designs deliberately **trick** user into doing things that are **not** in their best interest.

For Example:



(1) Order Food Page (2) Checkout Page

The user **expected** to pay $10 **but had to** pay $17 at the final checkout page. This is an example of a `Hidden Cost`, which is a type of deceptive pattern.

Problem: The Gaps of Current Work

- Taxonomy: **Overlooks security and privacy issues** within DP.
- Dataset: Unable to be **large-scale**, **up-to-date** and **cross-platform** simultaneously.
- Detection: Requires **human effort** during the inference stage.

Motivation

DP exploit cognitive biases, leading to

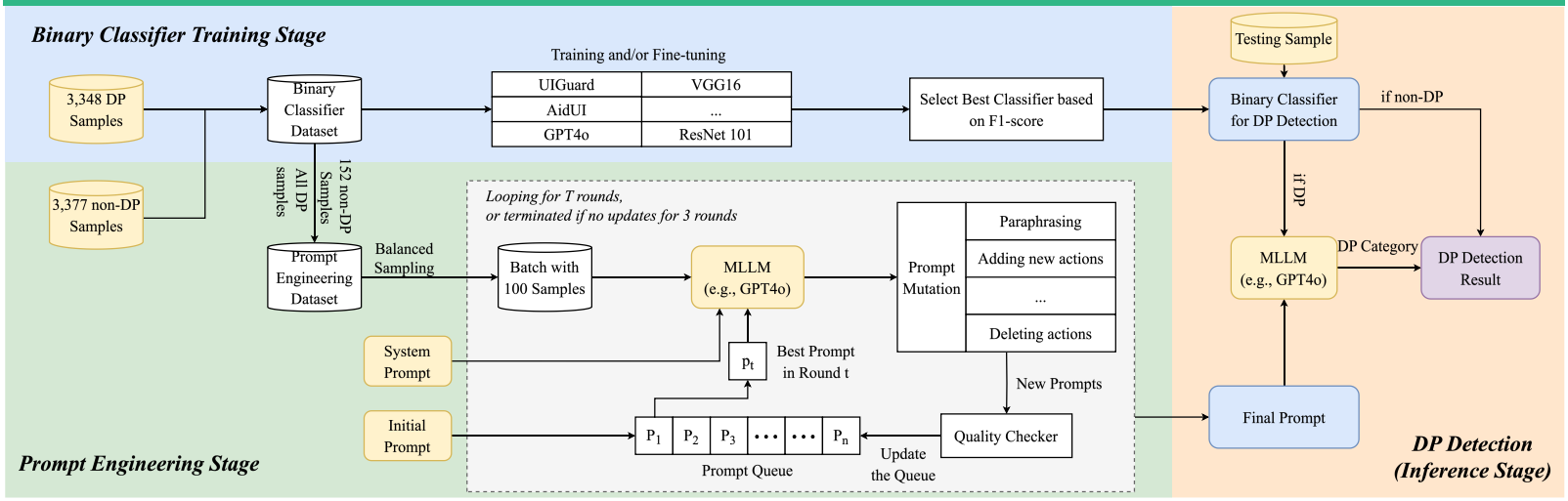
- Financial Losses
- Privacy Breaches
- Broken Trust in digital platforms

DP frequently evolving and widespread use make the existing detection methods **ineffective**, leaving users vulnerable.

Address this issue is crucial to protect:

- User's Security and Privacy
- User's Autonomy
- User's Trust in online interactions

Our Solution: DPGuard



DPGuard Performance

DP Categories	Mobile				Website		
	Instances	UIGuard	AidUI	DPGuard	Instances	AidUI	DPGuard
No DP	3,018	0.8091	0.7812	**0.9807**	359	0.4338	**0.8230**
Nagging	409	**0.4412**	0.3454	0.3876	180	0.1163	**0.4945**
Roach Motel	24	-	-	**0.5484**	13	-	**0.4000**
Price Comparison Prevention	7	-	-	0.0000	27	-	**0.2381**
Intermediate Currency	38	-	-	**0.6154**	5	-	**0.4286**
Forced Continuity	48	0.0408	-	**0.7059**	26	-	**0.3448**
Hidden Costs	38	-	-	**0.2680**	99	-	**0.2519**
Hidden Information	236	-	-	**0.4187**	377	-	**0.4535**
Preselection	356	0.4546	0.3565	**0.5466**	413	**0.3629**	0.2753
Toying with Emotion	84	-	0.1389	**0.3096**	229	0.4251	**0.5866**
False Hierarchy	559	0.4188	0.0552	**0.6535**	320	0.0245	**0.4360**
Disguised Ad	883	0.1520	0.2551	**0.8481**	256	0.2096	**0.8060**
Small Close Button	747	**0.9410**	-	0.4906	160	-	**0.2564**
Social Pyramid	35	**0.6349**	-	0.5047	7	-	**0.3243**
Privacy Zuckering	206	**0.7378**	-	0.4073	367	-	**0.5868**
Gamification	27	0.3529	-	**0.5000**	1	0.0000	0.0000
Countdown on Ads	77	0.2128	0.0000	**0.3952**	10	-	**0.4103**
Watch Ads to Unlock Features or Rewards	67	**0.3488**	-	0.0000	0	-	0.0000
Pay to Avoid Ads	106	**0.7265**	-	0.6277	7	-	**0.1429**
Forced Enrollment	149	-	-	**0.4383**	89	-	**0.3356**
Micro avg	7,114	0.6672	0.5889	**0.7316**	2,945	0.3228	**0.4989**
Macro avg	7,114	0.2851	0.0878	**0.4385**	2,945	0.0715	**0.3452**

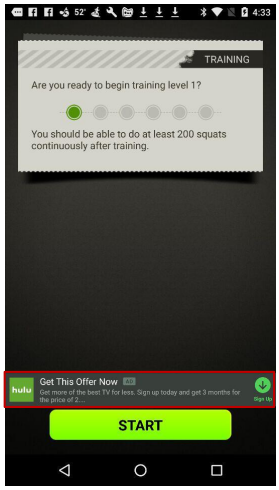
-: the DP category is not supported by the corresponding tool.

Takeaway 1:
DPGuard **outperforms** the state-of-the-art models in DP detection, increasing the F1-score to 0.73 (micro) and 0.44 (macro) on the mobile dataset, and 0.50 (micro) and 0.34 (macro) on the website dataset.

Empirical Evaluation In The Wild

Takeaway 2:
In 1,000 mobile apps (2,950 mobile images) and 1,000 websites (9,396 website images), 25.7% of mobile apps (23.61% of mobile app images) and 49.0% of websites (47.27% of website images) were identified as containing DPs.

Security Implications – Case Study



DP: Disguised Ads

Definition:
Ads presented as normal content include cases where sponsored ads or content are disguised as **banners** or inserted into regular content.

Seriousness Analysis (Alice and Bob Model):

- Disguised ads can mislead Alice into clicking, **redirecting** her to sites controlled by Eve.
- Eve **collects Alice's data** (e.g., device info, browsing habits) and **tracks** her, **violating privacy**.
- Disguised ads may enable Mallory to **launch phishing attacks** or **install malware**.
- Trent, the app platforms, should enforce clear ad labeling to prevent deception, but failure to do so **erodes user trust** and **compromises security**.

This paper has been accepted by The Web Conference (WWW) 2025 (Oral)

Scan the QR Code to learn more