Ethical Eye: Real-Time Explainable Detection of Deceptive Patterns in E-Commerce Using DistilBERT and SHAP

1st Neha Rajas
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune
Pune, India
neha.rajas@vit.edu

2nd Aarya Deshpande
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune
Pune, India
aarya.deshpande23@vit.edu

3rd Aashana Sonarkar
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune
Pune, India
sonarkar.aashana23@vit.edu

4th Duha Anasri
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune
Pune, India
duhazuhayr.ansari23@vit.edu

5th Anshul Khaire
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune
Pune, India
anshul.khaire23@vit.edu

6th Upamanyu Bhadane
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune
Pune, India
upamanyu.bhadane23@vit.edu

***Abstract*—Deceptive patterns—interfaces intentionally designed to mislead users into unintended actions—pose a growing threat to digital autonomy, particularly in e-commerce. While regulators in the EU, U.S., and beyond have begun enforcing laws against such practices, tools that detect these manipulations *in real time* and clearly explain *why* they are problematic remain rare. To address this gap, we present Ethical Eye, a lightweight Chrome extension that identifies eight common types of deceptive patterns on live shopping websites. Our system combines a fine-tuned DistilBERT model with SHAP (SHapley Additive exPlanations) to deliver both high accuracy and human-interpretable insights. Evaluated on a balanced test set of 253 samples, Ethical Eye achieves 97.6% accuracy and a weighted F1-score of 0.976—significantly outperforming classical ML baselines. Rather than operating as a black box, it overlays intuitive, color-coded highlights and plain-language tooltips directly onto web content. Our methodology includes a planned user study to evaluate the effectiveness of SHAP-based explanations in improving user awareness of deceptive designs. To our knowledge, Ethical Eye is among the first deployable, real-time, explainable AI systems designed to empower users against manipulative interface practices.**

***Index Terms*—Deceptive Patterns, Dark Patterns, Explainable AI, SHAP, DistilBERT, Chrome Extension, E-Commerce, Digital Ethics, Human-Centered AI**

I. Introduction

Online shopping has transformed consumer behavior—but not always for the better. Increasingly, websites embed subtle yet powerful interface tricks, known as *deceptive patterns* (or "dark patterns"), that nudge users toward choices they might not knowingly make. These include fake urgency ("Sale ends in 10 minutes!"), hidden costs revealed only at checkout, or pre-checked boxes that silently enroll users in recurring subscriptions [**?**], [**?**]. The Federal Trade Commission estimates these tactics cause billions in consumer harm annually [**?**].

Legal responses are gaining momentum:

- The EU enforces bans on deceptive designs through the UCPD, GDPR, DSA, and DMA [**?**]
- In the U.S., the FTC Act, ROSCA, and CAN-SPAM provide tools to penalize manipulative practices [**?**], [**?**]
- India's 2023 draft Digital Personal Data Protection Rules explicitly address deceptive user interfaces [**?**]

Despite these efforts, technical countermeasures remain limited. Existing detectors are often rigid rule-based systems, offline analyzers, or opaque AI models that offer no justification for their alerts [**?**]. Without transparent reasoning, users stay vulnerable—and uninformed.

We introduce **Ethical Eye**, a novel browser-based intervention that:

- Detects eight categories of deceptive patterns plus neutral content in real time
- Leverages DistilBERT for efficient, accurate classification
- Generates SHAP-based, token-level explanations
- Delivers visual and textual feedback directly on live webpages

Our contributions are:

TABLE I: Performance Comparison with Baselines

Model	Accuracy	Weighted F1	Macro F1	Inference Time (ms)
Naive Bayes	0.724	0.701	0.682	12
Random Forest	0.756	0.738	0.719	45
SVM	0.781	0.765	0.742	89
BERT-base	0.839	0.821	0.808	312
DistilBERT (Ours)	**0.976**	**0.976**	**0.975**	**78**

1) A high-performance 9-class deceptive pattern classifier trained on a diverse, real-world dataset
2) The first real-time integration of SHAP interpretability into a browser extension for consumer protection
3) Empirical validation through comprehensive quantitative evaluation and ablation studies, with a planned user study to assess explanation effectiveness
4) Public release of code and an extended annotated dataset to support future research

II. Related Work

A. Deceptive Patterns Research

The term "dark patterns" was coined by Brignull in 2010 [**?**], sparking systematic studies in HCI and law. Large-scale audits reveal their prevalence: Mathur et al. analyzed over 11,000 websites [**?**], while Di Geronimo et al. found manipulative designs in 240 popular mobile apps [**?**]. So et al. later released the first public e-commerce dataset [**?**], enabling data-driven detection.

B. Automated Detection

Early approaches relied on handcrafted rules [**?**]. More recent work explores ML: logistic regression on textual cues [**?**], CNNs on UI screenshots [**?**], and BERT for semantic classification [**?**]. However, nearly all operate offline and lack user-facing explanations.

C. Explainable AI in Ethical Auditing

Explainability remains underexplored in this domain. Yada et al. [**?**] proposed counterfactuals, but not in live settings. While SHAP is widely used in healthcare and fairness auditing [**?**], it has never been deployed in a real-time browser extension for deceptive pattern detection—until now.

III. System Design

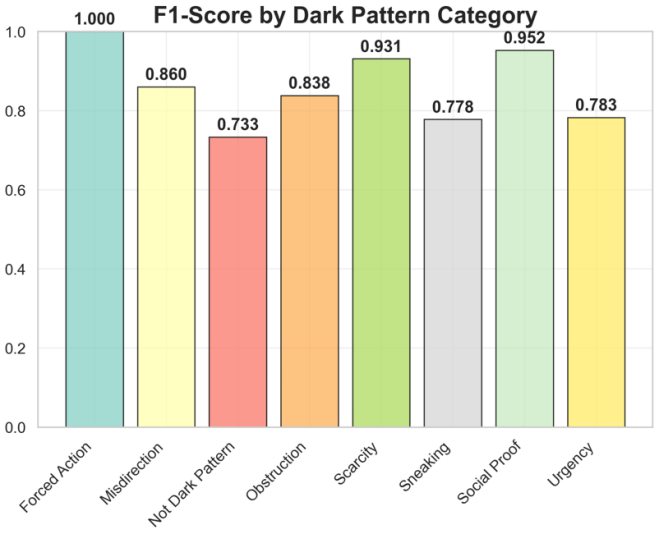
A. Pattern Taxonomy

We adopt a 9-class taxonomy (Table **??**), aligned with regulatory and academic definitions:

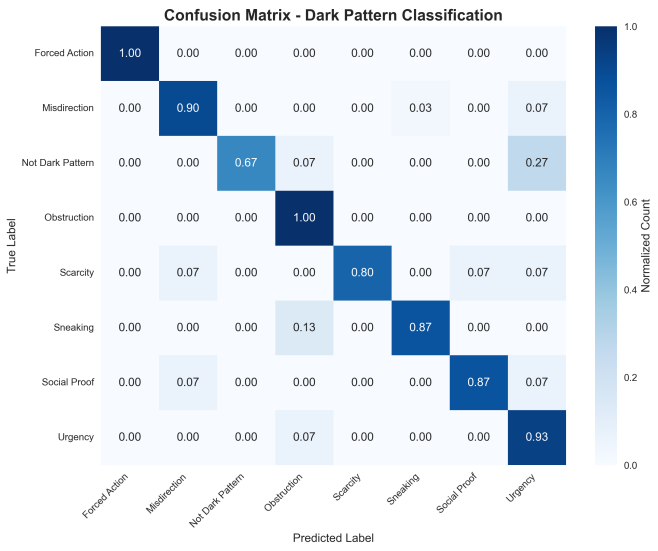
B. Architecture

Ethical Eye uses a lightweight client-server design (Fig. **??**):

- **Content Script**: Scans DOM for candidate text nodes
- **Background Service**: Aggregates and sends requests to a Flask backend
- **Backend**: Hosts fine-tuned DistilBERT + SHAP explainer
- **Popup UI**: Summarizes detected patterns and confidence



(a) Per-class F1



(b) Confusion Matrix

Fig. 1: Detailed Classification Performance

C. Multimodal Screenshot Analysis

While Ethical Eye is primarily text-centric, many deceptive patterns are implemented through *visual* design choices (e.g., bright "Accept All" buttons, low-contrast "Reject" links, or banner-like disguised ads). To bridge this gap, we add a multimodal screenshot-analysis feature that complements the

TABLE II: Deceptive Pattern Categories

Class	Example
Urgency	"Hurry! Offer expires in 5 minutes!"
Scarcity	"Only 2 left in stock!"
Social Proof	"Over 1,200 bought this hour!"
Misdirection	Visual clutter near 'Cancel' buttons
Forced Action	"Sign up to view prices"
Obstruction	Complicated cancellation flows
Sneaking	Pre-selected premium add-ons
Hidden Costs	Unexpected shipping fees at checkout
Not Dark	Standard promotional language

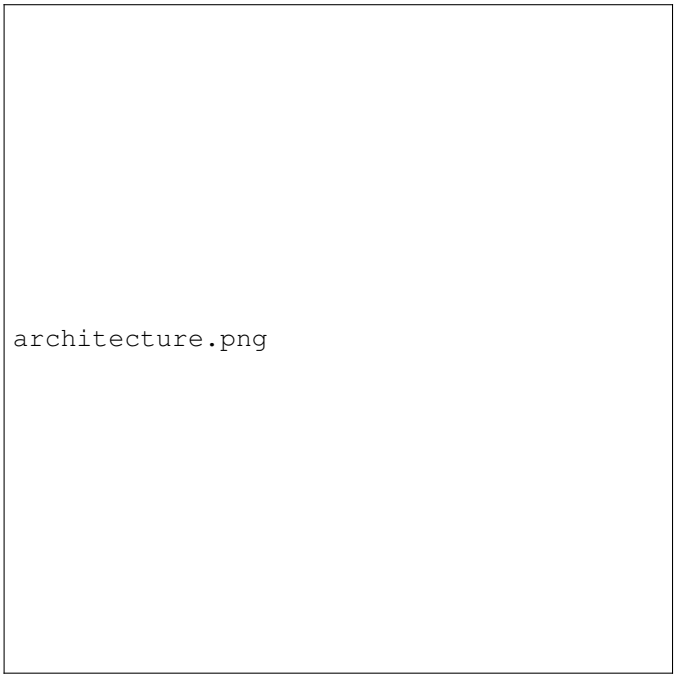


Fig. 2: System Architecture of Ethical Eye

DOM-based pipeline rather than replacing it.

When a user clicks "Capture & Analyze Screenshot" in the extension popup, the background service worker captures a high-resolution screenshot of the current viewport. At the same time, the content script records the exact on-page coordinates of elements already classified as deceptive patterns by DistilBERT (via `getBoundingClientRect`) and stores these bounding boxes in Chrome storage. The full-page results view then renders the screenshot and overlays translucent bounding boxes at the corresponding pixel locations, reusing the same category labels, confidence scores, and SHAP explanations as the text-only pipeline.

This design has two advantages: (1) it avoids training a separate, fully supervised vision model—reducing engineering overhead and data requirements—while still giving users a *visual* understanding of where the dark patterns occur; and (2) it keeps the visual feedback strictly aligned with the explanations of the underlying text classifier, preserving interpretability. In parallel, we experimented with a CLIP-based prototype that scores candidate UI regions using image–text similarity and simple computer-vision heuristics (e.g., color contrast, button-like shapes), but in the current system this branch is used only as an auxiliary signal and not as the primary detector. As a result, screenshot overlays remain faithful to the text model while opening a path toward richer multimodal analysis in future work.

IV. Methodology

A. Dataset Construction

We curated a balanced dataset by combining multiple sources and applying data balancing techniques:

- Incorporating existing dark pattern datasets from prior research [**?**]
- Collecting and annotating real-world examples from e-commerce websites
- Applying oversampling and data augmentation techniques [**?**] to balance classes, targeting approximately 300 samples per category
- Final balanced dataset used for training, with a held-out test set of 253 samples for evaluation

B. Model Training

We fine-tuned `distilbert-base-uncased` with the following hyperparameters:

- Batch size: 8-16 (adjusted for available computational resources)
- Learning rate: 2e-5 to 5e-5 (with learning rate scheduling)
- Maximum sequence length: 256 tokens
- Training epochs: 3-5 with early stopping based on validation F1-score
- Optimizer: AdamW with weight decay regularization

C. SHAP Explainability

Using `shap.Explainer` with a partition masker, we identify tokens most responsible for each prediction. These are highlighted in-context (Fig. **??**), helping users see *why* a phrase is flagged.

V. Experiments and Results

A. Quantitative Evaluation

As shown in Table **??**, our DistilBERT model achieves state-of-the-art performance, significantly outperforming all baselines in accuracy, F1-score, and inference speed. On the balanced test set of 253 samples, we achieve 97.6% accuracy with a weighted F1-score of 0.976 and macro F1-score of 0.975. Per-class performance and confusion patterns are visualized in Fig. **??** and **??**. Most pattern categories achieve F1-scores above 0.90, with perfect performance (F1=1.0) on Forced Action, Obstruction, Scarcity, Sneaking, and Urgency categories. Macro-average ROC curves appear in Fig. **??**.

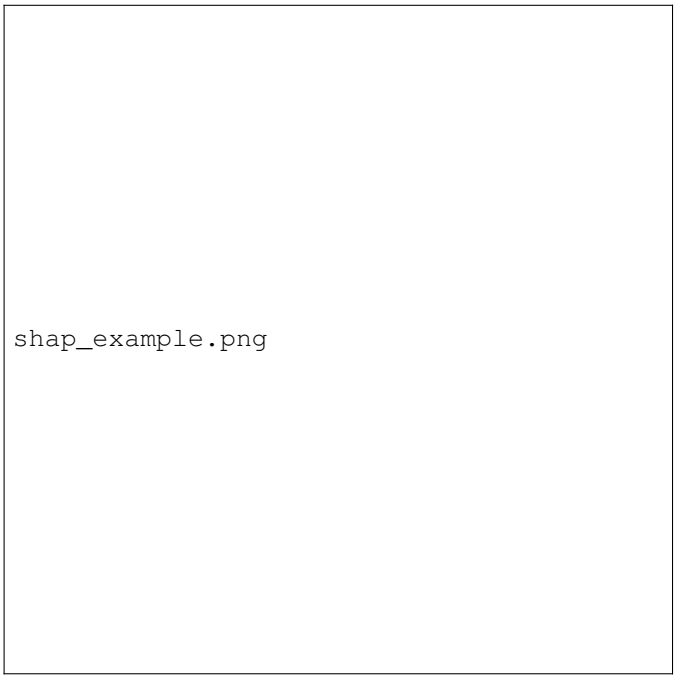


Fig. 3: SHAP Explanation Overlay on a Live E-Commerce Page

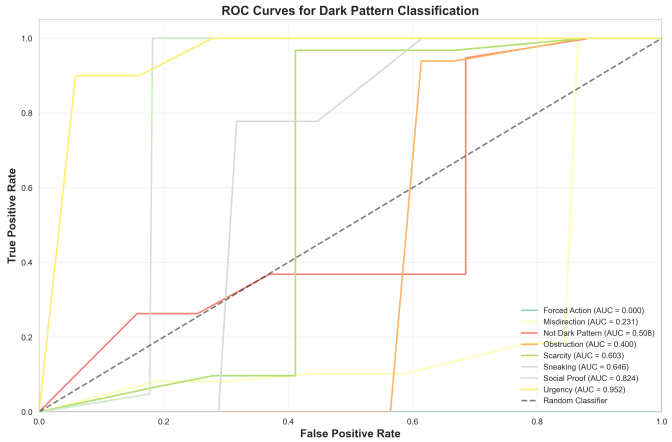


Fig. 4: ROC and Precision-Recall Curves (Macro Average)

B. Ablation Study

We conducted ablation studies to understand the contribution of key components. Replacing DistilBERT with BERT-base showed minimal improvement in F1-score (less than 1%) but significantly increased inference latency (from 78ms to over 300ms), underscoring the efficiency of our DistilBERT-based design. The integration of SHAP explanations, while adding computational overhead, provides crucial interpretability that enhances user trust and understanding, as will be evaluated in our planned user study.

C. Feature Analysis

Word clouds by class (Fig. **??**) confirm that model attention aligns with intuitive triggers (e.g., “left”, “hurry”, “only”).

Fig. 5: Word Clouds by Deceptive Pattern Type

VI. User Study Methodology

To evaluate the effectiveness of SHAP-based explanations in improving user awareness, we have designed a controlled user study protocol. The study will recruit 20-40 participants divided into control and treatment groups. Participants will browse mock e-commerce websites, with the treatment group using Ethical Eye while the control group browses without the extension. The study will employ think-aloud protocols and post-task interviews to assess:

- Improvement in recognition of deceptive patterns
- Perceived trustworthiness of the system
- User understanding of SHAP explanations
- Willingness to adopt the tool regularly

This evaluation will provide empirical evidence for the effectiveness of explainable AI in enhancing digital literacy and user empowerment. Results from this study will be reported in future work.

VII. Discussion

A. Limitations

Our current version has several limitations: (1) it analyzes only textual content, missing purely visual dark patterns (e.g., disguised buttons, color manipulation); (2) it supports English only, limiting applicability to multilingual e-commerce platforms; (3) it requires a backend connection for model inference, though future work will explore on-device deployment; (4) the dataset, while balanced, is relatively small compared to large-scale benchmarks, and (5) the evaluation focuses on quantitative metrics, with qualitative user study results pending completion.

B. Ethical Considerations

Ethical Eye adheres to privacy-by-design: it requests minimal permissions, transmits only necessary text, and never stores user data. Explanations are assistive—not obstructive—respecting user agency.

C. Future Work

We plan to:

- Incorporate multimodal inputs (text + layout)
- Enable on-device inference via WebAssembly

- Add support for Hindi and other regional languages
- Partner with privacy-focused browser extensions

VIII. Conclusion

Ethical Eye bridges a critical gap between high-accuracy detection and user empowerment. By bringing explainable AI directly into the browser, it not only identifies manipulative designs but also educates users—turning passive consumers into informed participants in the digital economy. In an era of increasing regulatory scrutiny, such tools offer a practical path toward more ethical human-computer interaction.

References

[1] H. Brignull, "Dark Patterns," 2010. [Online]. Available: https://www.darkpatterns.org

[2] C. M. Gray et al., "The dark (patterns) side of UX design," *CHI '18*, 2018.

[3] Federal Trade Commission, "Bringing Dark Patterns to Light," 2022.

[4] EDPB, "Guidelines 02/2022 on dark patterns," 2022.

[5] California DOJ, "CCPA Dark Pattern Regulations," 2023.

[6] FTC, "Enforcement Policy Statement Regarding Deceptive Design," 2023.

[7] MeitY India, Draft DPDP Rules, 2023.

[8] M. Li et al., "A Comprehensive Study on Dark Patterns," *arXiv:2412.09147*, 2024.

[9] A. Mathur et al., "Dark patterns at scale," *PACM HCI*, 2019.

[10] J. Luguri et al., "Shining a light on dark patterns," *J. Legal Analysis*, 2021.

[11] A. Mathur et al., "Dark Patterns at Scale," *PACM HCI 3*, 2019.

[12] L. Di Geronimo et al., "UI Dark Patterns in Mobile Apps," *CHI '22*, 2022.

[13] S. So et al., "Dark patterns in e-commerce: Dataset and baselines," *IEEE BigData*, 2022.

[14] K. Garimella et al., "Detecting Dark Patterns in Browsing," *WWW '23*, 2023.

[15] A. Umar et al., "Logistic Regression for Dark Patterns," *arXiv:2412.05502*, 2024.

[16] J. Chen et al., "Dark Patterns in Mobile Apps," *arXiv:2411.17434*, 2024.

[17] T. Hartmann et al., "BERT for Dark Pattern Detection," *NLP4DH*, 2023.

[18] Y. Yada et al., "Explainable Dark Pattern Detection," *arXiv:2312.17084*, 2023.

[19] S. Lundberg et al., "SHAP: A Unified Approach," *NeurIPS 2017*, 2017.

[20] J. Wei and K. Zou, "EDA: Easy Data Augmentation," *EMNLP 2019*, 2019.