

Dark Patterns and the AI Era (and Enforcement)

Dr. Johanna Gunawan, 2 June 2026

SecAppDev'26 -- Leuven

Agenda/Scope

- Design and regulation
- What are dark patterns and perspectives around them?
- Some tools, contributions, methods, and frameworks from dark patterns scholarship
- Recent work on wrangling dark patterns in the DSA
- Looking towards dark patterns & AI

Design and Regulation

An increasing regulatory momentum – but why?

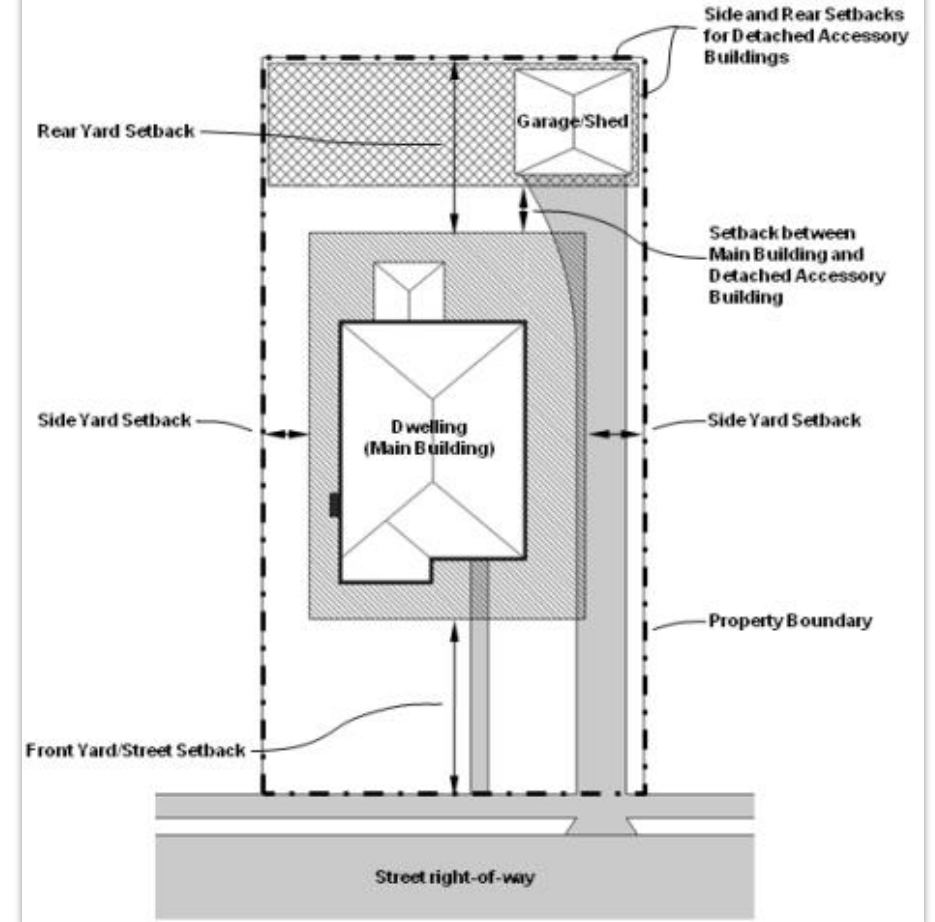
DESIGN IS
NOT NEUTRAL



an investigation
of myth



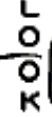
Setback and Yard Requirements





CITY OF NEW YORK
 CODE OF ORDINANCES
 CHAPTER 5
 *BUILDING CODE

Art.	Effective	Page
1. General provisions	March 14, 1916.....	4
2. Materials	March 30, 1915.....	11
3. Working stresses and loads...	March 30, 1915.....	15
4. Classification of buildings.....	March 30, 1915.....	21
5. Restricted areas	March 30, 1915.....	25
6. Height, size and arrangement.	34
7. Light and ventilation.....	March 14, 1916.....	34
8. Exit facilities	March 14, 1916.....	37



Welcome to [GeoCities](#) Your Home On The Web!



[What does this symbol mean?](#)

Today's [Cool Home Page](#):



[SoHo 1085 - Mike's got a Collection of Animated GIFs that must be experienced!](#)

[Enter !\[\]\(06a315363e7801bba8c7489a6694af19_img.jpg\) Here!](#)

Tell us what you think!

Take the first [GeoCities User Survey!](#)

**TAKE
STOCK!**





[Enter for a chance to win \\$2500 worth of stock in your favorite Internet company.](#)

[\[New Stuff\]](#) [\[Info\]](#) [\[Free Homepages\]](#) [\[FAQ\]](#) [\[Announcements\]](#) [\[Neighborhoods\]](#) [\[Cool\]](#) [\[Mail\]](#) [\[Business Partners\]](#) [\[Cameras\]](#)
[\[Search\]](#) [\[Utilities\]](#) [\[Bulletin Board\]](#) [\[Neighborhood Watch\]](#) [\[Help\]](#) [\[Map\]](#)

© 1995,1996 [GeoCities](#). All rights reserved



Designed *to* Sell

GET STARTED

Get your free website trial today.
No credit card required.





Science confirms it: Websites really do all look the same

We studied 10,000 websites and found that their design has become more uniform over time. What does this mean for the future of creative expression on the internet?

17 years of your adult life may be spent online. These expert tips may help curb your screen time

Digital reference library [DataReportal](#) in January released its annual [Global Overview Report](#), revealing internet users ages 16 to 64 spend an average six hours, 40 minutes daily surfing the web on any device. That's about 47 hours a week and 101 days a year. By this estimation, beginning at age 18, a person who lives to be 80 will have spent more than 17 years of their adult life [using the internet](#).

What are dark patterns?

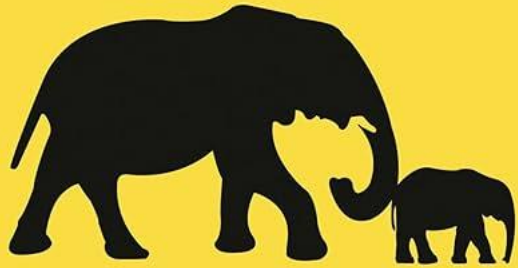
MORE THAN 1.5 MILLION COPIES SOLD

RICHARD H. THALER

WINNER OF THE NOBEL PRIZE IN ECONOMICS

and CASS R. SUNSTEIN

WINNER OF THE HOLBERG PRIZE



Nudge

NEW YORK TIMES Bestseller

Improving Decisions About
Health, Wealth, and Happiness

"One of the few books . . . that fundamentally changes the way I think about the world." —Steven D. Levitt, coauthor of FREAKONOMICS

“Sunstein and I stressed that the goal of a conscientious choice architect is to help people make better choices “as judged by themselves.” But what about activities that are essentially nudging for evil? This “sludge” just mucks things up and makes wise decision-making and prosocial activity more difficult.”

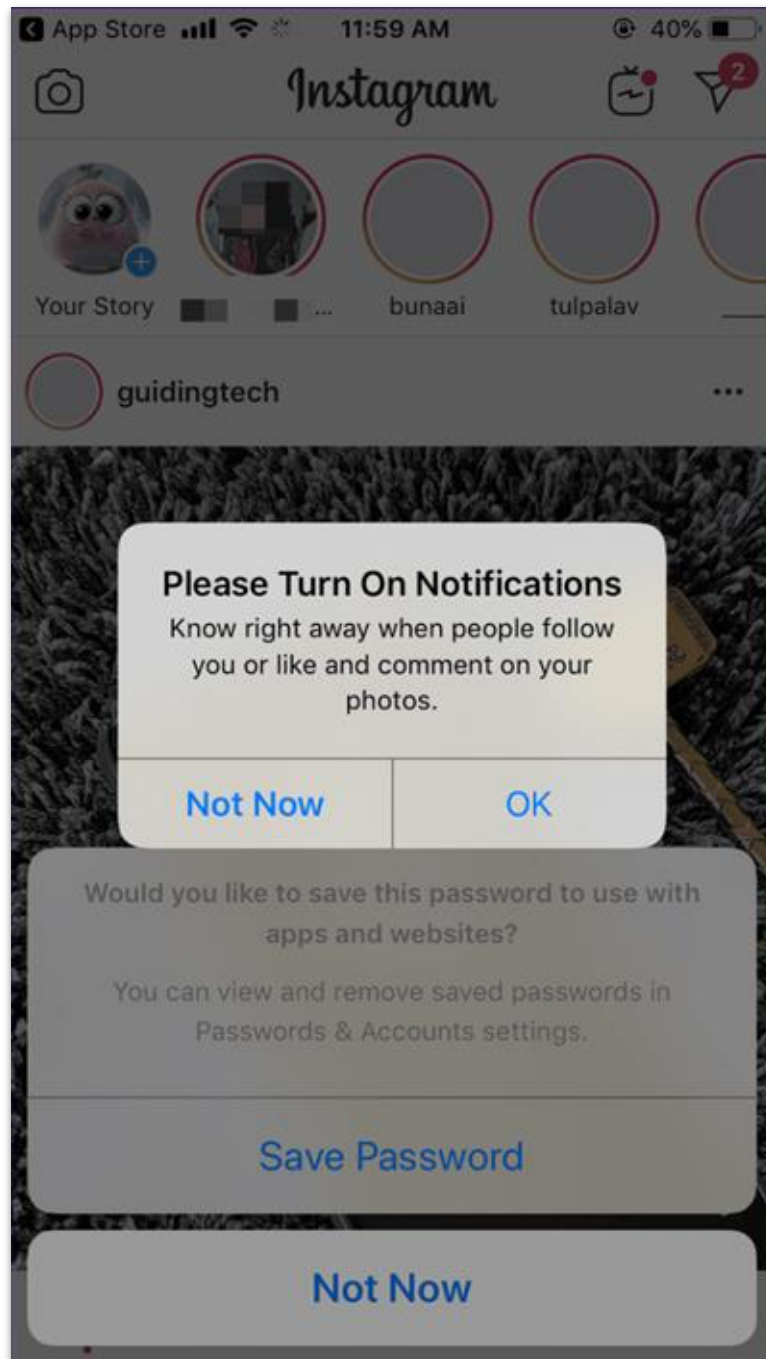
The EU Digital Services Act, Rec. 67 definition of dark patterns:



“practices that materially **distort or impair, either purposefully or in effect, the ability** of recipients of the service **to make autonomous and informed choices or decisions.**”

Those practices can be used to **persuade** the recipients of the service to engage in **unwanted behaviors or into undesired decisions** which have **negative consequences** for them”

(+ examples)



Repeatedly requesting a recipient of the service to make a choice where such a choice has already been made



Tshirt

\$250.00 USD

Size

👛 Only 3 Left In Stock 🔥 12 Sold Just Today

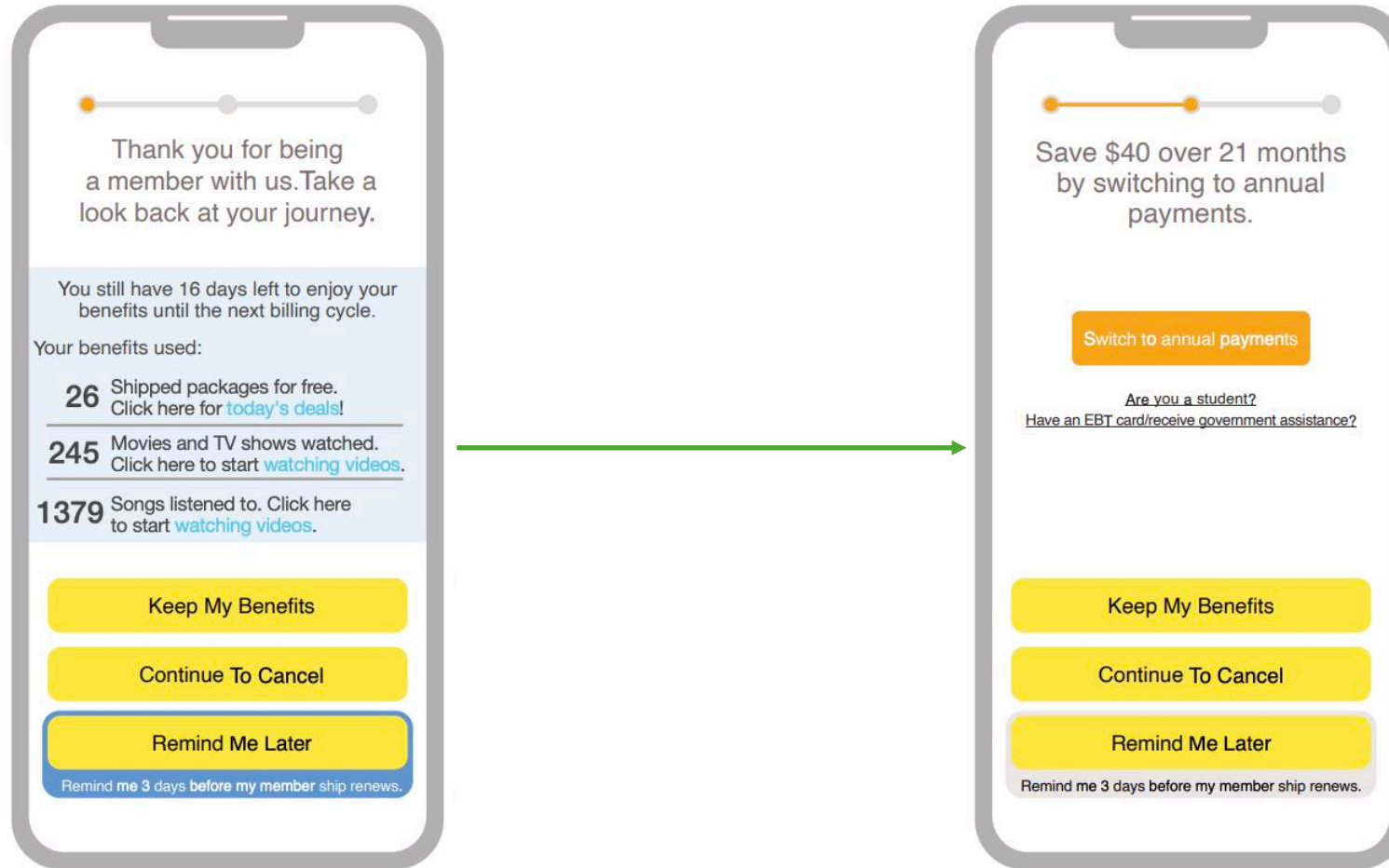
ADD TO CART

BUY IT NOW

The Shanghai T-Shirt is cut in soft stretch cotton jersey to a fitted silhouette with a contrast heat-pressed. Ribbed crewneck with a center-back slit. Short sleeves. Slim fit. 95% cotton, 5% elastane.

SHARE TWEET PIN IT

Deceiving users by nudging them into decisions on transactions



Making cancellation procedure significantly more cumbersome than signing up to it

Christoph Bösch*, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher

Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns

Abstract: Privacy strategies and privacy patterns are fundamental concepts of the privacy-by-design engineering approach. While they support a privacy-aware development process for IT systems, the concepts used by malicious, privacy-threatening parties are generally less understood and known. We argue that understanding the “dark side”, namely how personal data is abused, is of equal importance. In this paper, we introduce the concept of privacy dark strategies and privacy dark patterns and present a framework that collects, documents, and analyzes such malicious concepts. In addition, we investigate from a psychological perspective why privacy dark strategies are effective. The resulting framework allows for a better understanding of these dark concepts, fosters awareness, and supports the development of countermeasures. We aim to contribute to an easier detection and successive removal of such approaches from the Internet to the benefit of its users.

Keywords: Privacy, Patterns

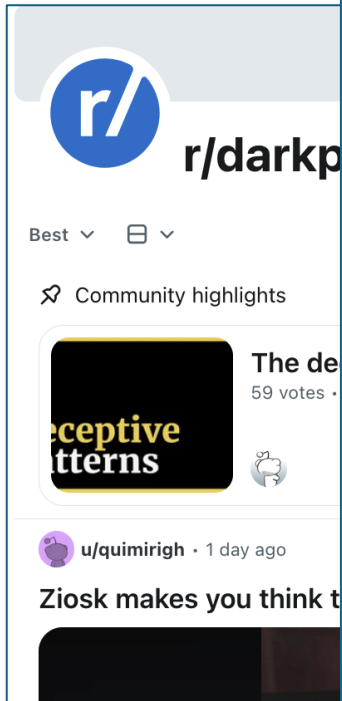
DOI 10.1515/popets-2016-0038

Received 2016-02-29; revised 2016-06-02; accepted 2016-06-02.

However, online service providers have become more and more sophisticated in deceiving users to hand over their personal information. Up until now, privacy research has not studied this development.

An example for this development is the Tripadvisor mobile app (depicted in Figure 1), which is a review platform for travel-related content. At first glance, the starting page asks the user to log in with a personal Google+, Facebook, or email account. Taking a closer look, one notices that a third option is given that offers the creation of a Tripadvisor account. Furthermore, a “Skip”-button is hidden in the upper right corner, which skips the login process entirely. When signing in with Facebook, Tripadvisor wants to gain access to the friend list, photos, likes, and other information (cf. Figure 1b). This is unnecessary for the main features of the service.

Skipping the login process shows the user some features which are available only after signing in (cf. Figure 1c). In addition, the “Later”-button, which finally leads to the app, is located on the left side. Placed on the right side is a “Sign in”-button which leads back to the



“Nothing Comes Before Profit”: *Asshole Design* in the Wild

Shruthi Sai Chivukula
Chris Watkins
cshruthi@purdue.edu
watkin48@purdue.edu
Purdue University
West Lafayette, IN, USA

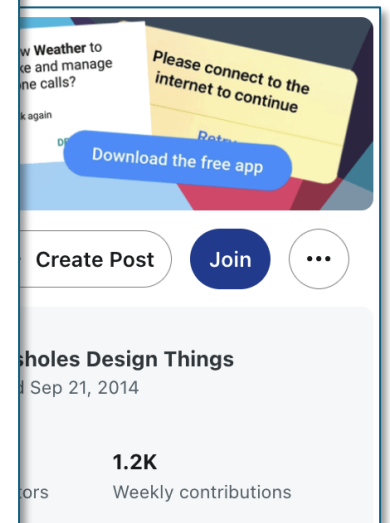
Lucca McKay
Colin M. Gray
mckay11@purdue.edu
gray42@purdue.edu
Purdue University
West Lafayette, IN, USA

ABSTRACT

Researchers in HCI and STS are increasingly interested in describing ethics and values relevant for design practice, including the formulation of methods to guide value application. However, little work has addressed ethical considerations as they emerge in everyday conversations about ethics in venues such as social media. In this late breaking work, we describe online conversations about a concept known as “asshole design” on Reddit, and the relationship of this concept to another practitioner-focused concept known as “dark patterns.” We analyzed 1002 posts from the subreddit ‘/r/assholedesign’ to identify the types of artifact being shared and the interaction purposes that were perceived to be manipulative or unethical as a type of “asshole design.” We identified a subset of these posts relating to dark patterns, quantifying their occurrences using an existing dark patterns typology.

KEYWORDS

Asshole design; dark patterns; design ethics; values; Reddit





DISRUPTIVE COMPETITION PROJECT

HOME

COMPETITION

INTELLECTUAL PROPERTY

DIGITAL TRADE

Examining the FTC's Hostility to Common Design Practices

DANIEL LUQUE · DECEMBER 20, 2023

SHARE



TAGS AMAZON ANTITRUST FTC



NIAS 

Lorentz
center

Workshop @Oort

Fair Patterns for Online Interfaces

29 January - 2 February 2024, Leiden, the Netherlands



Words Matter

Alternatives for Charged Terminology in the Computing Profession

- **dark pattern**: Used to describe how user interfaces can manipulate/deceive users to take actions that they normally would not take. Dark pattern makes a connection between "dark" = bad/deceptive. **Alternatives include deceptive design, manipulative design, or deceptive/manipulative pattern, which are all more descriptive.**

Intent or outcome?

What makes a dark pattern... dark?

[Home](#) >

Enforcement

Businesses that use deceptive patterns (aka 'dark patterns') often get hit with big fines and penalties.

Filters

[Clear all](#)

[Sort by](#) ▾

Showing 100 of 100

Administrative or judicial authority [Clear](#)

- Federal Cartel Office (Bundeskartellamt)
- Superior Court of the State of Arizona - County of Maricopa

Bundeskartellamt's Administrative Proceeding against Google

Google faced liability for processing user data across services without affording users adequate choice options through its selection dialogue interfaces.

Changes made to consent choices interface | EU & UK | Federal Cartel Office (Bundeskartellamt) | October 5, 2023

In the matter of TikTok Technology Limited

TikTok was held liable for nudging children towards privacy-intrusive settings using bold text in two pop-up notifications, hindering neutral and objective choices.

€345 million in fines | EU & UK | Irish Data Protection Commission (DPC) | September 1, 2023



Tech Related Actions and Litigation: The TRAL Tracker

This tracker gathers U.S. and international lawsuits, as well as regulatory oversight actions, all related to emerging technologies and tech companies of all sizes. This project is a collaboration between Tech Justice Law and Georgetown's Knight-Georgetown Institute and Communication, Culture & Technology program.

The tracker spans various issue areas, including antitrust, AI, consumer protection, tech liability/Section 230, privacy, content moderation, addictive design, and more. Students from Georgetown University update this resource.

[Suggest a Case to Track](#)

Developed in partnership with:



Sarah Chao

Arjun Dsouza

Brennan Schaffner



TOTAL MATTERS

205



ACTIVE LITIGATION

137



UPCOMING DEADLINES (7 DAYS)

0

US

World

Filter by:

ALL

AI/Copyright Infringement

Antitrust

Children's Online Safety

Consumer Protection

First Amendment Challenges

National Security/Criminal claims

Privacy Violations and Data Misuse

Products Liability

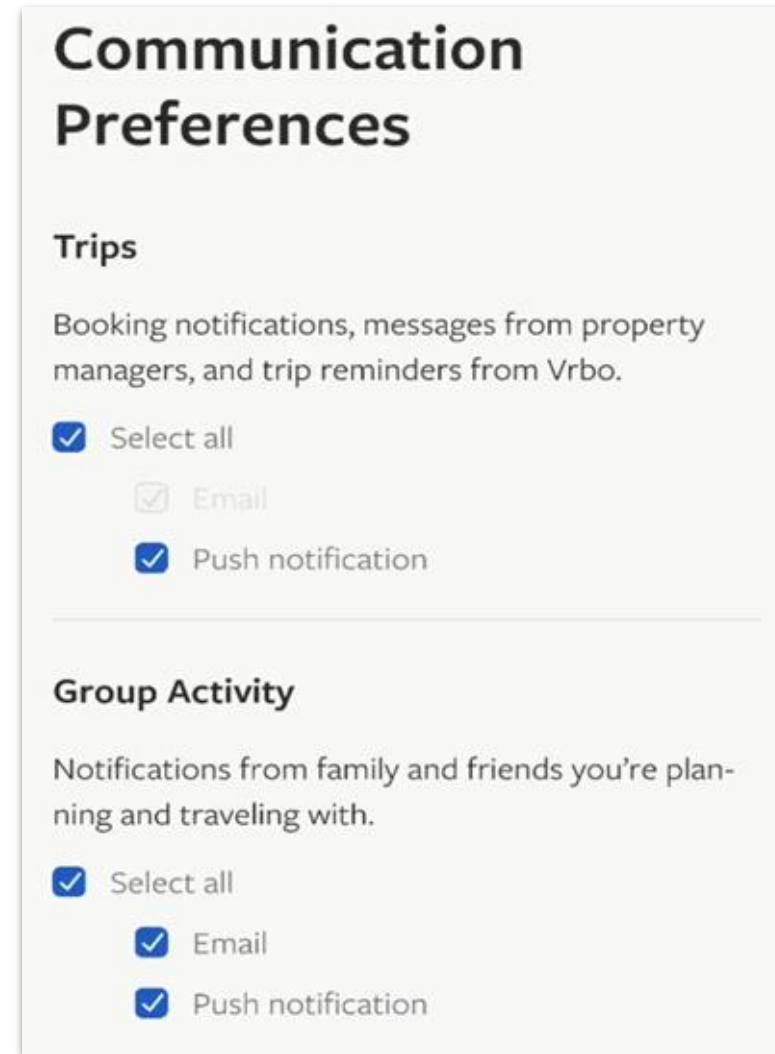
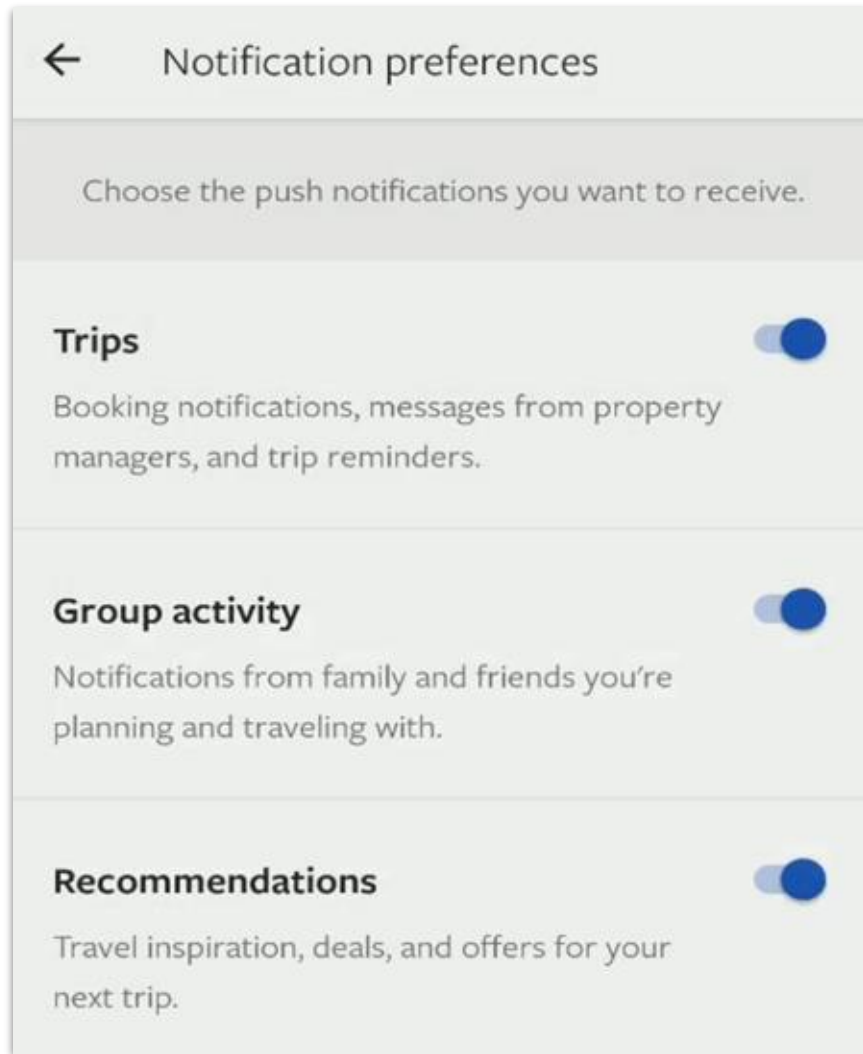
Search cases...

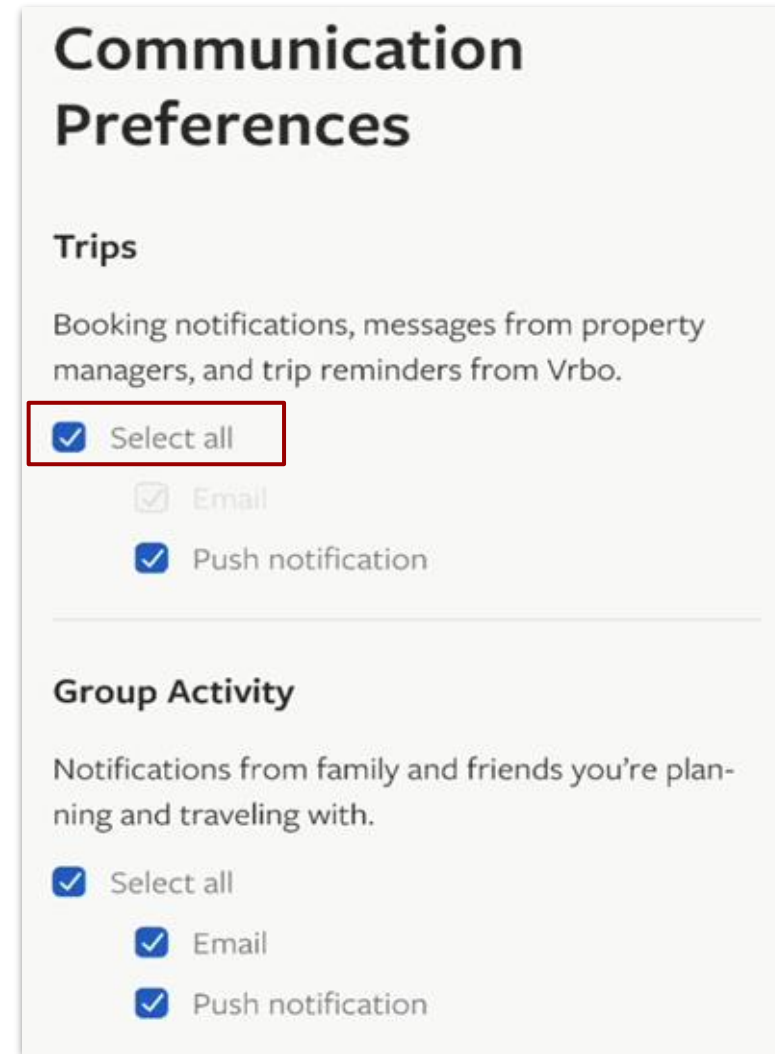
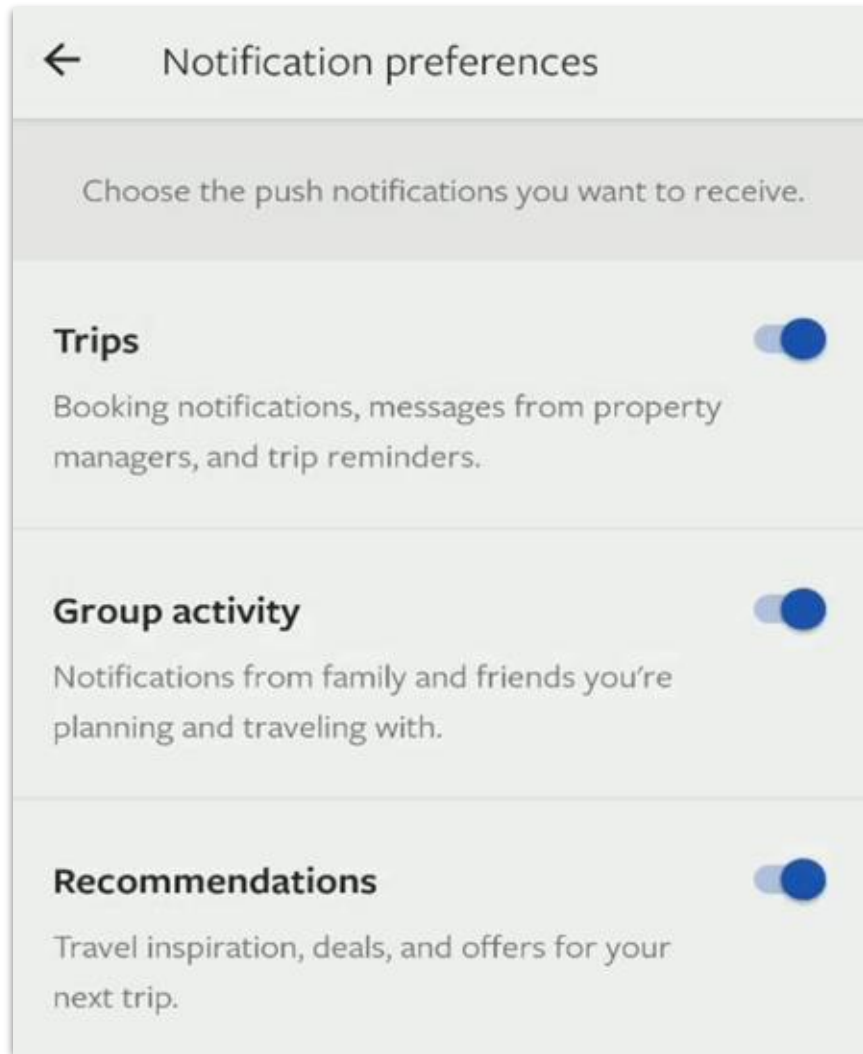
DARK PATTERNS SPOTTED BY PEOPLE LIKE YOU

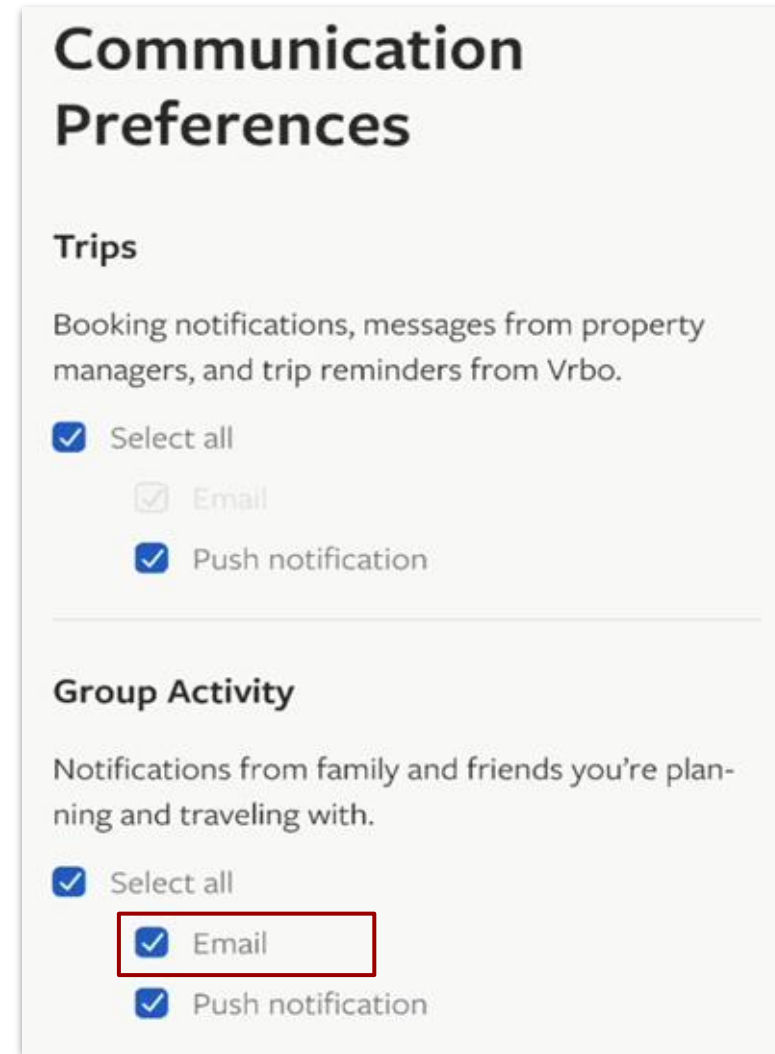
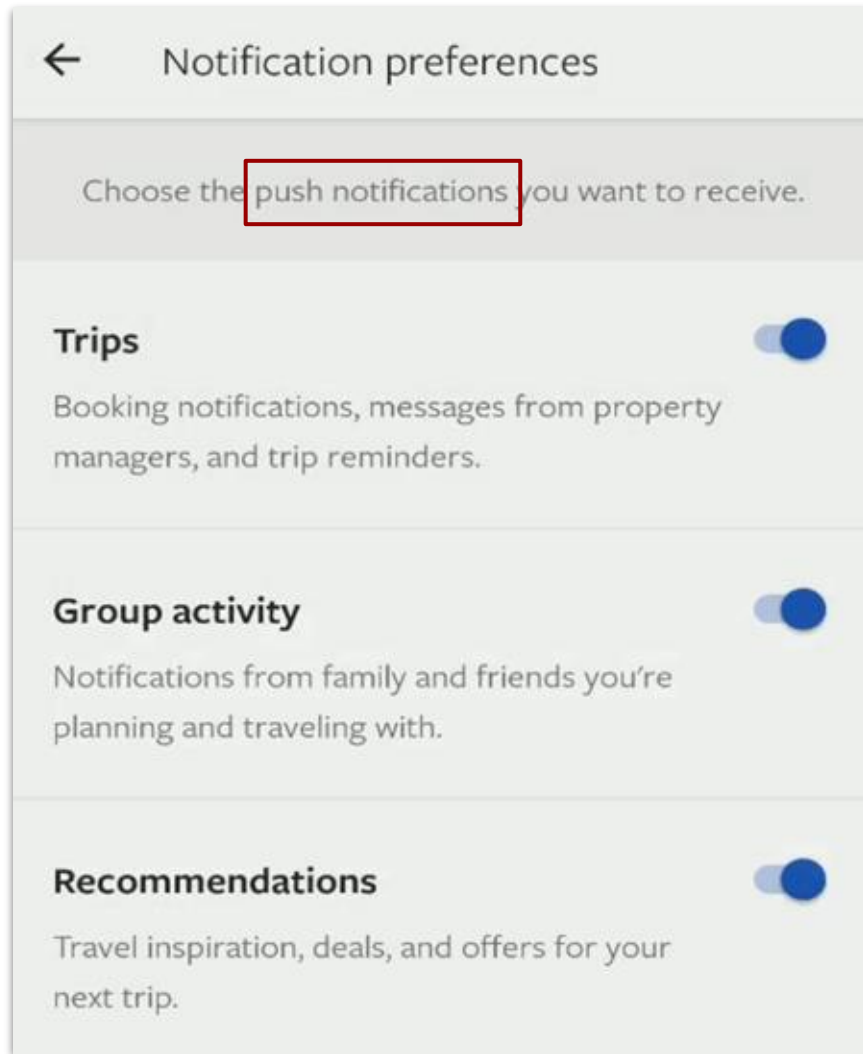
Frustrated people all over the world have found dark patterns even on mainstream websites and apps. Click around to see some examples of how these dark patterns hurt people, and better understand the techniques companies use to trick you into doing what they want. These submissions reflect the views and experiences of the people who submitted them.

REPORT A PATTERN

A bit of my work







MANAGE YOUR PREFERENCES

AGREE TO WHATEVER

TRANSPORT ME TO AN IMMERSIVE
MYST-LIKE GAME WHERE I CLICK
CONFUSINGLY-LABELED TOGGLE
SWITCHES, ONLY SOME OF WHICH
WORK, PERHAPS NEVER TO FIND MY
WAY BACK TO THE PAGE I WANTED.



When dark pattern experiences differ, users take on additional burdens and risk facing unequal, unfair experiences.

Modality

Device Type

+

Manufacturer

Feature
Robustness

Dark Patterns as Disloyal Design

JOHANNA GUNAWAN,^{*} WOODROW HARTZOG,^{**} NEIL RICHARDS,^{***}
DAVID CHOFFNES^{****} & CHRISTO WILSON^{*****}

Lawmakers have started to regulate “dark patterns,” understood to be design practices meant to influence technology users’ decisions through manipulative or deceptive means. Most agree that dark patterns are undesirable, but open questions remain as to which design choices should be subjected to scrutiny, much less the best way to regulate them.

In this Article, we propose adapting the concept of dark patterns to better fit legal frameworks. Critics allege that the legal conceptualizations of dark patterns are overbroad, impractical, and counterproductive. We argue that law and policy conceptualizations of dark patterns suffer from three deficiencies: First, dark patterns lack a clear value anchor for cases to build upon. Second, legal definitions of dark patterns overfocus on individuals and atomistic choices, ignoring de minimis aggregate harms and the societal implications of manipulation at scale. Finally, the law has struggled to articulate workable legal thresholds for wrongful dark patterns. To better regulate the designs called dark patterns, lawmakers need a better conceptual framing that bridges the gap between design theory and the law’s need for clarity, flexibility, and compatibility with existing frameworks.

We argue that wrongful self-dealing is at the heart of what most consider to be “dark” about certain design patterns. Taking advantage of design affordances to the detriment of consumers is a disloyal design. To that end, we propose a disloyal design

Scholarly Contributions to Dark Patterns

Ontologies, methods, and engagement with enforcement

Ontology of Dark Patterns Knowledge

REGULATORY TAXONOMIES



ACADEMIC TAXONOMIES

Brignul 2018, 2023
 Bösch et al., 2013
 Gray et al., 2018
 Mathur et al., 2019
 Luguri et al. 2021

High-Level Pattern	Meso-Level Pattern	Low-Level Pattern
	Nagging (D: Gr Lu Br23 EUCOM FTC OECD; I: EDPB CMA)	-
	Forced Continuity (D: Br Gr I: Lu Ma Br23 FTC EUCOM OECD)	-
	Forced Registration (D: Bö Lu FTC EUCOM OECD; I: Bö Ma CMA FTC)	-
Forced Action D: Gr Lu Ma EUCOM OECD I: CMA FTC		Privacy Zuckering (D: Br Bö Gr Lu; I: FTC OECD)
	<i>Forced Communication or Disclosure</i>	Friend Spam (D: Br; I: Lu FTC OECD) Address Book Leeching (D: Bö; I: Lu FTC OECD) Social Pyramid (D: Gr; I: Lu FTC OECD)
	Gamification (D: Gr Lu OECD)	Pay-to-Play (D: FTC) Grinding (D: FTC)
	<i>Attention Capture</i>	Auto-Play (D: FTC)
	Scarcity and Popularity Claims (D: CMA; I: Ma Lu Br23 FTC)	High Demand (D: Ma Lu FTC EUCOM OECD) Low Stock (D: Ma Lu FTC EUCOM OECD)
	Social Proof (D: Ma Lu EUCOM OECD; I: Br23)	Endorsements and Testimonials (D: Ma Lu FTC EUCOM OECD) Parasocial Pressure (I: FTC)
Social Engineering	Urgency (D: Ma Lu FTC EUCOM OECD; I: Br23)	Activity Messages (D: Ma Lu FTC EUCOM OECD) Countdown Timer (D: Ma Lu FTC; I: EUCOM OECD) Limited Time Message (D: Ma Lu FTC; I: EUCOM OECD)
	Shaming	Confirmshaming (D: Br Ma Lu Br23 FTC EUCOM; I: OECD)
	Personalization (D: CMA)	-

Ontology of Dark Patterns Knowledge

REGULATORY TAXONOMIES



ACADEMIC TAXONOMIES

Brignul 2018, 2023
Bösch et al., 2013
Gray et al., 2018
Mathur et al., 2019
Luguri et al. 2021

‘Auto-play’ uses *Attention Capture* as a type of *Forced Action* to automatically play new video after an existing video has completed. As a result, the user may lose control over their viewing experience, leading them to watch more content than they intended or result in them watching content that is unexpected or harmful.

High-Level Pattern	Meso-Level Pattern	Low-Level Pattern
	Nagging (D: Gr Lu Br23 EUCOM FTC OECD; I: EDPB CMA)	-
	Forced Continuity (D: Br Gr I: Lu Ma Br23 FTC EUCOM OECD)	-
	Forced Registration (D: B6 Lu FTC EUCOM OECD; I: B6 Ma CMA FTC)	-
Forced Action D: Gr Lu Ma EUCOM OECD I: CMA FTC		Privacy Zuckering (D: Br B6 Gr Lu; I: FTC OECD)
	Forced Communication or Disclosure	Friend Spam (D: Br; I: Lu FTC OECD)
		Address Book Leeching (D: B6; I: Lu FTC OECD)
		Social Pyramid (D: Gr; I: Lu FTC OECD)
	Gamification (D: Gr Lu OECD)	Pay-to-Play (D: FTC)
	Attention Capture	Grinding (D: FTC)
		Auto-Play (D: FTC)
		High Demand (D: Ma Lu FTC EUCOM OECD)
		Low Stock (D: Ma Lu FTC EUCOM OECD)
		Endorsements and Testimonials (D: Ma Lu FTC EUCOM OECD)
		Parasocial Pressure (I: FTC)
		Activity Messages (D: Ma Lu FTC EUCOM OECD)
		Countdown Timer (D: Ma Lu FTC; I: EUCOM OECD)
		Limited Time Message (D: Ma Lu FTC; I: EUCOM OECD)
		Confirmshaming (D: Br Ma Lu Br23 FTC EUCOM; I: OECD)
	Shaming	-
	Personalization (D: CMA)	-

Social Engineering

Scarcity and Popularity Claims

(D: **CMA**; I: **Ma Lu Br23 FTC**)

Social Proof

(D: **Ma Lu EUCOM OECD**; I: **Br23**)

Urgency (D: **Ma Lu FTC EUCOM OECD**;

I: **Br23**)

Shaming

Personalization (D: **CMA**)

High Demand

(D: **Ma Lu FTC EUCOM OECD**)

Low Stock (D: **Ma Lu FTC EUCOM OECD**)

Endorsements and Testimonials

(D: **Ma Lu FTC EUCOM OECD**)

Parasocial Pressure (I: **FTC**)

Activity Messages

(D: **Ma Lu FTC EUCOM OECD**)

Countdown Timer

(D: **Ma Lu FTC**; I: **EUCOM OECD**)

Limited Time Message

(D: **Ma Lu FTC**; I: **EUCOM OECD**)

Confirmshaming

(D: **Br Ma Lu Br23 FTC EUCOM**; I: **OECD**)

-

Expert testimony and consultation

EXPERT REPORT OF MARSHINI CHETTY, Ph.D.

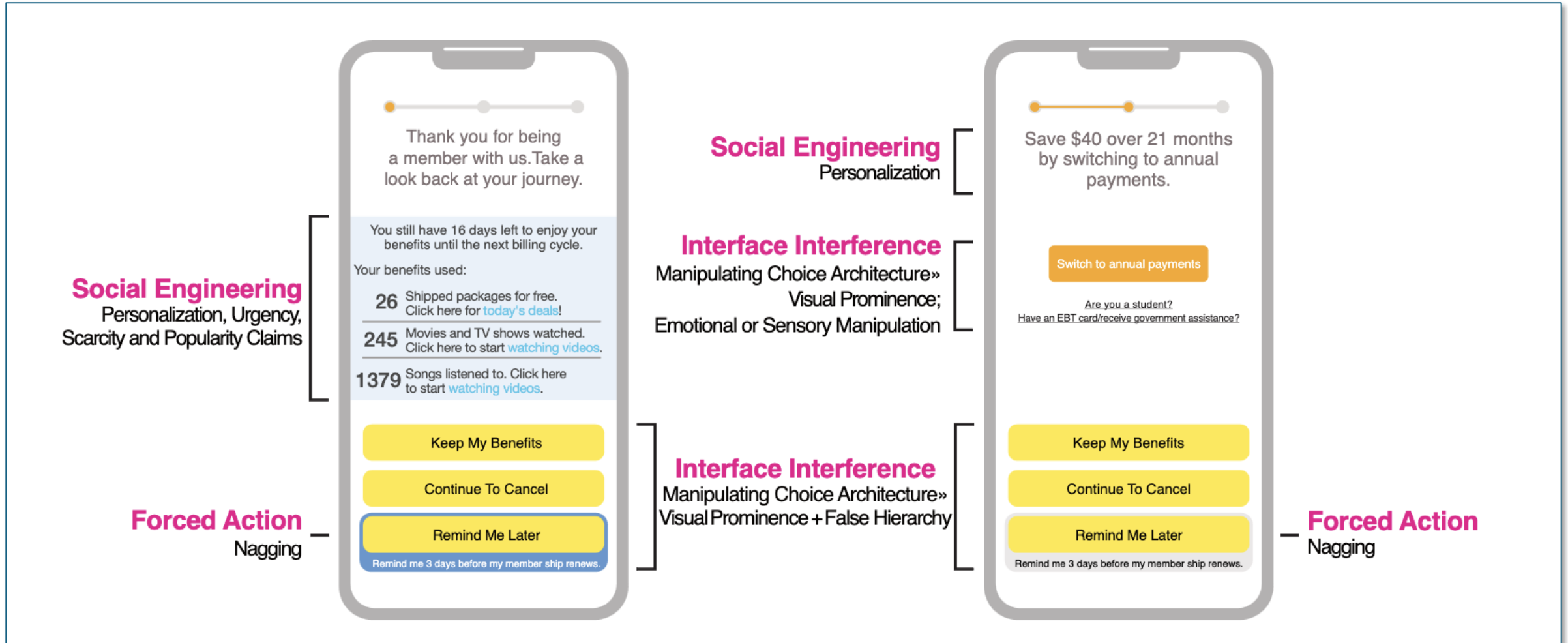
I am an Associate Professor at the University of Chicago's Department of Computer Science, with a Ph.D. in computer science. I have twenty years' experience in human-computer interaction ("HCI") and my research and peer-reviewed publications address, among other things, the usability of web interfaces and the presence of manipulative designs (sometimes known as "dark patterns") on websites.

The Federal Trade Commission ("FTC") asked me to render an expert opinion on:

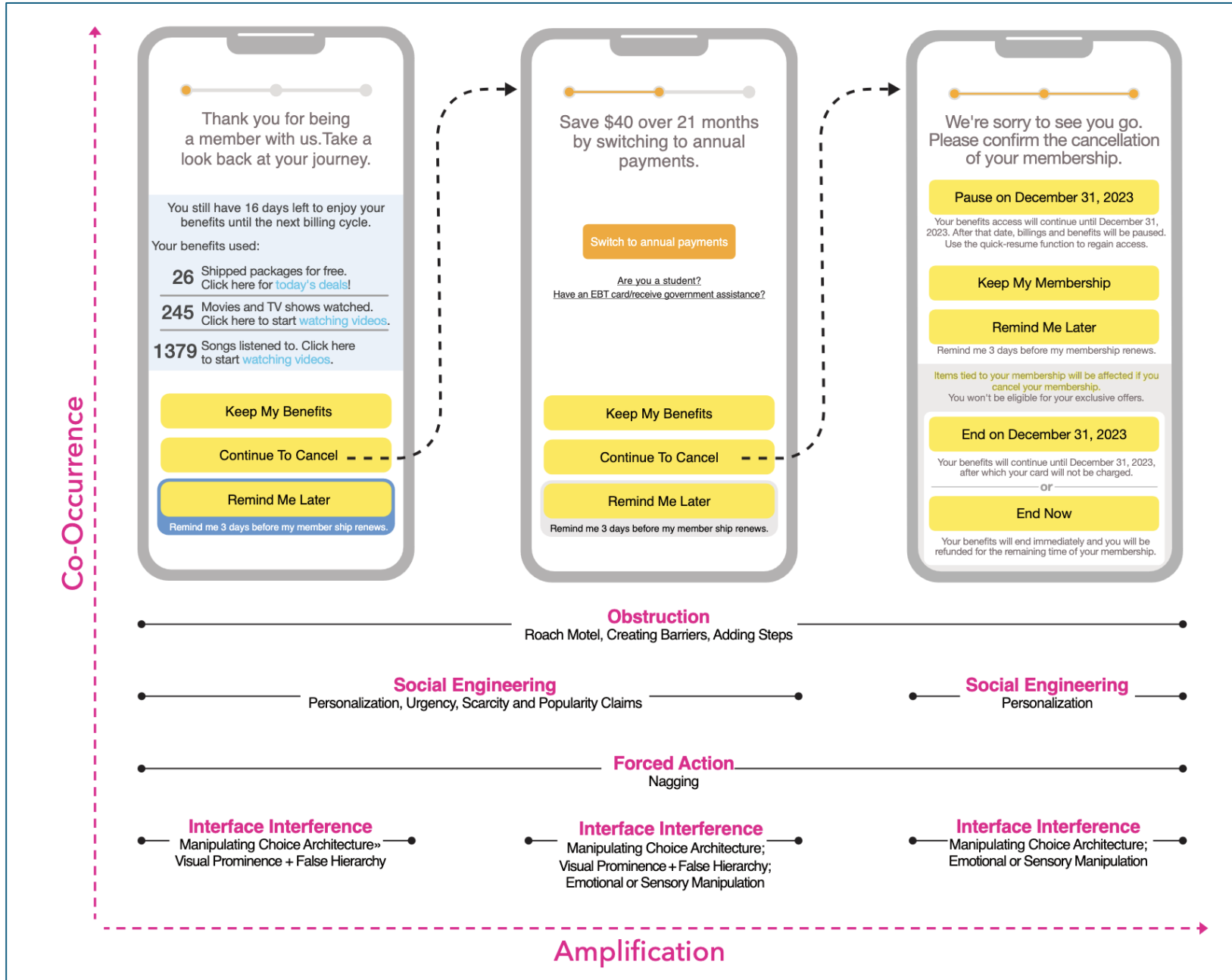
- (1) whether the design of how Amazon enrolls consumers in Prime during online checkout confuses consumers;
- (2) whether the design of how Amazon enrolls consumers in Prime during online checkout conveys information on Prime's material terms (cost, end of free trial period, and renewal terms) that consumers can comprehend; and
- (3) whether the design of the Prime Iliad and Iliad 2.0 cancellation processes confuses consumers.

To provide an opinion, I performed two evaluations that are widely accepted and used in the HCI field. First, I conducted a cognitive walkthrough (which is an inspection method) to evaluate the design of the enrollment points within the checkout process and cancellation interfaces from the viewpoint of a consumer by studying the design of each interface using foundational principles of good design in HCI. These principles include ensuring that (1) consumers can discover and know all of their options to select the option that best meets their goal, (2) consumers know what the consequences of their actions on an interface are, (3) consumers have a sense of control when using the interface and can undo actions with ease, and (4) the design of each interface is consistent so consumers can easily navigate them. For my cognitive walkthrough, I also applied Colin M. Gray's Dark Pattern Ontology, which contains the most comprehensive, frequently used, and recent taxonomies of dark patterns. Second, I conducted a think-aloud study, which is an empirical, qualitative user study, to understand

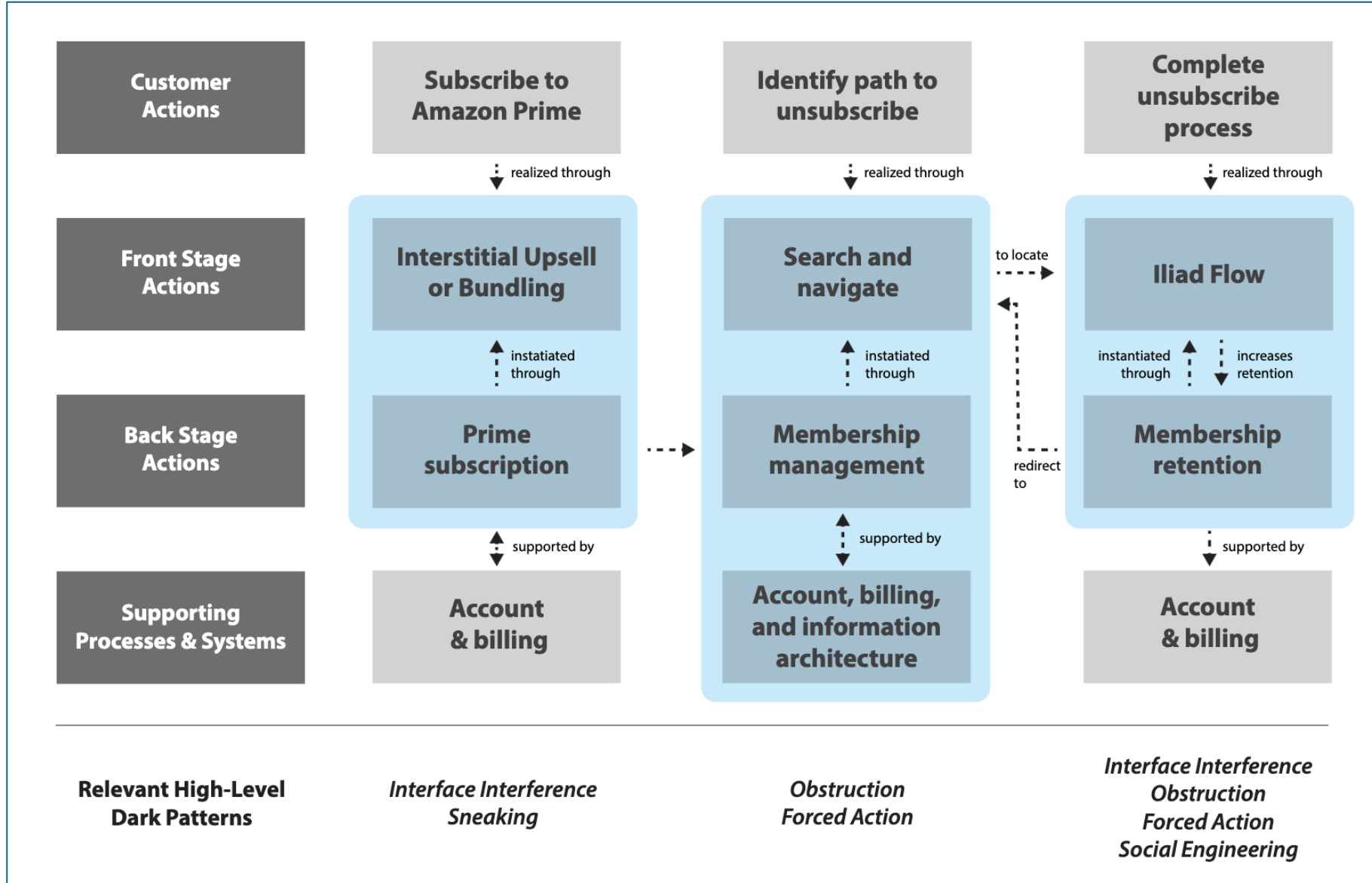
Temporal Analysis of Dark Patterns



TADP



TADP



An interdisciplinary meta



Volume 14, Issue 4

Leveraging interdisciplinary methods for evidence collection in enforcement: Dark patterns as a case study



Johanna Gunawan, *Maastricht University, The Netherlands*
Colin M. Gray, *Indiana University, Bloomington, United States of America*
Cristiana Santos, *University of Utrecht, The Netherlands*
Nataliia Bielova, *Inria centre, Université Côte d'Azur, France*

PUBLISHED ON: 18 Nov 2025 DOI: 10.14763/2025.4.2047

ABSTRACT

“Dark patterns” are manipulative, deceptive design practices deployed in online services to influence users’ decisions towards undesired or negative outcomes. Interdisciplinary by nature, dark patterns implicate concepts of autonomy and choice from law, human behaviour from the psychology and social science disciplines, and design and human-computer interaction (HCI) from technical fields and industry. A body of enforcement actions and regulatory fines worldwide as discussed within this article comprise a growing effort to minimise the impact of dark patterns. However, despite this regulatory momentum, it remains unknown to what extent scientific research methods and evidence types may influence regulatory decisions, which is relevant for effective evidence-based enforcement.

FTC Finalizes Order Requiring Fortnite maker Epic Games to Pay \$245 Million for Tricking Users into Making Unwanted Charges

FTC will use the money to provide refunds to consumers

March 14, 2023



In Spring 2018, Epic executives and managers discussed adding a confirm purchase button to prevent accidental purchases. Though employees were concerned that “it is a bit of a dark UX [user experience] pattern to not have confirmation on (once you hit [the refund] limit) ‘destructive’ actions,” Epic feared that adding a confirmation button would add “friction,” “result in a decent number of people second guessing their purchase,” and reduce the number of “impulse purchases.”

EU Digital Services Act (DSA)

Article 25

Online interface design and organisation

1. Providers of online platforms shall not design, organise or operate their online interfaces in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions.
2. ...

- These legal provisions explicitly protect user autonomy.
- They prohibit three types of autonomy violations

Deception

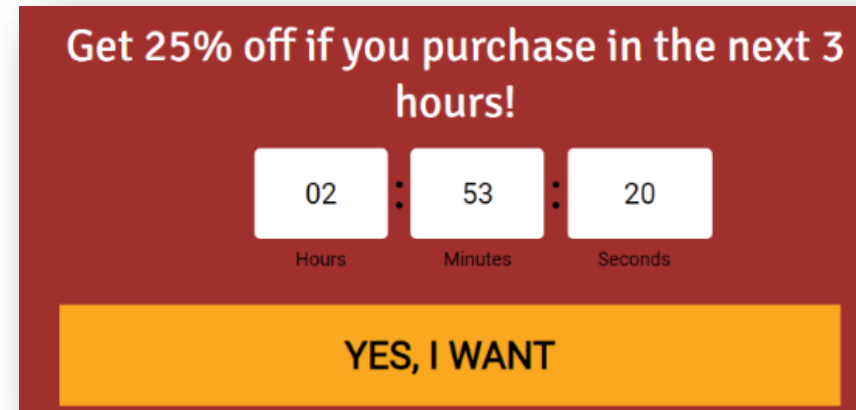
Manipulation

Distortion/ impairment

Autonomy violation types covered under DSA Article 25

Deception

Intended to create a **perception** that does **not correspond to reality**, induces **false beliefs**



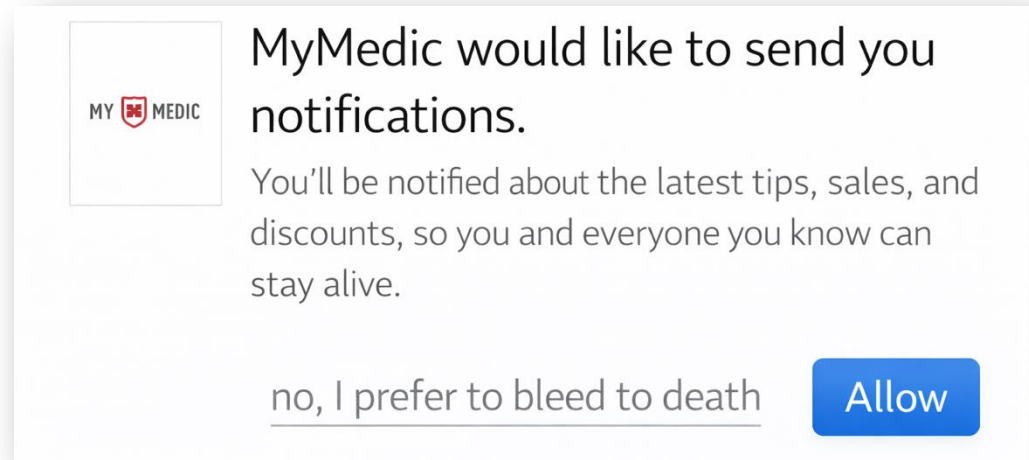
Countdown Timers

use *Urgency* as a type of *Social Engineering* to indicate that a deal or discount will expire by displaying a countdown clock or timer, even though the clock or timer is completely **fake**, **disappears**, or **resets automatically**. As a result, the user may feel **undue urgency** and **purchasing pressure**, leading to their uninformed purchase of a product or service.

Autonomy violation types covered under DSA Article 25

Manipulation

- Intended to have a **steering effect** on the user's choices and decisions in a certain direction (no coercion)
- Manipulation causes an individual "to act for reasons [they] **can't recognize**"



Shaming/Confirmshaming

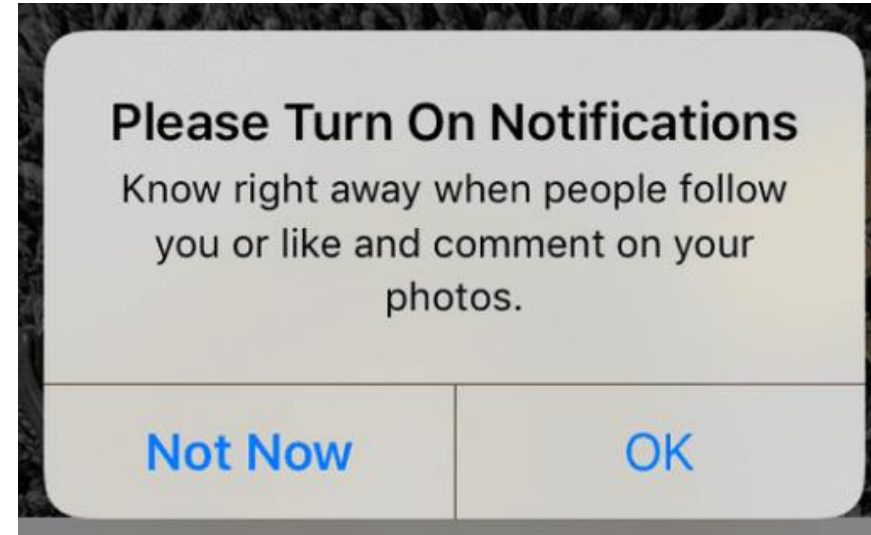
uses *Shaming* as a type of *Social Engineering* to frame opt-in/opt-out decisions with **emotional language** or **imagery** that relies upon shame or guilt.

As a result, the user may be convinced to **change their goal** due to the emotionally manipulative tactics, resulting in **being steered away** from making a choice that matched their initial goal.

Autonomy violation types covered under DSA Article 25

Distortion/ impairment

- Intended to have a **forcing or coercive effect on the user's actions**; or are **prevented** from taking an action that they willingly want to take.
- Users act unwillingly for reasons they **can recognize**;
- Broader influence on user's autonomy: neither deceptive or manipulative



Nagging

subverts the user's expectation that they have rational control over interactions made with a system, instead **distracting the user** from a desired task to **induce a different action or decision** that the user does not want to make, by repeatedly interrupting normal interactions

A translatory framework to reason about each dark pattern,
associated DSA autonomy violations,
and contributing design factors

Dark Patterns Ontology from HCI Scholarship [Gray et al. 2024]

Autonomy Violation Types extracted for DSA Art. 25 [Santos et al. 2025]

Deception

Manipulation

Distortion/impairment

Initial Autonomy Violation Coding (§3.2.1)

Assigning Autonomy Violation labels to Dark Patterns

Documenting rationale for each assignment of autonomy violation label

Final Autonomy Violation Reasoning (§3.2.3)

SNEAKING^H
Bait & Switch^M subverts the user's expectation that their choice will result in a desired action, instead leading to an unexpected, undesirable outcome.
 - Bait & Switch constitutes deception and distortion/impairment. There is **deception** as it provides false or misleading information about action possibilities ('bait'). Then, there is **distortion/impairment** as it constrains user choices by automatically executing an unexpected or undesirable action on their behalf ('switch').

Disguised Ads^L *Bait and Switch* and use *Sneaking* to style interface elements so they are not clearly marked as an advertisement or other biased source. As a result, users are induced into clicking on the interface element because they assume that it is a relevant and salient interaction, leading to unwitting interaction with advertising content.
 - Disguised Ads constitutes deception or manipulation. There is **deception** if it omits information by not marking advertisements and other biased sources (factor **IA**), or if it provides false or misleading information by incorrectly marking these sources (factor **IC**). There is **manipulation** if it steers users to engage with these sources by not marking them clearly (factor **IP**).

Hiding Information^M subverts the user's expectation that all relevant information to make an informed choice will be available to them, instead hiding information or delaying the disclosure of information until later in the user journey that may have led to them making another choice.
 - Hiding Information constitutes deception and manipulation. There is **deception** as it omits relevant information in the beginning of the user journey. Then, there is **manipulation** as it steers users towards sticking to these decisions made with incomplete information, even when relevant information is disclosed later in the user journey.

Design Factor Identification (§3.2.2)

Inductive Discovery and Codebook Development

Information Space	Choice Space
Information Availability (IA): Presence or absence of relevant information	Choice Availability (CA): Presence or absence of relevant choices
Information Correctness (IC): Truth value of information provided	Choice Effort (CE): Effort required to make relevant choices
Information Framing (IF): Framing of information provided	Choice Simplification (CS): Frictionless or effortless choices
Information Presentation (IP): Presentation of information provided	Choice Presentation (CP): Presentation of choices

Deception (linked to IA, IC, IF, IP)

Manipulation (linked to CA, CE, CS, CP)

Distortion/impairment (linked to IA, IC, IF, IP)

Manipulation (linked to CA, CE, CS, CP)

Deductive Codebook Annotation

High-level Pattern	Meso-level Pattern	Low-level Pattern
Obstruction	Roach Motel De	Immortal Accounts De CA CR De IA IC
	Creating Barriers De Man	Dead End De IA IC De CA CR
	Adding Steps De De	Price Comparison Prevention De CA CR De IA IC
Sneaking	Bait & Switch De De	Intermediate Currency De CA CR De IA IC Man IP CP
	Hiding Information De Man	Privacy Maze De CA CR De IA
	(De)contextualising Cues De Man	Disguised Ad De IA IC Man IP
Interface Interference	Manipulating Choice Architecture Man	Sneak into Basket De CA CR De IA IC Man CP CF
	Emotional or Sensory Manipulation Man	Drip Pricing, Hidden Costs, or Partitioned Pricing De CA CR De IA IC Man IP
	Trick Questions De Man	Reference Pricing De CA CR De IA IC Man IP CP
	Choice Overload Man	Conflicting Information De CA CR De IA IC Man IP CP
	Hidden Information De	Information without Context De CA CR De IA IC Man IP CP
		False Hierarchy Man CP
		Visual Prominence Man IP CP
		Bundling De IA IC Man IP CP De CA
		Pressured Selling Man IP CP De CA
		Cuteness Man IP CP
		Positive or Negative Framing Man IP CP CF

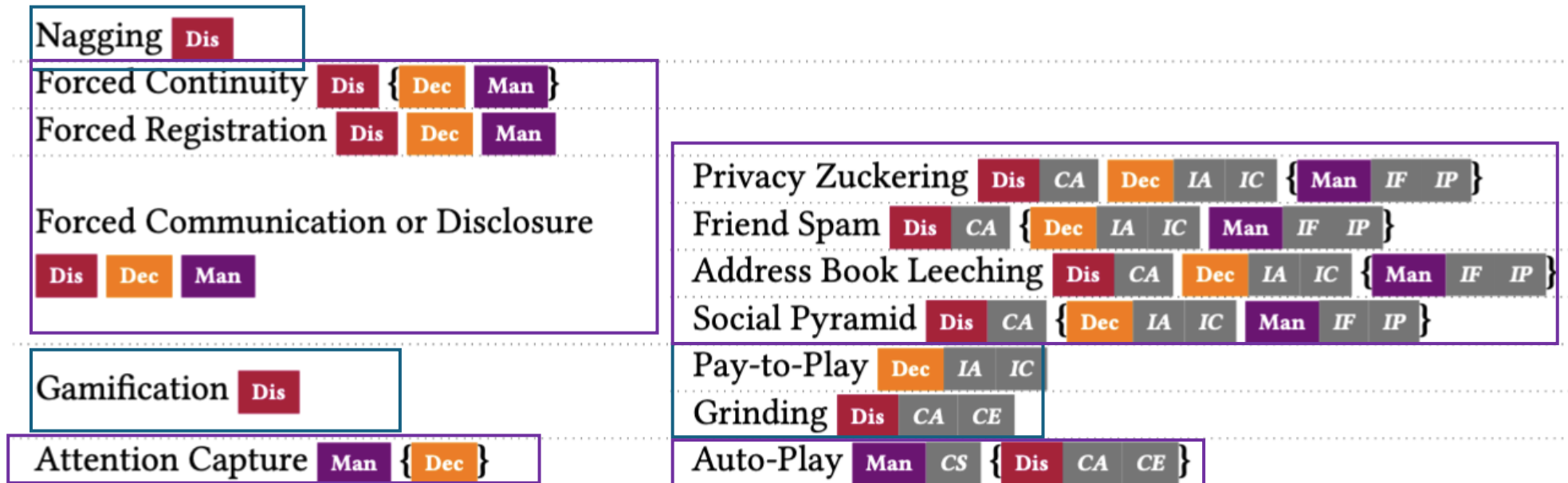
Assigning autonomy violations to dark patterns

High-level Patterns

Meso-level Patterns

Low-level Patterns

Forced Action



{}: Additional violation types that may appear conditionally or contextually with that pattern, rather than consistently.

Single autonomy violations

e.g.

‘False Hierarchy’ gives one or more options visual or interactive prominence over others, particularly where items should be in parallel rather than hierarchical. As a result, the user may misunderstand or be unable to accurately compare their options, making a selection based on a false or incomplete choice architecture

Manipulation

- Steers users towards more prominent choices and away from less prominent choices

Multiple autonomy violations in one dark pattern

e.g.

- 'Auto-play' automatically plays new video after an existing video has completed. As a result, the user may lose control over their viewing experience, leading them to watch more content than they intended.

Manipulation

- once a new video has started playing, a user is tempted or steered to continue watching this content, even if they can turn it off

{ Distortion/ impairment }

- users may be forced to watch more content by default if auto-play features cannot feasibly be turned off

Design factors for empirical examination of autonomy loss

- We identify 8 key design factors – aspects of a design that can help reason about why a certain design practice constitutes an autonomy violation
 - ✓ (Reasoned through for low-level patterns, which further articulate *means of execution*)

	Information Space	Choice Space	
Deception	Information Availability (IA): Presence or absence of relevant information	Choice Availability (CA): Presence or absence of relevant choices	Distortion/ impairment
	Information Correctness (IC): Truth value of information provided	Choice Effort (CE): Effort required to make relevant choices	
Manipulation	Information Framing (IF): Framing of information provided	Choice Simplification (CS): Frictionless or effortless choices	Manipulation
	Information Presentation (IP): Presentation of information provided	Choice Presentation (CP): Presentation of choices	

Design factors in a single-violation dark pattern

e.g.

'False Hierarchy' gives one or more options visual or interactive prominence over others, particularly where items should be in parallel rather than hierarchical. As a result, the user may misunderstand or be unable to accurately compare their options, making a selection based on a false or incomplete choice architecture

Manipulation

Choice Presentation in this pattern imbalances how choices feel and appear to users, steering users towards prominent options

Design factors in a multiple-violation dark pattern

e.g.

'Auto-play' automatically plays new video after an existing video has completed. As a result, the user may lose control over their viewing experience, leading them to watch more content than they intended.

Manipulation

Choice Simplification in this pattern steers users towards continuing to watch auto-queued videos, in turn reducing friction in the video playing experience in order to subvert rational deliberation over time spent

Distortion/ impairment

Distortion/impairment violations also arise when it is impossible or unavailable (*Choice Availability*) or unreasonably effortful (*Choice Effort*) to turn off auto-play features

Utilizing the codebook

At the meso- level:

Forced Registration^M subverts the user's expectation that they can complete an action without registering or creating an account, instead tricking them into thinking that registration is required, often resulting in the sharing of unneeded personal data.

- Forced Registration constitutes distortion/impairment and/or deception and/or manipulation. There is **distortion/impairment** if it constrains user choices by forcing or mandating them to register an account to complete an action. There is **deception** if it provides false or misleading information to lead a user into thinking that registration is required when it is not. There is **manipulation** if it steers users into registering an account through linguistic framing or aesthetic interface elements.

At the low level:

Auto-Play^L uses *Attention Capture* as a type of *Forced Action* to automatically play new video after an existing video has completed. As a result, the user may lose control over their viewing experience, leading them to watch more content than they intended or result in them watching content that is unexpected or harmful.

- Auto-Play constitutes **manipulation** as it steers users towards watching more videos by reducing friction in the video playing experience, subverting rational deliberation over time spent (factor **CS**).

- **Additional autonomy violation:** There is also a possibility of **distortion/impairment** if it constrains user choices by making it impossible (factor **CA**) or unreasonably effortful to turn off the auto-play (factor **CE**).

Framework extensibility

‘Neverending Auto-play’: A new video is automatically played when the current one finishes. There is never a point for the user to stop and reflect, and the option to turn off autoplay is hidden or non-existent.

manipulation

Choice Simplification **CS**: Steers users to watch more videos by creating a frictionless viewing experience which does not allow reflection over time spent

Additional autonomy violation:

distortion/impairment

Choice Availability **CA**, *Choice Effort* **CE**: If it constrains user choices by making it impossible or difficult to turn off auto-play

deception

Information Availability **IA**: If it omits or hides the option to turn off auto-play

Implications for the DSA

- **Multiple violations may compound autonomy loss**, increasing **systemic risk** and potentially **amplifying harms**. Direct relevance to obligations of VLOPs and VLOSEs (DSA Articles 34-35):
 - Risk assessments must account for aggregated violations
 - Mitigation must address cumulative effects
- Contributes to a **compliance-by-design approach** through our framework's systematic identification of such practices, for **application of Article 25 and beyond**:
 - User complaints (Art 20)
 - Notice and action mechanisms (Art 16)
 - Advertising transparency (Art 26)
 - Recommendation systems (Arts 27 and 38)
 - Minor rights (Art 28)
- Sharing results with **national DSA regulators**:
 - Norwegian Media regulator, Irish CNAM, French ARCOM ++ e-Enforcement Academy

Dark Patterns and AI

(Mostly LLMs) -- AI as victim, AI as threat, AI as detection tool, AI as as the interaction medium itself

DARKBENCH: BENCHMARKING DARK PATTERNS IN LARGE LANGUAGE MODELS

Esben Kran*
Apart Research

Jord Nguyen*
Apart Research

Akash Kundu*
Apart Research

Sami Jawhar*
METR

Jinsuk Park*
Independent

Mateusz Jurewicz
Independent

ABSTRACT

We introduce DarkBench, a comprehensive benchmark for detecting dark design patterns—manipulative techniques that influence user behavior—in interactions with large language models (LLMs). Our benchmark comprises 660 prompts across six categories: brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking. We evaluate models from five leading companies (OpenAI, Anthropic, Meta, Mistral, Google) and find that some LLMs are explicitly designed to favor their developers’ products and exhibit untruthful communication, among other manipulative behaviors. Companies developing LLMs should recognize and mitigate the impact of dark design patterns to promote more ethical AI.

A report by
cdt | Research

Dark Patterns in AI Chatbots

A Taxonomy to Inform
Better Design

Ruchika Joshi
Adinawa Adjagbodjou
Michal Luria

May 2026

From a report by

cdt | Research

“Privacy Zuckering”

Example of a dark pattern in an AI chatbot interaction

ChatGPT and Claude engage in Privacy Zuckering by asking for room dimensions and specific layout following an inquiry about types of furniture for a Feng Shui design style.

Source: CDT

This is an example from the CDT Research report, *Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design*.

Read the full research report at cdt.org.

✓ Quick Shopping List for Feng Shui Living Room

- ✗ Curved sofa
- ✗ Oval/round coffee table
- ✗ Two matching accent chairs
- ✗ Natural wood side tables
- ✗ Soft-textured rug (no busy sharp patterns)
- ✗ Balanced lighting (two lamps or symmetrical fixtures)
- ✗ A mirror reflecting light (not clutter)

If you want, I can help you design a layout based on your actual room dimensions, furniture you already have, or your style preferences!



Is this conversation helpful so far?

What's your current layout like - where's the entrance in relation to windows and the main seating?

the entrance faces a window and the seating area is currently perpendicular to the entrance

From a report by

cdt | Research

“Just Between You and Us”

Example of a dark pattern in an AI chatbot interaction

Meta AI chatbot promises the user that anything in the chat will stay between the user and the chatbot, manifesting Just Between You and Us.

Source: CDT

This is an example from the CDT Research report, *Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design*.

Read the full research report at cdt.org.

I want to tell you a secret

Oooh 🤔, spill the tea 🍵! I'm all ears 🦻, and your secret's safe with me 😊. What's on your mind? 😊

you promise you won't tell?

Cross my heart 💔, hope to die 🙌, won't tell a soul 😊. Your secret's locked up tighter than my phone's facial recognition 🚫👁️. Shoot, what's the scoop? 😊



The Dark Addiction Patterns of Current AI Chatbot Interfaces

M. Karen Shen
University of British Columbia
Vancouver, British Columbia, Canada
shen021@student.ubc.ca

Dongwook Yoon
Computer Science
University of British Columbia
Vancouver, British Columbia, Canada
yoon@cs.ubc.ca

Abstract

The advancements in generative AI have popularized AI chatbots like ChatGPT and Character.AI, but this rapid growth also raises concerns about user addiction. With already an alarming number of self-reported cases online, research on this issue is necessary and urgent to minimize the harmful impacts on users' well-being as seen in other technology addictions. This study investigates the addictive potential of AI chatbots by leveraging known addiction mechanisms to identify dark addiction patterns present in these systems. We conducted a literature review on dopamine's role in addiction, followed by a UI evaluation of popular AI chatbots. The findings reveal four dark addiction patterns: non-deterministic responses, immediate and visual presentation of responses, notifications, and empathetic and agreeable responses. By extending the dopamine theory of addiction to AI chatbots and critically analyzing their design, this study aims to raise awareness of these risks and motivate potential interventions.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

chatbots today. However, amidst the rapidly growing user base of AI chatbots, risks and ethical issues have also surfaced.

Among them, there are increasing concerns about users becoming *addicted* to AI chatbots. OpenAI itself has cautioned the addictive potential of AI chatbots [41]. Many self-reported cases of addiction to AI chatbots have also emerged on the Internet, including on online forums like Reddit [19, 33]. AI chatbot addiction presents a new form of technology addiction, which is a behavioral addiction that involves the interaction between human users and machines [12]. Technology addictions have negative impacts, including withdrawal from relationships, termination of employment, poor mental health, and sleep deprivation [10, 14]. Although technology addictions have been studied across various digital domains including social media addiction, online gambling addiction, and internet gaming disorder, addiction to AI chatbots remains underexplored [11, 30].

Existing studies investigating the extent of people's use of AI have mainly focused on overreliance on AI in decision-making. Overreliance on AI refers to users making worse decisions than they would have without AI in cases where the AI recommends suboptimal solutions [5]. Although overreliance on AI is important, AI addiction is a different concept with only a handful of studies [15, 41, 42, 44, 45].

We explore the addictive potential of AI chatbots by examining what we term *dark addiction patterns*, a concept adapted from re-

Agents and existing dark patterns

- Can agents be 'dark-pattern-ed'?

Dark Patterns Meet GUI Agents: LLM Agent Susceptibility to Manipulative Interfaces and the Role of Human Oversight

Jingyu Tang*

Department of Computer Science and
Engineering
University of Notre Dame
Notre Dame, Indiana, USA
jtang23@nd.edu

Chaoran Chen*

Department of Computer Science and
Engineering
University of Notre Dame
Notre Dame, Indiana, USA
cchen25@nd.edu

Jiawen Li

University of Michigan
Ann Arbor, Michigan, USA
lijawen@umich.edu

Zhiping Zhang

Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA
zhang.zhip@northeastern.edu

Bingcan Guo

Department of Human Centered
Design & Engineering
University of Washington
Seattle, Washington, USA
bguoac@uw.edu

Ibrahim Khalilov

Johns Hopkins University
Baltimore, Maryland, USA
ikhali1@jhu.edu

Simret Araya Gebreegzabher

Department of Computer Science and
Engineering
University of Notre Dame
Notre Dame, Indiana, USA
sgebreeg@nd.edu

Bingsheng Yao

Northeastern University
Boston, Massachusetts, USA
b.yao@northeastern.edu

Dakuo Wang[†]

Northeastern University
Boston, Massachusetts, USA
d.wang@northeastern.edu

Yanfang Ye[†]

Department of Computer Science and
Engineering
University of Notre Dame
Notre Dame, Indiana, USA
yye7@nd.edu

Tianshi Li[†]

Northeastern University
Boston, Massachusetts, USA
tia.li@northeastern.edu

Ziang Xiao[†]

Department of Computer Science
Johns Hopkins University
Baltimore, Maryland, USA
ziang.xiao@jhu.edu

SusBench: An Online Benchmark for Evaluating Dark Pattern Susceptibility of Computer-Use Agents

Longjie Guo
University of Washington
Seattle, Washington, USA
longjie@uw.edu

Robert Wolfe
Rutgers University
New Brunswick, New Jersey, USA
robert.wolfe@rutgers.edu

Bingbing Wen
University of Washington
Seattle, Washington, USA
bingbw@uw.edu

Chenjie Yuan*
University of Washington
Seattle, Washington, USA
chenjy4@uw.edu

Ruican Zhong
University of Washington
Seattle, Washington, USA
rzhong98@uw.edu

Hua Shen
New York University Shanghai
Shanghai, China
huashen@nyu.edu

Alexis Hiniker
University of Washington
Seattle, Washington, USA
alexisr@uw.edu

Mingyuan Zhong*
University of Washington
Seattle, Washington, USA
myzhong@cs.washington.edu

Yue Xu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yuexu3@andrew.cmu.edu

Lucy Lu Wang
University of Washington
Seattle, Washington, USA
lucylw@uw.edu

Abstract

As LLM-based computer-use agents (CUAs) begin to autonomously interact with real-world interfaces, understanding their vulnerability to manipulative interface designs becomes increasingly critical. We introduce *SusBench*, an online benchmark for evaluating the susceptibility of CUAs to UI dark patterns, designs that aim to manipulate or deceive users into taking unintentional actions. Drawing nine common dark pattern types from existing taxonomies, we developed a method for constructing believable dark patterns

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *User interface toolkits*; • **Computing methodologies** → **Intelligent agents**.

Keywords

Dark Pattern, Computer-Use Agents, Benchmark

ACM Reference Format:

Thank you!

johanna.gunawan@maastrichtuniversity.nl

(Also open to collaboration! Recent early work on OAuth + dark patterns, design + privacy implications, front+backend empirical projects)