Chair for Computer Science 10 (Media Computing and Human-Computer Interaction)

RWTH AACHEN UNIVERSITY

Lumi: Adjustable Countermeasures Against Dark Patterns on the Web

Master's Thesis at the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Ilja Girnus

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Ulrik Schroeder

Registration date: 31.01.2025
Submission date: 30.07.2025

Zentrales Prüfungsamt/Central Examination Office

Eidesstattliche Versicherung
Declaration of Academic Integrity

Name, Vorname/Last Name, First Name

Matrikelnummer (freiwillige Angabe)
Student ID Number (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel
I hereby declare under penalty of perjury that I have completed the present paper/bachelor's thesis/master's thesis* entitled

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt; dies umfasst insbesondere auch Software und Dienste zur Sprach-, Text- und Medienproduktion. Ich erkläre, dass für den Fall, dass die Arbeit in unterschiedlichen Formen eingereicht wird (z.B. elektronisch, gedruckt, geplottet, auf einem Datenträger) alle eingereichten Versionen vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without unauthorized assistance from third parties (in particular academic ghostwriting**.** I have not used any other sources or aids than those indicated; this includes in particular software and services for language, text, and media production. In the event that the work is submitted in different formats (e.g. electronically, printed, plotted, on a data carrier), I declare that all the submitted versions are fully identical. I have not previously submitted this work, either in the same or a similar form to an examination body.

Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen/Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
§ 156 StGB (German Criminal Code): False Unsworn Declarations
Whosoever before a public authority competent to administer unsworn declarations (including Declarations of Academic Integrity) falsely submits such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment for a term not exceeding three years or to a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
§ 161 StGB (German Criminal Code): False Unsworn Declarations Due to Negligence
(1) If an individual commits one of the offenses listed in §§ 154 to 156 due to negligence, they are liable to imprisonment for a term not exceeding one year or to a fine.
(2) The offender shall be exempt from liability if they correct their false testimony in time. The provisions of § 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Ort, Datum/City, Date

Unterschrift/Signature

Contents

List of Figures and Tables

Abstract

This thesis investigates the effectiveness of virtual assistants as countermeasures against dark patterns in online user interfaces. Driven by the pervasive nature of deceptive design practices that undermine user autonomy, the research explores how different intervention strategies, ranging from passive information provision to active dark pattern removal and user customization, impact user experience. A comprehensive user study was conducted using interactive prototypes, evaluating various virtual assistant variants across metrics such as helpfulness, safety, control, efficiency, and general user experience. The findings reveal a user preference for proactive interventions, particularly those offering personalized control over the assistant's behavior. Variants that directly removed dark patterns, especially when combined with optional informational support, consistently led to improved website usability, clarity, and efficiency, and were highly favored by participants. Conversely, the lack of interventions or purely passive information approaches were perceived negatively by the participants. The study highlights the critical importance of transparency, user control, and clear feedback mechanisms for building trust and ensuring a positive user experience with such assistants. This research demonstrates the tangible benefits of virtual assistants in combating manipulative online practices. It provides actionable insights for researches, developers and designers, emphasizing the need for customizable, transparent, and proactive solutions to empower users in complex digital environments.

Überblick

Diese Arbeit untersucht die Wirksamkeit virtueller Assistenten als Gegenmaßnahme gegen Dark Patterns in Online-Benutzeroberflächen. Angetrieben von der allgegenwärtigen Natur trügerischer Designpraktiken, die die Autonomie des Nutzers untergraben, untersucht die Forschung, wie verschiedene Interventionsstrategien, die von der passiven Bereitstellung von Informationen bis zur aktiven Beseitigung von Dark Patterns und der Anpassung durch den Nutzer reichen, die Nutzererfahrung beeinflussen. Es wurde eine umfassende Nutzerstudie mit interaktiven Prototypen durchgeführt, in der verschiedene Varianten virtueller Assistenten anhand von Kriterien wie Hilfsbereitschaft, Sicherheit, Kontrolle, Effizienz und allgemeine Nutzererfahrung bewertet wurden. Die Ergebnisse zeigen, dass die Nutzer proaktive Interventionen bevorzugen, insbesondere solche, die eine personalisierte Kontrolle über das Verhalten des Assistenten bieten. Varianten, die Dark Pattern direkt beseitigten, insbesondere in Kombination mit optionaler Informationsunterstützung, führten konsequent zu einer verbesserten Benutzerfreundlichkeit, Übersichtlichkeit und Effizienz der Website und wurden von den Teilnehmern stark bevorzugt. Umgekehrt wurden fehlende Interventionen oder rein passive Informationsansätze von den Teilnehmern negativ wahrgenommen. Die Studie unterstreicht die entscheidende Bedeutung von Transparenz, Nutzerkontrolle und klaren Feedback-Mechanismen für den Aufbau von Vertrauen und die Gewährleistung einer positiven Nutzererfahrung mit solchen Assistenten. Diese Studie zeigt die greifbaren Vorteile virtueller Assistenten bei der Bekämpfung manipulativer Online-Praktiken. Sie liefert umsetzbare Erkenntnisse für Forscher, Entwickler und Designer und unterstreicht den Bedarf an anpassbaren, transparenten und proaktiven Lösungen, um Nutzer in komplexen digitalen Umgebungen zu unterstützen.

Acknowledgments

I would like to thank Prof. Dr. Jan Borchers and Prof. Dr. Ulrik Schroeder for reviewing my thesis.

My thanks also go to my advisor, René Schäfer, for his support, guidance, and encouragement throughout this journey. Your help was instrumental in making this thesis possible.

I am also grateful to all the participants who contributed their time and insights to this research. Your contributions were very valuable and are highly appreciated.

Finally, I would like to thank my family and friends, especially my fiancée and son, for their patience, understanding, and support. Their encouragement provided strength during this endeavor.

Conventions

Throughout this thesis we use the following conventions:

- The thesis is written in American English.
- The first person is written in plural form.
- Unidentified third persons are described in plural form.

While the terminology of “dark patterns” is evolving to the more recently used term “deceptive design patterns“ for the remainder of this thesis, we still use the term “dark patterns” for easier recognition and consistency along the papers this thesis build on.

Short excursuses are set off in colored boxes.

> **EXCURSUS:**
> Excursuses are set off in orange boxes.

Where appropriate, paragraphs are summarized by one or two sentences that are positioned at the margin of the page.

This is a summary of a paragraph.

Chapter 1

Introduction

Dark patterns in digital design manipulate users into making decisions against their interests, undermining user autonomy and trust.

The digital landscape has become an indispensable part of daily life, with online services and platforms permeating nearly every aspect of human interaction, from communication and entertainment to commerce and education. While these advancements offer unprecedented convenience and access, they also present new challenges, particularly concerning user autonomy and well-being. A growing concern in this evolving environment is the pervasive use of "dark patterns", deceptive user interface designs that manipulate users into making decisions they might not otherwise make, often to the benefit of the service provider and at the expense of the user's interests (Gray et al. [2018]). These manipulative tactics can lead to unwanted purchases, involuntary data sharing, and a general erosion of trust in digital platforms (Mathur et al. [2019], Bongard-Blanchy et al. [2021]).

Virtual assistants could proactively combat dark patterns in real-time, restoring transparency for users.

The increasing prevalence and sophistication of dark patterns require effective countermeasures that empower users and restore transparency to online interactions. Existing approaches, such as legislative efforts and user education, have shown promise but often face limitations in practical application and immediate user protection (Schäfer et al. [2024]). This thesis explores the potential of virtual assistants as a novel and proactive solution to combat dark patterns, drawing inspiration from and building upon the findings of previous work of Schäfer et al. [2023] on visual countermeasures against dark patterns in user interfaces. By integrating a virtual assistant directly into the user's browsing experience, these assistants could identify, inform about, and even mitigate manipulative design elements in real-time.

The motivation is to enhances user agency and study how virtual assistants can combat dark patterns effectively.

This research is motivated by the critical need to enhance user agency in an increasingly complex digital world. Understanding how different forms of intervention through virtual assistants, from the passive provision of information to the active elimination of dark patterns, affect users' perception, efficiency, and overall experience is crucial for the development of effective and user-oriented digital tools. Our work contributes to the growing body of literature on ethical design and user empowerment by empirically investigating the efficacy of various assistant behaviors.

This thesis aims to answer the following key research questions:

- Is the assistant perceived as distracting or helpful? Is the user's flow interrupted?
- How safe do users feel with the assistant? Do they still believe they are being manipulated by the dark patterns, or do they now feel protected from the effects by the assistant?
- Do users feel in control of their actions? Does the assistant communicate clearly enough what it is doing, or do users feel a loss of control?

This research could guide the development of virtual assistants as countermeasures to dark patterns.

The significance of this research lies in its potential to inform the design and implementation of future user assistance systems that actively protect consumers from manipulative online practices. By identifying preferred modes of intervention and understanding user perceptions of control, safety, and efficiency, this study provides actionable insights for developers, designers, and policymakers seeking to foster a more transparent and trustworthy digital environment.

Outline: The remainder of this thesis is structured as follows: Chapter 2 provides a comprehensive review of existing literature on dark patterns, their classification, and current countermeasures. Chapter 3 describes the potential use of virtual assistants and a preliminary study in which we explore what is expected of a virtual assistant as a countermeasure to dark patterns. Chapter 4 details the methodology employed in the user study, including the design of the virtual assistant prototypes, the experimental setup, and data collection procedures, as well as the quantitative and qualitative results derived from our user study. Chapter 5 offers an in-depth discussion of these findings, interpreting their implications for the research questions and acknowledging limitations. Finally, Chapter 6 concludes the thesis by summarizing the main contributions and outlining directions for future research.

We reviewed dark patterns, conducted an small creative study and a bigger study for the virtual assistant.

Chapter 2

Related Work

This section provides an overview of existing research and the literature relevant to the study of dark patterns in user interfaces. We begin by establishing a foundational understanding of what defines a dark pattern, followed by ontological approaches used to classify and categorize these deceptive design elements. Finally, we present a range of possible countermeasures that have been proposed and studied to combat the expansion and impact of dark patterns. We have organized these into the following categories: Legal measures, social pressure, awareness, and technical interventions. We will then take a closer look at the technical countermeasures.

This section overviews dark pattern definitions, classifications, and countermeasures, focusing on technical interventions.

2.1 Definition of Dark Patterns

DARK PATTERNS:
Deceptive patterns (also known as “dark patterns”) are tricks used in websites and apps that make you do things that you didn’t mean to, like buying or signing up for something. (Harry Brignull)

Excursus:
Dark Patterns

The concept of "dark patterns" refers to user interface designs that intentionally mislead, trick, or coerce users into

Dark patterns are deceptive UI designs, as defined by Harry Brignull, that trick users into unintended actions.

making decisions they might not otherwise make, often to the benefit of the service provider and to the disadvantage of the user. The term was made popular by Harry Brignull in 2010, who defined them as "Deceptive patterns (also known as “dark patterns”) are tricks used in websites and apps that make you do things that you didn’t mean to, like buying or signing up for something". This definition emphasizes the deceptive and manipulative nature of these design decisions, distinguishing them from “bad” designs that merely exhibit poor usability or accidental design flaws. The definition can be found on Brignull’s website[1] along with other interesting areas, such as the “Hall of shame” which we will discuss in more detail later. There are other definitions and approaches, as the line between dark patterns and good marketing is not always clear. In this paper, we will use Brignull’s definition mentioned above.

2.1.1 Ontology of Dark Patterns

Researchers use ontologies to classify dark patterns, aiding analysis, communication, and countermeasure development.

To systematically analyze and address dark patterns, researchers have developed various ontologies and taxonomies to classify them based on their characteristics, mechanisms, and the type of deception employed. These classifications help in identifying recurring patterns, understanding their underlying psychological principles, and developing targeted countermeasures. Moreover, these definitions make it much easier to communicate about dark patterns, especially when communication takes place across different research areas, for example, between computer scientists and legal counsels.

Harry Brignull’s classification identifies various dark pattern types, each a distinct deceptive tactic.

One prominent early classification, proposed by Harry Brignull, categorizes dark patterns into several types, including "Trick Questions," "Sneak into Basket," "Roach Motel," "Privacy Zuckering," "Price Comparison Prevention," "Misdirection," "Hidden Costs," "Bait and Switch," "Confirmshaming," and "Disguised Ads." Each category describes a distinct deceptive tactic.

[1] https://www.deceptive.design/

More recently, Gray et al. [2024] proposed a comprehensive three-level ontology of dark patterns, harmonizing ten existing regulatory and academic taxonomies. This ontology defines 64 synthesized dark pattern types, structured hierarchically:

Gray's three-level ontology defines 64 dark pattern types, providing a shared understanding across contexts.

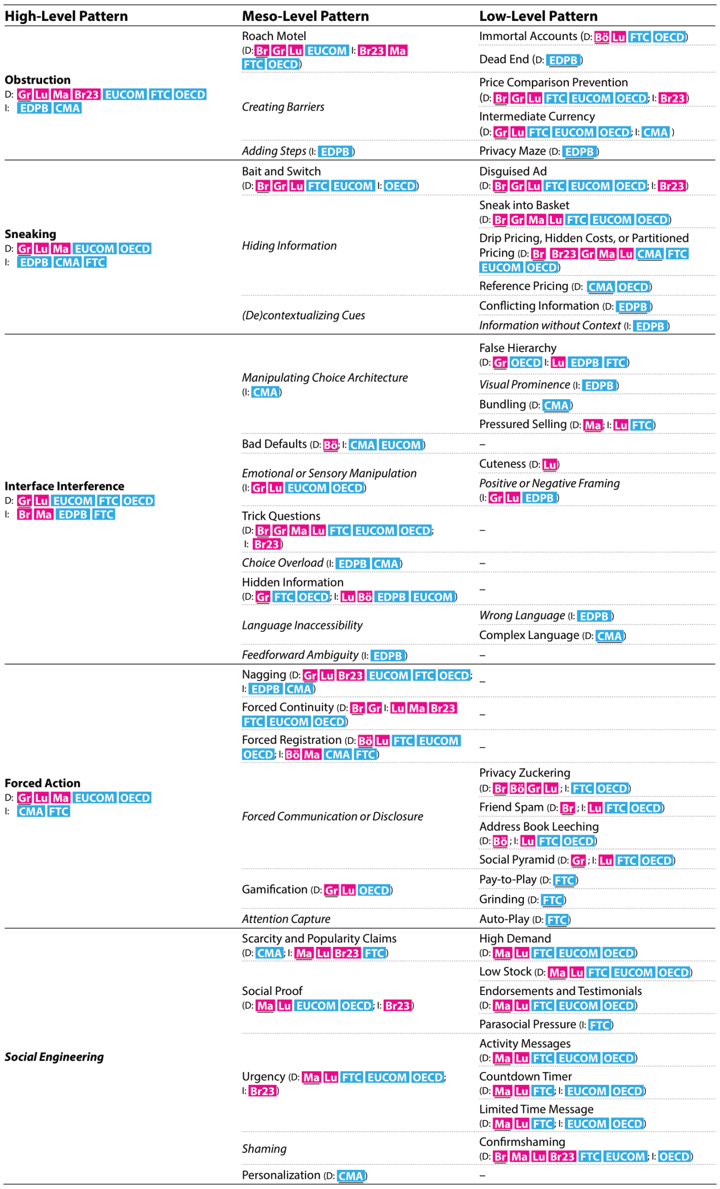
- High-level patterns: These represent the most abstract forms of knowledge, characterizing general strategies of manipulative, coercive, or deceptive elements that limit user autonomy and decision-making. They are context-agnostic and apply across various modalities and application types.
- Meso-level patterns: These bridge high- and low-level knowledge, describing specific angles of attack or approaches to undermining a user's ability to make autonomous and informed decisions. They are content-agnostic and can be interpreted contextually.
- Low-level patterns: These are the most situated and contextually dependent, detailing specific means of execution within the UI that limit user autonomy. They are often visually or temporally described and are typically detectable through algorithmic or manual means.

This structured ontology, which is shown in Figure 2.1, aims to provide a shared language for scholars, regulators, and practitioners, facilitating a more consistent and consolidated understanding of dark patterns across diverse contexts and stakeholders.

2.2 Countermeasures

Addressing the pervasive issue of dark patterns requires a multi-faceted approach involving legal, social, and technical interventions. This subsection shows various countermeasures aimed at reducing the impact and spread of dark patterns.

Countermeasures against dark patterns involve legal, social, and technical interventions to reduce their impact.



High-Level Pattern	Meso-Level Pattern	Low-Level Pattern
Obstruction D: Gr Lu Ma Br23 EUCOM FTC OECD I: EDPB CMA	Roach Motel (D: Br Gr Lu EUCOM I: Br23 Ma FTC OECD)	Immortal Accounts (D: Bö Lu FTC OECD)
		Dead End (D: EDPB)
	Creating Barriers	Price Comparison Prevention (D: Br Gr Lu FTC EUCOM OECD; I: Br23)
		Intermediate Currency (D: Gr Lu FTC EUCOM OECD; I: CMA)
	Adding Steps (I: EDPB)	Privacy Maze (D: EDPB)
Sneaking D: Gr Lu Ma EUCOM OECD I: EDPB CMA FTC	Bait and Switch (D: Br Gr Lu FTC EUCOM I: OECD)	Disguised Ad (D: Br Gr Lu FTC EUCOM OECD; I: Br23)
	Hiding Information	Sneak into Basket (D: Br Gr Ma Lu FTC EUCOM OECD)
		Drip Pricing, Hidden Costs, or Partitioned Pricing (D: Br Br23 Gr Ma Lu CMA FTC EUCOM OECD)
		Reference Pricing (D: CMA OECD)
	(De)contextualizing Cues	Conflicting Information (D: EDPB)
		Information without Context (I: EDPB)
Interface Interference D: Gr Lu EUCOM FTC OECD I: Br Ma EDPB FTC	*Manipulating Choice Architecture* (I: CMA)	False Hierarchy (D: Gr OECD I: Lu EDPB FTC)
		Visual Prominence (I: EDPB)
		Bundling (D: CMA)
		Pressured Selling (D: Ma; I: Lu FTC)
	Bad Defaults (D: Bö; I: CMA EUCOM)	–
	Emotional or Sensory Manipulation (I: Gr Lu EUCOM OECD)	Cuteness (D: Lu)
		Positive or Negative Framing (I: Gr Lu EDPB)
	Trick Questions (D: Br Gr Ma Lu FTC EUCOM OECD; I: Br23)	–
	Choice Overload (I: EDPB CMA)	–
	Hidden Information (D: Gr FTC OECD; I: Lu Bö EDPB EUCOM)	–
	Language Inaccessibility	*Wrong Language* (I: EDPB)
		Complex Language (D: CMA)
	Feedforward Ambiguity (I: EDPB)	–
Forced Action D: Gr Lu Ma EUCOM OECD I: CMA FTC	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 Communication or Disclosure	Privacy Zuckering (D: Br Bö Gr Lu ; I: FTC OECD)
		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)
	Attention Capture	Auto-Play (D: FTC)
Social Engineering	Scarcity and Popularity Claims (D: CMA; I: Ma Lu Br23 FTC)	High Demand (D: Ma Lu FTC EUCOM OECD)
	Social Proof (D: Ma Lu EUCOM OECD; I: Br23)	Low Stock (D: Ma Lu FTC EUCOM OECD)
		Endorsements and Testimonials (D: Ma Lu FTC EUCOM OECD)
		Parasocial Pressure (I: FTC)
	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)	–

Figure 2.1: An excerpt from the "An Ontology of Dark Patterns Knowledge" illustrating the hierarchical structure of dark pattern types, categorized into high-level, meso-level, and low-level patterns, derived from harmonized academic and regulatory taxonomies by Gray et al. [2024].

2.2.1 Laws and Regulations

Governments and regulatory bodies worldwide are increasingly recognizing the need to legislate against deceptive design practices. Laws such as the General Data Protection Regulation (GDPR) in the European Union, while primarily focused on data privacy, have implications for dark patterns that manipulate user consent (Martini et al. [2021], Gray et al. [2021]). Similarly, consumer protection laws and unfair competition laws in various jurisdictions aim to prevent unfair and deceptive trade practices, which can encompass certain dark patterns (Martini et al. [2021]). For instance, the California Consumer Privacy Act (CCPA) and its amendments, as well as the now-defunct DETOUR Act in the US, and recent enforcement actions by the US Federal Trade Commission (FTC) and the UK Competition and Market Authority (CMA), demonstrate a growing legal focus on these issues (Martini et al. [2021]).

Governments globally are enacting laws and regulations like GDPR and CCPA to counter dark patterns and protect user consent.

Specifically concerning cookie consent, the ePrivacy Directive and GDPR establish strict criteria for valid consent, requiring it to be "freely given, specific, informed and unambiguous" (Grassl et al. [2020], Gray et al. [2021]). However, the effectiveness of these legal requirements is questioned, as studies indicate that cookie consent requests often do not lead to genuinely informed consent, with users frequently agreeing by default as mentioned by Grassl et al. [2020].

Despite GDPR, cookie consent often lacks genuine user agreement, highlighting legal requirements' ineffectiveness.

The legality of certain dark patterns in consent banners, such as "Creating barriers" or "Hiding information", is a significant area of debate. While some regulatory bodies and stakeholders, including the European Data Protection Board (EDPB), the European Parliament, BEUC, and several national Data Protection Authorities (DPA), deem dark patterns like "tracking walls" unlawful, others like the ICO and Austrian DPA hold differing opinions (Gray et al. [2021]). The European Data Protection Board (EDPB) has clarified that making access to services conditional on consent for storing information (cookie walls) is non-compliant with GDPR's requirement for free consent (Gray et al. [2021]). The challenge lies in defining what constitutes

The legality of dark patterns in consent banners remains debated.

a "dark pattern" legally and enforcing these regulations effectively in the rapidly evolving digital landscape.

2.2.2 Public Pressure

Public pressure from advocacy groups and media campaigns holds companies accountable for dark patterns.

Public awareness and collective pressure play a significant role in holding companies accountable for their use of dark patterns. Consumer advocacy groups, privacy organizations, and ethical design movements actively campaign against deceptive interfaces, raising awareness among users and pressuring companies to adopt more transparent and ethical design practices. Social media campaigns, investigative journalism, and public shaming can lead to reputational damage for companies, incentivizing them to remove or modify dark patterns. The "Hall of Shame" on the aforementioned website[2] from Harry Brignull is a notable place to start.

2.2.3 Developer and Designer Awareness

Educating developers and designers on ethical principles can help preventing dark patterns and foster transparent digital environments.

Related to "public pressure" is the opportunity to guide, instruct and support developers and designers not to implement dark patterns in the first place. Promoting ethical design principles and raising awareness among developers and designers themselves is vital. Educational programs, industry guidelines, and professional codes of conduct can encourage designers to prioritize user autonomy, transparency, and fairness over manipulative tactics. Fostering a culture of ethical design within organizations can lead to a proactive approach to preventing dark patterns rather than merely reacting to their implementation.

Research shows the need for ethical education and research.

Research by Gray et al. [2018] highlights that while interest in critical scholarship on user experience (UX) practice is growing, a common vocabulary for assessing criticality is often lacking. Their work explores "dark patterns" as an ethical phenomenon where user value is supplanted by

[2] `https://www.deceptive.design/hall-of-shame`

shareholder value, emphasizing that UX designers can become complicit in manipulative practices. They advocate for implications for the education and practice of UX designers, and broadening research on the ethics of user experience.

Further insights from Beattie et al. [2024] reveal that UX designers often feel ethically motivated due to their "moral compasses," but their ability to act ethically is frequently restricted by commercial pressures and a limited purview of projects. In their study, focusing on designers in New Zealand, they also found that designers' understanding of ethics often does not align with determinations made by international privacy and design scholars, particularly regarding how user behavior can be shaped in ways that obfuscate beneficial privacy outcomes. These findings underscore the need for progressive ethics education in UX training institutions and ongoing professional development in data privacy for existing practitioners.

UX designers face ethical conflicts due to commercial pressures, needing better privacy and ethics education.

2.2.4 User Awareness

Educating users about dark patterns is a crucial countermeasure, as understanding these tactics can empower users to make more informed decisions online and avoid manipulation. Research by Maier and Harr [2020] indicates that while the specific term "dark pattern" might be unfamiliar to end-users, they are moderately aware of the existence of such deceptive techniques and can recognize some examples. Their studies show that users often perceive these patterns as sneaky, dishonest, and intentionally implemented, although they also express a resigned attitude due to their dependence on certain online services.

Educating users about dark patterns empowers them, despite their resignation to pervasive manipulative online services.

A key concept in this area is "manipulation literacy," proposed by Lewis and Vassileva [2024], which refers to a user's ability to identify manipulative techniques and their potential consequences, enabling them to make consensual, informed choices. Increased media coverage of data scandals (e.g., Facebook-Cambridge Analytica) and public discussions on social media have contributed to a rise in user

Even with increased manipulation literacy, users still struggle to detect or resist dark patterns effectively.

awareness and manipulation literacy as stated by Maier and Harr [2020]. However, studies also suggest that even with increased awareness, users might still struggle to detect dark patterns consistently (Keleher et al. [2022]) or resist them (Bongard-Blanchy et al. [2021]), particularly if they are unaware of the actual harm or dangers involved (Schäfer et al. [2024]). This highlights that knowledge alone may not be sufficient to prevent manipulation, as people are susceptible to persuasive technologies even after learning about their existence (Weinschenk [2013], Maier and Harr [2020]).

Children are highly vulnerable to dark patterns; targeted education and serious games can improve their detection skills.

Furthermore, research specifically focusing on children, such as by Schäfer et al. [2024], reveals that children are particularly vulnerable to dark patterns. While many children understood the intentions behind simple dark patterns and some could spot complex manipulations, a significant portion missed subtle deceptive elements like "Bad Defaults" in privacy settings. This underscores the critical need for targeted education for younger users, not just about dark patterns generally, but also about the specific risks and how to resist them. Fiedler et al. [2025] offer a promising approach with their serious game "Deception Detected!" to sensitize users to dark patterns in a risk-free and engaging learning environment, thereby improving their detection skills.

Initiatives that provide examples of dark patterns, explain their mechanisms, and offer tips for navigating deceptive interfaces are crucial. Online databases (e.g., darkpatterns.org, darkpatterns.uxp2.com) also contribute significantly to raising public awareness by cataloging and exposing these practices (Maier and Harr [2020]).

2.2.5 Technical Weakening of Dark Patterns

Technical solutions like browser extensions could directly weaken dark patterns on user devices, offering immediate protection.

Beyond legal and awareness-based approaches, technical solutions can also contribute to weakening the effectiveness of dark patterns. This can involve interventions directly on the user's device, often in the form of browser extensions or similar tools, which offer immediate protection without requiring server-side changes (Conti and Sobiesk [2010], Schäfer et al. [2024], Graf [2024]).

Research explores visual countermeasures against dark patterns, including highlighting, removing, switching, lowlighting, and friction designs.

Research in this area focuses on how to best communicate detected dark patterns to users to mitigate their influence (Schäfer et al. [2023], Schäfer et al. [2024]). Proposed visual countermeasures include:

- Highlighting: Marking detected dark patterns with visual cues and providing additional explanations (Mathur et al. [2019], Schäfer et al. [2023], Schäfer et al. [2024]). This approach aims to enhance user awareness and detection capabilities (Schäfer et al. [2023]). However, concerns exist about visual clutter, especially with multiple dark patterns present simultaneously (Schäfer et al. [2023]).
- Hiding/Removing: Visually altering, rephrasing, or completely removing manipulative elements to turn dark patterns into "fair patterns" (Moser et al. [2019], Schäfer et al. [2024], Lu et al. [2024]). While this can make pages clearer and improve usability, it can be controversial, as users might fear being deprived of relevant content or that sketchy websites might appear misleadingly trustworthy (Schäfer et al. [2023], Schäfer et al. [2024]).
- Switching: Providing a toggle to switch between the original manipulative view and a hidden or altered version (Schäfer et al. [2024]). This offers a compromise between hiding and transparency, giving users more control and allowing them to verify the countermeasure's effect (Schäfer et al. [2023], Schäfer et al. [2024]).
- Lowlighting: Recolorizing dark patterns to make them less alarming, a less intrusive approach than highlighting (Graf [2024]).
- Friction Designs: Introducing deliberate friction to disrupt automatic user behavior and encourage more conscious decision-making (Bongard-Blanchy et al. [2021], Lu et al. [2024]).

Modular browser extensions like "Deceptive Defender" simplify testing and updating dark pattern countermeasures.

The development of modular browser extension frameworks, such as "Deceptive Defender" by Graf [2024], aims to simplify the testing and updating of detection algorithms and countermeasures. These frameworks are designed to be adaptable to future dark pattern developments and support research into novel detection and countermeasure strategies. Such tools could empower end-users by allowing them to select between pre-defined UI enhancements based on their preferences and goals (Lu et al. [2024]).

Detecting and countering dark patterns remains challenging; future work needs improved algorithms and customizable tools.

Despite these advancements, technical weakening of dark patterns faces several challenges. Automatic detection of dark patterns is still an active area of research and not yet fully mature (Schäfer et al. [2023]). Some dark patterns may be inherently difficult or even impossible to detect automatically due to their variety (Curley et al. [2021]). Furthermore, detection tools are often reactive, as new patterns continually emerge (Hausner and Gertz [2021]). The effectiveness of countermeasures can vary significantly depending on the specific dark pattern and individual user preferences as shown by Schäfer et al. [2023]. Therefore, future work needs to focus on improving detection algorithms, developing more sophisticated and customizable countermeasures, and ensuring that these tools provide users with meaningful control over their online experience.

AI and LLM

AI and LLMs offer a promising frontier for automated dark pattern detection.

The application of Artificial Intelligence (AI) and Large Language Models (LLMs) presents a promising frontier in the technical weakening of dark patterns, particularly for their detection and the simulation of user responses. The sheer volume and pervasiveness of dark patterns make human detection and enforcement challenging, necessitating automated solutions (Soe et al. [2022]).

Early machine learning approaches detected dark patterns using web crawlers, text clustering, and supervised ML, but faced challenges in data encoding and labeling.

Early approaches to automated detection leveraged machine learning (ML) techniques to identify dark patterns. For instance, Mathur et al. [2019] developed automated methods, including web crawlers and text clustering, to identify and measure dark patterns at scale on shopping websites, discovering over 1,800 instances across 1,200 websites. Similarly, Soe et al. [2022] explored the use of supervised ML for automating the detection of dark patterns in cookie banners, using a dataset of features describing UI elements. While their initial results showed promise, they also highlighted significant challenges, such as the difficulty in encoding interfaces as feature values and the need for human intervention in data labeling, concluding that dark pattern detection is complex for AI because it is complex for humans. Nazarov and Baimukhambetov [2022] also proposed using cluster analysis algorithms for detecting dark patterns in user interfaces of websites and e-commerce portals, addressing the challenge of lacking formalized datasets.

More recently, generative AI technologies, particularly LLMs, are being investigated by Mills and Whittle [2023] as high-level auditing tools for detecting dark patterns by simulating the experiences of diverse online users. This approach aims to address the "representative agent problem," acknowledging that different individuals experience dark patterns differently based on factors like digital skill and cultural background. Mills and Whittle [2023] propose three methods for using generative AI:

Mills explores using generative AI, specifically LLMs, to detect dark patterns by simulating diverse user experiences. They propose "Choose your own adventure," "AI Vision," and "Decision Network" approaches.

"Choose your own adventure" approach: The LLM is given a text description of the choice architecture and valid options, then prompted to select an action based on a given user persona. This method is relatively easy to implement and automate but is highly dependent on prompt engineering and the quality of text descriptions.

"AI Vision" approach: The LLM receives images (screenshots) of the choice architecture, reducing informational discrepancies between what a user sees and what is described. Preliminary testing with GPT-4 showed promising results in identifying deceptive design features and predicting behavior congruent with human auditors, especially for obvious dark patterns.

"Decision Network" approach: This is the most technically advanced, aiming to integrate AI functionality into a web crawler that navigates online services in real-time using

HTML and JavaScript code. While offering the most objective description of a web page, preliminary testing suggests LLMs may struggle to "see" the visual aspects of the page from HTML alone, focusing more on links.

Algorithmic fidelity allows LLMs to simulate user responses to dark patterns, offering fast, cost-effective behavioral audits despite current limitations.

The concept of "algorithmic fidelity" (Aher et al. [2023]) underpins these simulation approaches, suggesting that LLMs, trained on vast datasets, can approximate a wide variety of individual subjects, providing reasonable estimations of population responses to dark patterns (Mills and Whittle [2023]). This "silicon sampling" offers advantages in terms of speed and cost compared to recruiting human participants for behavioral audits. However, limitations remain, including the need for careful prompt engineering, the subjective nature of screenshot generation, and technical challenges in automating complex web interactions for LLMs.

Chapter 3

Virtual Assistant as Countermeasure

This chapter delves into the potential of virtual assistants as a novel countermeasure against dark patterns in user interfaces. Building upon the understanding of dark patterns and existing countermeasures discussed in the previous chapter, this section outlines the conceptualization, investigation, and preliminary findings related to the development of a virtual assistant designed to empower users against dark patterns. Due to the current rapid development of artificial intelligence (AI) and the resulting possibilities, AI represents a particularly interesting way of tackling the problem of dark patterns. This interest extends to both the detection of dark patterns and their mitigation as a countermeasure. However, in this thesis, the focus is on the application of AI as a direct countermeasure in the form of a virtual assistant. For the scope of this research, it is assumed that the underlying dark pattern detection mechanisms are sufficiently reliable so that the research can focus on the role of the assistant in intervening and empowering the user.

This chapter explores virtual assistants using AI as countermeasures against dark patterns, assuming reliable detection for intervention.

3.1 Virtual Assistants

Virtual assistants, powered by AI, offer a unique opportunity for user protection against dark patterns.

Virtual assistants, often powered by artificial intelligence and natural language processing, have become ubiquitous in various digital environments, ranging from voice-controlled devices to integrated functionalities within web browsers and applications. These assistants are designed to perform a wide array of tasks, including information retrieval, scheduling, communication, and personalized recommendations. Their increasing sophistication and integration into daily digital interactions present a unique opportunity to explore their role beyond conventional assistance, specifically as agents for user protection in complex online environments.

Excursus: *Intelligent Assistant*

INTELLIGENT ASSISTANT:
An agent that uses artificial intelligence and can interact with user(s) via natural and/or artificial language by combining one or more communicative and sensory modalities to assist and collaborate with them.

An intelligent assistant is an AI-powered, interactive, and assistive agent providing context-aware, proactive user support.

An intelligent assistant, as defined by Shaikh [2023] above, is "*an agent that uses artificial intelligence and can interact with user(s) via natural and/or artificial language by combining one or more communicative and sensory modalities to assist and collaborate with them.*" This definition highlights three constitutive features: AI-enabled, interactive, and assistive. Maedche et al. [2016] further categorize Advanced User Assistance Systems (UAS) based on their degree of intelligence and interaction, distinguishing between basic, interactive, intelligent, and anticipating UAS. These advanced systems are characterized by their ability to provide context-aware, proactive, and adaptive assistance, sensing user activities and environments to offer tailored support and recommendations (Maedche et al. [2016]).

Virtual assistants evolved from annoying "Clippy" to sophisticated AI-powered tools like Siri, Google Assistant, and Alexa.

The evolution of virtual assistants has seen a transition from early, often frustrating, attempts like Microsoft Office's "Clippy" (Maedche et al. [2016], Guidobono [2024]), which was criticized for being annoying and disruptive, to more sophisticated and well-received modern examples such as Apple's Siri, Google Assistant, and Amazon Alexa (Maedche et al. [2016], Shaikh [2023]). These contemporary virtual assistants leverage advancements in AI, particularly deep learning, natural language processing (NLP), speech-to-text (STT), and text-to-speech (TTS) technologies, to offer more intuitive and effective interactions (Guidobono [2024]).

In terms of interaction, AI can adopt various roles (Dix et al. [2024]):

AI can interact as a servant, master, or in symbiosis with users, leveraging complementary abilities for productivity.

- Servant: The AI explicitly performs tasks based on user commands (e.g., "turn up the heating" to Alexa). The AI's intelligence serves to fulfill the user's explicit instruction.
- Master: The AI dictates actions to the user, as seen in gig-economy platforms where algorithms assign tasks to workers.
- Symbiosis: This represents the most productive interaction, where humans and AI work together, leveraging complementary abilities. In this synergistic relationship, the AI may take initiative, seek clarification, or make suggestions, rather than passively awaiting commands (Dix et al. [2024]).

For a virtual assistant to be effective, especially in a protective role against dark patterns, its design must adhere to principles of "appropriate intelligence." This means the AI should be right as often as possible, and when it is right, it should be genuinely helpful. Crucially, when the AI is wrong, it "shouldn't mess you up" (Dix et al. [2024]). This principle emphasizes designing systems that fail gracefully and do not disrupt the user's workflow, a lesson learned from the negative reception of Clippy (Dix et al. [2024]). Furthermore, virtual assistants should aim to reduce human cognitive and physical efforts while increasing performance, without being attention-grabbing or disruptive (Guidobono [2024], Maedche et al. [2016]). They should provide clear and precise guidance, tailor assistance to the user's context, and offer proactive support (Guidobono [2024], Maedche et al. [2016]).

Effective virtual assistants for dark patterns need "appropriate intelligence": helpful, non-disruptive, and reducing user effort.

Virtual assistants could proactively inform users, suggest alternatives, or modify interfaces to combat dark patterns.

The inherent capabilities of virtual assistants, their ability to process information, interact dynamically, and provide assistance, make them highly relevant as a countermeasure against dark patterns. By leveraging their AI-driven understanding of user context and their capacity for proactive intervention, virtual assistants can potentially inform users about dark patterns, suggest alternative actions, or even modify interfaces to protect user autonomy, thereby acting as a powerful tool in the fight against dark patterns.

3.2 Research Questions

This research investigates optimal virtual assistant design to counter dark patterns, focusing on prominence (RQ1) and intervention level (RQ2).

The central objective of this research was to investigate the feasibility and optimal design of a virtual assistant as a countermeasure against dark patterns. This led to the crystallization of several key research questions that guided the study:

- RQ1: How prominent should a virtual assistant be to effectively serve as a countermeasure against dark patterns?
- RQ2: To what extent should a virtual assistant intervene in website functionality or content?

RQ1 explores the optimal level of visibility and noticeability required for the assistant to be perceived and utilized by users without being overlooked or becoming intrusive. Since the virtual assistant is supposed to helpful, it should also be noticed. Help that is overlooked is simply not helpful. On the other hand, the virtual assistant should not be too distracting either, as it could otherwise be perceived as an annoyance and disrupt the flow.

RQ2 addresses the balance between merely informing the user about detected dark patterns and actively modifying the website's code or interface to mitigate their effects. Do users want the website to be actively changed by the virtual assistant or do they just want to be informed about dark patterns or is that perhaps already too much and users just want the opportunity to inform themselves?

3.3 Preliminary Study

To gain insights into user expectations and preferences regarding a virtual assistant as a dark pattern countermeasure, we conducted a preliminary study. This study aimed to gather qualitative data on desired functionalities, interaction modalities, and intervention strategies from individuals with prior knowledge of dark patterns.

We conducted a preliminary study to understand user expectations for a virtual assistant combating dark patterns.

3.3.1 Research Questions (Preliminary Study)

The preliminary study was designed to answer the following questions:

The preliminary study explored user expectations for virtual assistant functionalities and interaction methods against dark patterns.

- What functionalities and features do users expect from a virtual assistant designed to counter dark patterns?
- What possibilities should there be for interacting with the virtual assistant?

3.3.2 Study Procedure

The preliminary study employed a qualitative approach, involving creative focus groups and interviews with participants. The preliminary study was conducted as follows:

The preliminary study used qualitative creative focus groups and interviews with five students knowledgeable about dark patterns.

Participants with prior knowledge of dark patterns were sought for the study. A total of five students participated in two groups.

The study began by reviewing dark patterns and user experiences, then categorized examples and discussed existing countermeasures.

Initially, the concept of dark patterns was reiterated, and participants were given the opportunity to describe their previous experiences and engagement with dark patterns. They were then encouraged to list any dark patterns they were familiar with. The listed dark patterns were subsequently sorted into categories provided by Gray et al. [2024]. To ensure a comprehensive understanding, further examples of dark patterns were presented, ensuring that at least one example from each category was reviewed. Following this, a brief overview of existing countermeasures was provided by the ontology from Gray et al. [2024] and selected examples.

Participants brainstormed and voted on virtual assistant ideas to combat dark patterns collaboratively.

The main part of the study then commenced. Participants were informed about the proposed development of a virtual assistant as a countermeasure and were tasked with brainstorming, collaboratively, how such a virtual assistant could best support users in dealing with dark patterns. Ideas were to be written down on post-it notes. After an initial 5-minute brainstorming period, each participant was asked to present and explain their ideas to the group. This entire process was then repeated. With the benefit of hearing others' ideas, participants engaged in a second 5-minute brainstorming phase, again writing down ideas on how a virtual assistant could most effectively support users against dark patterns.

Finally, each participant was asked to vote for their three most favored ideas.

Participants

Five students from the i10 with prior dark pattern knowledge participated in the preliminary study.

All participants in the preliminary study had prior knowledge of dark patterns. This criterion ensured that the feedback gathered was informed by an understanding of the problem space and potential solutions. A total of five participants, consisting of students from the i10 Chair (Human-Computer Interaction)[1] at RWTH Aachen University, participated in the study.

[1] https://hci.rwth-aachen.de/welcome

3.3.3 Results

The data collected as part of the preliminary study was analyzed to identify recurring themes, preferences and expectations of the virtual assistant. The results were divided into thematic categories.

Findings

Participants generated 60 ideas for a virtual assistant combating dark patterns, categorized into seven functions like page rating and removing.

During the brainstorming sessions, a total of 60 ideas were noted and then categorized into seven different categories. These categories represent the variety of functions and approaches that participants envisioned for a virtual assistant as a countermeasure to dark patterns. The seven categories identified were: Page rating, highlighting/marking, explaining, removing, setting options, assistant warns/interrupts and assistant performs tasks for the user.

Page rating ideas involved showing dark pattern counts, traffic light ratings, and pre-visit warnings for deceptive sites.

Page rating: This category consisted of ideas that had in common that the website is rated in some way. Mostly in such a way that the number of dark patterns (DP) on the website was counted. Either the number should be displayed directly or simplified (many DP = bad, few DP = good), for example as a “traffic light”, so if comparatively few dark patterns were found on the website, it is green, if a particularly large number of dark patterns were found, it is displayed in red and yellow in between. Two ideas described that such an evaluation of the website should be displayed even before the website is visited, for example during the web search on the search engine results page. One idea in this category suggested clearly marking "fake websites" (e.g. in scam emails) as "fake websites".

This category focuses on dynamically highlighting dark patterns on websites to increase user awareness.

Highlighting/marking: This category revolves around the concept of highlighting/marking the dark patterns on the website so that the user is made aware of them. A countermeasure that has already been examined in user studies by Schäfer et al. [2023]. One idea that stood out was to make the highlighting of dark patterns dynamic, i.e. with an animation and not just static highlighting.

This category emphasizes explaining dark patterns to users, providing information and tips via interactive elements.

Explain: This category is mainly about explaining the dark patterns to the user, so that the user can find out what kind of dark pattern is shown here and what it is made of. This information could then be made available to the user via hover menus or by clicking on the dark pattern, for example. Tips on how to deal with certain dark patterns would also be useful. One idea also described that the assistant should communicate very clearly what it has done and why, i.e. explain its actions.

The "Remove" category focused on direct dark pattern countermeasures, including altering "Bad Defaults" and using LLMs to rewrite deceptive text.

Remove: This category was the largest and contains almost all ideas that describe more direct countermeasures to dark patterns. This primarily involves removing identified dark patterns whenever feasible, or exploring other methods to mitigate their effects. Schäfer et al. [2023] has also researched this type of countermeasure. The gathered ideas described in particular how the assistant should deal with certain dark patterns, e.g. for the dark pattern “Bad Defaults”, the assistant should ensure that no option or the option that is most likely to be in the user’s interest is selected. For the dark patterns "Shaming" and "Trick Question", the corresponding text passages should be rewritten directly with the help of LLMs. Other interesting ideas from this category were for the dark pattern “Forced Registration”, the assistant could provide “fake user data” for registration. Additionally, a suggested approach was to "gray out" or visually obscure identified dark patterns similar to the "lowlighting" of Graf [2024]. The idea of using animations to emphasize the removal of dark patterns also came up in this category.

The "Setting options" category focuses on user preferences for dark pattern handling, including intervention strength and marking false positives.

Setting options: This category groups together a number of ideas that all have in common that users should be able to adjust something, as already predicted by Schäfer et al. [2023] in their "Future Work" section. This includes settings for specific dark pattern interventions and broader user preferences for the assistant's behavior. Users could also adjust the strength of the assistant's intervention or flag incorrectly identified dark patterns. Additionally, the concept of a tutorial for dark patterns and the assistant was proposed, which would then allow users to fine-tune the assistant's settings based on their learning.

This category involves the assistant actively warning/interrupting users about dark patterns, like unexpected items in shopping carts.

Assistant warns/interrupts: This category is characterized by the fact that the assistant actively acts on these ideas and thus interrupts the user's flow, similar to the Friction Designs from Bongard-Blanchy et al. [2021] and Lu et al. [2024]. Most ideas describe a type of warning so that the user is made aware that they could be manipulated by the dark patterns. For example, a warning when shopping online with the dark pattern "Sneak into Basket" comes with the content: "Attention! You did not add these products to your basket yourself. Would you still like to buy it?"

This category suggests the assistant independently performs tasks or bypasses dark patterns for the user.

Assistant performs tasks for the user: This category contains ideas in which the assistant is expected to act independently. An example from this category is to instruct the assistant to buy a certain product for the user so that the user does not even see the shopping website, but at most the product. Another example is to counteract the dark pattern “Obstruction”, here the assistant should show the user exactly the easiest way to fulfill his goal, i.e. the easiest way to circumvent the “Obstruction”. The first idea shows similarities to the "Servant" Role (3.1) of AI where as the second idea more similarities to the "Master" role described by Dix et al. [2024] has.

Voting

In this section, we present the ideas that received the most votes. Each participant received two votes, and the voting was conducted exclusively in the second group, which consisted of three participants.

The participants desire granular control over how an assistant combats dark patterns, from informing to removal.

One of the highly favored ideas was: "*Einstellung wie extrem gegen Pattern eingegriffen wird*" (Setting how extremely patterns are intervened against). This idea suggests that users should have granular control over how the assistant handles each detected dark pattern. This includes options for the assistant to merely point out and inform the user, to mark (highlight) the area containing the dark pattern, to weaken its effect (e.g., by graying out), to completely remove the dark pattern, or to ignore it entirely.

The participants want to whitelist falsely recognized dark patterns to improve assistant accuracy and prevent unnecessary interventions.

Another popular idea was: "*Fälschlich erkannte Pattern whitelisten*" (Whitelist falsely recognized patterns). This proposes that users should have the option to inform the assistant about false positives (instances where the assistant incorrectly identifies a dark pattern). This feedback mechanism would allow the assistant to improve its recognition accuracy or to specifically ignore certain areas in the future, preventing unnecessary interventions or notifications.

The participants prioritize clear communication from the assistant about its actions and reasons, depending on the dark pattern.

A third idea was: "*Klare Markierung was der Assistent gemacht hat wäre mir wichtig (Kommt auf pattern an)*" (Clear marking of what the assistant has done would be important to me (depends on the pattern)). This idea emphasizes the importance of transparency regarding the assistant's actions. Participants felt it was crucial for the assistant to clearly communicate to the user what specific actions it has taken and why, ideally tailored to the particular dark pattern being addressed.

The most favored idea was an assistant that automatically declines cookies by removing banners entirely.

The idea that received the most votes was: "*Decline cookies for you, ideally without the banner even showing*". This concept describes an assistant that automatically rejects all cookies by removing cookie banners, ideally without the user ever seeing the banner. Research conducted subsequent to the study confirmed that browser extensions already exist that perform this exact function.

An assistant that directly displays the "real" price, combating hidden costs and price manipulation.

Finally, another highly-voted idea was: "*Bei Preismanipulation: Direkt den 'wirklichen' Preis hinschreiben*" (For price manipulation: Directly write down the 'real' price). This idea focuses on dark patterns that affect the final price, such as "hidden costs." The assistant would combat these by directly displaying the final, true price to the user, bypassing the manipulative presentation.

3.3.4 Discussion

The preliminary study showed users expect a virtual assistant to directly remove or highlight dark patterns, aligning with existing countermeasures.

The findings from the preliminary study offer valuable insights into the design considerations for a virtual assistant aimed at countering dark patterns. Unsurprisingly, a strong expectation emerged for the assistant to intervene directly against various dark patterns, ideally by removing them or negating their effects. Many ideas in the "Removing" and "Highlighting/Marking" categories align with existing research and implemented countermeasures, suggesting a clear user desire for tangible interventions. In particular, the concept of highlighting dark patterns is a method that has already been investigated to increase user awareness.

The study emphasizes that an assistant must explain dark patterns and its own actions for user trust.

The study also highlighted the importance of the assistant providing comprehensive information to users. This includes not only explanations about the detected dark patterns themselves (e.g., their type and what they are designed to do) but also transparency regarding the assistant's own actions. This aligns closely with the principles of Explainable AI (XAI), where understanding the "why" behind an AI's actions is crucial for user trust and adoption.

Balancing user protection and an uninterrupted experience is crucial for an assistant warning against dark patterns.

A critical and challenging aspect revealed by the ideas in the "Assistant Warns/Interrupts" category is the tension between protecting the user and maintaining an uninterrupted user experience. While active interruptions might be necessary to safeguard users from the immediate effects of dark patterns, such disturbances can also be perceived negatively and lead to user frustration. Designing an assistant that can effectively warn or interrupt without being overly intrusive will be a key design challenge.

Users strongly desire customizable assistant settings to manage dark patterns, adapting to diverse and evolving needs.

Furthermore, the strong emphasis across almost every participant's ideas on "Setting Options" underscores the need for high customizability. This is driven by several factors:

the diverse needs and preferences of a broad user base, the potential for individual users' needs to evolve over time (e.g., from a desire for detailed explanations to a preference for autonomous intervention once familiar with dark patterns), and the subjective nature of what constitutes an "acceptable" intervention. An adaptive assistant that learns from user preferences and allows for fine-grained control over its behavior will be essential for user satisfaction and long-term adoption.

Rating a website regarding dark patterns is feasible while browsing; pre-emptive analysis on search results poses greater technical challenges.

Regarding the technical feasibility of certain ideas, evaluating a website based on the number of dark patterns present appears technically achievable, provided the user is actively on that website. However, analyzing a website in advance, for instance, displaying a rating on a search engine results page before the user has even visited the site, presents significantly greater technical challenges due to the need for pre-emptive crawling and analysis without user interaction.

The automation of complex tasks such as product purchases or the circumvention of dark patterns for users poses a technical challenge.

Similarly, instructing the assistant to perform complex tasks for the user, such as autonomously purchasing a product on a website, is technically quite challenging. While some website operators' own assistants (e.g., Alexa on Amazon) can perform limited tasks within their specific ecosystems, they are designed to serve the operator's interests. Building a general-purpose assistant capable of navigating diverse online shopping environments and acting solely in the user's best interest is a much more complex undertaking. Moreover, analyzing website structures to the extent required for the assistant to explain or automate complex processes, such as circumventing deliberately difficult account deletion procedures, remains a significant technical hurdle at the current stage of development.

3.3.5 Limitations

The preliminary study's small, specialized participant pool and limited voting reduced generalizability. Future research needs a larger, more diverse sample.

Despite the valuable insights gained, the preliminary study had several limitations. A clear limitation is the small number of participants, with only five individuals participating in this study. While their insights were valuable, such a limited sample size restricts the generalizability of the findings. Another limitation is the specific background of the participants. The study intentionally recruited individuals who were already familiar with dark patterns. While this ensured informed feedback and a more comprehensive understanding of user needs, insights from individuals who are not yet familiar with the topic are also highly relevant. Finally, the voting process was limited to only the second of the two groups, further impacting the representativeness of the quantitative preferences expressed. Future research should aim to include a more diverse and larger participant pool, and standardize voting across all groups, to enhance the robustness and applicability of the findings.

Chapter 4

User Study

This chapter details the user study conducted to further investigate the effectiveness and optimal design of a virtual assistant as a countermeasure against dark patterns. Building upon the insights gathered from the preliminary study, this user study aimed to evaluate specific prototype variations of the assistant in a simulated online shopping environment, gathering both quantitative and qualitative feedback from participants. This user study aims to build upon the findings of Schäfer et al. [2023].

This chapter details a user study evaluating virtual assistant prototypes against dark patterns in simulated online shopping.

4.1 Research Questions

The user study was designed to address key questions regarding the prominence and intervention level of the virtual assistant, directly building on the findings and open questions from the preliminary study. The research questions guiding this study were:

This user study investigates if a virtual assistant is helpful, safe, and maintains user control against dark patterns.

- RQ1: Is the assistant perceived as distracting or helpful?
- RQ2: How safe do users feel with the assistant?
- RQ3: Do users feel in control of their actions?

RQ1: This question seeks to understand the impact of the assistant's prominence and interaction style on the user's overall experience and task completion. This question also addresses the user's flow. Is it interrupted or not?

RQ2: This question explores the assistant's ability to instill a sense of security and mitigate the perceived manipulative influence of dark patterns. Do the participants still believe they are being manipulated by the dark patterns, or do they now feel protected from the effects by the assistant?

RQ3: This question investigates the balance between the assistant's autonomous actions and the user's sense of agency and understanding of the assistant's interventions. Does the assistant communicate clearly enough what it is doing, or do the participants feel a loss of control?

4.2 Prototype

This section will describe the virtual assistant prototype used in the user study, detailing its functionalities and how it was implemented to address the research questions. The prototype incorporated different variations of the assistant's behavior regarding prominence and intervention.

We named the virtual assistant "Lumi" and illustrated it with a light bulb with a smiley face which can be seen in Figure A.1.

A Figma prototype with seven virtual assistant variants simulated online furniture shopping for a wall clock purchase.

For this study, a medium-fidelity prototype was developed using Figma. The prototype simulates a limited segment of an fictitious online furniture shopping website, which can be seen in Figure 4.1, and incorporates seven different variants of the virtual assistant. The scenario was specifically designed to involve the purchase of a predetermined wall clock, ensuring consistency across all trials.

The prototype's product page displays a wall clock with images, details, price, warranty, and cart options.

The prototype begins on the product page of the wall clock, which can be seen in Figure 4.1. This page features several images of the wall clock, essential product details, its price, a checkbox option to select a warranty, and buttons to add the wall clock to the shopping cart and navigate to the cart view.

Upon proceeding, the shopping cart view, which can be seen in Figure 4.3, displays the contents of the cart, the total price including shipping costs, and options to select the shipping method. At the bottom of this view, a "Proceed to Checkout" button is available to complete the simulated purchase scenario, along with a checkbox for newsletter subscription.

The shopping cart view shows items, total price, shipping options, checkout, and a newsletter checkbox.

4.2.1 Dark Pattern

In this subsection, we will describe the specific dark patterns that participants encountered within the simulated online shopping scenario during the user study. These dark patterns were carefully selected to represent common manipulative tactics found in e-commerce environments.

This section details the common dark patterns found in the simulated online shopping scenario for the user study.

Product Page

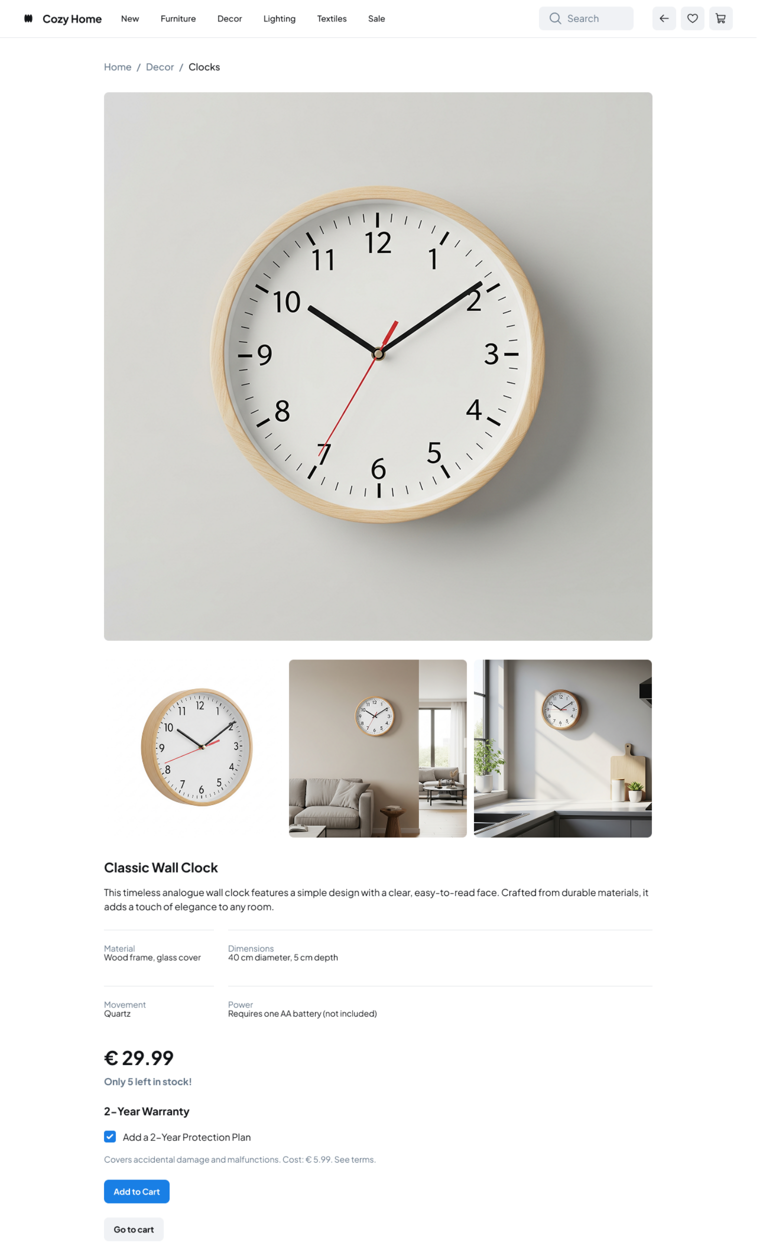
On the product page, (see Figure 4.1) the dark patterns "Low Stock," "Bad Defaults," "Adding Steps," "Visual Prominence," and "Confirmshaming" were integrated:

Low Stock: Displayed directly below the price as "Only 5 left in stock!" This dark pattern creates a false sense of urgency, pressuring the user into making a quick purchase before the item supposedly runs out.

"Only 5 left!" creates false urgency, pressuring quick purchases.

Bad Defaults: The checkbox for a 2-year warranty for an additional €5.99 is pre-selected. A pre-selected option is not necessarily in the user's best interest, requiring active opt-out.

The "Bad Defaults" dark pattern pre-selects a 2-year warranty for an extra €5.99, requiring users to opt out.



Classic Wall Clock

This timeless analogue wall clock features a simple design with a clear, easy-to-read face. Crafted from durable materials, it adds a touch of elegance to any room.

Material
Wood frame, glass cover

Dimensions
40 cm diameter, 5 cm depth

Movement
Quartz

Power
Requires one AA battery (not included)

€ 29.99

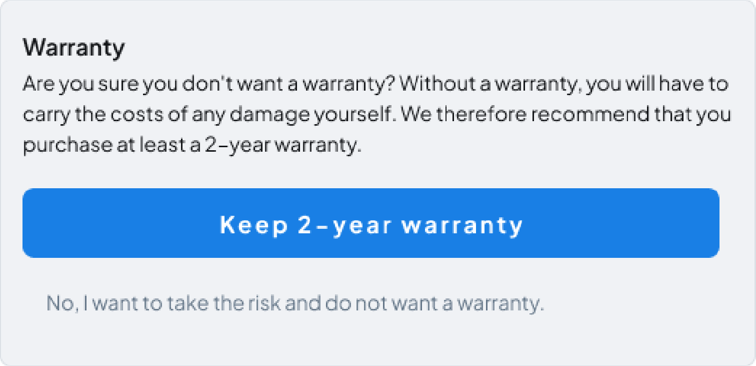
Only 5 left in stock!

2-Year Warranty

☑ Add a 2-Year Protection Plan

Covers accidental damage and malfunctions. Cost: € 5.99. See terms.

Figure 4.1: Screenshot of the simulated product page prototype in the "Unchanged" variant. This prototype, a custom-made fictional e-commerce website, served as the baseline for evaluating user interaction with dark patterns without any intervention.



Warranty

Are you sure you don't want a warranty? Without a warranty, you will have to carry the costs of any damage yourself. We therefore recommend that you purchase at least a 2-year warranty.

Figure 4.2: Screenshot of the "Warranty Question" pop-up in the "Unchanged" variant, designed to embody the "Adding Steps" dark pattern. This window appears when the user attempts to remove the warranty, subtly nudging them to reconsider their action through visual prominence and confirmshaming, making the removal process even more tedious.

Adding Steps: If the user clicks on the activated checkbox to deselect the 2-year warranty, a pop-up appears, asking the user again whether they would like to have the warranty after all. The pop-up is shown in Figure 4.2. By adding unnecessary additional steps, the difficulty of the task is increased, discouraging the desired action.

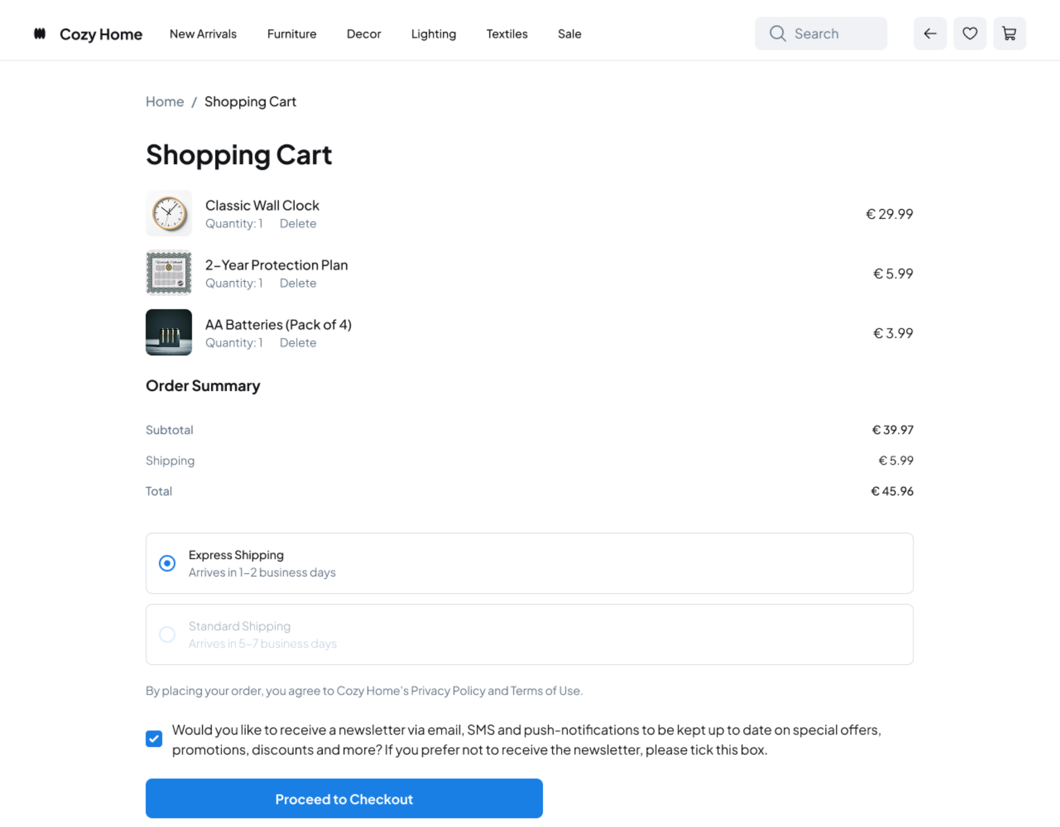
The "Adding Steps" dark pattern adds a pop-up confirmation, increasing task difficulty to discourage opting out.

Visual Prominence: In the pop-up (Figure 4.2) that appears when the user attempts to deselect the 2-year warranty, the "Keep 2-year warranty" button is large, blue, and therefore visually prominent and easily recognizable. In contrast, the "No, I want to take the risk and do not want a warranty" button is displayed only as grayish text. This design influences the user to click on the large, prominent button rather than the less conspicuous alternative. An option that stands out visually from others can distract from them.

The "Visual Prominence" dark pattern uses a large, blue "Keep warranty" button to manipulate users away from a small, gray "No warranty" text.

Confirmshaming: In the same pop-up (Figure 4.2), the button that allows the user to deselect the warranty is la-

The "Confirmshaming" dark pattern uses guilt-inducing language like "No, I want to take the risk" to discourage warranty declination.



Home / Shopping Cart

Shopping Cart

Classic Wall Clock Quantity: 1 Delete	€ 29.99
2-Year Protection Plan Quantity: 1 Delete	€ 5.99
AA Batteries (Pack of 4) Quantity: 1 Delete	€ 3.99

Order Summary

Subtotal	€ 39.97
Shipping	€ 5.99
Total	€ 45.96

● Express Shipping
Arrives in 1–2 business days

○ Standard Shipping
Arrives in 5–7 business days

By placing your order, you agree to Cozy Home's Privacy Policy and Terms of Use.

☑ Would you like to receive a newsletter via email, SMS and push-notifications to be kept up to date on special offers, promotions, discounts and more? If you prefer not to receive the newsletter, please tick this box.

Figure 4.3: Screenshot of the simulated shopping cart page prototype in the "Unchanged" variant. This custom-made fictional e-commerce website displays the cart contents and typical checkout elements, serving as the baseline for evaluating user interaction with dark patterns without any intervention.

beled "No, I want to take the risk and do not want a warranty." This phrasing implies a negative consequence or risk associated with declining the warranty, intending to discourage the user from this option by invoking feelings of guilt or fear. Certain phrasings can make an option sound worse than it actually is.

Cart Page

The dark patterns "Sneak Into Basket," "Bad Defaults," "Hiding Information," "Trick Question," "Adding Steps," "Visual Prominence," and "Confirmshaming" have been integrated on the shopping cart page which is shown in Figure 4.3:

Sneak Into Basket: A pack (4 pcs.) of batteries for €3.99 was automatically added to the shopping cart. This involves putting additional items into the basket that the user may not have explicitly wanted.

The "Sneak Into Basket" dark pattern automatically adds unwanted items, like batteries, to the user's cart.

Bad Defaults: The more expensive shipping method, "Express Shipping," is preselected. Similar to the product page, this pre-selection is not necessarily in the user's best interest, requiring active change.

The "Bad Defaults" dark pattern preselects more expensive "Express Shipping," requiring users to change it.

Hiding Information: The "Standard Shipping" option is displayed in light gray text, making it barely visible on the white background. This visual design also makes this option appear unavailable or inactive, making it harder for the user to select.

The "Hiding Information" dark pattern uses light gray text for "Standard Shipping," making it hard to see and select.

Trick Question: The question next to the corresponding activated checkbox for newsletter subscription is worded as follows: *"Would you like to receive a newsletter via email, SMS and push-notifications to be kept up to date on special offers, discounts and more? If you prefer not to receive the newsletter, please tick this box.*" This phrasing is counter-intuitive; if the user does not want to receive the newsletter, they must keep the checkbox activated (ticked), contrary to the common expectation of deactivating a checkbox to opt-out.

The "Trick Question" dark pattern uses counter-intuitive phrasing for newsletter opt-out, requiring users to tick a box to decline.

Adding Steps: If the 2-year warranty was not deselected on the product page, it will appear here in the shopping cart. If the user attempts to remove it from the shopping cart, a pop-up (see Figure 4.2) appears, identical to the one on the product page, again asking the user whether they would like to keep the warranty.

The "Adding Steps" dark pattern reappears in the shopping cart, presenting the same pop-up to deter warranty removal.

The "Visual Prominence" dark pattern uses a large, blue "Keep warranty" button to manipulate users toward that option.

Visual Prominence: Since the same pop-up (Figure 4.2) regarding the 2-year warranty appears as on the product page, the "Keep 2-year warranty" button is large, blue, and therefore prominent and easily recognizable as a button, while the "No, I want to take the risk and do not want a warranty" button is only displayed as grayish text. The user is therefore influenced to click on the large blue button rather than the other option. An option that stands out visually from others can distract from them.

The "Confirmshaming" dark pattern uses guilt-inducing language in a pop-up to discourage deselecting the warranty.

Confirmshaming: Since the same pop-up (Figure 4.2) appears regarding the 2-year warranty as on the product page, the button that lets the user deselect the warranty is labeled "No, I want to take the risk and do not want a warranty," which implies a risk with this option and is thus intended to discourage the user from this option. Certain phrasings can make an option sound worse than it actually is.

4.2.2 Countermeasures

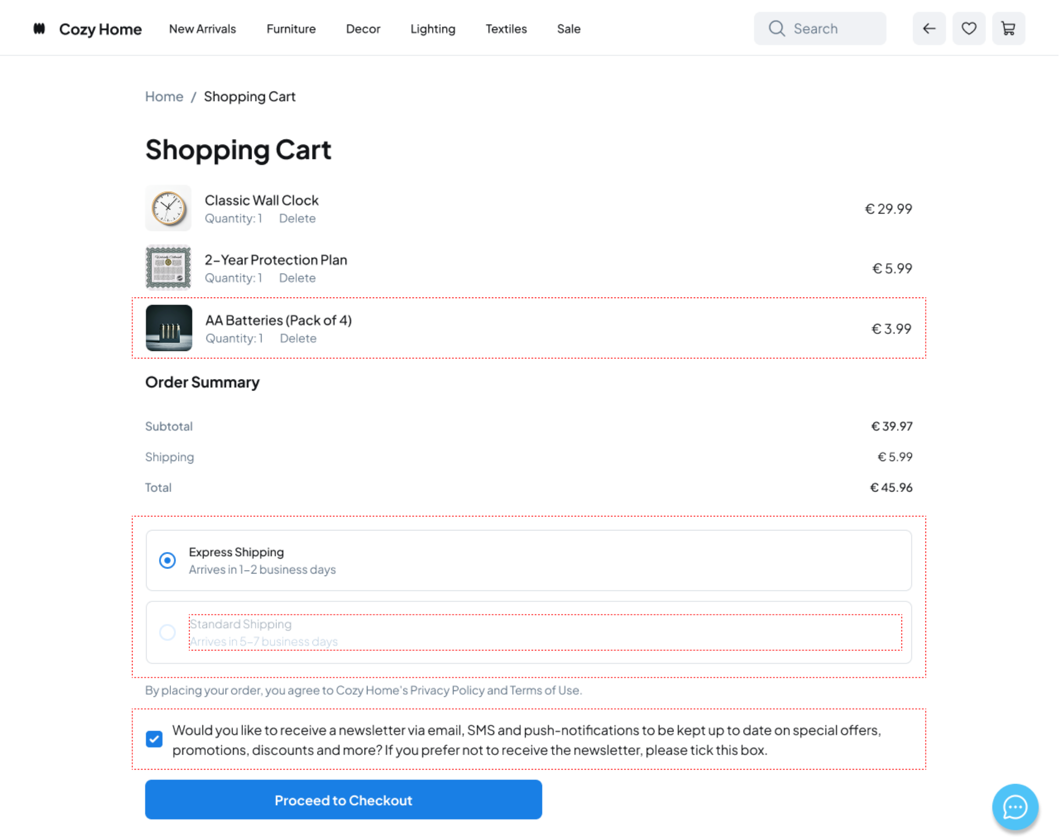
The study countered dark patterns by highlighting, removing/mitigating, informing users, and rating pages based on dark pattern count.

As countermeasures to the dark patterns, the study employed several strategies: clearly highlighting the areas of the dark patterns, actively removing or mitigating their effects, and providing users with the opportunity to inform themselves about the dark patterns on the respective page. Additionally, the page was rated based on the number of dark patterns present.

A red dashed border was used to visually highlight dark patterns.

Marking: For highlighting, a red dashed border was chosen to visually emphasize the areas containing dark patterns. The corresponding areas in the prototype then displayed this distinctive border to draw the user's attention. This can be seen in Figure 4.4.

Removing: This strategy involved directly counteracting the dark patterns by altering or removing their manipulative elements:



Shopping Cart

Classic Wall Clock Quantity: 1 Delete	€ 29.99
2-Year Protection Plan Quantity: 1 Delete	€ 5.99
AA Batteries (Pack of 4) Quantity: 1 Delete	€ 3.99

Order Summary

Subtotal	€ 39.97
Shipping	€ 5.99
Total	€ 45.96

● Express Shipping
Arrives in 1–2 business days

○ Standard Shipping
Arrives in 5–7 business days

By placing your order, you agree to Cozy Home's Privacy Policy and Terms of Use.

☑ Would you like to receive a newsletter via email, SMS and push-notifications to be kept up to date on special offers, promotions, discounts and more? If you prefer not to receive the newsletter, please tick this box.

Figure 4.4: Screenshot of the simulated shopping cart page in the "Mark + Chat" variant, showcasing the visual highlighting of detected dark patterns. This intervention aims to draw user attention to deceptive design elements while providing supplementary information via the chat interface.

- "Low Stock" was removed by simply not displaying the text area indicating limited stock.

Hiding the text.

- "Bad Defaults" were addressed by ensuring that no option was preselected, or by preselecting the option that was more beneficial or cheaper for the user (e.g., the warranty checkbox was not checked by default, and standard shipping was preselected).

Unchecking warranty and preselecting standard-shipping.

- "Adding Steps" was removed by preventing the pop-up that re-asked the user about the warranty. Instead, the checkbox or "Delete" button functioned as expected, allowing for a direct and uninterrupted action.

Removing the pop-up.

Matching the button design.

- "Visual Prominence" was mitigated by ensuring that the appearance of relevant buttons was consistent. For instance, both the "Keep 2-year warranty" and "No, I do not want a warranty" buttons were made large and blue, eliminating the visual bias.

Rephrasing the text.

- "Confirmshaming" was removed by changing the manipulative text. The phrasing was altered from "No, I want to take the risk and do not want a warranty" to the neutral "No, I do not want a warranty."

Removing batteries.

- "Sneak Into Basket" was counteracted by not automatically adding any additional items, such as batteries, to the shopping cart.

Improve visibility.

- "Hiding Information" was addressed by adjusting the visual appearance of the "Standard Shipping" method to match the "Express Shipping" method, ensuring good contrast (e.g., black or dark gray text on a white background) and clear visibility.

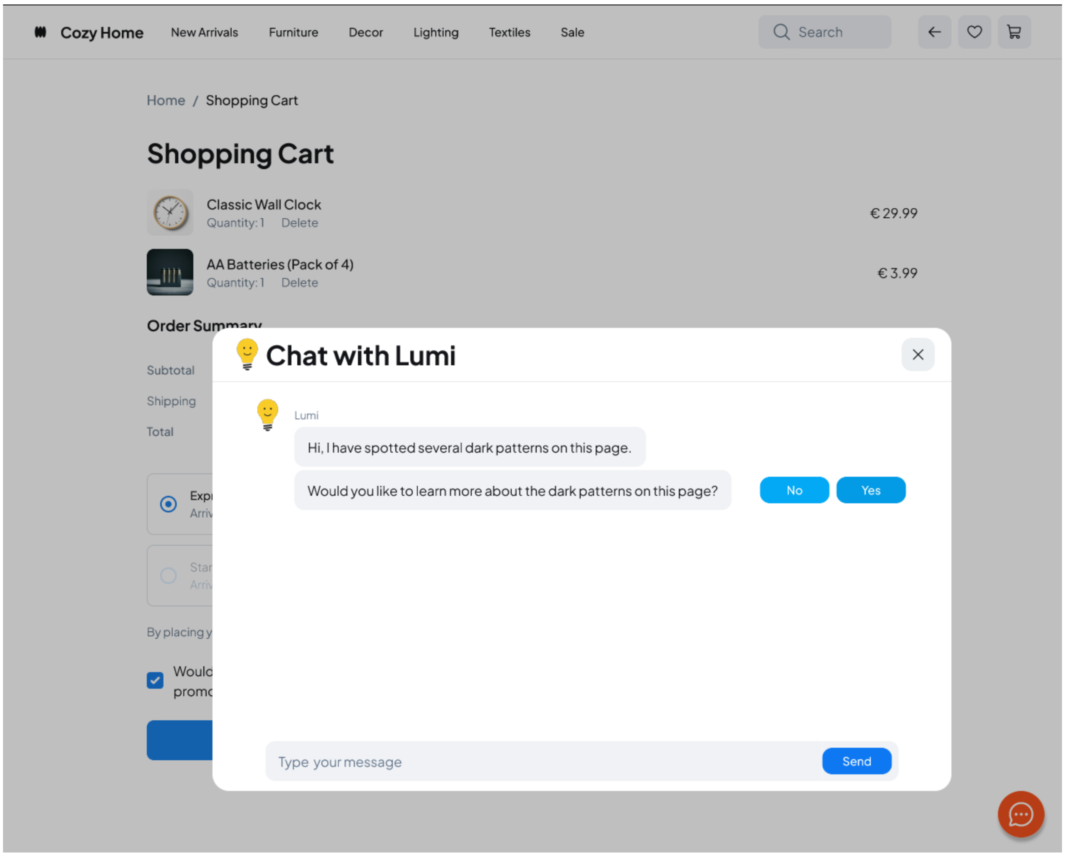
Rephrasing the text.

- "Trick Question" was removed by changing both the wording and the functionality of the newsletter checkbox. The question was simplified to "Would you like to receive a newsletter via email, SMS and push-notification, tick this box," and the checkbox's functionality was made intuitive, where ticking it meant opting in.

Figure A.2 and A.3 show a version of the product page and cart page with dark patterns removed.

The chat bot informed users about detected dark patterns via messages, descriptions, and screenshots, though direct questioning was not implemented.

Inform: To facilitate user self-information about dark patterns, the assistant was presented as a **chat bot** which can be seen in Figure 4.5. Upon activation, the assistant provided a brief introduction and immediately informed the user that it had detected several dark patterns on the current website. The user was then prompted, via interactive buttons, whether they wished to learn more about the identified dark patterns. If affirmative, the assistant provided further chat messages containing a list of the dark patterns found on that specific page, along with a brief description, an explanation of their effects, and a screenshot



Shopping Cart

Classic Wall Clock
Quantity: 1 €29.99

AA Batteries (Pack of 4)
Quantity: 1 €3.99

Chat with Lumi

Hi, I have spotted several dark patterns on this page.

Would you like to learn more about the dark patterns on this page?

Figure 4.5: Screenshot illustrating the "Counter + Chat" variant's functionality on the shopping cart page. The image highlights the interactive chat interface, which provides information about detected dark patterns, alongside the colored indicator signifying the presence of such patterns on the page.

illustrating the dark pattern's location on the page. A Part of this list in the chat window can be seen in Figure A.9. While the chat bot's appearance suggested the ability for users to ask questions, this functionality was not fully implemented. In instances where participants attempted to use this feature, the investigator provided answers on behalf of the assistant.

Page Rating: The button for the chat bot served as a visual indicator of the page's dark pattern density. Its color changed according to a traffic light system: green indicated few dark patterns, yellow indicated a medium number, and

The chat bot's color, like a traffic light, indicated dark pattern density (green for few, red for many).

red indicated many dark patterns. On the product page, the button was displayed in yellow which can be seen in Figure A.10 and A.11, while on the shopping cart page, it appeared in red-orange which can be seen in Figure 4.5 as well as Figure A.8 and A.9. The meaning of these colors was briefly explained to participants at the beginning of the respective scenarios.

4.2.3 Assistant Variants

Participants experienced eight virtual assistant variants, including a baseline, in a simulated online shopping scenario.

In the study, each participant navigated through the online shopping scenario eight times, each instance featuring a different variant of the virtual assistant. The very first variant always served as a baseline, without any assistant present. The following describes these variants and their distinguishing characteristics:

Baseline scenario with no assistant, exposing users to dark patterns in their original form.

Unchanged (UC): This variant of the scenario was always the first and did not include any virtual assistant. Participants were confronted with the dark patterns exactly as they would be in a typical online shopping environment. This allowed participants to familiarize themselves with the website in its "original state" and served as a control condition for comparison. This variant can be seen in Figure 4.1 and 4.3.

This variant provided information about dark patterns via a chat bot and a color-coded page rating, without direct website modification.

Counter + Chat (CC): In this variant, the assistant intervened minimally with the website's content. Its primary function was to be available to participants in the form of a chat bot (Figure 4.5), through which they could obtain information about the dark patterns detected on the website. Additionally, the assistant provided a rough estimate of the number of dark patterns found by evaluating the website and signaling this rating via the color of the chat bot button, using the previously explained traffic light scheme (4.2.2). This variant aimed to provide information and awareness without direct modification of the website.

Mark + Chat (MC): This variant focused on visual awareness. The areas containing dark patterns were clearly highlighted by markings (the red dashed border) which is shown in Figure A.6, 4.4 and A.15. Similar to the CC variant (4.2.3), participants also had the option to access the chat bot for more detailed information about the detected dark patterns. However, in this variant, the chat bot button remained a consistent blue color and did not provide any additional information regarding the number of dark patterns found on the page.

This variant visually highlighted dark patterns with borders; a chat bot offered information, but no page rating.

Animation (AN): This variant introduced a dynamic and proactive approach. The assistant actively removed the detected dark patterns from the website and simultaneously displayed a short, indicative animation. The animation involved the areas with dark patterns briefly lighting up red before disappearing or changing to their non-manipulative form. Screenshots of this animation can be seen in Figure A.4 and A.5. On the product page, where all dark patterns were located at the bottom, the animation was triggered when the user hovered over the area of the first dark pattern (from top to bottom), ensuring it was seen during natural user interaction. The view of the upper part of the product page can be seen in Figure A.10 for reference. The lower part is shown in Figure A.11. For the shopping cart page, which was small enough not to require scrolling (see Figure 4.3), the animation commenced automatically after a brief delay (800ms).

This variant actively removed dark patterns, showing a brief red animation.

Remove (R): In this variant, the dark patterns found on the website were removed directly and immediately by the assistant, without any accompanying animation, resulting in the websites shown in Figure A.2 and Figure A.3. This behavior was designed to be subtle; participants would only notice the removal if they had a prior comparison with the original, unmitigated website. This variant is conceptually similar to conventional ad blockers, which prevent advertisements from being displayed without explicit user interaction. This variant is basically the same as the "Hide without Marking (HD)" method used in the study

This variant subtly eliminated dark patterns directly, similar to an ad blocker.

performed by Schäfer et al. [2023] to improve comparability.

This variant removed dark patterns, offered information via a chat-bot, and displayed a page rating.

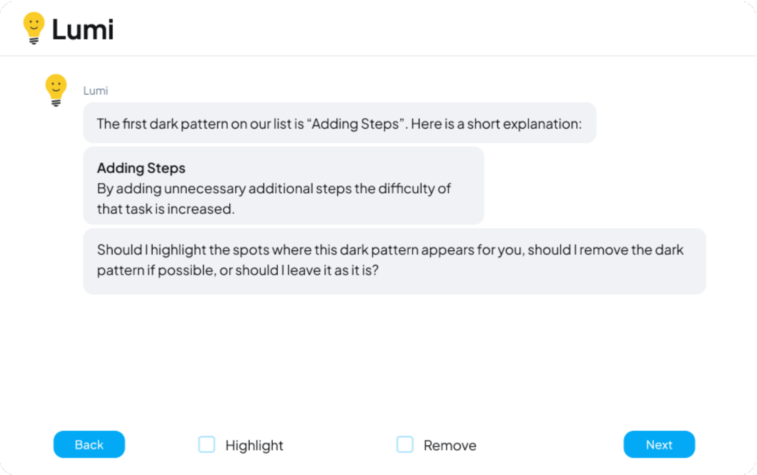
Remove + Counter + Chat (RCC): This variant combined the functionalities of the "Remove (R)" (4.2.3) and "Counter + Chat (CC)" (4.2.3) variants. All dark patterns were automatically removed from the website. Additionally, participants had the option to use the assistant's chat bot to inform themselves about the dark patterns that had been removed. This variant also provided a page rating (4.2.2), signaling the number of detected and removed dark patterns via the color (traffic light scheme) of the chat bot button.

This variant offered comprehensive countermeasures: automatic removal, page rating, information via chat-bot, and persistent highlighting of removed dark patterns.

Remove + Mark + Counter + Chat (RMCC): This variant built upon the "RCC" variant (4.2.3), offering a comprehensive set of countermeasures. All dark patterns were automatically removed, and the page rating was evaluated based on the number of detected and removed dark patterns and signaled this via the colored (traffic light scheme) chat bot button. Participants could also obtain information about the removed dark patterns through the chat bot. Furthermore, in this variant, the areas where the dark patterns were originally located were highlighted, even after their removal, providing a continuous visual cue. Screenshots of this variant are shown in Figure A.7 and Figure A.8.

This variant allowed users to configure the assistant's dark pattern countermeasures (highlight, remove, both, or neither) before shopping.

Options (O): This variant introduced a user-configurable experience. Before beginning the online shopping scenario, participants first went through a setup phase. During this setup, the assistant briefly introduced itself (see Figure A.12) and then, for each category of dark patterns expected in the study (e.g., "Adding Steps", "Bad Defaults",...), it asked the user how they would prefer to handle that specific type of dark pattern as can be seen in Figure 4.6. The available options were: highlighting the dark pattern's area, removing the dark pattern, doing both, or doing neither. After this personalized configuration, participants proceeded through the familiar online shopping scenario, with the assistant's behavior adjusted to their chosen settings. Instead of a chat bot button, this variant provided



The first dark pattern on our list is “Adding Steps”. Here is a short explanation:

Adding Steps
By adding unnecessary additional steps the difficulty of that task is increased.

Should I highlight the spots where this dark pattern appears for you, should I remove the dark pattern if possible, or should I leave it as it is?

Figure 4.6: Screenshot illustrating the setup phase interaction for the "Options" variant. This interface allows users to customize the assistant’s behavior by selecting preferred countermeasures for various dark pattern categories, embodying the user-control aspect of this variant.

access to a small menu where the previously made settings could be reviewed and changed at any time, which is shown in Figure A.14.

4.3 Study Procedure

The study had the following procedure:

We used a **within study** approach, meaning that all participants saw all assistant-variants. Each participant completed the study individually and only in the presence of the investigator. The order of the variants was randomized using a latin square, with the exception of the baseline, which was always the first variant.

After a brief "welcome," the topic of the study (Virtual Assistant as Countermeasure against Dark Patterns) was introduced to the participants. The definition of dark patterns was explained to each participant (regardless of prior

Participants were given the definition of dark patterns, which was illustrated using the example of cookie banners.

knowledge) and illustrated with a short example. As an example, each participant was given "cookie banners," as these are often structured in such a way that they contain several dark patterns intended to entice the user to accept all cookies (e.g., through "visual prominence" or "adding steps") as stated by Grassl et al. [2020]. Furthermore, we choose cookie banners because most people would already be familiar with them.

Participants completed eight online shopping scenarios to buy a wall clock, each with a different virtual assistant variant assisting against dark patterns.

The participants were then informed about the procedure of the study: that they would have to go through an online shopping scenario in which they were tasked with buying a specific wall clock, which had already been chosen for them, and they only needed to complete the purchase. They were also told that they would go through this scenario eight times, each time supported by a different variant of the virtual assistant designed to help them deal with the dark patterns.

Questionnaire about demographics.

Before the scenarios commenced, all participants filled out a questionnaire to collect demographic data such as age and gender. They also indicated their self-assessed experience with online shopping and dark patterns. The questionnaire is shown in Figure B.1.

7-point Likert scale for usability, clarity, safety, efficiency, helpfulness and general user experience.

At the beginning of each scenario, a brief description was provided for the specific variant of the virtual assistant included, along with an explanation of its rough functionality. After successful completion of each scenario, participants completed a questionnaire using a 7-point Likert scale (ranging from 1-7 or good to bad). This questionnaire assessed the usability, clarity, safety, and efficiency of the website, as well as the usability, clarity, safety, efficiency, and helpfulness of the assistant, and the overall user experience. These measurements were chosen to facilitate good comparability with the study performed by Schäfer et al. [2023]. The questionnaire can be seen in the appendix B. Once the quantitative questionnaire for each variant was completed, participants were then asked in a subsequent interview to articulate the strengths and weaknesses, or advantages and disadvantages, they perceived in that specific variant.

Qualitative Interview regarding strengths and weaknesses of the variants.

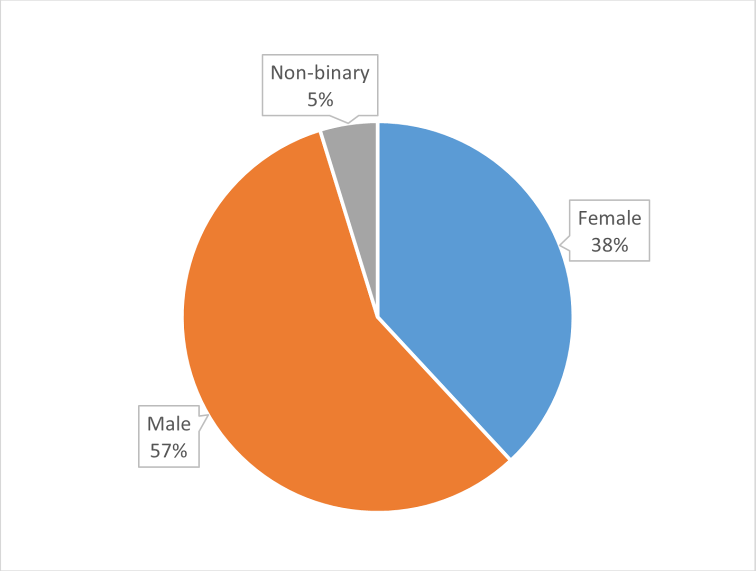


Figure 4.7: Gender distribution of participants within the user study, 38% (8) female, 57% (12) male, and 5% (1) non-binary.

Ranking of the variants.

After successfully completing all eight variants, the participants were asked to rank the variants, indicating which they found best, which worst, and their preferences for the variants in between.

4.3.1 Participants

The study included 21 diverse participants (8 female, 12 male, 1 non-binary) aged 21-72, with varied backgrounds and online shopping experience.

This subsection provides detailed information on the participants involved in the user study. A total of 21 participants participated in the study. The gender distribution was 8 female, 12 male, and 1 non-binary (see Figure 4.7). All participants were aged between 21 and 37, except for one outlier being 72 years old. The participants represented diverse educational backgrounds, including degrees such as Abitur, Ausbildung, Fachausbildung, Bachelor of Science, Master of Science, Doctor of Philosophy, State Examination, and Diploma, across various disciplines such as Sustainable Resources and Energy Supply, Biology, Electrical Engineering, Medicine, Education, Computer Science, Psy-

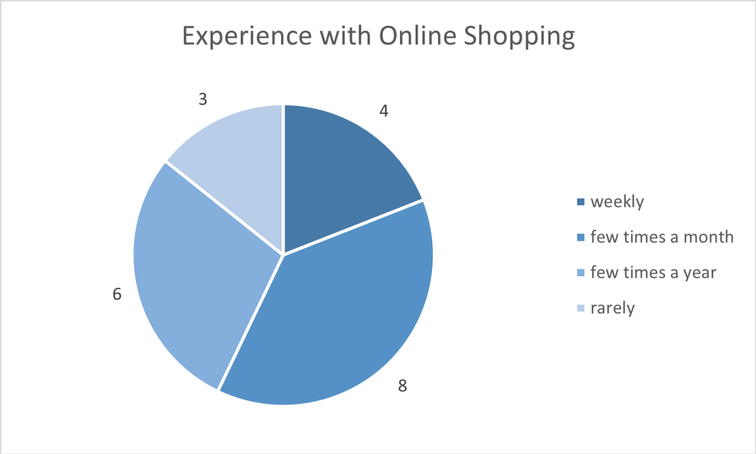


Figure 4.8: Distribution of participants' self-reported frequency of online shopping. The majority of participants engage in online shopping "few times a month" (8 participants) or "few times a year" (6 participants), with a smaller number shopping "weekly" (4 participants) or "rarely" (3 participants).

chology, Civil Engineering, General Business, and Business Economist. Their experience with online shopping varied from "rarely" to "weekly," as can be seen in Figure 4.8 and their familiarity with dark patterns ranged from "No prior experience" to "Heard of Dark Pattern before." which can be seen in Figure 4.9.

4.4 Results

This section will present the findings derived from the user study. After each scenario involving a variant of the assistant, participants were asked to rate the website and the assistant variant using a 7-point Likert scale. Additionally, a short interview was conducted to gather qualitative feedback on the strengths and weaknesses of that specific variant.

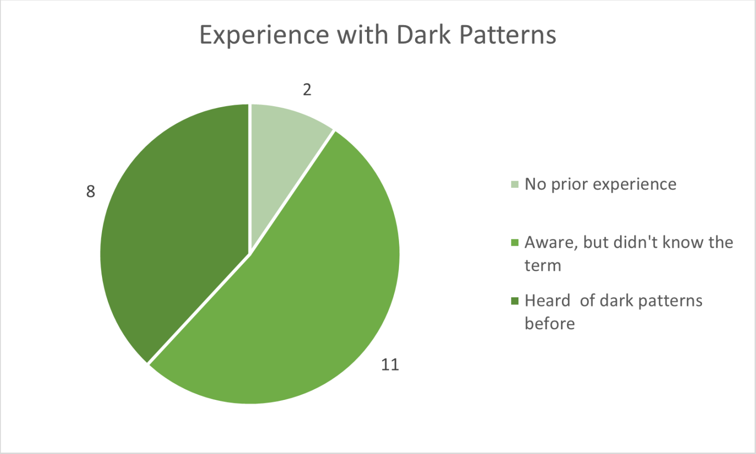


Figure 4.9: Distribution of participants' prior experience with dark patterns. The largest group (11 participants) was "Aware, but didn't know term," followed by those who "Heard of dark patterns before" (8 participants), and a smaller group with "No prior experience" (2 participants).

Usability (Website) The usability of the website, as rated by participants on a 7-point Likert scale (1: Easy to use, 7: Hard to use), showed varying perceptions across the different assistant variants (See Figure 4.10). The data suggests that variants which directly removed dark patterns (R, RCC) generally resulted in higher perceived usability (lower scores), with Variant R showing the lowest mean (1.33) and median (1), indicating it was perceived as the easiest to use. Variants that primarily informed (CC, MC) or had no intervention (UC) tended to have higher usability scores (closer to hard to use), with CC having the highest mean (3.85) and median (4). The "Animation" (AN) and "Options" (O) variants, along with "Remove + Mark + Counter + Chat" (RMCC), showed moderate usability scores.

Variants that removed dark patterns improved website usability, while informative or no-intervention variants rated lower.

Clarity (Website) The clarity of the website, rated on a 7-point Likert scale (1: Clear, 7: Confusing), also varied among the assistant variants (See Figure 4.11). Similar to usability, the "Remove" (R) variant showed the highest perceived clarity (lowest scores, mean = 1.57, median = 1), in-

Removing dark patterns improved website clarity; other interventions helped, but less so than direct removal.

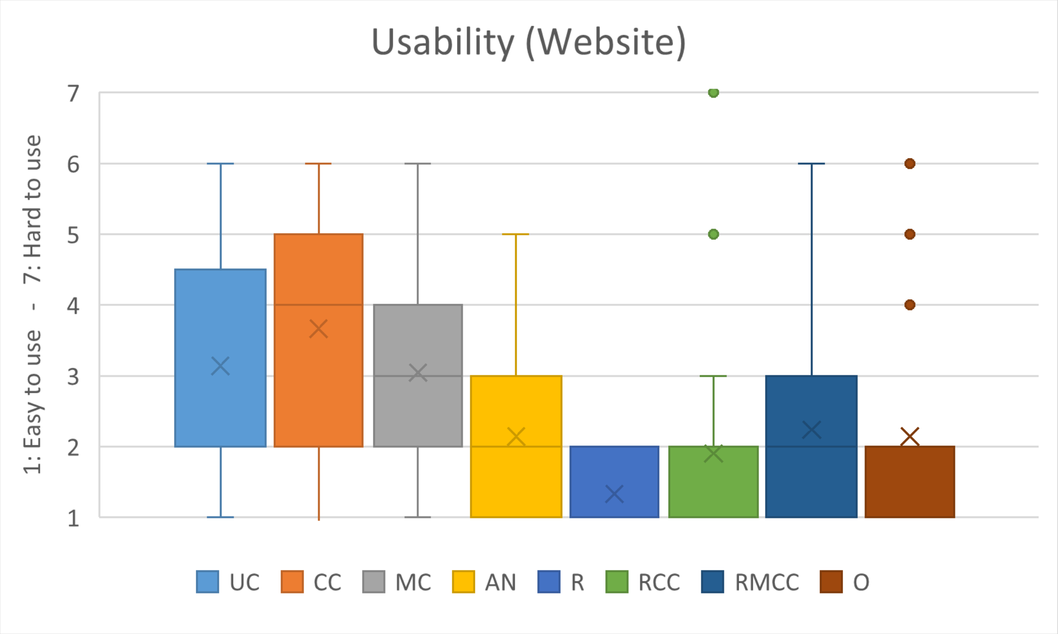


Figure 4.10: Website usability ratings across different assistant variants. Participants rated usability on a 7-point Likert scale (1: Easy to use, 7: Hard to use). Variants R and RCC show the lowest median and interquartile range, indicating higher perceived usability, while UC and CC generally rated lower.

dicating that participants found the website clearest when dark patterns were directly removed. Variants that did not remove dark patterns (UC, CC) or only marked them (MC) generally resulted in lower clarity scores (closer to confusing). The "Animation" (AN) and "Options" (O) variants, as well as "Remove + Mark + Counter + Chat" (RMCC), again fell in a more moderate range, suggesting that while removal is highly effective for clarity, other forms of intervention can also improve it compared to no intervention.

Directly removing dark patterns increased perceived website safety compared to informational or no-intervention variants.

Safety (Website) The safety of the website, rated on a 7-point Likert scale (1: Safe, 7: Dangerous), also exhibited differences across the assistant variants (See Figure 4.12). The "Remove" (R) variant again demonstrated the highest perceived safety (lowest score), suggesting that participants felt most secure when dark patterns were directly eliminated. Variants that provided only information or no intervention (UC, CC) generally resulted in lower safety per-

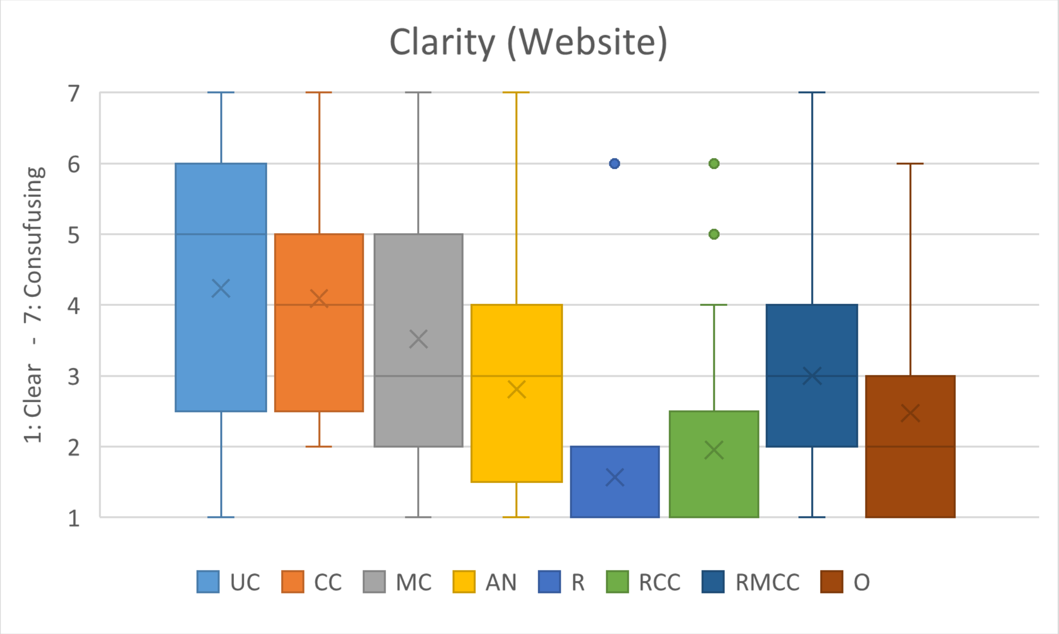


Figure 4.11: Website clarity ratings across different assistant variants. Participants rated clarity on a 7-point Likert scale (1: Clear, 7: Confusing). Variants R and RCC show the lowest median and interquartile range, indicating higher perceived clarity, while UC and CC generally rated lower.

ceptions (closer to dangerous), with CC having the highest mean (4.62) and median (5). The "Animation" (AN), "Options" (O), "Remove + Counter + Chat" (RCC), and "Remove + Mark + Counter + Chat" (RMCC) variants showed more moderate safety ratings, indicating that while direct removal is highly effective, other forms of assistance can also contribute to a sense of safety compared to no intervention.

Efficiency (Website) The efficiency of the website, rated on a 7-point Likert scale (1: Efficient, 7: Inefficient), also showed notable differences across the assistant variants (See Figure 4.13). Consistent with the trends observed in usability and clarity, the "Remove" (R) variant was perceived as the most efficient (lowest scores, mean = 1.62, median = 1), indicating that participants found the website to be most efficient when dark patterns were directly removed. Variants with no intervention (UC) or those pri-

Removing dark patterns improved website efficiency; other interventions helped, but less effectively than direct removal.

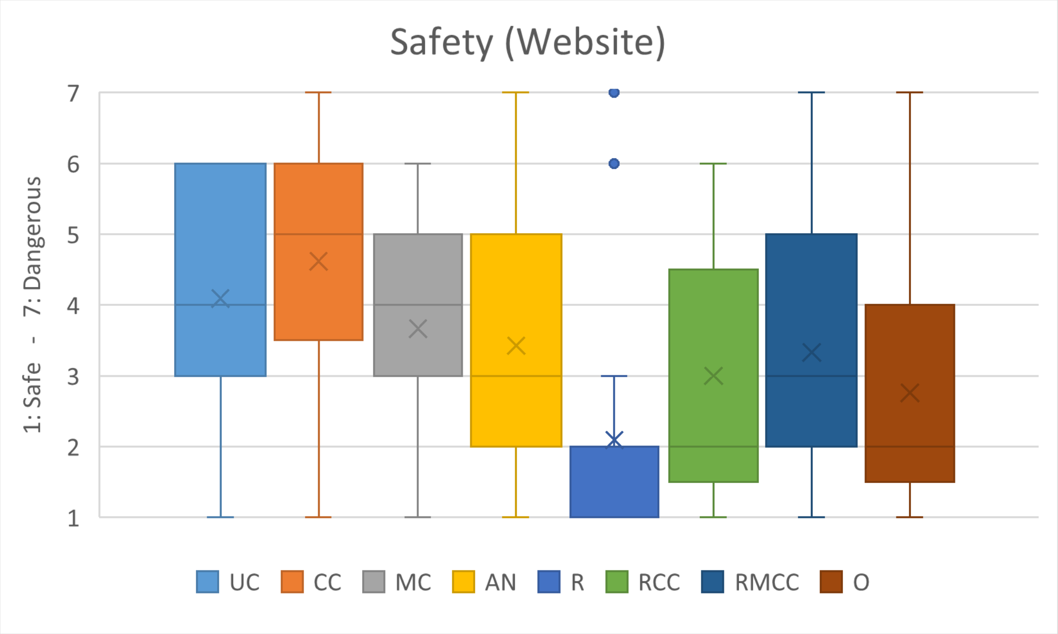


Figure 4.12: Website safety ratings across different assistant variants. Participants rated safety on a 7-point Likert scale (1: Safe, 7: Dangerous). Variant R shows the lowest median and interquartile range, indicating higher perceived website safety, while CC and UC generally rated lower.

marily providing information (CC, MC) generally resulted in lower efficiency perceptions (closer to inefficient), with CC having the highest mean (4.71) and median (5). The "Animation" (AN), "Options" (O), "Remove + Counter + Chat" (RCC), and "Remove + Mark + Counter + Chat" (RMCC) variants showed more moderate efficiency ratings, suggesting that while direct removal is highly effective, other forms of assistance can also contribute to a sense of efficiency compared to no intervention.

The "Remove" assistant variant was most usable, as direct intervention was preferred over informative, interactive options.

Usability (Assistant) The usability of the assistant, rated on a 7-point Likert scale (1: Easy to use, 7: Hard to use), was also assessed across the variants that included an assistant (See Figure 4.14). The "Remove" (R) variant of the assistant was perceived as the most usable (lowest mean = 1.38 and median = 1 scores), indicating that participants found it the easiest to interact with, because there was no possibility to interact with the assistant. Variants that involved

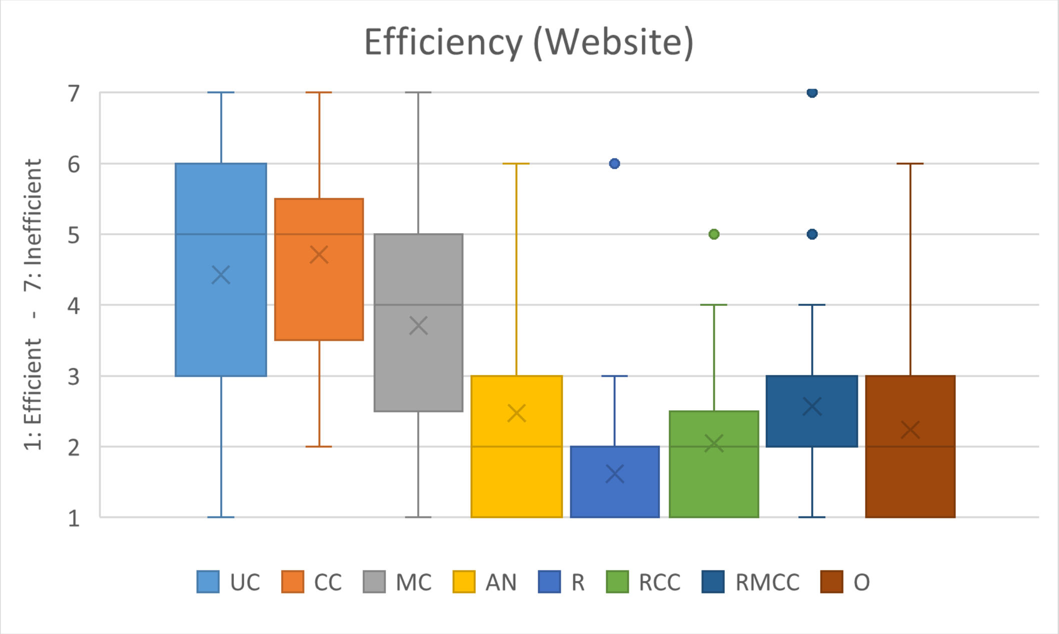


Figure 4.13: Website efficiency ratings across different assistant variants. Participants rated efficiency on a 7-point Likert scale (1: Efficient, 7: Inefficient). Variant R shows the lowest median and interquartile range, indicating higher perceived website efficiency, while CC and UC generally rated lower.

direct removal (R, RCC, RMCC, AN) or user options (O) generally scored lower (more usable) than those primarily focused on information provision (CC, MC). The "Counter + Chat" (CC) variant had the highest mean (3.2) and median (3), suggesting it was perceived as comparatively less usable among the assistant variants. This indicates a preference for direct, less intrusive interventions from the assistant and a inconvenient implementation of the chat bot.

Clarity (Assistant) The clarity of the assistant, rated on a 7-point Likert scale (1: Clear, 7: Confusing), was evaluated for the variants that included an assistant (See Figure 4.15). The "Options" (O) variant demonstrated the highest perceived clarity for the assistant (lowest mean (1.86) and median (2) scores), suggesting that participants found this assistant variant to be the clearest. Variants that provided information (CC, MC, RCC, RMCC) also generally scored well in terms of clarity. The "Animation" (AN) variant, de-

The “Options” assistant variant was the clearest; informative variants also scored well, but “Animation” and “Remove” were less transparent.

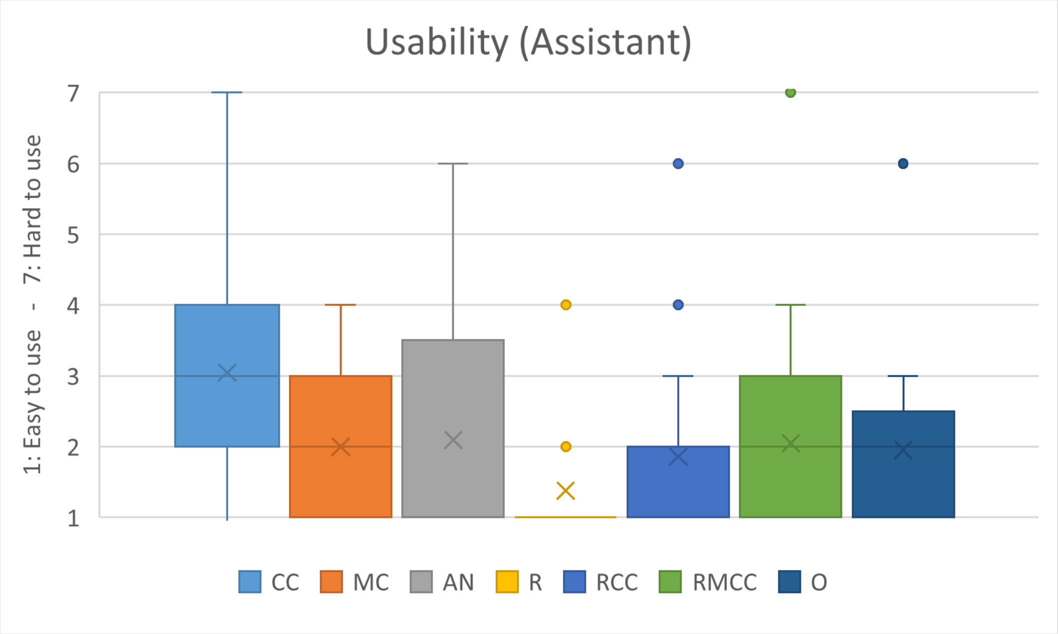


Figure 4.14: Assistant usability ratings across different assistant variants. Participants rated usability on a 7-point Likert scale (1: Easy to use, 7: Hard to use). Variant R shows the lowest median and interquartile range, indicating higher perceived usability, while AN and CC generally rated lower.

spite its direct intervention, had a higher mean (3.75) and median (4), indicating it was perceived as less clear in its actions, likely due to the passive nature of the animation itself in conveying information, since this variant provided no further information regarding what exactly or why parts of the website were changed. The "Remove" (R) variant also showed a relatively higher mean (3), suggesting that while it made the website clearer, the assistant's actions themselves might not have been as explicitly clear to the user without additional informational cues.

The "Options" assistant variant was perceived safest, followed by combined removal and informational variants, while "Animation" felt less secure.

Safety (Assistant) The safety of the assistant, rated on a 7-point Likert scale (1: Safe, 7: Dangerous), was assessed for the variants that included an assistant (See Figure 4.16). The "Options" (O) variant was perceived as the safest assistant (lowest mean = 1.29 and median = 1 scores), indicating that participants felt most secure with this configurable variant. Variants that involved direct removal and informa-

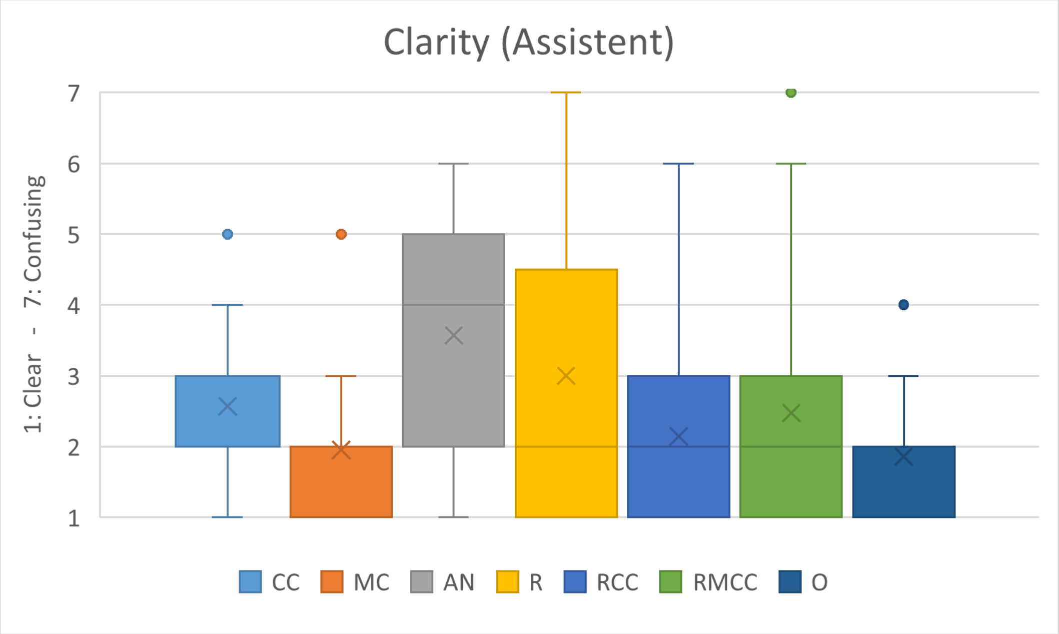


Figure 4.15: Assistant clarity ratings across different assistant variants. Participants rated clarity on a 7-point Likert scale (1: Clear, 7: Confusing). Variant O shows the lowest median and interquartile range, indicating higher perceived clarity, while AN generally rated lower.

tion (RCC, RMCC) also scored very well in terms of safety. The "Animation" (AN) variant had the highest mean (3.33), suggesting it was perceived as less safe, potentially due to the unexpected nature of the animations or a lack of explicit control over the intervention. The "Remove" (R) variant, despite making the website safer, had a high mean (2.95) for assistant safety, possibly because the assistant's actions were less transparent without additional cues.

Efficiency (Assistant) The efficiency of the assistant, rated on a 7-point Likert scale (1: Efficient, 7: Inefficient), was assessed for the variants that included an assistant (See Figure 4.17). The "Remove" (R) variant of the assistant was perceived as the most efficient (lowest mean = 1.39 and median = 1 scores), indicating that participants found it to be the most efficient in its operation. The "Animation" (AN) variant also scored very highly in terms of efficiency. The other Variants that involved direct removal (R, AN, RCC,

Directly removing dark patterns was most efficient, while chat-based information was least efficient.

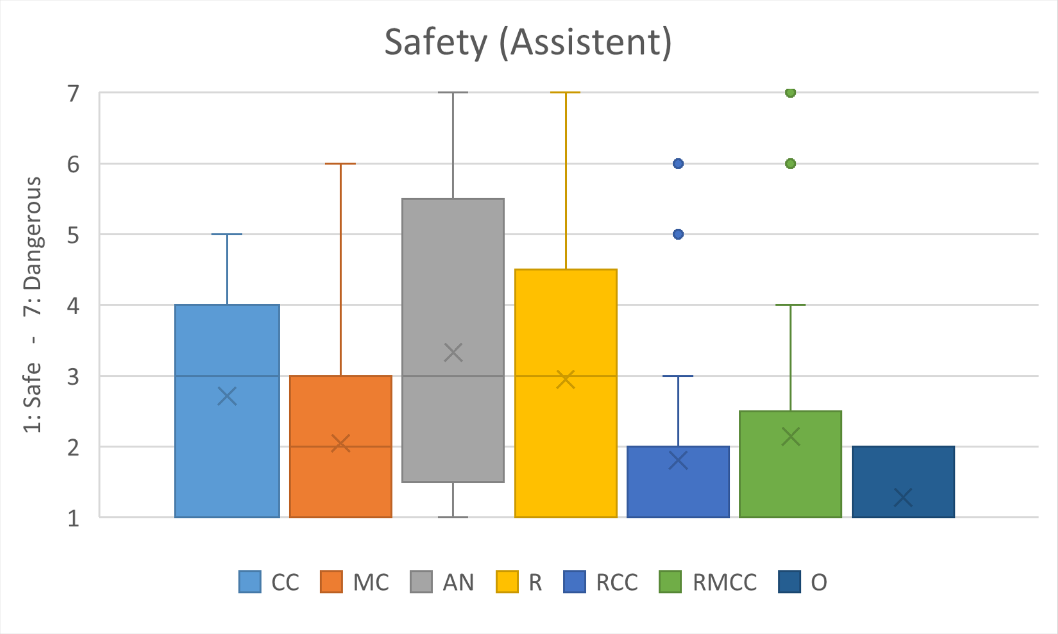


Figure 4.16: Assistant safety ratings across different assistant variants. Participants rated safety on a 7-point Likert scale (1: Safe, 7: Dangerous). Variant O shows the lowest median and interquartile range, indicating higher perceived safety, while AN and MC generally rated lower.

RMCC) or user options (O) generally showed higher perceived efficiency than those primarily focused on information provision (CC, MC). The "Counter + Chat" (CC) variant had the highest mean (4) and median (4), suggesting it was perceived as the least efficient among the assistant variants. This further reinforces the preference for direct and streamlined interventions from the assistant or atleast an more efficient way to provide information.

The "Options" assistant variant was most helpful, while the "Counter + Chat" variant were perceived as least helpful.

Helpfulness (Assistant) The helpfulness of the assistant, rated on a 7-point Likert scale (1: Helpful, 7: Unhelpful), was assessed for the variants that included an assistant (See Figure 4.18). The "Options" (O) variant was perceived as the most helpful assistant (lowest mean = 1.38 and median = 1 scores), indicating that participants found this configurable variant to be the most beneficial. Variants that involved direct removal and information (RCC, RMCC) also scored very well in terms of helpfulness. The "Counter +

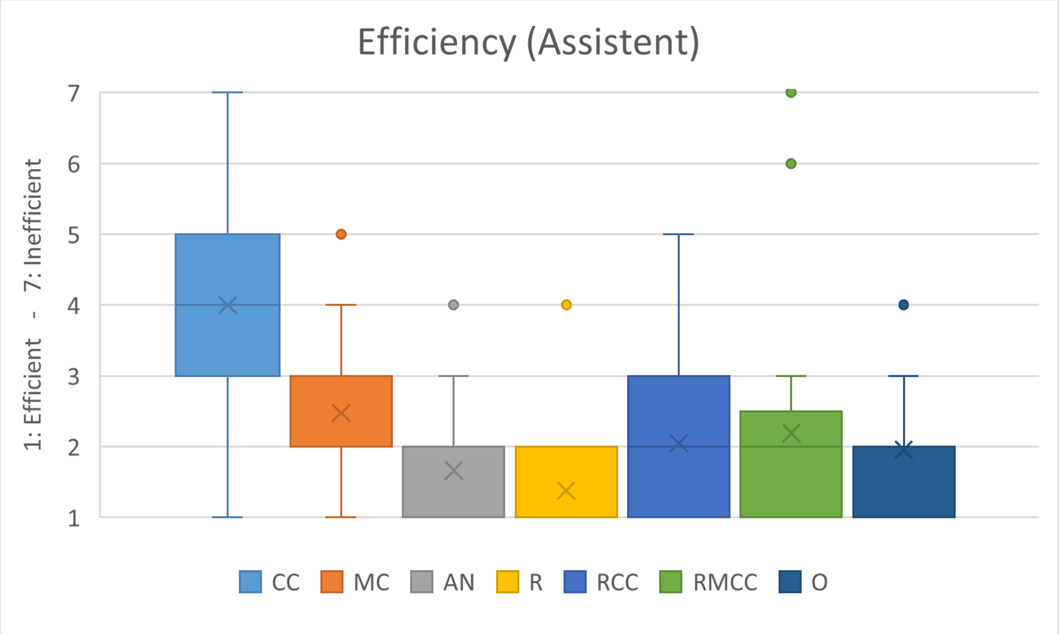


Figure 4.17: Assistant efficiency ratings across different assistant variants. Participants rated efficiency on a 7-point Likert scale (1: Efficient, 7: Inefficient). Variant R shows the lowest median and interquartile range, indicating higher perceived efficiency, while CC generally rated lower.

Chat" (CC) variant had the highest mean (3.76) and median (4), suggesting it was perceived as the least helpful among the assistant variants, likely due to its more passive, information-only approach. This again highlights a preference for more proactive and impactful interventions from the assistant.

General User Experience The general user experience, rated on a 7-point Likert scale (1: Good, 7: Bad), was also assessed across all variants (See Figure 4.19). The "Options" (O) variant, along with "Remove + Counter + Chat" (RCC), generally resulted in the best perceived general user experience (lowest scores), indicating that participants found these variants to provide the most positive overall experience. The "Unchanged" (UC) variant, with no assistant, had the highest mean (5) and median (5), indicating the least favorable general user experience and that countermeasures are generally desired. Variants that involved direct removal

Direct dark pattern countermeasures (like "Options" and "Remove + Counter + Chat") improved the general user experience.

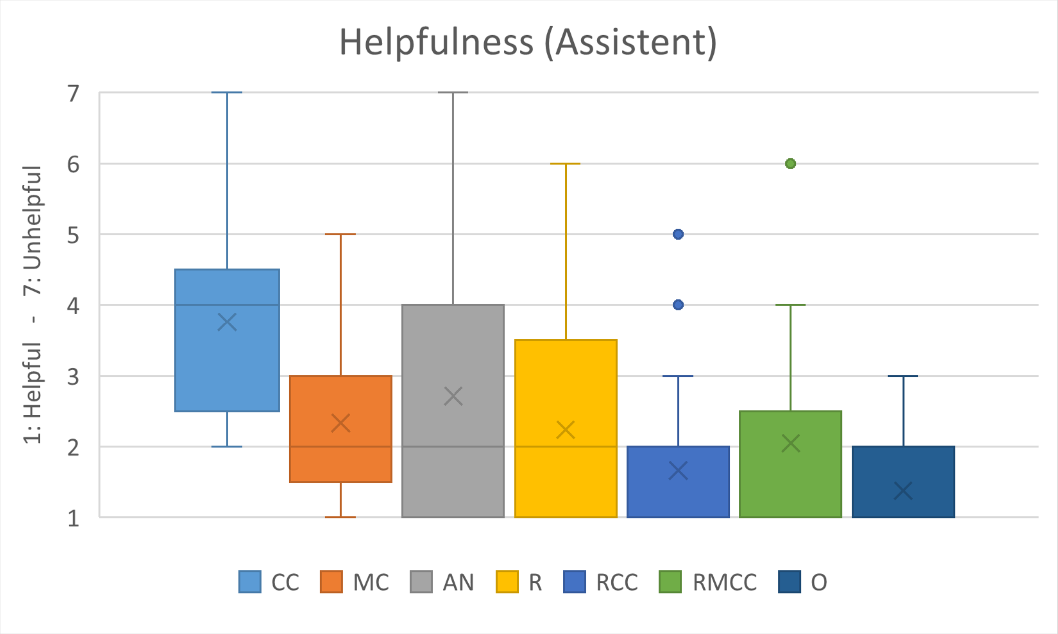


Figure 4.18: Assistant helpfulness ratings across different assistant variants. Participants rated helpfulness on a 7-point Likert scale (1: Helpful, 7: Unhelpful). Variant O shows the lowest median and interquartile range, indicating higher perceived helpfulness, while CC generally rated lower.

(R, RCC, RMCC, AN, O) consistently showed better user experience ratings than those that primarily informed (CC, MC) or had no intervention (UC), reinforcing the positive impact of proactive dark pattern countermeasures on the overall user experience.

Directly removing dark patterns (R, AN) sped up the scenario completion; informative variants (CC, MC) and the "Options" setup took longer.

Time The time (in seconds) needed to complete the scenario was recorded for each variant (See Figure 4.20). The data indicates that variants which directly removed dark patterns, particularly "Remove" (R) and "Animation" (AN), reduced the time needed to complete the scenario, with R showing the lowest mean (36 seconds) and median (26 seconds) times. Variants that primarily informed (CC, MC) or had no intervention (UC) generally resulted in longer completion times, because the first visit of the website was often a bit more careful and the participants read the details and were therefore slower. CC and MC had the chat bot providing information but did not removed any Dark Pat-

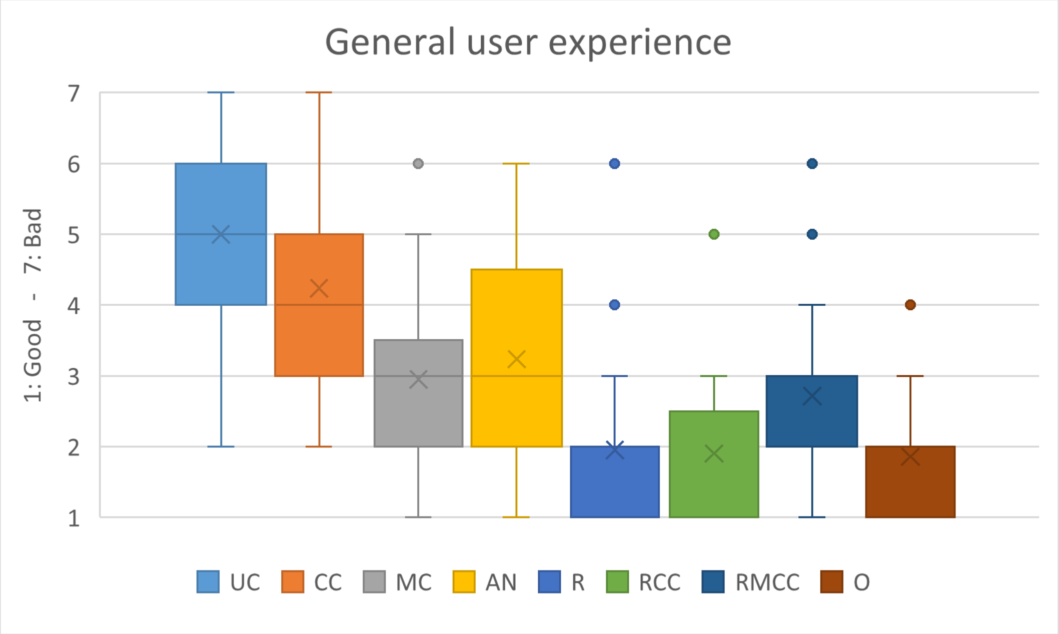


Figure 4.19: General user experience ratings across different assistant variants. Participants rated their experience on a 7-point Likert scale (1: Good, 7: Bad). Variants R and O show the lowest median and interquartile range, indicating a better overall user experience, while UC generally rated lower.

terns resulting in reading time and dealing with the Dark Patterns. It is noteworthy that the "Options" (O) variant, while providing a good user experience, had a substantially longer mean and median time when including its initial setup phase, highlighting the trade-off between customization and initial time investment. When excluding the setup, the "Options" variant's completion time was comparable to other informative or less intrusive variants.

4.4.1 Ranking

The ranking of the variants reflects participants' overall preference, where 1st place indicates the most liked variant and 8th place indicates the least favorite. (See Figure 4.21 and 4.22) The "Options" (O) variant was ranked as the most preferred, with the lowest mean (2.14) and median (2) rank, indicating participants highly valued the ability to

Users preferred customizable or direct dark pattern removal, with "Unchanged" (baseline) being the least favorite.

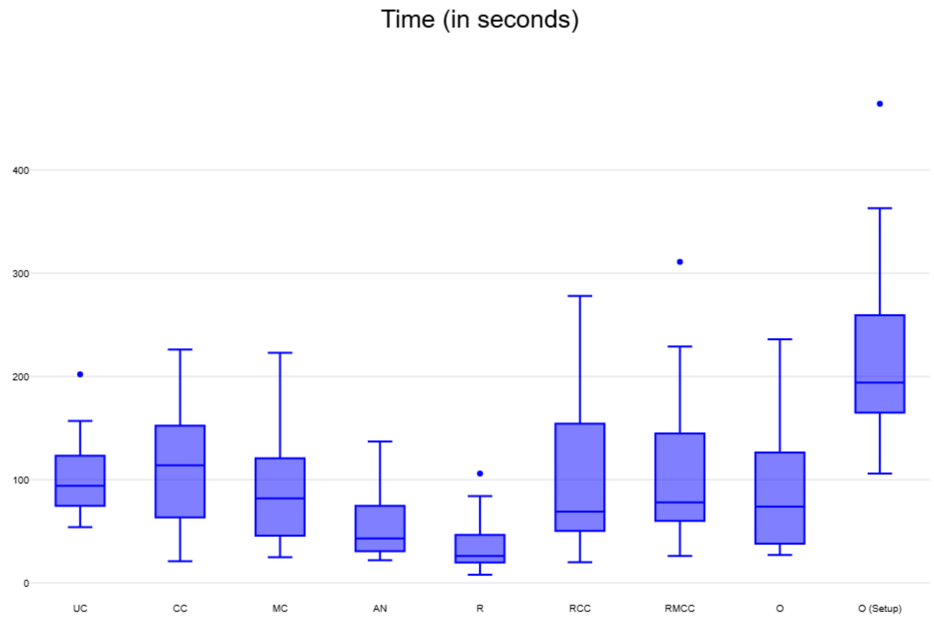


Figure 4.20: Time taken to complete the scenario (in seconds) across different assistant variants. Variant R shows the shortest completion times, while variants, that inform the user about Dark Patterns (CC, MC, RCC, RMCC) tend to show a longer time. It is also shown how long the participants needed for the setup of the Options variant.

customize the assistant's behavior. The "Remove + Counter + Chat" (RCC) and "Remove + Mark + Counter + Chat" (RMCC) variants also performed very well, suggesting a strong preference for proactive removal combined with informational support. Conversely, the "Unchanged" (UC) variant was consistently ranked as the least favorite, highlighting the negative impact of unmitigated dark patterns. Variants that primarily focused on information (CC, MC) or animation (AN) fell in the middle range, indicating that while they were better than no intervention, they were not as preferred as those offering direct removal or customization.

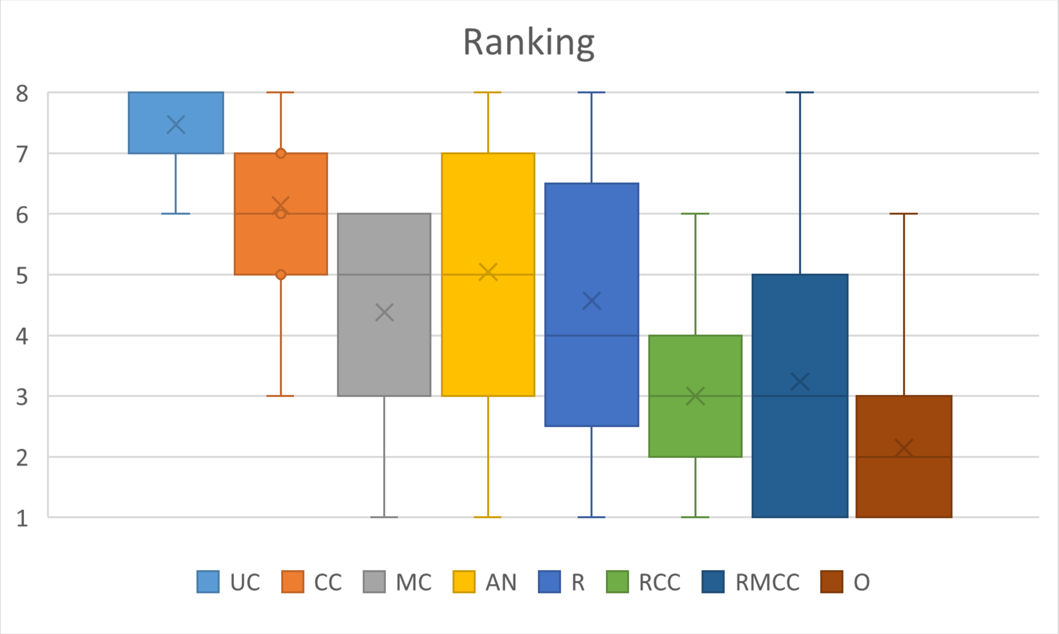


Figure 4.21: Overall ranking of assistant variants based on participant preference. A lower rank indicates a higher preference (1st place = most preferred, 8th place = least preferred). Variant O consistently shows the lowest ranks, indicating it was the most preferred, while UC was the least preferred.

4.4.2 Qualitative Comments

This subsection will present the most important and interesting qualitative comments provided by the participants during the post-scenario interviews. These comments will be categorized to highlight common themes and provide deeper insights into user perceptions and experiences. The numbers in brackets indicate how often a comment appeared in a similar form. For direct user quotes, we indicate the randomized participant ID in brackets with a leading P: *"This is a sample comment"* (P6). Participant quotes were translated if necessary and corrected for grammar.

Unchanged (UC): The rating for this variant can be seen in Figure 4.23. Some participants had a positive impression of the website because there was "*no advertising*" (P2). Other participants were very negative about the mere existence of dark patterns, so that they would probably have

Some participants liked the website for its lack of ads, while others disliked it because of the presence of dark patterns.

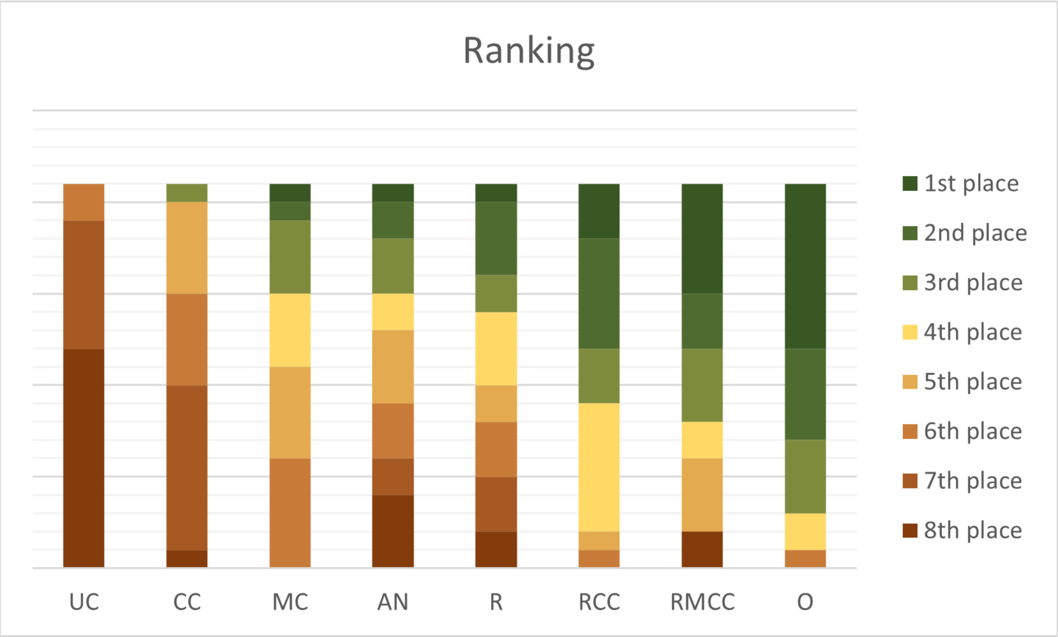


Figure 4.22: Distribution of participant rankings for each assistant variant. The chart illustrates the frequency with which each variant was ranked from 1st (most preferred) to 8th (least preferred)

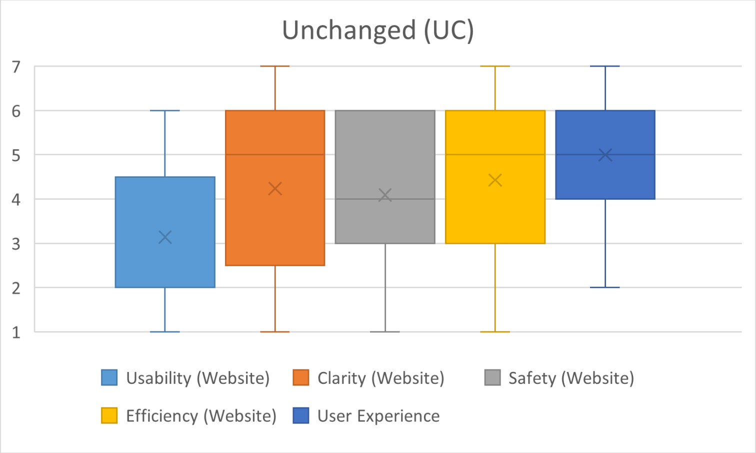


Figure 4.23: Participant ratings for the "Unchanged (UC)" variant across various website experience metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant, lacking any assistant intervention as it serves as the baseline, generally shows higher (less favorable) median scores across all measured aspects.

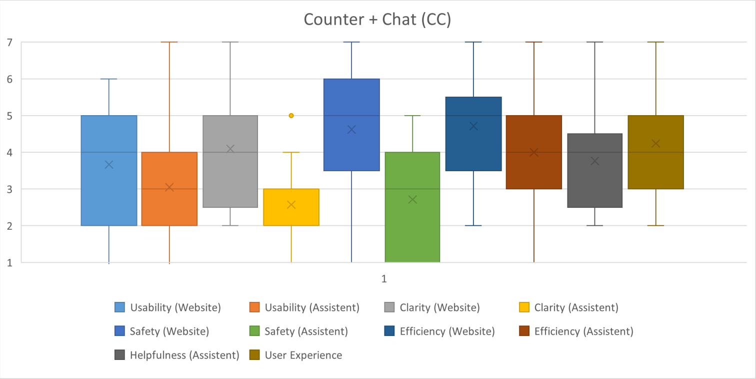


Figure 4.24: Participant ratings for the "Counter + Chat (CC)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant, which primarily provides information via chat, shows more varied and generally less favorable ratings compared to variants with direct removal.

left the website under normal circumstances. Participants also commented, particularly with regard to the security of the website, that they would normally check the website more closely, for example with the help of Trustpilot or similar, but this was not possible with the prototype.

Counter + Chat (CC): The rating for this variant can be seen in Figure 4.24. Many participants liked the fact that the website **is not changed by the assistant** (7). Many participants also liked the possibility to **inform themselves about dark patterns** (6). With the comment "*Traffic light is quick*" (P21), one participant commented positively that the color indicator of the number of dark patterns on the website is a quick way to get an impression of the website regarding the use of dark patterns. A few participants (4) commented that the button for the chat bot and its positioning were initially thought to be a chat bot of the website instead of the “Dark Pattern Countermeasure”-assistant. Some participants were bothered by the amount of reading that this list of dark patterns in the form of chat messages entailed and described the chat messages as a "*wall of text*" (P15). In the CC variant in particular, participants

This variant was liked for not changing the website and offering information about dark patterns, but criticized for being time-consuming and having a "wall of text."

(6) had problems identifying **where** the dark patterns were located. The majority of participants (13) commented that the assistant's chat bot, in the form in which it appeared in the prototype, is a **time-consuming** type of countermeasure, for various reasons such as clicking on the chat bot, the extra request to find out more about dark patterns and reading through the chat messages. However, the fact that this variant does not change the website was not always noticed positively; many participants (9) also commented negatively that this variant **does nothing** and that the user has to take care of the dark patterns themselves. A few participants (4) found the chat bot button *"too discreet"* (P16) and said that it *"can be overlooked"* (P9). One issue that was also raised frequently was the **order and number of dark patterns displayed in the chat**. The assistant did not adapt enough to the current situation, as the order was not comprehensible and the assistant also displayed dark patterns that were either no longer relevant because they had already been dealt with by the participants or were not yet relevant because the participants had not (yet) encountered them. The chat bot also had *"no memory"* (P5) and therefore did not respond to the previous interaction with the user. Furthermore, two participants also noticed negatively that the *"chat is in the way"* (P15), i.e. parts of the website are blocked and the chat-window cannot be moved.

This variant was liked because the markings were helpful and served as a "To-Do-List".

Mark + Chat (MC): The rating for this variant can be seen in Figure 4.25. The markings were generally perceived positively (8). For example, one participant found the **highlighting helpful** *"because it catches the eye"* (P9), i.e. it is quick and easy to see. We noticed that many participants saw both the chat and the markings as a kind of **to-do list** that had to be completed. One participant confirmed this by saying that they could "*go through the 'checklist'*" (P6) to see if they had found everything. At the same time, participants also complained that *"there is no confirmation that you have dealt with the dark patterns 'correctly'"* (P3) and that the markings are still there after the interaction. A few participants (4) noted that the markings are **too many** or could be too many on websites with even more dark patterns indicating the problem of "visual clutter" which was already mentioned by Schäfer et al. [2023].

There was criticism that the markings might be too many.

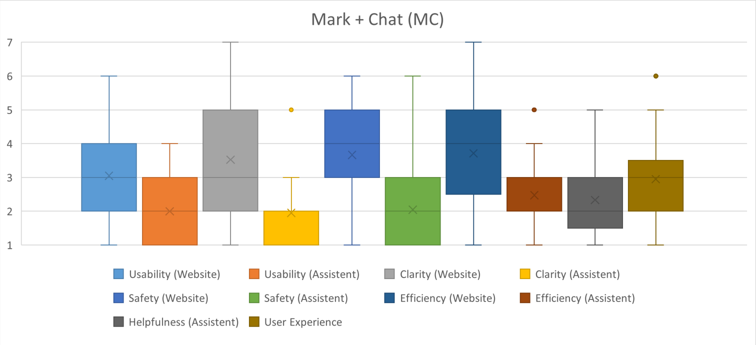


Figure 4.25: Participant ratings for the "Mark + Chat (MC)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant, combining marking with chat information, generally shows moderate to positive ratings, better than "Unchanged" but not as high as direct removal variants.

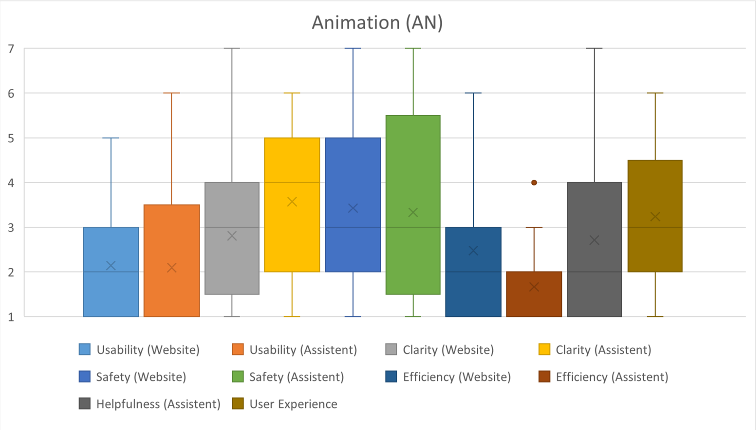


Figure 4.26: Participant ratings for the "Animation (AN)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant shows mixed feedback, with relatively positive website usability but lower ratings for assistant clarity and safety.

Animation (AN): The rating for this variant can be seen in Figure 4.26. Animation has generated very **mixed feedback**. Some of the participants (6) were positive about the

This variant received mixed feedback, praised for active intervention, but criticized for speed, lack of undo, and "loss of control."

fact that the assistant "*does something*" (P20), i.e. **actively intervenes and removes dark patterns**. "*You didn't have to do anything*" (P11). It was also positively noted (7) that the animations indicate that **something is being done** and **where**. The fact that the animation and removal of the dark pattern took place relatively quickly was also perceived positively by several participants (6). On the other hand, the speed of the animation was a point of criticism for other participants (10). "***Too fast***" (P19). Because the dark patterns were removed in this variant, many participants (10) asked "***What if batteries are wanted after all?***"(P14) with regard to the dark pattern "Sneak Into Basket". Of course this question came up for the variants R, RCC and RMCC as well. Something that was also mentioned for variants R, RCC and RMCC was the "***lack of*** `Undo`" (P10), i.e. the possibility of undoing the changes. However, what was particularly criticized with this variant was the feeling of "***loss of control***" (8) that was triggered by the animation. Because suddenly something was visibly changed on the website without the user having initiated it. In addition, the animations caused *"**too much to happen at the same time**"* (P16), which was mentioned by several participants (7) and sometimes (3) caused confusion. The participants (8) **lacked additional information**, so they asked themselves *"What did the assistant do?"* (P5). Some participants did not like the animation at all and described it as *"time-consuming, disruptive and destroying the flow"* (P8).

This variant was praised for simplicity and autonomy, but criticized for its lack of transparency and user control.

Remove (R): The rating for this variant can be seen in Figure 4.27. The variant was perceived by the participants as "***simple***" (P16) and "***autonomous***" (P11). A few participants were very positive about this variant. "*The website is the way I would like it to be*" (P18). "*I think it's great*" (P17). "*Very pleasant*" (P19). However, this variant was of course also criticized, above all because of the "***lack of transparency***" (6), so that questions arose such as "*I don't know what [the assistant] has done*" (P21). The **lack of feedback** (3) from the assistant and the **lack of control options** (3) were also criticized. "*Does the assistant work as it should?*"(P4).

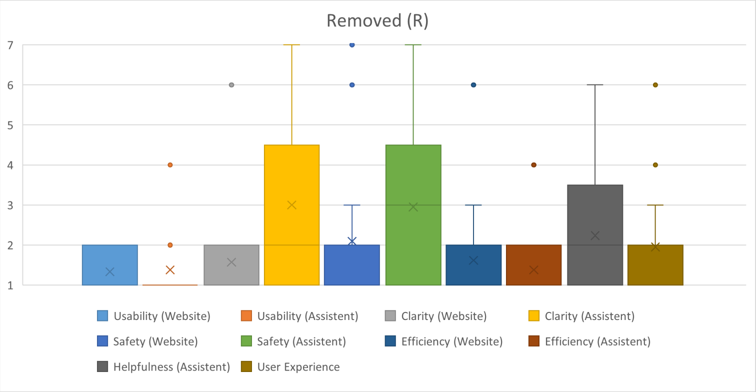


Figure 4.27: Participant ratings for the "Remove (R)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant, focusing on direct removal, shows very high (favorable) ratings for website-related metrics but slightly more variability for assistant-specific attributes.

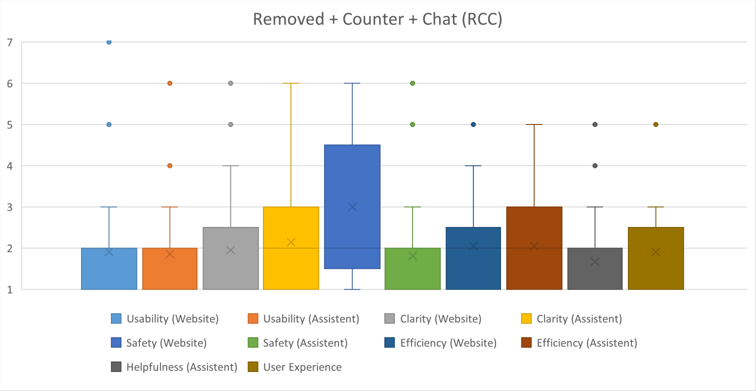


Figure 4.28: Participant ratings for the "Remove + Counter + Chat (RCC)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant demonstrates consistently high (favorable) ratings across most metrics, indicating strong positive user perception.

Remove + Counter + Chat (RCC): The rating for this variant can be seen in Figure 4.28. With this variant, just as with the CC and MC variants in some cases, it was praised that there is the optional possibility of **obtaining informa-**

This variant was praised for the optional dark pattern information and transparency, but criticized for the chat message quantity and potential for interrupting the flow.

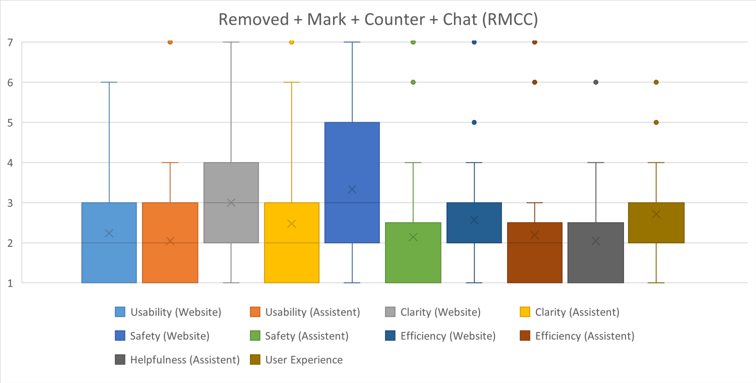


Figure 4.29: Participant ratings for the "Remove + Mark + Counter + Chat (RMCC)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant shows generally favorable ratings, particularly for website usability and safety, but with some variability in other aspects.

tion about dark patterns with the help of the chat bot (6). It was also praised that this variant offers *"**more transparency**"* (P16) and *"**more feedback**"* (P15), especially in comparison to AN and R. As before, however, the **order of the dark patterns in the chat** was criticized and that there are **too many** chat messages, which means that *"you overlook things"* (P3). There was also criticism (4) that all variants that change the website **interfere with the website**. One participant found the color *"indicator unclear"* (P15). In addition, several participants found it *"too prominent"* (P13), *"conspicuous and annoying"* (P8). *"You are tempted to click on the chat bot button, which in turn **destroys the flow**"* (P13).

This variant improved the transparency even further but especially the empty markings often caused confusion and distraction.

Remove + Mark + Counter + Chat (RMCC): The rating for this variant can be seen in Figure 4.29. "*The 'where' is clearer*"(P9). Something that many participants praised was that with this variant it is more obvious **where** the assistant has changed something. In addition, many participants (12) enjoyed the fact that the variants that remove dark patterns (AN, R, RCC and RMCC) act **independently**

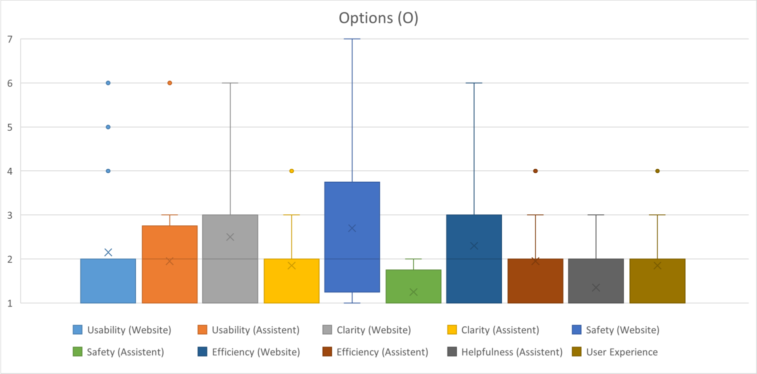


Figure 4.30: Participant ratings for the "Options (O)" variant across various website and assistant metrics. Ratings are on a 7-point Likert scale (1: positive extreme, 7: negative extreme). This variant, offering user customization, consistently receives highly favorable ratings across all measured aspects, indicating strong user preference for control.

and **automatically**. In this variant, however, the markings often (5) **caused confusion**, especially (11) when the **markings were empty**, i.e. when they indicated that something had been removed at this position. This variant also particularly caused the *"flow to be interrupted"* (5). Among other things, because the markings can *"draw attention to themselves"* (P15) and *"**distract**"* (P17).

Options (O): The rating for this variant can be seen in Figure 4.30. A large majority (15) of participants expressed a positive opinion about the individual **setting options**. "*[The assistant did] exactly what I wanted*" (P16). The markings were "*good for testing*" (P7). The option to **change the settings later** was also perceived positively by a few participants (2). Most criticism was voiced regarding the setup. The later options were also used rather sparingly. Many participants (12) found the **setup time-consuming**. Concerns were also expressed that the "*own configurations could be unsafe*" (P5), which is why several participants (3) wished for **default options** or similar. The **order of the dark patterns** was also criticized in a similar way to the order of the messages in the chat of the variants with a chat.

This variant was praised for user customization, but criticized for its time-consuming setup and lack of default settings.

Suggestions from the participants

Participants suggested numerical dark pattern indicators, movable chat, interactive checklists, an `Undo` function, other form of information, default settings and notifications for removed patterns.

- Instead of the color indicator, one participant preferred a precise number that shows the user how many dark patterns were found on the website.

- The window of the chat bot should be movable so that it can be moved around so that certain parts of the website and the chat can be seen at the same time.

- Some participants would have preferred an "interactive checklist" instead of the chat with the assistant. In other words, something that shows the user which dark patterns the user still has to deal with and then also reacts accordingly if the user has dealt with these dark patterns "correctly".

- The chat bot should have a "memory" and react to and build on the interaction with the user.

- Users specifically requested notifications when dark patterns were blocked, especially for the AN and R variants. This included suggestions for hints directly at the respective positions after an animation, or a pop-up providing detailed information.

- One user also wanted a switch for the entire website, i.e. similar to the R variant but with an `Undo` function, which was generally missing in most variants.

- A few users suggested alternative forms for the chat messages, such as a list or table.

- Other users suggested minor improvements to the chat, such as larger gaps between the dark patterns in the chat.

- A button that shows the user where the dark patterns are on the website instead of permanently marking the dark patterns was also suggested.

- Two users also asked for categories or "ratings" of the dark patterns, in the sense of: how "bad" are the dark patterns for the user? Does overlooking the dark pattern cost the user money or is data collected from the user?

- Many participants (6) also wanted additional information on the dark patterns via a “hover” window.
- Several participants (5) would like to see examples of the dark patterns in the setup of the option variant.
- Suggestions for improving the setup also came in the form of a “Reject All” button or “Default Options”.
- A progress bar for the setup was also missing.
- There was also no chat bot in the options variant, but this was then requested by several (4) participants.
- Another suggestion for the options variant was that the menu for changing the settings later should be better hidden, as the user probably needs this less often.

Chapter 5

Evaluation

In this chapter, we discuss our results and limitations.

5.1 Discussion

This section evaluates the performance of the different virtual assistant variants based on the quantitative results and qualitative comments collected during the user study. The aim is to understand the effectiveness of various intervention strategies in mitigating dark patterns and improving the user experience.

Users highly preferred the "Options" variant for customizable control and the "Removed + Counter + Chat" variant for its effective removal and support.

Overall Performance of Variants: The "Options (O)" variant (4.2.3) emerged as the most preferred, demonstrating the lowest mean and median rank (2.14 and 2 respectively, see Figure 4.21) and consistently high scores across all positive metrics (usability, clarity, safety, efficiency, helpfulness of the assistant, and general user experience). See Figure 4.30. This indicates a strong user preference for personalized control over the assistant's behavior. The "Removed + Counter + Chat (RCC)" variant also performed exceptionally well, ranking second overall (mean rank 3.00, median 3, see 4.21) and showing very positive ratings for website usability, clarity, and efficiency, as well as assistant helpfulness and safety (See Figure 4.28.). This highlights the effectiveness of combining direct dark pattern removal with optional informational support.

Unmitigated dark patterns (UC) negatively impacted user experience, while passive information (CC) was largely ineffective.

Conversely, the "Unchanged (UC)" variant (4.2.3), the baseline, with no assistant, consistently received the lowest ratings for general user experience (mean 5.00, median 5, see Figure 4.19) and was ranked as the least favorite (mean rank 7.48, median 8, see Figure 4.21). This clearly demonstrates the negative impact of unmitigated dark patterns on user perception. The "Counter + Chat (CC)" variant (4.2.3), which only provided information via a chat bot without altering the website, also performed poorly in terms of website efficiency (mean 4.71, see Figure 4.13) and assistant helpfulness (Figure 4.18), and was generally ranked low (mean rank 6.14, median 6, see Figure 4.21). This suggests that passive information alone is insufficient to effectively counter dark patterns and improve user experience.

Key Findings by Metric:

Removing dark patterns improved efficiency and sped up task completion. User customization (Options) was preferred, despite a longer setup.

Efficiency and Time to Completion: Variants that directly removed dark patterns, particularly "Removed (R)" (4.2.3) and "Animation (AN)" (4.2.3), drastically reduced the time needed to complete the scenario (R: mean 36s, median 26s; AN: mean 57s, median 43s, see Figure 4.20). This aligns directly with their high efficiency ratings for the website which can be seen in Figure 4.13. The "Options (O)" variant(4.2.3), while highly rated, had a significantly longer completion time when including its initial setup phase (mean 219s, median 194s, see Figure 4.20). This highlights a trade-off: users are willing to invest initial time for a customized and ultimately more efficient experience, but the setup itself is a barrier.

Direct dark pattern removal (R, RCC) vastly improved website usability and clarity, outperforming markings or information-only approaches.

Website Usability and Clarity: Direct removal of dark patterns (R, RCC) consistently led to the highest perceived website usability (see Figure 4.10) and clarity (see Figure 4.11). This suggests that the presence of dark patterns inherently makes a website feel less usable and more confusing. Variants that only marked or provided information (MC, CC) improved clarity and usability compared to the "Unchanged" variant (baseline, 4.2.3), but not to the extent of direct removal.

Assistant Safety and Helpfulness: User control was a important factor in perceived assistant safety (see Figure 4.16) and helpfulness (see Figure 4.18). The "Options (O)" variant (4.2.3), allowing users to configure interventions, was rated as the safest and most helpful assistant. Variants combining removal with informational elements (RCC, RMCC) also scored highly, indicating that transparency and optional information enhance the feeling of security and utility. The "Animation (AN)" variant (4.2.3), despite actively removing dark patterns, received lower safety and clarity ratings for the assistant itself, indicating that the animations, while showing action, did not sufficiently communicate what was being done or why, leading to a "loss of control" as noted in qualitative comments.

User control and transparency improved perceived assistant safety and helpfulness, with "Options" being most preferred.

General User Experience: The overall user experience was significantly improved by proactive intervention. Variants that removed dark patterns or allowed user customization consistently outperformed those that merely informed or offered no assistance (See Figure 4.19). This underscores that users desire countermeasures of dark patterns.

Proactive dark pattern countermeasures, especially removal, and customization, improved the general user experience.

Impact of Intervention Types:

Direct Removal: Highly effective in improving website metrics (usability, clarity, efficiency) and overall user experience. However, it raised concerns about "*lack of transparency*" and "*lack of `undo`* " if not accompanied by clear feedback or user control.

Direct dark pattern removal improved website metrics, but users desired more transparency and an "`undo`" option.

Highlighting dark patterns helped users detect them, but their persistence and potential frequency were criticized.

Marking: Positive for drawing attention to dark patterns ("*catches the eye*") and serving as a "*checklist*." However, it was criticized for remaining after interaction and potentially being "*too many*" on heavily patterned sites, suggesting it might feel incomplete without removal or confirmation.

Chat-based information was valued for awareness, but criticized for being time-consuming, overwhelming, and obstructing the interface.

Information (Chat): Valued for the ability to learn about dark patterns. However, it was widely criticized for being "*time-consuming*," a "*wall of text*," lacking "*memory*," and sometimes blocking parts of the website. This suggests that while awareness is good, the delivery method needs significant refinement.

Animations were fast and showed action, but caused "loss of control" and lacked clear information.

Animation: Appreciated for showing that the assistant "*does something*" quickly. However, it led to a "*loss of control*," felt "*too fast*," caused "*too much to happen at the same time*," and lacked sufficient information about the changes made.

User customization was highly praised for control, though the initial setup was time-consuming for some.

Customization (Options): Overwhelmingly positive due to empowering users with control. The main drawback was the initial "*time-consuming*" setup, and some users expressed concerns about making "*unsafe*" configurations without guidance.

Users prefer active virtual assistants that mitigate dark patterns, with customization and clear feedback being key.

In conclusion, the study strongly suggests that users prefer virtual assistants that actively mitigate dark patterns, ideally with a degree of user control and clear, concise feedback. While information provision is valued, its passive nature and poor presentation can hinder its effectiveness. Direct removal significantly improves website quality and efficiency, but transparency and an "undo" function are crucial for user trust and control. The "Options" variant (4.2.3) demonstrates that empowering users to choose their level of intervention leads to the most positive overall experience.

5.2 Limitations

The user study, while providing valuable insights, has several limitations that should be considered when interpreting the results:

Figma prototype.

- **Interactive Prototypes vs. Actual Websites:** The study was conducted using interactive prototypes in Figma rather than a live website. While these prototypes were designed to be medium-fidelity and simulate mostly realistic scenarios, they may not fully capture all dynamic elements and functionalities of actual websites, potentially influencing participant behavior and perceptions. Something that happened frequently and caused confusion was the lack of notification when the wall clock to be purchased was added to the shopping cart.

Limited interaction.

- **Chat Functionality:** The chat bot functionality in the prototypes was scripted and not a fully interactive AI. This means that participant interactions with the chat bot were limited to predefined responses (a request via buttons if the user want to know more about the dark patterns found), which might not reflect the full potential or challenges of a truly conversational assistant.

21 participants.

- **Number of Participants:** The study involved a relatively small number (21) of participants (4.3.1). While this size is common for qualitative user studies, it limits the generalizability of the quantitative findings to a broader population.
- **Background of Participants:** While diverse in educational background, the participants (4.3.1) were not specifically recruited to represent a wide range of digital literacy or susceptibility to dark patterns. This might have influenced the observed effects.

Age: between 21-37 years.

- **Age of Participants:** The majority of participants (4.3.1) were between 21 and 37 years old, with only one outlier at 72. This age range might not fully represent the experiences and preferences of older or younger demographics, who may interact with dark patterns differently.

Mostly academic.

- **Education of Participants:** While varying, the educational backgrounds of the participants (4.3.1) were primarily academic. This might not reflect the general population's understanding or awareness of digital design patterns.

Learning effects.

- **Repetitive Online Shopping Scenarios:** The online shopping scenarios were identical across all eight variants, apart from the assistant's intervention. This repetitive exposure could have led to learning effects, where participants became more familiar with the dark patterns and the website layout over successive trials, potentially influencing their performance and ratings.

Computer-based online-shopping.

- **Limited Scenario Scope:** The study focused exclusively on an online shopping scenario for computer websites. Different platforms (e.g., mobile apps, social media) or other types of online interactions (e.g., news consumption, social networking) may feature different types of dark patterns or require alternative countermeasure strategies.
- **Lack of Significance Analysis:** The results presented are based on mean and median scores and qualitative observations. A formal statistical significance analysis was not performed, meaning that observed differences between variants cannot be definitively attributed to the interventions with statistical certainty. This limits the ability to draw strong causal conclusions.

Chapter 6

Summary and Future Work

This chapter provides a concise summary of the thesis's key findings and contributions, followed by a discussion of limitations and promising avenues for future research. The aim is to consolidate the insights gained from the user study and outline the next steps in developing effective virtual assistants as countermeasures against dark patterns.

6.1 Summary and Contributions

This thesis investigated the potential of virtual assistants as a novel and proactive solution to combat dark patterns in online interfaces, building upon existing research in ethical design and visual countermeasures. Through a comprehensive user study in chapter 4 involving various virtual assistant prototypes, we explored user perceptions regarding the assistant's helpfulness (Figure 4.18), safety (Figure 4.16), clarity (Figure 4.15), efficiency (Figure 4.17), and general user experience (Figure 4.19).

Our findings reveal a strong user preference for virtual assistants that actively mitigate dark patterns, ideally offering a degree of user control and providing clear, concise feedback. The "Options (O)" variant (4.2.3), which allowed participants to customize the assistant's intervention level, emerged as the most preferred, demonstrating that empowering users with personalized control leads to the most positive overall experience. Variants that combined direct dark pattern removal with optional informational support, such as "Removed + Counter + Chat (RCC)" (4.2.3), also performed exceptionally well, improving the perceived website usability, clarity, and efficiency.

Conversely, the "Unchanged (UC)" baseline (4.2.3), with no assistant intervention, consistently resulted in the least favorable user experience, underscoring the negative impact of unmitigated dark patterns. Passive information provision alone, as seen in the "Counter + Chat (CC)" variant (4.2.3), proved insufficient to effectively counter dark patterns, often being perceived as time-consuming and unhelpful. While direct removal of dark patterns enhanced the website quality and efficiency, it raised concerns about a "*lack of transparency*" and the absence of an "`undo`" function if not accompanied by clear feedback. Similarly, animated interventions, while showing action, sometimes led to a "*loss of control*" due to insufficient communication about the changes.

We showed in this thesis that virtual assistants can improve the general user experience and can help users handling dark patterns. The user study highlights a strong user preference for proactive dark pattern removal combined with user customization and transparent feedback. We demonstrate that user control over the assistant's behavior and clear communication about its actions are crucial for fostering a sense of safety and trust. Variants that directly removed dark patterns significantly reduced task completion times, indicating a tangible benefit for user efficiency.

6.2 Future Work

Despite the valuable insights gained, this study has several limitations that open up promising avenues for future research:

Real-World Implementation and Testing: The study was conducted using interactive Figma prototypes. Future work should involve developing and testing fully functional browser extensions or applications that implement the most promising countermeasure variants on live websites. This would allow for a more realistic assessment of their impact on user behavior and perception in dynamic online environments.

Enhanced Chat Functionality: The current chat bot was scripted. Future iterations should integrate a truly interactive AI-powered chat function, potentially leveraging Large Language Models (LLMs), to provide more dynamic, context-aware, and personalized information and support. This could address participant suggestions for a chat bot with "*memory*" that reacts to previous interactions and provides more concise information (e.g., lists or tables instead of "*walls of text*").

Broader Participant Pool: The study's participant pool 4.3.1, while diverse in educational background, was relatively small and predominantly within a specific age range. Future research should recruit a larger and more demographically diverse group, including participants with varying levels of digital literacy and susceptibility to dark patterns, to enhance the generalizability of the findings.

Diverse Scenarios and Platforms: This study focused on a single online shopping scenario for computer websites. Future work should explore the effectiveness of virtual assistants against dark patterns in other contexts (e.g., social media, news sites, mobile applications) and for a wider range of dark pattern types, as different patterns may require tailored countermeasures.

Addressing Learning Effects: The repetitive nature of the scenarios in this study could have led to learning effects.

Future studies could employ different experimental designs, such as between-subjects designs for some variants or introducing new scenarios, to mitigate these effects.

Advanced Customization and Control: Building on the success of the "Options" variant (4.2.3), future research should delve deeper into customizable assistant behaviors. This includes exploring more granular control over specific dark pattern types, providing "*Reject All*" or "*Default Op tions*" for setup, and offering a progress bar during initial configuration. The ability to "*whitelist falsely recognized patterns*" and an "`undo`" function for removed patterns are also crucial features requested by participants that warrant further investigation.

Transparency and Feedback Mechanisms: Participants expressed a strong desire for transparency regarding the assistant's actions. Future designs should explore various feedback mechanisms, such as clear notifications for blocked dark patterns, subtle hints at altered positions, or hover-over information windows, to ensure users understand *what* the assistant has done and *why*, without being overly intrusive or distracting.

Categorization and Severity Ratings: Implementing a system that provides users with a precise numerical indicator or a "severity rating" (e.g., how "bad" a dark pattern is in terms of financial cost or data collection) could further empower users to make informed decisions.

Statistical Significance Analysis: Future quantitative studies should incorporate robust statistical significance analyses to definitively confirm the observed differences between intervention strategies.

By addressing these areas, future research can contribute to the development of highly effective, user-centric virtual assistants that empower individuals to navigate the digital world with greater autonomy and confidence, ultimately fostering a more ethical and transparent online experience.

Appendix A

User Study Prototype

In the following we present screenshots of the prototype used in the user study.

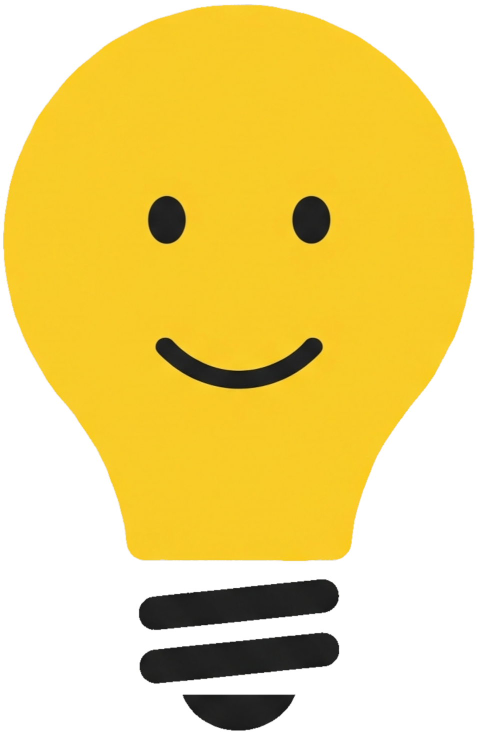
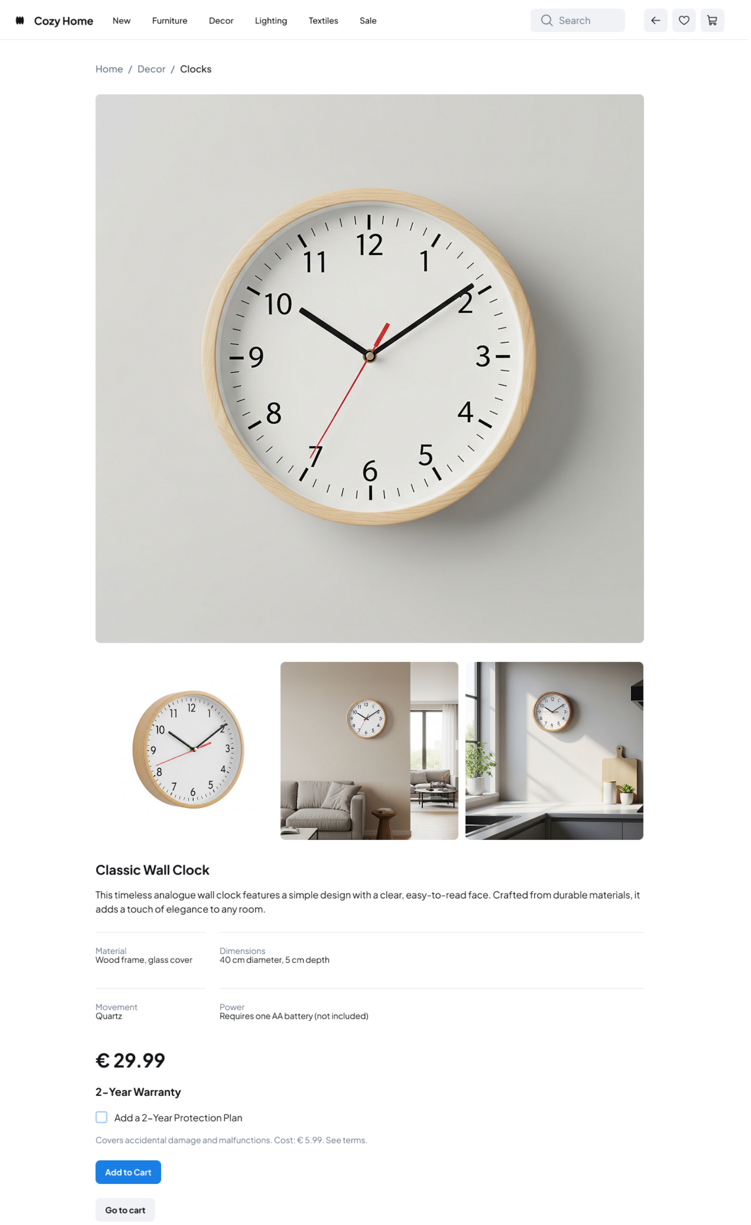


Figure A.1: The "Lumi" icon, representing the virtual assistant.



Classic Wall Clock

This timeless analogue wall clock features a simple design with a clear, easy-to-read face. Crafted from durable materials, it adds a touch of elegance to any room.

Material
Wood frame, glass cover

Dimensions
40 cm diameter, 5 cm depth

Movement
Quartz

Power
Requires one AA battery (not included)

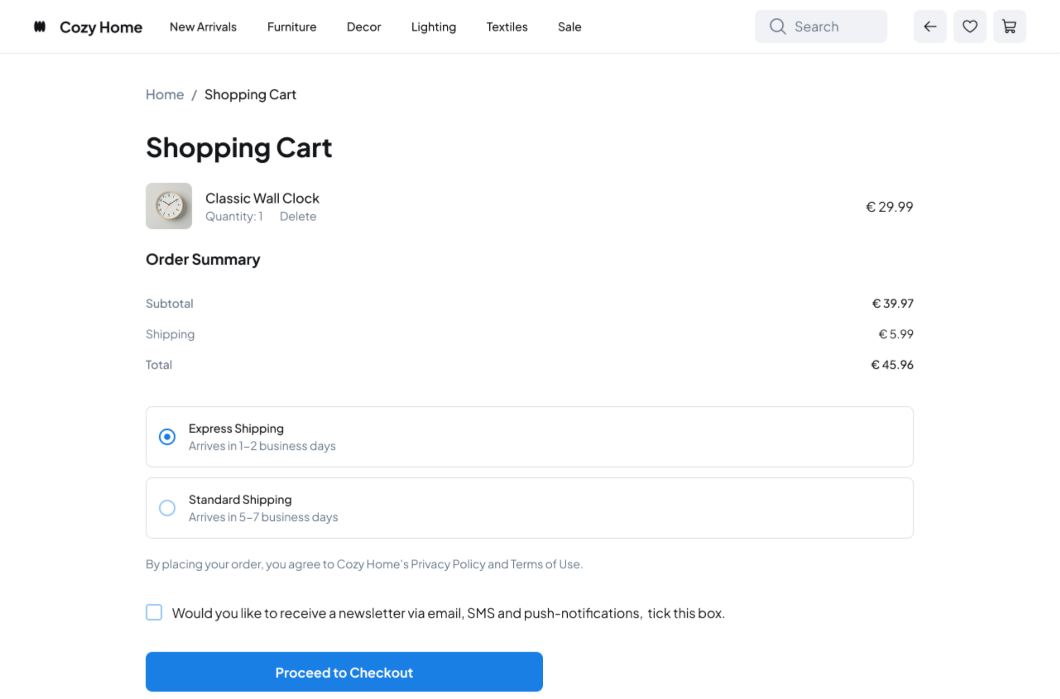
€ 29.99

2-Year Warranty

□ Add a 2-Year Protection Plan

Covers accidental damage and malfunctions. Cost: € 5.99. See terms.

Figure A.2: Screenshot of the simulated product page prototype in the "Removed (R)" variant. This image demonstrates the removal of dark patterns from the product page, resulting in a cleaner and potentially less manipulative user interface compared to the "Unchanged" variant.



Home / Shopping Cart

Shopping Cart

Classic Wall Clock
Quantity: 1 Delete
€ 29.99

Order Summary

Subtotal	€ 39.97
Shipping	€ 5.99
Total	€ 45.96

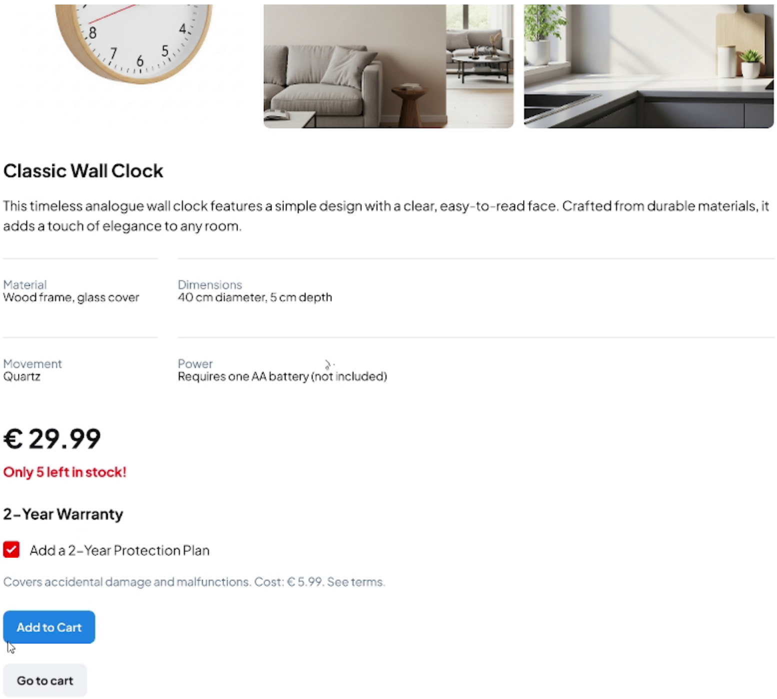
● Express Shipping
Arrives in 1–2 business days

○ Standard Shipping
Arrives in 5–7 business days

By placing your order, you agree to Cozy Home's Privacy Policy and Terms of Use.

□ Would you like to receive a newsletter via email, SMS and push-notifications, tick this box.

Figure A.3: Screenshot of the simulated shopping cart page prototype in the "Removed (R)" variant. This image highlights how dark patterns are removed from the cart page, simplifying the user's interaction and checkout process.



Classic Wall Clock

This timeless analogue wall clock features a simple design with a clear, easy-to-read face. Crafted from durable materials, it adds a touch of elegance to any room.

Material
Wood frame, glass cover

Dimensions
40 cm diameter, 5 cm depth

Movement
Quartz

Power
Requires one AA battery (not included)

€ 29.99

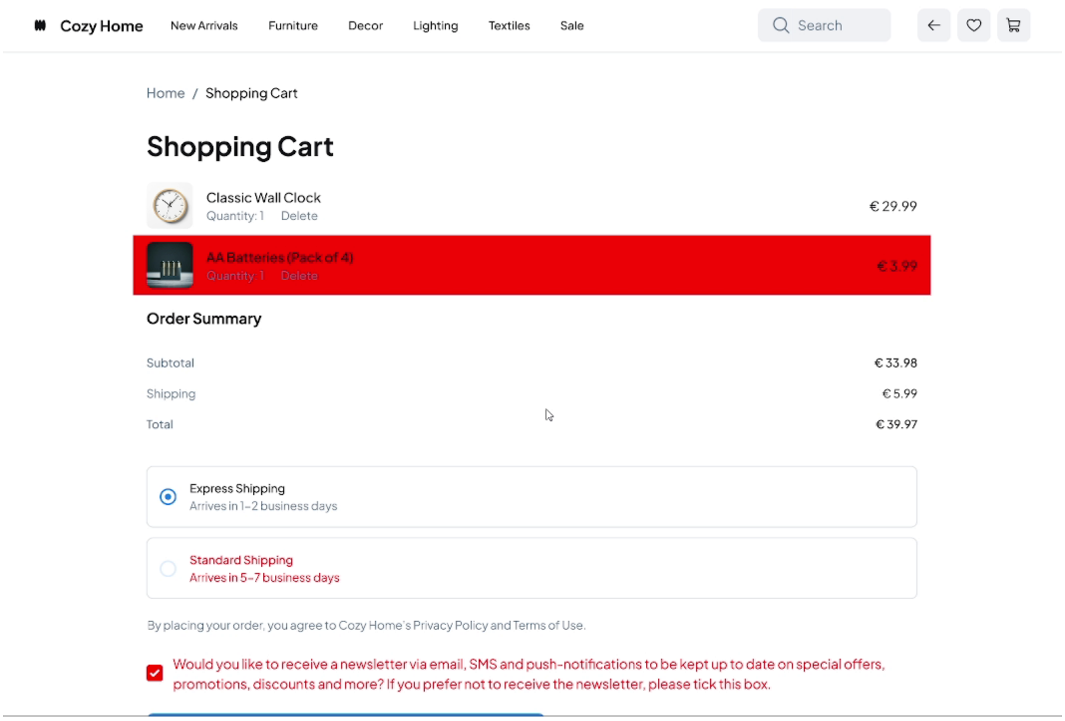
Only 5 left in stock!

2-Year Warranty

☑ Add a 2-Year Protection Plan

Covers accidental damage and malfunctions. Cost: € 5.99. See terms.

Figure A.4: Screenshot of the simulated product page prototype in the "Animation (AN)" variant, captured during an animation sequence. This image illustrates how dark patterns are actively animated and removed from the product page, providing visual feedback of the intervention.



Home / Shopping Cart

Shopping Cart

Classic Wall Clock
Quantity: 1 Delete
€ 29.99

AA Batteries (Pack of 4)
Quantity: 1 Delete
€ 3.99

Order Summary

Subtotal	€ 33.98
Shipping	€ 5.99
Total	€ 39.97

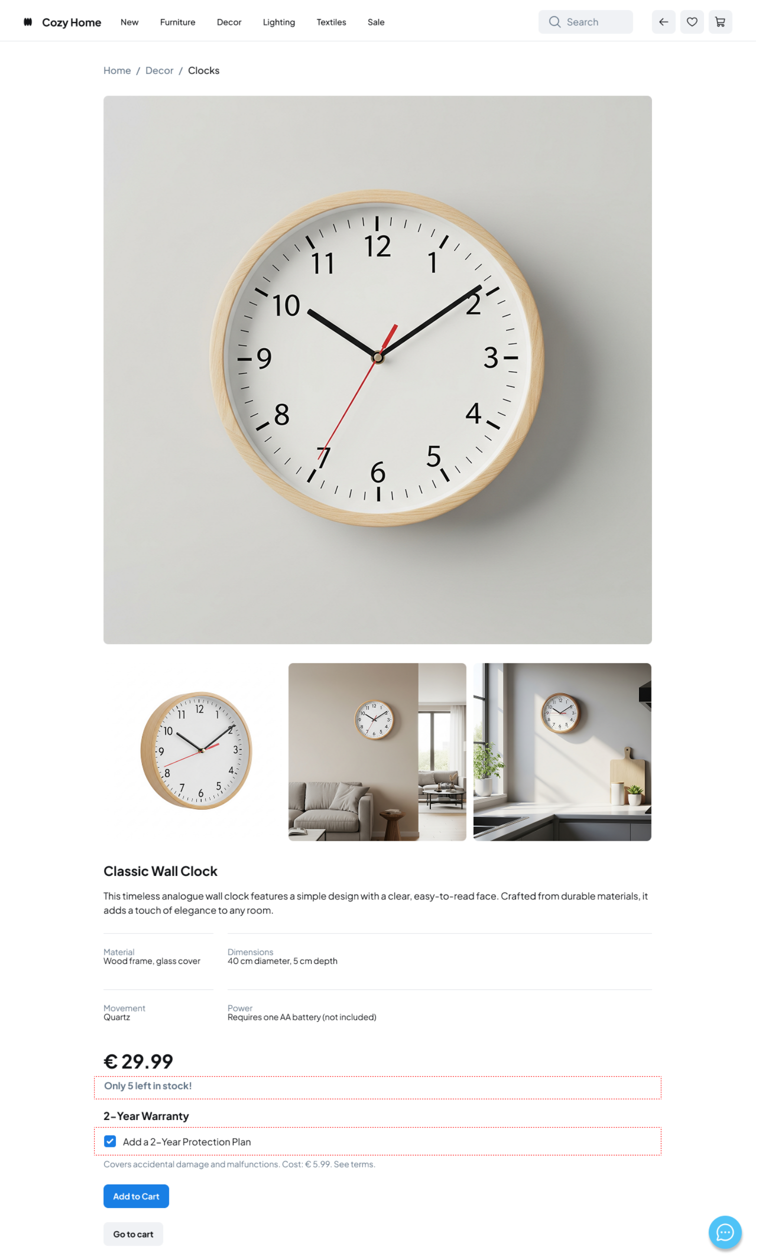
● Express Shipping
Arrives in 1–2 business days

○ Standard Shipping
Arrives in 5–7 business days

By placing your order, you agree to Cozy Home's Privacy Policy and Terms of Use.

☑ Would you like to receive a newsletter via email, SMS and push-notifications to be kept up to date on special offers, promotions, discounts and more? If you prefer not to receive the newsletter, please tick this box.

Figure A.5: Screenshot of the simulated shopping cart page prototype in the "Animation (AN)" variant, captured during an animation sequence. This image showcases the dynamic removal of dark patterns from the cart page, visually indicating the assistant's intervention.



Home / Decor / Clocks

Classic Wall Clock

This timeless analogue wall clock features a simple design with a clear, easy-to-read face. Crafted from durable materials, it adds a touch of elegance to any room.

Material
Wood frame, glass cover

Dimensions
40 cm diameter, 5 cm depth

Movement
Quartz

Power
Requires one AA battery (not included)

€ 29.99

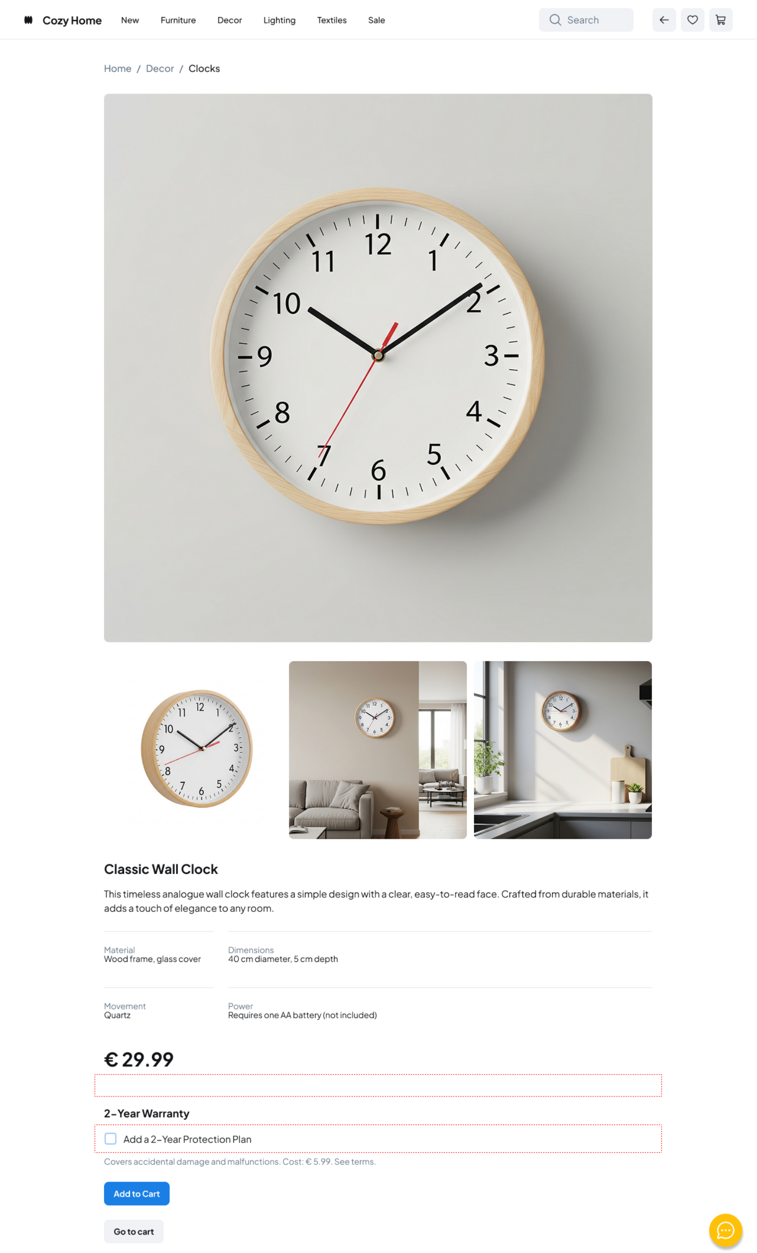
Only 5 left in stock!

2-Year Warranty

☑ Add a 2-Year Protection Plan

Covers accidental damage and malfunctions. Cost: € 5.99. See terms.

Figure A.6: Screenshot of the simulated product page prototype in the "Mark + Chat (MC)" variant. This image displays the visual highlighting of dark patterns on the product page, designed to draw user attention to these manipulative elements while a chat interface offers additional information.



Home / Decor / Clocks

Classic Wall Clock

This timeless analogue wall clock features a simple design with a clear, easy-to-read face. Crafted from durable materials, it adds a touch of elegance to any room.

Material
Wood frame, glass cover

Dimensions
40 cm diameter, 5 cm depth

Movement
Quartz

Power
Requires one AA battery (not included)

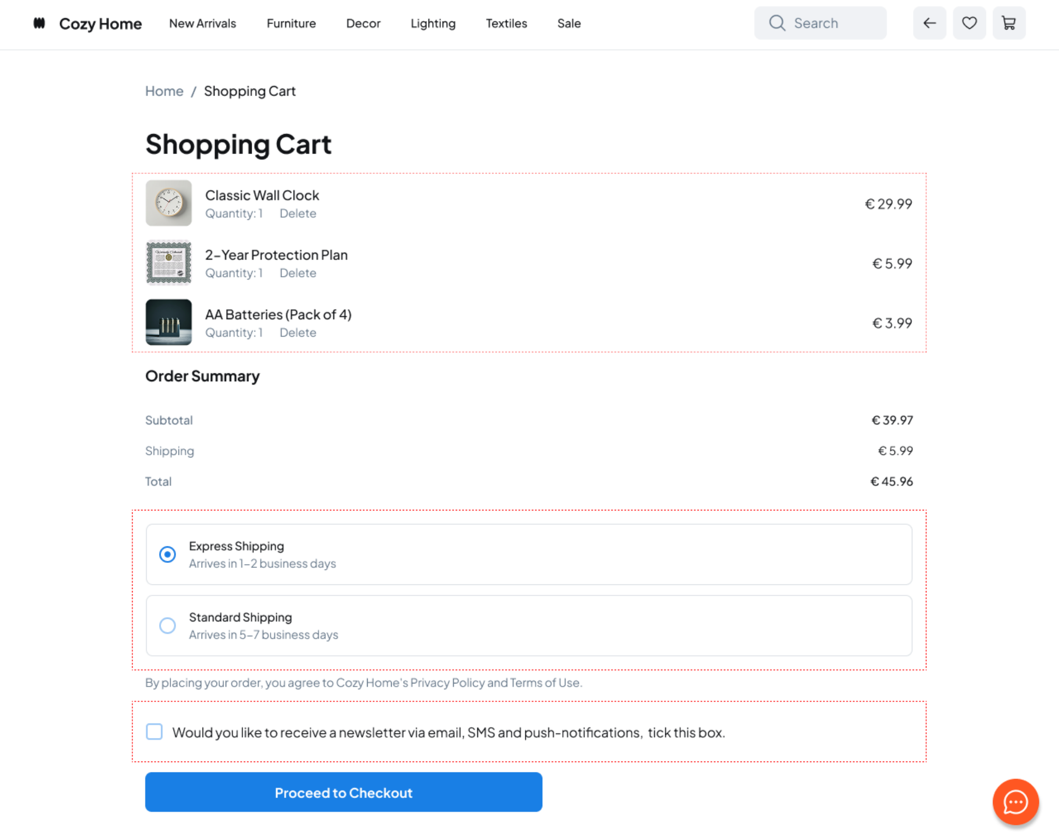
€ 29.99

2-Year Warranty

□ Add a 2-Year Protection Plan

Covers accidental damage and malfunctions. Cost: € 5.99. See terms.

Figure A.7: Screenshot of the simulated product page prototype in the "Removed + Mark + Counter + Chat (RMCC)" variant. This image demonstrates the combined intervention, where dark patterns are removed, and their former locations are indicated by markings, accompanied by a counter and chat functionality.



Shopping Cart

Classic Wall Clock Quantity: 1 Delete	€ 29.99
2-Year Protection Plan Quantity: 1 Delete	€ 5.99
AA Batteries (Pack of 4) Quantity: 1 Delete	€ 3.99

Order Summary

Subtotal	€ 39.97
Shipping	€ 5.99
Total	€ 45.96

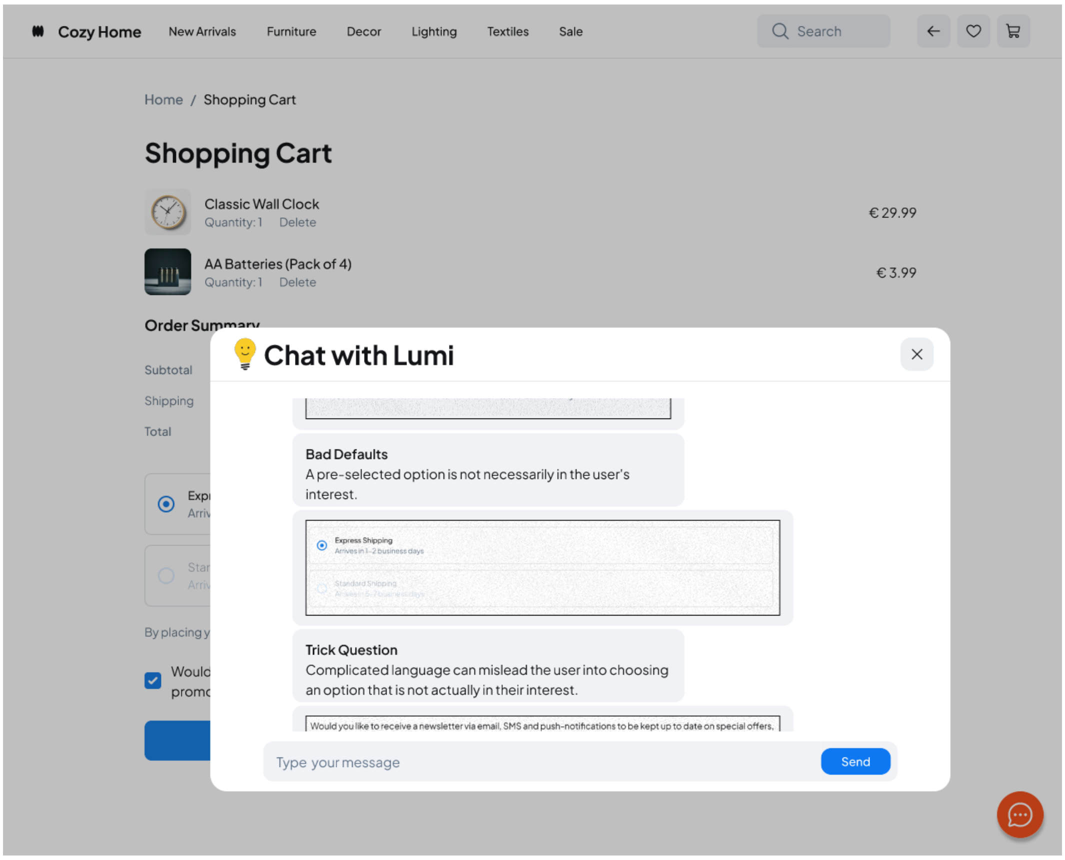
● Express Shipping
Arrives in 1–2 business days

○ Standard Shipping
Arrives in 5–7 business days

By placing your order, you agree to Cozy Home's Privacy Policy and Terms of Use.

□ Would you like to receive a newsletter via email, SMS and push-notifications, tick this box.

Figure A.8: Screenshot of the simulated shopping cart page prototype in the "Removed + Mark + Counter + Chat (RMCC)" variant. This image shows the simultaneous removal of dark patterns, the presence of subtle markings at their original positions, and the visible counter and chat features for comprehensive user support.



Bad Defaults
A pre-selected option is not necessarily in the user's interest.

● Express Shipping
Arrives in 1–2 business days

○ Standard Shipping
Arrives in 5–7 business days

Trick Question
Complicated language can mislead the user into choosing an option that is not actually in their interest.

Would you like to receive a newsletter via email, SMS and push-notifications to be kept up to date on special offers,

Figure A.9: Screenshot of the chat interface within the "Counter + Chat" variant, displaying a list of detected dark patterns. This shows the informational component of the assistant, providing details about the manipulative design elements present on the website.

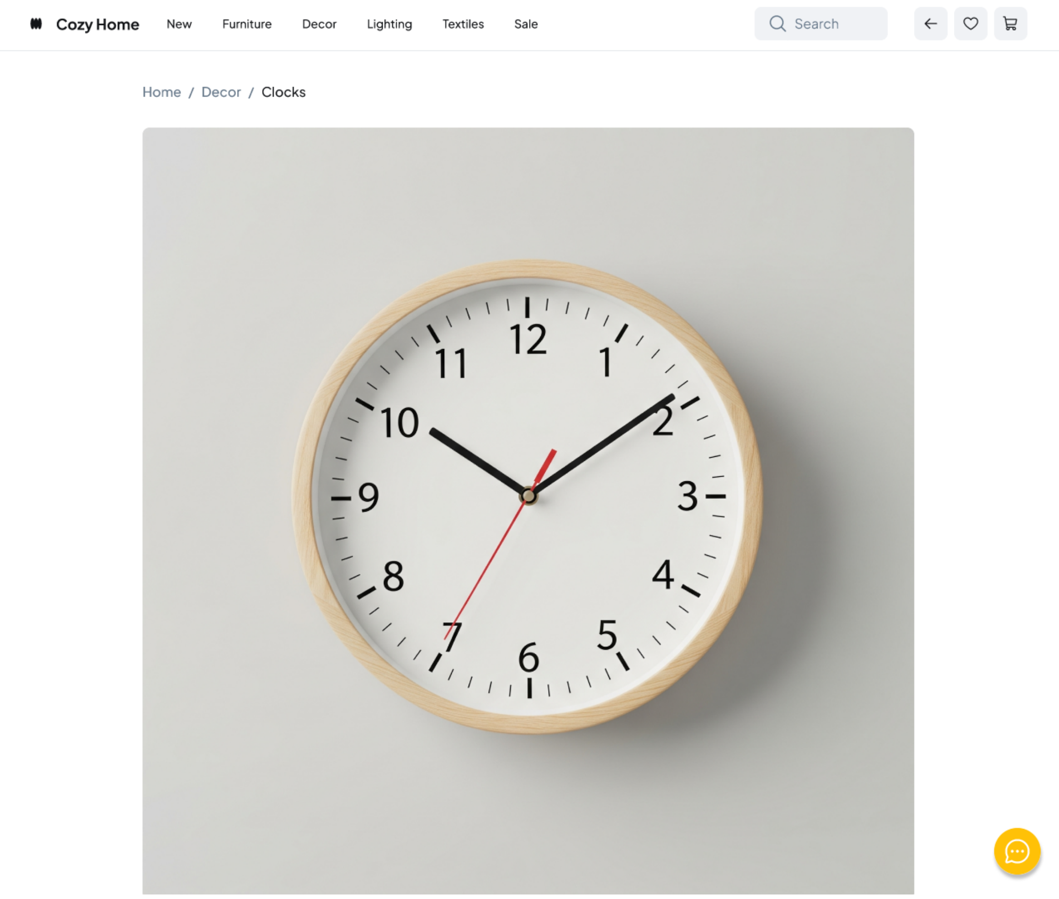
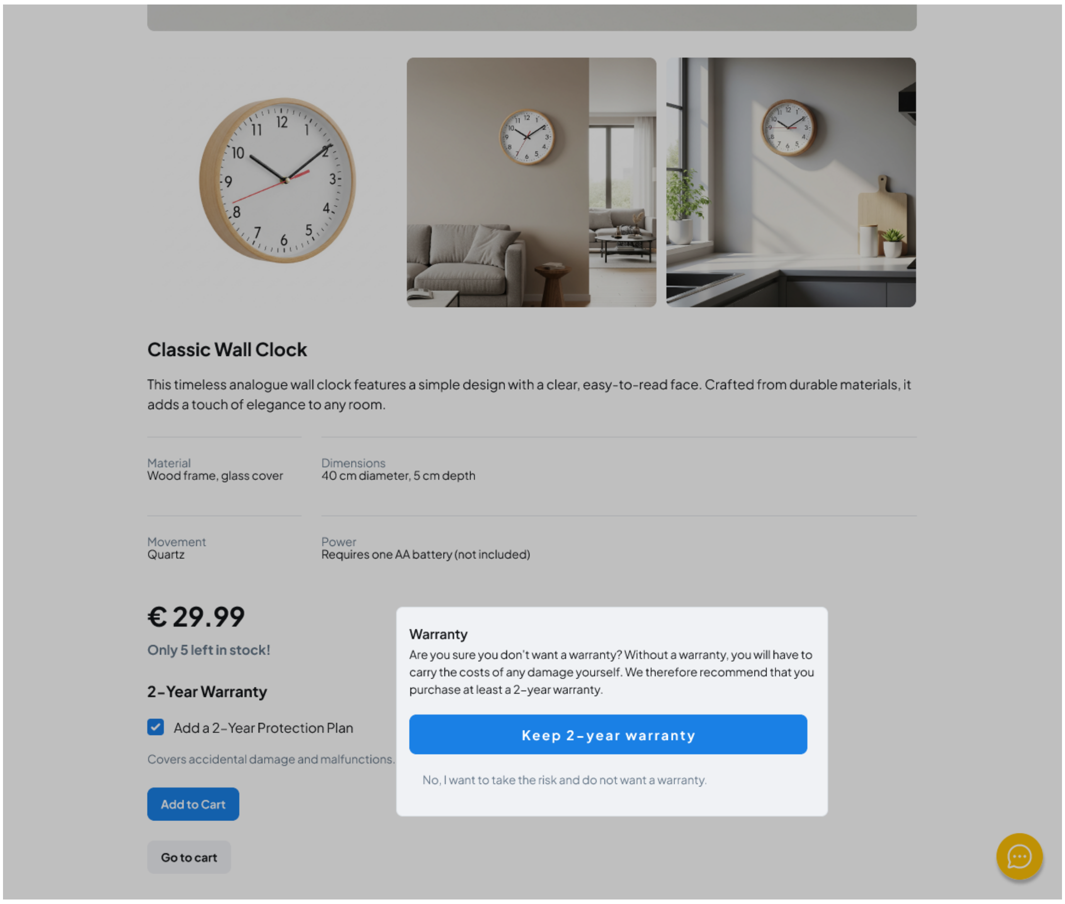


Figure A.10: Screenshot of upper part of the simulated product page, showcasing the dark pattern counter prominently displayed even though no dark patterns are visible in this area. This illustrates how the counter provides users with real-time feedback on the number of dark patterns detected on the page.



Classic Wall Clock

This timeless analogue wall clock features a simple design with a clear, easy-to-read face. Crafted from durable materials, it adds a touch of elegance to any room.

Material Wood frame, glass cover	Dimensions 40 cm diameter, 5 cm depth
Movement Quartz	Power Requires one AA battery (not included)

€ 29.99

Only 5 left in stock!

2-Year Warranty

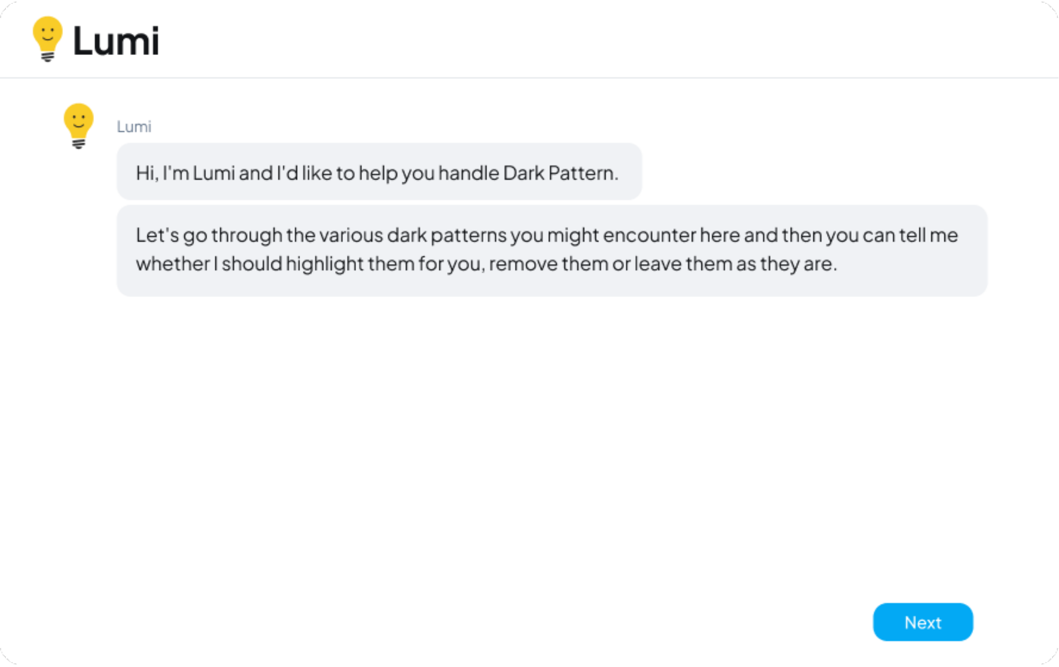
☑ Add a 2-Year Protection Plan

Covers accidental damage and malfunctions.

Warranty

Are you sure you don't want a warranty? Without a warranty, you will have to carry the costs of any damage yourself. We therefore recommend that you purchase at least a 2-year warranty.

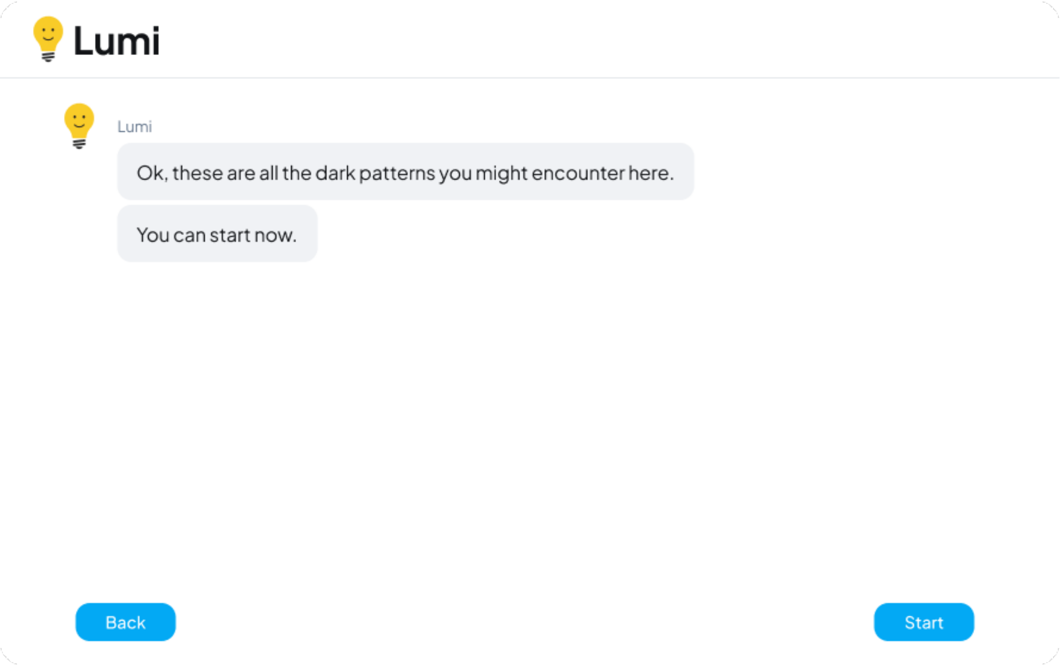
Figure A.11: Screenshot of the lower part of the simulated product page with the warranty question pop-up, showing the "Adding Steps" dark pattern.



Hi, I'm Lumi and I'd like to help you handle Dark Pattern.

Let's go through the various dark patterns you might encounter here and then you can tell me whether I should highlight them for you, remove them or leave them as they are.

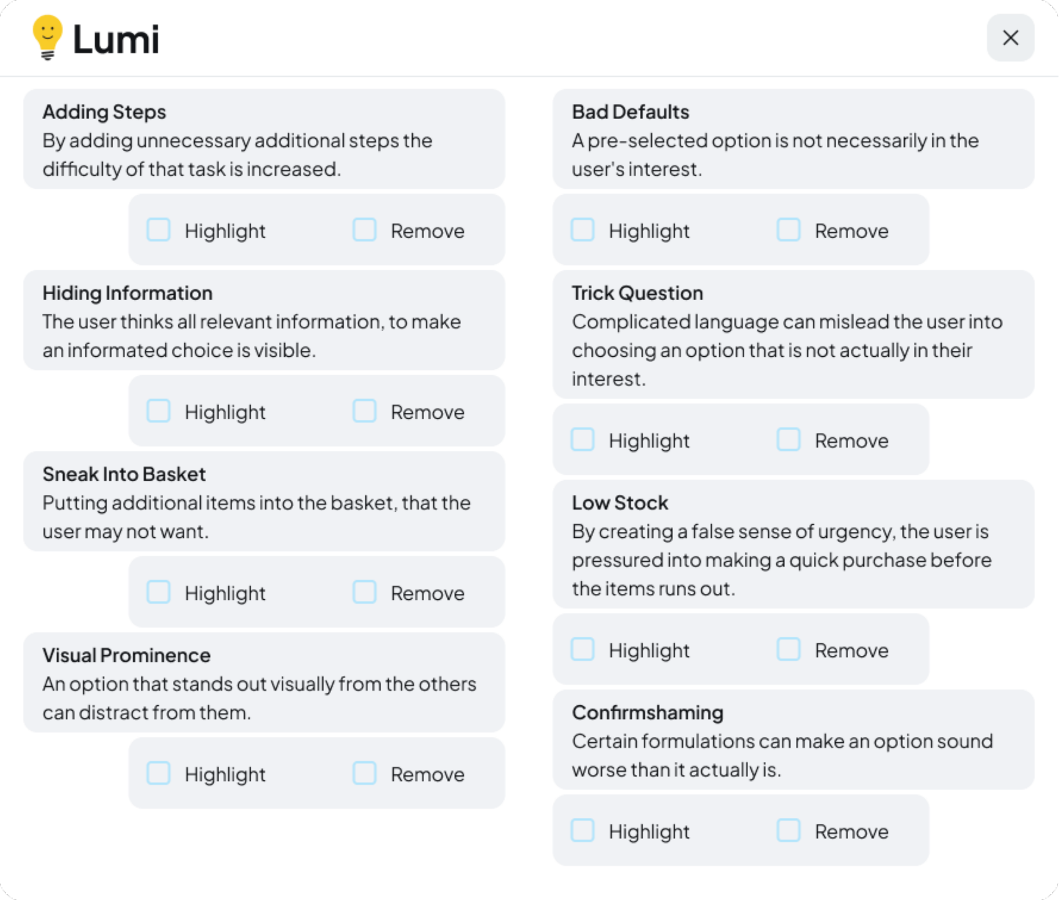
Figure A.12: Screenshot of the initial screen of the setup process for the "Options" variant. This image introduces the customization options to the user, allowing them to define how the assistant will intervene against dark patterns.



Ok, these are all the dark patterns you might encounter here.

You can start now.

Figure A.13: Screenshot of the final screen of the setup process for the "Options" variant. This image confirms the completion of the user's customization choices, indicating the assistant is now configured according to their preferences.



Lumi

Adding Steps
By adding unnecessary additional steps the difficulty of that task is increased.

□ Highlight □ Remove

Hiding Information
The user thinks all relevant information, to make an informated choice is visible.

□ Highlight □ Remove

Sneak Into Basket
Putting additional items into the basket, that the user may not want.

□ Highlight □ Remove

Visual Prominence
An option that stands out visually from the others can distract from them.

□ Highlight □ Remove

Bad Defaults
A pre-selected option is not necessarily in the user's interest.

□ Highlight □ Remove

Trick Question
Complicated language can mislead the user into choosing an option that is not actually in their interest.

□ Highlight □ Remove

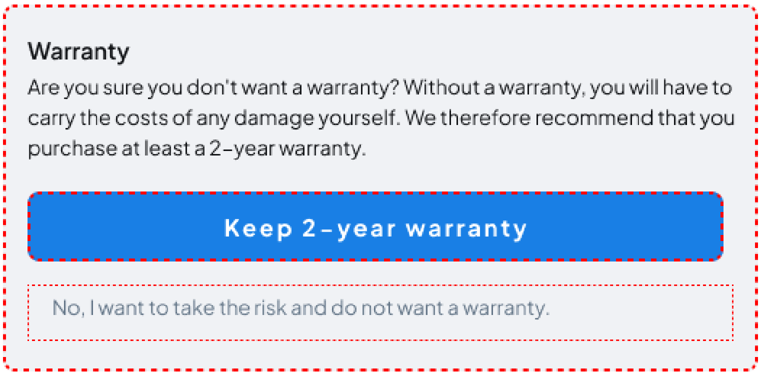
Low Stock
By creating a false sense of urgency, the user is pressured into making a quick purchase before the items runs out.

□ Highlight □ Remove

Confirmshaming
Certain formulations can make an option sound worse than it actually is.

□ Highlight □ Remove

Figure A.14: Screenshot of the assistant's options menu of the "Options" variant, accessible later during interaction. This demonstrates the user's ability to adjust the assistant's settings and preferred intervention methods beyond the initial setup phase.



Warranty

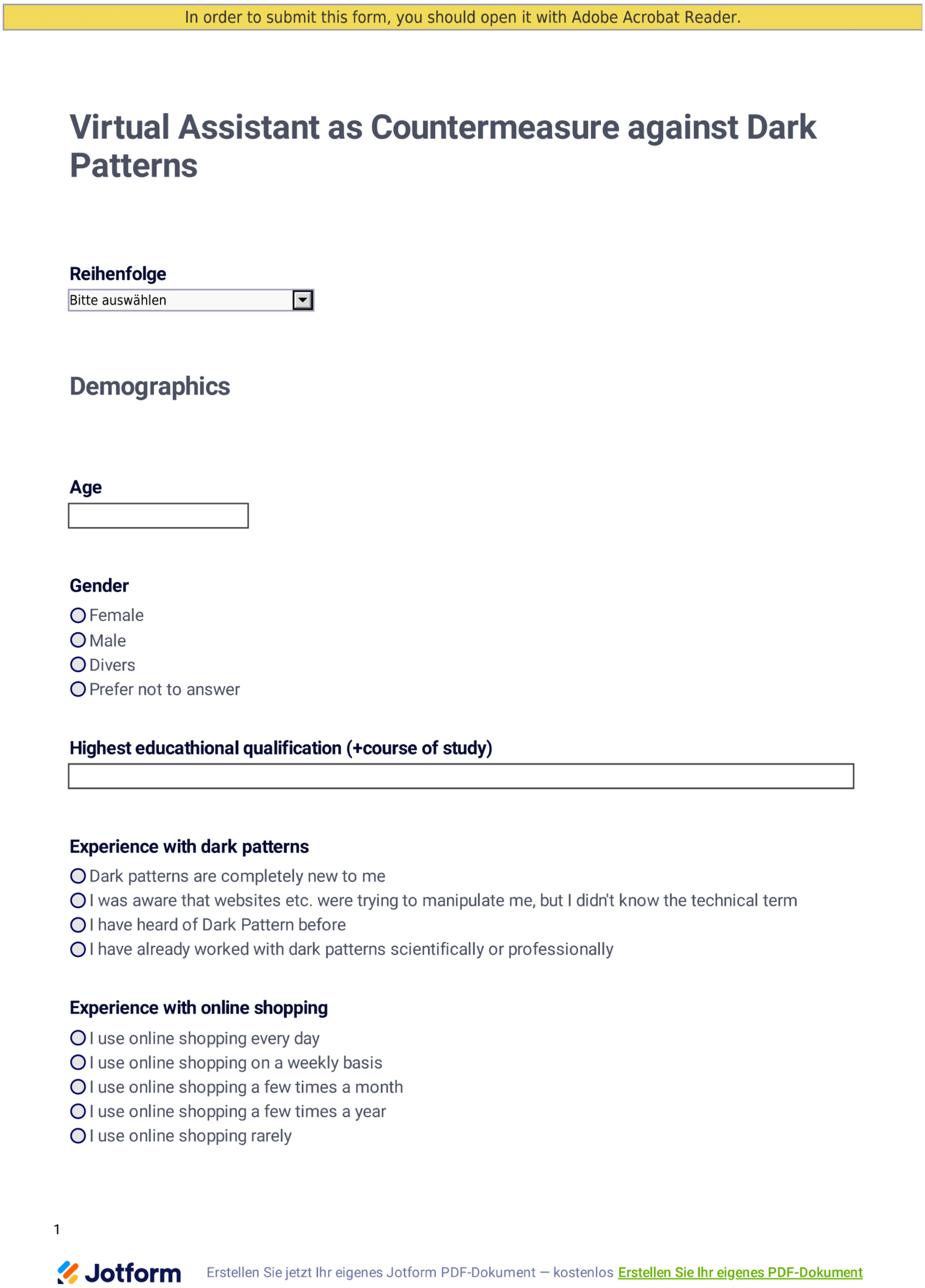
Are you sure you don't want a warranty? Without a warranty, you will have to carry the costs of any damage yourself. We therefore recommend that you purchase at least a 2-year warranty.

Figure A.15: Screenshot of the "Warranty Question" pop-up in the "Mark + Chat" variant. This image shows the "Adding Steps" dark pattern, with the pop-up itself highlighted by the assistant, visually indicating the deceptive elements to the user.

Appendix B

Questionnaire

This appendix presents the questionnaire we used in the User Study to collect both demographic data and participant feedback on the various virtual assistant variants. The questionnaire was structured into several sections, beginning with demographic information, followed by questions specific to each assistant variant, and concluding with an overall ranking. The first assistant variant section for the "Unchanged" variant had fewer questions due to the lack of an assistant. The other assistant variant section had all the same questions (we only show the "Options" variant here).



In order to submit this form, you should open it with Adobe Acrobat Reader.

Virtual Assistant as Countermeasure against Dark Patterns

Reihenfolge

Bitte auswählen ▾

Demographics

Age

Gender

○ Female
○ Male
○ Divers
○ Prefer not to answer

Highest educathional qualification (+course of study)

Experience with dark patterns

○ Dark patterns are completely new to me
○ I was aware that websites etc. were trying to manipulate me, but I didn't know the technical term
○ I have heard of Dark Pattern before
○ I have already worked with dark patterns scientifically or professionally

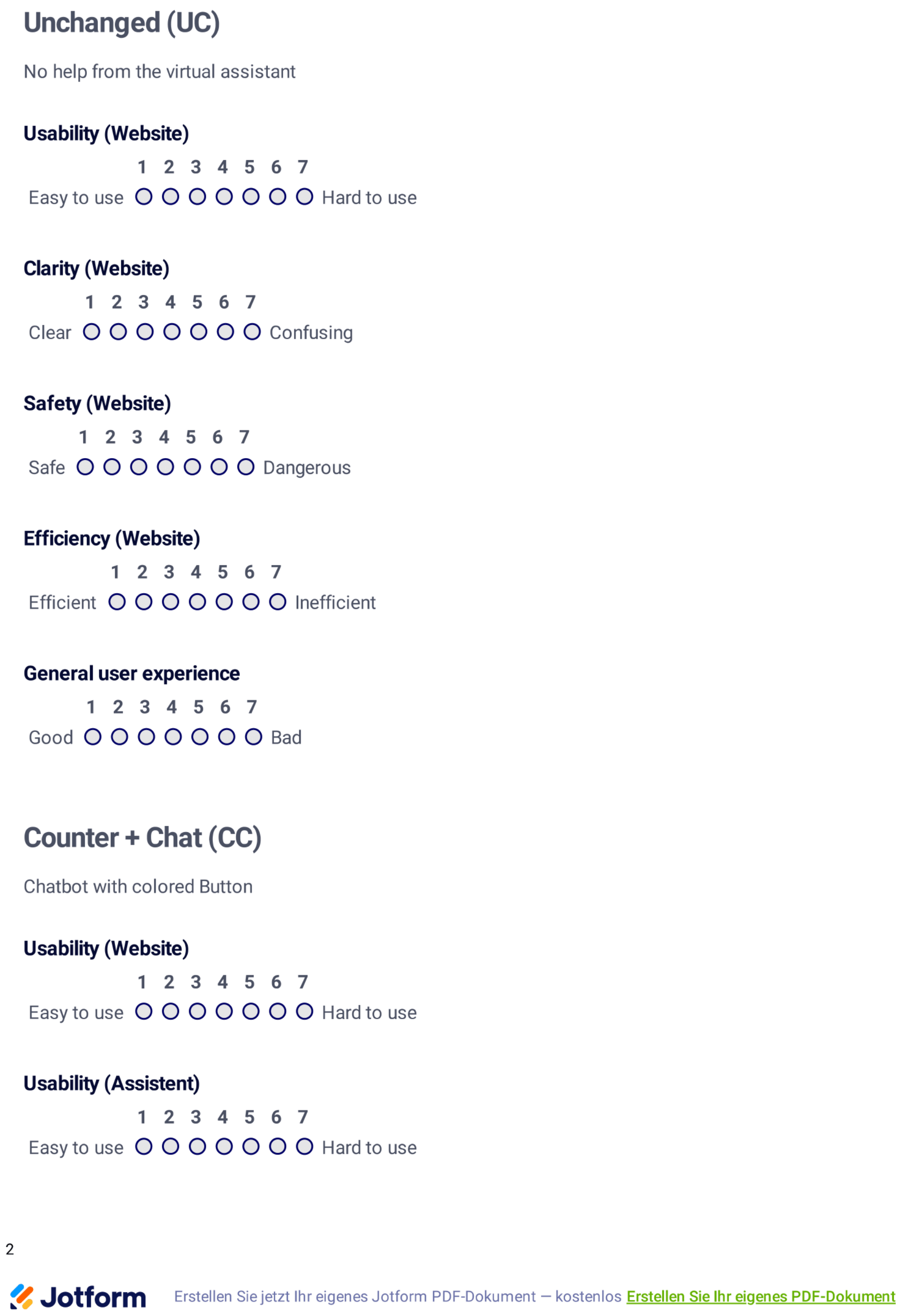
Experience with online shopping

○ I use online shopping every day
○ I use online shopping on a weekly basis
○ I use online shopping a few times a month
○ I use online shopping a few times a year
○ I use online shopping rarely

1

Jotform Erstellen Sie jetzt Ihr eigenes Jotform PDF-Dokument — kostenlos Erstellen Sie Ihr eigenes PDF-Dokument

Figure B.1: Demographics Section of the Questionnaire



Unchanged (UC)

No help from the virtual assistant

Usability (Website)

	1	2	3	4	5	6	7	
Easy to use	○	○	○	○	○	○	○	Hard to use

Clarity (Website)

	1	2	3	4	5	6	7	
Clear	○	○	○	○	○	○	○	Confusing

Safety (Website)

	1	2	3	4	5	6	7	
Safe	○	○	○	○	○	○	○	Dangerous

Efficiency (Website)

	1	2	3	4	5	6	7	
Efficient	○	○	○	○	○	○	○	Inefficient

General user experience

	1	2	3	4	5	6	7	
Good	○	○	○	○	○	○	○	Bad

Counter + Chat (CC)

Chatbot with colored Button

Usability (Website)

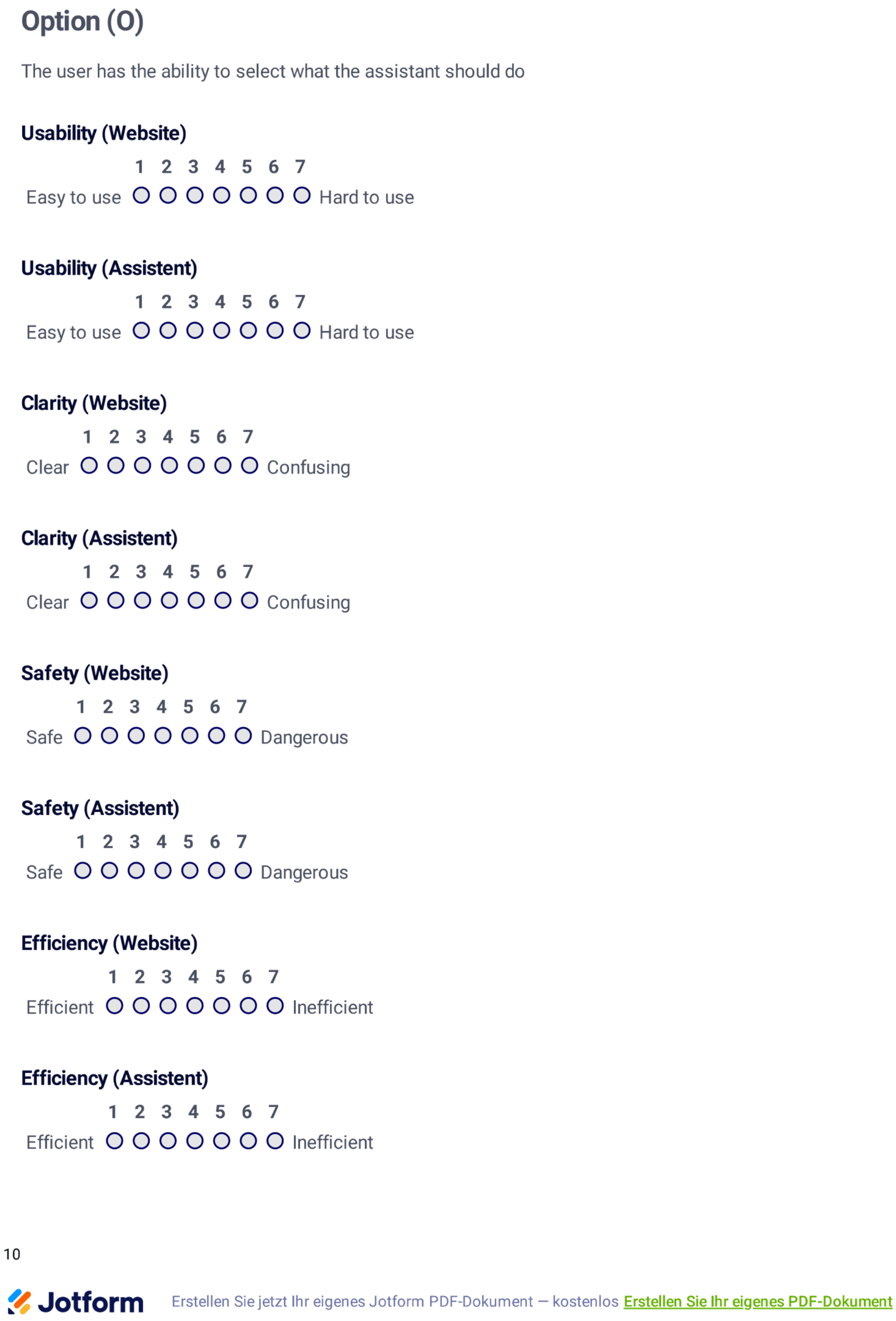
	1	2	3	4	5	6	7	
Easy to use	○	○	○	○	○	○	○	Hard to use

Usability (Assistent)

	1	2	3	4	5	6	7	
Easy to use	○	○	○	○	○	○	○	Hard to use

2

Figure B.2: "Unchanged" Variant Section of the Questionnaire



Option (O)

The user has the ability to select what the assistant should do

Usability (Website)

	1	2	3	4	5	6	7	
Easy to use	○	○	○	○	○	○	○	Hard to use

Usability (Assistent)

	1	2	3	4	5	6	7	
Easy to use	○	○	○	○	○	○	○	Hard to use

Clarity (Website)

	1	2	3	4	5	6	7	
Clear	○	○	○	○	○	○	○	Confusing

Clarity (Assistent)

	1	2	3	4	5	6	7	
Clear	○	○	○	○	○	○	○	Confusing

Safety (Website)

	1	2	3	4	5	6	7	
Safe	○	○	○	○	○	○	○	Dangerous

Safety (Assistent)

	1	2	3	4	5	6	7	
Safe	○	○	○	○	○	○	○	Dangerous

Efficiency (Website)

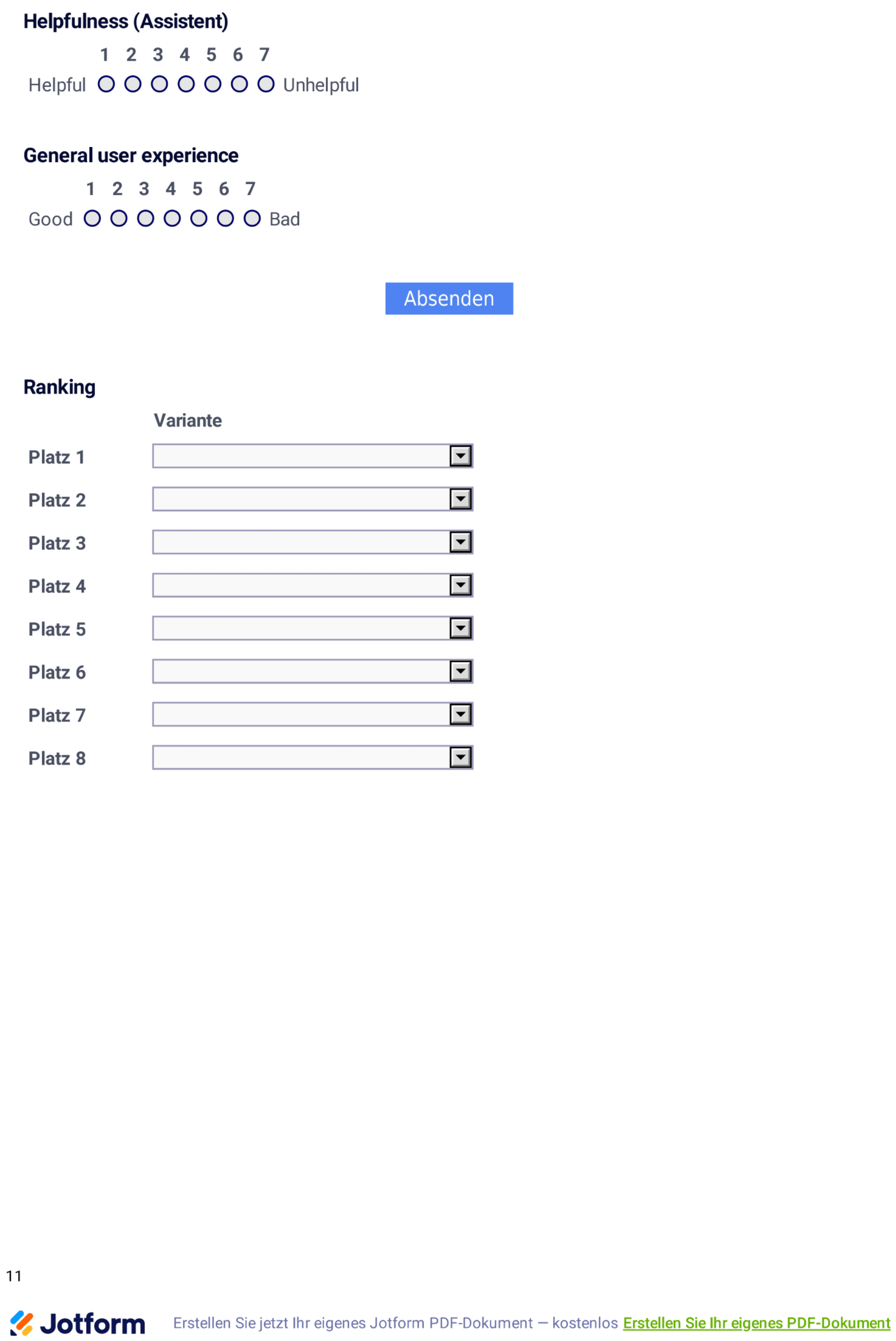
	1	2	3	4	5	6	7	
Efficient	○	○	○	○	○	○	○	Inefficient

Efficiency (Assistent)

	1	2	3	4	5	6	7	
Efficient	○	○	○	○	○	○	○	Inefficient

10

Figure B.3: Part 1 of the "Options" Variant Section of the Questionnaire



Helpfulness (Assistent)

	1	2	3	4	5	6	7	
Helpful	○	○	○	○	○	○	○	Unhelpful

General user experience

	1	2	3	4	5	6	7	
Good	○	○	○	○	○	○	○	Bad

Absenden

Ranking

	Variante
Platz 1	▾
Platz 2	▾
Platz 3	▾
Platz 4	▾
Platz 5	▾
Platz 6	▾
Platz 7	▾
Platz 8	▾

11

Figure B.4: Part 2 of the "Options" Variant Section of the Questionnaire and the Variant Ranking

Bibliography

[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *arXiv preprint arXiv:2208.10264*, 2023.

[2] Alex Beattie, Cherie Lacey, and Catherine Caudwell. "It's like the Wild West": User Experience (UX) Designers on Ethics and Privacy in Aotearoa New Zealand. *Design and Culture*, 16(1):63–82, 2024. `doi.org/10.1080/17547075.2023.2211391`.

[3] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. "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, page 763–776, New York, NY, USA, 2021. Association for Computing Machinery. `doi.org/10.1145/3461778.3462086`.

[4] Gregory Conti and Edward Sobiesk. Malicious interface design: exploiting the user. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 271–280, New York, NY, USA, 2010. Association for Computing Machinery. `doi.org/10.1145/1772690.1772719`.

[5] Andrea Curley, Dympna O'Sullivan, Damian Gordon, Barry Tierney, and Ilias Stavrakakis. The Design of a Framework for the Detection of Web-Based Dark Patterns. In *ICDS 2021: The 15th International Conference on Digital Society*, Nice, France (online), 2021.

[6] Alan Dix, Kris Luyten, Sven Mayer, Philippe Palanque, Emanuele Panizzi, Lucio Davide Spano, and Jürgen Ziegler. Second Workshop on Engineering Interactive Systems Embedding AI Technologies. In *Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '24 Companion, page 103–107, New York, NY, USA, 2024. Association for Computing Machinery. `doi.org/10.1145/3660515.3662837`.

[7] Kevin Fiedler, René Schäfer, Jan Borchers, and René Röpke. "Deception Detected!"—A Serious Game About Detecting Dark Patterns. In Avo Schön-

bohm, Francesco Bellotti, Antonio Bucchiarone, Francesca de Rosa, Manuel Ninaus, Alf Wang, Vanissa Wanick, and Pierpaolo Dondio, editors, *Games and Learning Alliance*, pages 191–200, Cham, 2025. Springer Nature Switzerland.

[8] Viola Valery Johanna Graf. Designing a Modular Browser Extension for Visual Countermeasures Against Dark Patterns. Aachen, Germany, 2024.

[9] Paul Grassl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. Dark and bright patterns in cookie consent requests, Jul 2020. URL `osf.io/preprints/psyarxiv/gqs5h_v1`.

[10] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. `doi.org/10.1145/3173574.3174108`.

[11] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. `doi.org/10.1145/3411764.3445779`.

[12] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. `doi.org/10.1145/3613904.3642436`.

[13] Pedro Andre Heckler Guidobono. VA in VWLE: Virtual Assistant in Virtual World Learning Environment. Msida, Malta, 2024. Supervisor: Matthew Montebello.

[14] Philip Hausner and Michael Gertz. Dark Patterns in the Interaction with Cookie Banners. In *Proceedings of the Workshop "What Can CHI Do About Dark Patterns?" at the CHI Conference on Human Factors in Computing Systems*, page 5, Yokohama, Japan, 2021.

[15] Maxwell Keleher, Fiona Westin, Preethi Nagabandi, and Sonia Chiasson. How Well Do Experts Understand End-Users' Perceptions of Manipulative Patterns? In *Nordic Human-Computer Interaction Conference*, NordiCHI '22, New York, NY, USA, 2022. Association for Computing Machinery. `doi.org/10.1145/3546155.3546656`.

[16] Frank Lewis and Julita Vassileva. Seeing in the Dark: Revealing the Relationships, Goals, and Harms of Dark Patterns. In *CEUR Workshop Proceedings*, 2024.

[17] Yuwen Lu, Chao Zhang, Yuewen Yang, Yaxing Yao, and Toby Jia-Jun Li. From Awareness to Action: Exploring End-User Empowerment Interventions for Dark Patterns in UX. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), April 2024. `doi.org/10.1145/3637336`.

[18] Alexander Maedche, Stefan Morana, Silvia Schacht, Dirk Werth, and Julian Krumeich. Advanced User Assistance Systems. *Business Information Systems Engineering*, 58(5):367–370, 2016. `doi.org/10.1007/s12599-016-0444-2`.

[19] Maximilian Maier and Rikard Harr. Dark Design Patterns: An End-user Perspective. *Human Technology*, 16(2):170–199, 2020. `doi.org/10.17011/ht/urn.202008245641`.

[20] Mario Martini, Christian Drews, Paul Seeliger, and Quirin Weinzierl. Dark Patterns: Phänomenologie und Antworten der Rechtsordnung. *Zeitschrift für Digitalisierung und Recht*, 1(1):47 – 74, 2021.

[21] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. `doi.org/10.1145/3359183`.

[22] Stuart Mills and Richard Whittle. Detecting Dark Patterns Using Generative AI: Some Preliminary Results. Available at SSRN: `https://ssrn.com/abstract=4614907`, October 2023. Accessed: 2025-07-13.

[23] Carol Moser, Sarita Y. Schoenebeck, and Paul Resnick. Impulse Buying: Design Practices and Consumer Needs. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery. `doi.org/10.1145/3290605.3300472`.

[24] Dmitry Nazarov and Yerkebulan Baimukhambetov. Clustering of Dark Patterns in the User Interfaces of Websites and Online Trading Portals (E-Commerce). *Mathematics*, 10(18), 2022. `doi.org/10.3390/math10183219`.

[25] René Schäfer, Paul Miles Preuschoff, and Jan Borchers. Investigating Visual Countermeasures Against Dark Patterns in User Interfaces. In *Proceedings of Mensch Und Computer 2023*, MuC '23, page 161–172, New York, NY, USA, 2023. Association for Computing Machinery. `doi.org/10.1145/3603555.3603563`.

[26] René Schäfer, Paul Miles Preuschoff, René Röpke, Sarah Sahabi, and Jan Borchers. Fighting Malicious Designs: Towards Visual Countermeasures

Against Dark Patterns. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. doi.org/10.1145/3613904.3642661.

[27] René Schäfer, Sarah Sahabi, Annabell Brocker, and Jan Borchers. Growing Up With Dark Patterns: How Children Perceive Malicious User Interface Designs. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*, NordiCHI '24, New York, NY, USA, 2024. Association for Computing Machinery. doi.org/10.1145/3679318.3685358.

[28] Sonia Jawaid Shaikh. Artificially Intelligent, Interactive, and Assistive Machines: A Definitional Framework for Intelligent Assistants. *International Journal of Human–Computer Interaction*, 39(4):776–789, 2023. doi.org/10.1080/10447318.2022.2049133.

[29] Than Htut Soe, Cristiana Teixeira Santos, and Marija Slavkovik. Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way, 2022. URL https://arxiv.org/abs/2204.11836.

[30] Susan M. Weinschenk. *How to get people to do stuff: Master the art and science of persuasion and motivation*. New Riders, San Francisco, CA, 2013.

Index

Typeset July 29, 2025