University of Bremen

MITIGATING DARK PATTERNS THROUGH RESPONSIBLE DESIGN

Ethical Design Considerations for
User-Centred Technologies

submitted by

THOMAS MILDNER
Bremen, Germany

defended on

June 21st, 2024

Supervised by

Prof. Dr. Rainer Malaka

External Examiner

Prof. Dr. Shruthi Sai Chivukula

DISSERTATION

of the University Bremen,
Submitted to Faculty 3
Mathematics and Computer Science
in fulfilment of the requirements for the degree of

Doctor of Engineering (Dr.-Ing.)

Mitigating Dark Patterns Through Responsible Design
Ethical Design Considerations for User-Centred Technologies

First edition.

University of Bremen,
Bremen, Germany, 2024

For My Family,

Thank you for bringing me up in an environment reminiscent of the privileges and education available to me; providing me with a moral compass that continues to guide me.

— & —

Katharina,

Your endless kindness continuously inspires me to be better. Your loving support allows me to look forward without worries, but excitement.

Acknowledgements

Before I began this journey, I believed that pursuing a PhD was supposed to be a one-person effort. Today, I know that this couldn't be further from the truth. Looking back, I was never alone and had the unbelievable fortune of a rich companionship who joined me for parts or even the entire adventure. I am wholly grateful for every single one of you.

My first thanks go out to you, Rainer, for giving me the freedom to choose my own path. Your feedback gave me the support of someone who shared the same beliefs about our responsibility with technology while helping me consider different angles for my work. Your trust and confidence enabled me to base my work on my core values and empowered me to always aim highest. Second, I would like to thank you, Shruthi. I am extremely happy that you accepted the role of my external examiner. Your work inspired mine in countless ways, and I am hopeful that we get to collaborate in the future. Third, I want to thank the entire digital media lab, its alumni, and my fellow students, for offering a space for enlightening conversations and exciting collaborations. While I am thankful to the entire team, I want to give special thanks to Caro, Daniel, Dasha, Evropi, Mehrdad, Nadine, Nima, Susanne, and Vino. Furthermore, Evgenie, Irmgard, Philipp, and Svenja, whose constant support is a foundation for our lab. Also, I want to thank the LSC Digital Public Health, as well as my Bachelor's and Master's students, whose great work contributed to my research. I wish you the brightest future!

I am still fascinated by the amazing people who engage in value-driven topics such as ethical design. During my PhD, I had the opportunity to work with such people, who dream big and work hard to make the world a better place. Jasmin and Paweł, you have always kept an open door and supported me through your immense methodological and technical knowledge. Colin, I remember reading your 2018 paper even before I began my PhD and feeling star-struck when meeting you at the first dark pattern workshop in 2021. A few years later, I have so much gratitude for our shared projects. I also want to reach back to where my academic journey began: In Dublin, where I met you, John, and later Phil. Not only did you help our paper's clarity through your great writing but it was often through early conversations with you that sparked new research ideas. I am very thankful for what I have learned from all of you, but I am even more thankful for our friendship.

Most importantly, I want to thank my parents. You created an environment fostering knowledge and culture of all kinds. More so, you taught me important values and equipped me with a moral compass. This also includes my Grandmother and late Grandfather. You have always given me a second place I can call home. Time with you never passed without lessons for life. Katharina, I cannot thank you enough for your patience and faith in me. Your compassion and kindness have been a beacon when I was in doubt. You constantly bring the best out of me. And, lastly, the person without whom I wouldn't be where I am today: Gian-Luca, our friendship is a pillar of constant support and joy. Many conversations began jokingly or ridiculously but led to fantastic projects. I could not have wished for a better roommate, companion, and friend. Thank you!

Bremen, May, 2024 — Thomas Mildner

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are my original work. This entails both textual as well as graphical content in support of this manuscript. Assistive and generative Artificial Intelligence (AI) tools have been used to improve the clarity, grammar, and spelling accuracy of my writing. The tools used were limited to DeepL[1], Grammarly[2], and ChatGPT version 3.5[3]. No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education. I hold the University of Bremen harmless against any third-party claims with regard to copyright violation, breach of confidentiality, defamation and any other third-party right infringement.

Faculty/Institute	Faculty 3: Mathematics & Computer Science
Degree	Doctor of Engineering (Dr.-Ing.)
Title	Mitigating Dark Patterns through Responsible Design — Ethical Design Considerations for User-Centred Technologies
Candidate (Id.)	Thomas Mildner (6089194)
Signature of Student	_____
Date	Monday 24th June, 2024

[1] https://www.deepl.com/translator
[2] https://www.grammarly.com/
[3] https://chatgpt.com/

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACM Association for Computing Machinery 18
AI Artificial Intelligence 4
CCPA California Consumer Privacy Act 7
CEC CUI Expectation Cycle 57, 151
CMA Competition and Market Authority 30
CPRA California Privacy Rights Act 30
CUI Conversational User Interface . . . . . 7, 81, 109, 151, IX, XI
DA Data Act 30
DMA Digital Markets Act 30
DSA Digital Service Act 7
EDPB European Data Protection Board 30
EU European Union 29, 82
FBM Fogg Behavior Model 68
FTC Federal Trade Commission 30
GDPR General Data Protection Regulation 7
GUI Graphical User Interface 24, 109
HCI Human-Computer Interaction . . . . . 5, 81, 109, 135, IX
LLM Large Language Model 78
OECD Organisation for Economic Co-operation and Development 30
SNS Social Networking Service . . . . . 5, 81, 99, 117, 135, 209, IX
TADP Temporal Analysis of Dark Patterns 35, 195
UCD User-Centred Design 17
UI User Interface 3, 99, 209
UX User Experience 33
VSD Value Sensitive Design 3

Abstract

Every digital interface is the result of intentions, incentives, and design philosophies. They can lead to user-friendly and empowering technologies or exploiting and manipulative ones that persuade users into engaging in actions they may regret later. The past decade of Human-Computer Interaction (HCI) research has explored the latter phenomenon, describing them as deceptive design strategies or "dark patterns". In this vein, studies have fostered a growing taxonomy of related instances in various domains, including, but not limited to, online shopping sites, digital games, and mobile applications.

Nonetheless, research gaps remain: Existing findings offer space for synthesised frameworks and open avenues for transdisciplinary work. Moreover, underlying mechanisms of dark patterns have yet to be studied to grasp their implications on users. Extracted knowledge can lead to the development of tools to understand dark patterns better and mitigate their effects.

Addressing these gaps, this thesis includes nine publications. These investigate the roots of dark patterns in design, their consequences for users, and opportunities through guidelines. Four of these focus on implications of Social Networking Service (SNS) platforms, two on Conversational User Interface (CUI) technologies, while three contribute to our general understanding of dark patterns in transdisciplinary contexts. By considering the perspectives of users, the thesis further explores how people develop expectations when engaging with these interfaces and studies how design supports or breaks their expectations. Based on these findings, the thesis promotes an alignment of design with their users' expectations through truthful representation of functional capabilities. Furthermore, it describes ethical caveats leading to unethical design or dark patterns if disregarded. Finally, the thesis follows design from its development into the real world, observing ethical implications once it leaves its creator's desk.

Drawing from a series of qualitative and quantitative research studies, included in this cumulative thesis, its contributions are threefold: It explores dark patterns in SNSs and adds domain-specific types to the related scholarship. It describes users' perception of dark patterns and spotlights difficulties in protecting themselves from nefarious SNS and CUI interfaces. Lastly, it contributes to design theory by revealing where dark patterns manifest and how responsible design can be used to mitigate them. Together, the contributions span the three angles — design, users, and guidelines — of the *Responsible Design Triangle*. This model reflects the interrelationships and dependencies between the angles.

Zusammenfassung

Jede digitale Benutzungsoberfläche ist das Ergebnis von Absichten, Anreizen und Designphilosophien. Sie können zu nutzungsfreundlichen und befähigenden Technologien führen, oder zu ausnutzenden und betrügerischen, die ihre Nutzer:innen verleiten, Handlungen auszuführen, die sie später bereuen könnten. In den vergangenen zehn Jahren hat das Forschungsfeld des Mensch-Computer-Interaktion (HCI) letztere Phänomene genauer untersucht und diese als betrügerische Designstrategien oder "Dark Patterns" zu beschrieben. Zugrundeliegende Studien haben seither zu einer wachsenden Taxonomie verwandter Fälle in diversen Bereichen beigetragen. Diese umfassen, sind aber nicht beschränkt auf, Online-Shopping Webseiten, digitale Spiele und mobile Anwendungen.

Dennoch bestehen Forschungslücken: Bisherige Ergebnisse bieten Raum für übergreifende Rahmenwerke und eröffnen Wege für transdisziplinäre Arbeiten. Ferner braucht es Studien, um die zugrundeliegenden Mechanismen von Dark Patterns zu untersuchen und ihre Implikationen auf das Wohlbefinden von Nutzer:innen zu verstehen. Diese Forschungsergebnisse können dazu dienen Werkzeuge zu entwickeln, um Dark Patterns aufzudecken und ihre Folgen zu mindern.

Um diese Lücken zu schließen, umfasst diese Dissertation neun Publikationen. Diese untersuchen die Wurzeln von Dark Patterns, ihre Konsequenzen auf Nutzer:innen sowie die Möglichkeiten wirkungsvoller Richtlinien. Vier von ihnen fokussieren sich dabei auf Soziale Netzwerke (SNS), zwei auf Conversational User Interface (CUI) Technologien, während drei zu unserem allgemeinen Verständnis von Dark Patterns beitragen und Wege für transdisziplinäre Kollaborationen bereiten. Mit Berücksichtigung der Perspektive von Nutzer:innen, erforscht diese Dissertation zudem, wie Menschen Erwartungen im Umgang mit Benutzungsoberflächen entwickeln und untersucht, wie Design ihre Erwartungen unterstützt oder auch enttäuscht. Basierend auf diesen Ergebnissen, stellt diese Dissertation Zusammenhänge zwischen Design und den Erwartungen ihrer Nutzer:innen dar und fokussiert sich auf die wahrheitsgetreue Repräsentation von funktionalen Kapazitäten. Darüber hinaus beschreibt sie ethische Fallstricke, die zu unethischem Design oder Dark Patterns führen können. Schließlich folgt die Arbeit Design entlang ihres Entwicklungsprozesses und beobachtet die ethischen Implikationen eines Designs, sobald es den Schreibtisch seines Schöpfenden verlässt.

Basierend auf einer Reihe qualitativer und quantitativer Studien, welche Teile dieser kumulativen Dissertation sind, leistet diese folgende wissenschaftliche Beiträge: Sie erkundet Dark Patterns in SNS und erweitert die zugehörige Forschung durch domänenspezifische Typen. Sie beschreibt wie Nutzer:innen Dark Patterns wahrnehmen und wirft ein Licht auf ihre Schwierigkeiten im Umgang mit bösartigen SNS und CUI Benutzungsoberflächen. Zuletzt trägt sie zur Designtheorie bei, indem sie aufdeckt, wo Dark Patterns auftreten und wie verantwortungsbewusstes Design verwendet werden kann, um ihr Auftreten und ihre Auswirkungen zu mindern. Zusammen umspannen diese Beiträge die drei Winkel — Design, Nutzer:innen und Richtlinien — des *Responsible Design Triangle*. Dieses Modell spiegelt die Wechselbeziehungen und Abhängigkeiten zwischen diesen Winkeln wider.

PART I

INTRODUCTION

CHAPTER 1

Motivation: Introducing the Responsible Design Triangle

By its nature, design steers users' perceptions to communicate its functionalities. This communication often happens nonverbally. For example, digital interfaces can utilise perceptible cues to highlight content to nearby people and, thus, capture their attention. Moreover, additional interface elements can be deployed to provoke specific actions, guiding users through complex User Interfaces (UIs). To foster effective communication, timely and transparent feedback is critical to inform about the consequences of interactions. People, who engage with the interface, interpret the given cues to construct plans for interactions (Norman, 2013). This plan is constructed around our expectations, which are guided by the design communicating its capabilities as extrinsic influence while drawing from intrinsic motivations and goals. Importantly, it is not in the hands of the designer to control the interaction in most cases after the design left their desk (Verbeek, 2006). Any inability to make informed decisions, in the advent of sufficient communication, raises questions about the design's responsibility. While in cases of bad design, people will simply be frustrated and turn away from any further interaction, in others, they may not even be aware of any harm done at the time when their choices become actions. Designs, which exploit how they communicate their capabilities to manipulate user choices against their best interest, carry strong ethical implications that require consideration.

The main focus of this dissertation lies on unethical designs that share a common dictation of actions by leveraging users' choice architecture without concerning their best interests as well as ethical design empowering users to engage with technologies autonomously. The former encompasses so-called "dark patterns"[1], which is the concern of a large portion of publications included in this dissertation. Dark patterns describe a particular group of unethical designs, which limit peoples' autonomy by restricting or obfuscating available choices and, thereby, decreasing their ability to make informed decisions (Mathur *et al.*, 2021), measured by their effect and not necessarily their intention (European Parliament, 2022). Throughout this dissertation, I will attempt to use the terms (un)ethical design and dark patterns as precisely as the given contexts of the following sections and chapters afford.

To explain what I[2] mean when writing about unethical design, I want to outline my understanding of the term. First, the idea of infusing design with ethics is not new (Findeli, 1994; Devon and Van de Poel, 2004; Shilton, 2013), with a cornerstone residing in Value Sensitive

[1] The term "dark pattern" has been critiqued to reinforce discriminatory language ("dark" referring to evil or bad). However, at the time of writing this thesis, no agreement on an alternative term has been made. For further reading on the term and its background, I refer to Section 2.1 of this dissertation.

[2] In this thesis, I will use first-person singular pronouns to refer to my personal opinions, arguments, and statements while using first-person plural pronouns, when respectfully referring to the publications included in this thesis, which are the product of multiple authors' work.

Design (VSD) (Friedman *et al.*, 2013) which builds on the idea of user autonomy (Friedman and Nissenbaum, 1997). Following these concepts, ethical design should be responsible when guiding the user through its practicalities by providing sufficient information to allow informed decision-making and empowering autonomous interactions. Ethical design should also be sustainable, reflecting its individual and societal impacts (Verbeek, 2006). This is in line with *ethics by design* principles for Artificial Intelligence (AI) (Mittelstadt, 2019), which find support from regulatory bodies (European Commission, 2021) and academia (d'Aquin *et al.*, 2018).

From a normative lens, the primary purpose of design can be diluted to delivering solutions for problems to its users. Naturally, people using a design could assume that it is their problem being solved. However, some designs present examples that question this stance as they seem to benefit the party that implemented it. Interested in the design of captivating systems, Seaver (2015) offers a metaphor for the unethical implementation of algorithmic recommender systems by linking traps and their anthropological history as human nature. From an ethical standpoint, the metaphor is powerful as it detects a victim in the trapped person and a perpetrator who set the trap. Peter-Paul Verbeek has contributed further valuable and thought-provoking work to the discourse of ethical and responsible design (Verbeek, 2005; Verbeek, 2006). Particularly interesting to my work is his theory of technology mediation. In sum, the theory places different roles in the designer, technology, and user. The designer *inscribes* their ideas into the technology, which *mediates* its functionalities, and is finally *interpreted* by the user. This theory displaces the ability to deploy moral or ethical values into technologies from the designer, as much of the interactions are not under their control. Once the design enters the real world, its implications are experienced by users, regardless of original intentions. Nonetheless, it remains within the designer's responsibility to create technologies in such a way that they are harmless to users' well-being.

To describe unethical design strategies in online interfaces, Harry Brignull (Brignull, 2010) described so-called "dark patterns" as design strategies that steer, covert, obfuscate, or in any way manipulate user decisions. While there is also a victim to dark patterns, similar to Seaver's trapping metaphor, the role of the perpetrator could be connected to the practitioner, as the originator of the design, but is not as apparent as there is no necessity for intent. The origin of the term in design theory concerns the design itself, detachable from any actors. Unlike the intentions behind trapping something, the opaqueness of consequences and potentially harmful outcomes are, after all, what makes a dark pattern "dark".

Dark patterns have been identified in a range of (digital) technologies and media including websites (Gunawan *et al.*, 2021; Gray *et al.*, 2024a) — online shopping sites in particular (Mathur *et al.*, 2019) — and mobile applications (Di Geronimo *et al.*, 2020; Gray *et al.*, 2018). Many trick users into purchasing unwanted products or hide additional costs within (e-)commerce settings (Mathur *et al.*, 2019). Others target users' personal data and, thus, risk infringing peoples' privacy (Gray *et al.*, 2021b). Figure 1.1 displays an example for a *Nagging* dark pattern that restricts user choices through reoccurring notifications. Overall, there are various motivations behind deploying dark patterns (Gray *et al.*, 2018), but they all harm users somehow. Research on the topic has fostered a growing collection of individual types and strategies that fit underneath the umbrella term "dark pattern", as further studies spotlighted

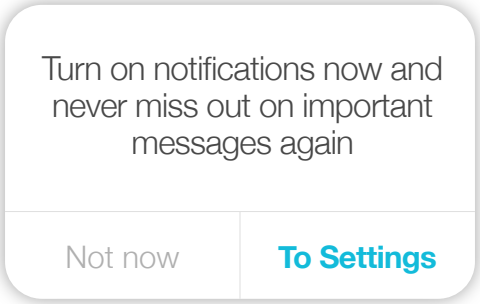


Fig. 1.1 Example of a *Nagging* dark pattern first proposed by Gray *et al.* (2018). The pattern prompts users to turn on their notification but does not give any options to dismiss the notification permanently.

difficulty among users to recognise dark patterns and avoid them (Di Geronimo *et al.*, 2020) — even if properly informed on the matter (Bongard-Blanchy *et al.*, 2021).

1.1 The Responsibility of Common Practice

To understand the difficult spot designers may find themselves in when designing interfaces, we have to reflect on the responsibility of steering users' perceptions to understand a design's capabilities. In a nutshell, design could be considered a language to communicate possible interactions effectively: A cup may be too hot to the touch, but an attached handle allows me to drink the freshly brewed cappuccino inside without burning my hand. Following traditional terms of Human-Computer Interaction (HCI), design *affords* interactions, optimised by well-placed *signifiers* that make them discoverable. Optimally, a design communicates responsive *feedback* for users to understand the results of their interactions (Norman, 2013). I refer back to a book first published almost fifty years ago to set a normative lens for design, in which its nature to steer people's perception is neither good nor bad. It is normal. But, it yields the responsibility to respect users' autonomy, making design challenging. Fortunately, designers can rely on *design patterns* as common solutions to universal design problems (Alexander *et al.*, 1977).Unfortunately, common practice is not always best practice — which changes depending on the perspective (Gray and Chivukula, 2019; Chivukula *et al.*, 2023).

As reflected in various studies contributing to this thesis, I found Social Networking Services (SNSs) to be particularly interesting in this regard. SNSs include platforms such as Facebook, Instagram, or X[3], but also video streamers like YouTube, TikTok, and Twitch. Their frequently described benefits regarding social connectedness (Ahn and Shin, 2013; Sinclair and Grieve, 2017) contrast studies spotlighting harms on users' well-being (Shakya and Christakis, 2017; Twenge *et al.*, 2018). As free-to-use services, they partially depend on selling their users' data to advertisers and other third parties (Enders *et al.*, 2008; Sindermann *et al.*, 2024). Deployed UI tricks that coerce users into consent raise questions about unethical design mechanisms that restrict informed decisions. Adjacent work shows (Brignull, 2023; Gray *et al.*, 2018)

[3] During the writing of this thesis, the social network formerly known as Twitter changed its name to X in the year 2023. Publications included in this dissertation refer to the platform as Twitter as they were published before this change.

that unethical design can swiftly result in unaccounted outcomes for the user (Hansen and Jespersen, 2013) instead of guiding them through functionalities as they expect. The lack of responsibility in the design of digital interfaces, resulting in harm for many users, motivated my research and, ultimately, this thesis.

1.2 Understanding the Actors — Framing the Research Questions

To effectively counter dark patterns, it is paramount to understand where and how they manifest. Aside from usability criteria, we should consider the impact of our designs by assessing their consequences and empowering users to make informed decisions. Design needs to be sustainable both for the service provider, to continue their products, and for the user, to benefit from their choices. Following HCI traditions, this thesis approaches the topic from three main angles. First, I investigate the design of systems and applications to understand how dark patterns occur in UIs. Second, I consider the users' perceptions and how they recognise and engage in unethical practices. Finally, I promote the relevance of design guidelines to inform the development of technologies together with ethical design considerations. Based on these three angles, this thesis addresses the high-level research question:

RQ: How can the responsibility of designs and their impacts be distributed between actors to protect users from deceptive, unethical design practices and dark patterns?

The publications included in this thesis span these three angles, providing answers to the high-level as well as more granulated research questions. At the beginning of this thesis, I provoked the idea that it is in the nature of design to steer users' perceptions. Following this train of thought, a good designer could be defined as *a person with a deep understanding of people's cognition and perception; who is able to effectively alter peoples' choice architecture by precisely creating things that convey a planned goal which is easily carried out.* Although this definition bridges the suggested notion of steering perception and usability, it leaves room for exploitation. Designing simple objects, such as a cup's handle, with the purpose of shielding its users from burning themselves may not pose ethical implications. However, design that binds users to services, especially when legally binding agreements are established, must enable the user to make informed decisions. Dark patterns, however, are designs that misguide users into making decisions that they may not have done if made aware of any potential repercussions. To adhere designers to certain ethical caveats, I suggest extending my previous definition towards the following:

A good designer is a person with a deep understanding of people's cognition and perception; who is able to effectively alter peoples' choice architecture by precisely creating things that convey a planned goal which is easily carried out, ***reflecting implications and empowering informed decision-making.***

The studies and experiments included in this cumulative thesis extend the contemporary discourse by transferring identified dark pattern types to ubiquitous technologies such as

SNSs and Conversational User Interfaces (CUIs). Thus, we cultivate an understanding of the extensive deployment of dark patterns in daily interactions while identifying gaps that we were able to fill through additional empirical studies. Prior work has considered established harms of dark patterns, for instance, in e-commerce (Brignull, 2010; Mathur *et al.*, 2019) and consent banners (Gray *et al.*, 2021b). These domains describe tensions between commercial incentives built on economic drivers to maximise profit (or approaches to collect users' personal data motivated by surveillance capitalism (Zuboff, 2023)) and users' best interests and their agency to make informed decisions.

While this strand of work has progressed to describe instances of dark patterns in various environments, the users' perspective has only been considered through a few studies (Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021). Continuing this important effort, the work featured here studies users' abilities to recognise dark patterns to identify ways to protect them. Thereby, our work adds to prior findings suggesting a difficulty among users to effectively distinguish dark patterns from harmless interfaces. Moreover, we build on the five high-level dark pattern characteristics proposed by Mathur *et al.* (2019), later extended to six characteristics (Mathur *et al.*, 2021), and demonstrate a scale to easily assess the malice of interfaces, aiding regulative bodies to protect users where harmful designs are detected. Moreover, by equipping regulatory bodies with the relevant knowledge, users can be provided the necessary protection where they cannot protect themselves.

While some of the service provider's decisions can be explained through mal-intent, particularly in online environments, others may have mistakingly adopted pre-existing but dark pattern-infested interfaces which they may have simply copied without realising its ethical implications. Here, templates for consent banners that require compliance with, for instance, Europe's General Data Protection Regulation (GDPR) (Commission, 2016), its recent Digital Service Act (DSA) (European Parliament, 2022), or the California Consumer Privacy Act (CCPA) (California State Legislature, 2018), added difficulties for web developers who did not previously require the necessary legal expertise. Yet, they demanded quick responses from big and small online services alike (Gray *et al.*, 2021b). Naturally, those unfamiliar with the legal requirements followed common practices — which, partially, contained numerous dark patterns (Maier and Harr, 2020; Gray *et al.*, 2021b). Although dark patterns were not necessarily deployed intentionally, the harm expressed to their users remains unchanged. My work considers these difficulties and elaborates on the ethical caveats of design while utilising knowledge about human cognition, particularly cognitive biases, to support practitioners in reflecting on the impact of their design.

Investigating dark patterns from three angles — design, users, and guidelines, our studies emphasise the relevance of design integrity and the responsibility of practitioners. To approach the aforementioned high-level research question systematically, this thesis first addresses three deductive research questions based on the contributions of the included publications. Each explores a particular aspect of the high-level research question while the union of their individual answers will offer implications for the high-level research question.

RQ1: ***How can design be used to create and break expectations that lead to dark patterns?*** Research has recorded dark patterns in a range of interface designs ranging from online shopping sites (Mathur *et al.*, 2019; Gray *et al.*, 2018) to mobile applications (Di Geronimo *et al.*, 2020; Gunawan *et al.*, 2021). Generally, dark patterns can be encountered wherever users make decisions. They emerge as strategies that manipulate graphical and text-based elements or elevate certain choices over others without the users' best interest in mind. Some decisions are positively encouraged, while others are restricted. From a design and HCI perspective, this opens the research question of how design creates expectations in users and how breaking them results in dark patterns.

RQ2: ***To what degree are users able to identify dark patterns in interfaces to safeguard themselves?*** Only through understanding users' ability to recognise dark patterns and safeguard themselves are we able to formulate effective countermeasures against dark patterns. Although the landscape of recorded dark patterns across domains has gotten relatively dense, only a few studies have considered the users' point of view. In this regard, studies show that users are able to differentiate between interfaces containing dark patterns and those that do not while suggesting difficulties among study participants (Bongard-Blanchy *et al.*, 2021; Di Geronimo *et al.*, 2020). The second research question continues these efforts while taking a more fine-grained approach to investigate users' perspectives.

RQ3: ***Which ethical design considerations are necessary to avoid the implementation of dark patterns?*** Work in the field of HCI can look back at various design guidelines and best practices aiding designers' work. Yet, only a few consider ethical caveats and take responsibility to empower informed decision-making. While these topics are not necessarily new (Friedman and Nissenbaum, 1997), practitioners, such as designers[4], continue to use practices that pose potential ethical conflicts (Brynjarsdottir *et al.*, 2012). Gaining a better understanding of the design aspects that lead to harm becomes a relevant step in the combat against dark patterns. Answers to this research question can eventually inform ethically aligned design.

1.3 The Responsible Design Triangle

Building on prior research, the contributions of this thesis are threefold: Firstly, it expands our theoretical knowledge about the interactions through which dark patterns manifest. To that end, our work expands contemporary dark pattern typologies with novel types captured in SNS interfaces and problematises instances in CUI interactions. Secondly, the thesis provides a deeper, empirical understanding of users' ability to effectively recognise dark patterns and differentiate between harmless and problematic interfaces. Through these studies, we demonstrate the necessity to take the burden of safeguarding away from users and point toward a need for better regulatory protection. Thirdly, the thesis delves into the intricate

[4] Throughout this thesis and included publications, I mention practitioners and designers frequently. I refer to practitioners as professionals involved anywhere during the development of digital interfaces. These include designers but also extend to other roles like software developers and engineers, project leads, and executives. Part of this group, designers have a special role in conceptualising and creating interactions while optimising usability and user experience.

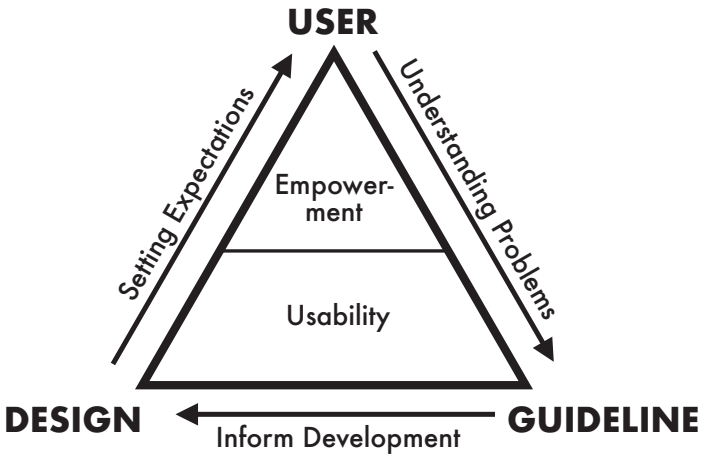


Fig. 1.2 The *Responsible Design Triangle* is a user-centred framework connecting three angles to enable responsible design. On top, most importantly, is the user. It is up to a design to set realistic expectations and communicate the consequences of interactions transparently. If problems arise, they must be understood while severe issues are distilled into guidelines. Guidelines should inform the design through sensible approaches that allow practitioners to develop responsible interactions. Internally, the triangle holds user empowerment over usability.

relationship between human cognition and dark patterns. By investigating the exploitative factors enabling dark patterns, we identify ethical caveats and provide a theoretical foundation to conceptualise how responsible design can mitigate harming interactions.

Situating these three contributions as angles within the same scenario, I propose the *Responsible Design Triangle*, shown in Figure 1.2. Contributing to the domain of HCI, particularly ethical design and dark pattern research, this framework connects *design*, *users*, and *guidelines* to inform ethically *good* design. Of course, good design is subjective; for this thesis, I define the term by drawing from my previous definition of a good designer: *Good design utilises peoples' cognition and perception to effectively alter their choice architecture to convey a planned goal, which is easily carried out, while implications are transparent and informed decision-making is empowered.* Within the *Responsible Design Triangle*, I define the three angles as follows:

Design is the embedded purpose or intent for any *thing*, perceivable for its users to present a solution for a specific problem. This thesis mainly considers design within the context of digital technologies. Although included work also partially considers practitioners' and designers' views, most studies focus on artefacts.

Users are first of all humans who engage with (digital) technologies through their design. Guided by their perceptions of the design, users develop expectations, formulate goals, and carry out actions.

Guidelines are documented recommendations and practices that follow certain ideals to achieve subsequent goals. As this thesis is written in the context of HCI, I will mainly consider design-related guidelines that prompt considerations for designers. However, policies and regulations would also fit this angle's perspective.

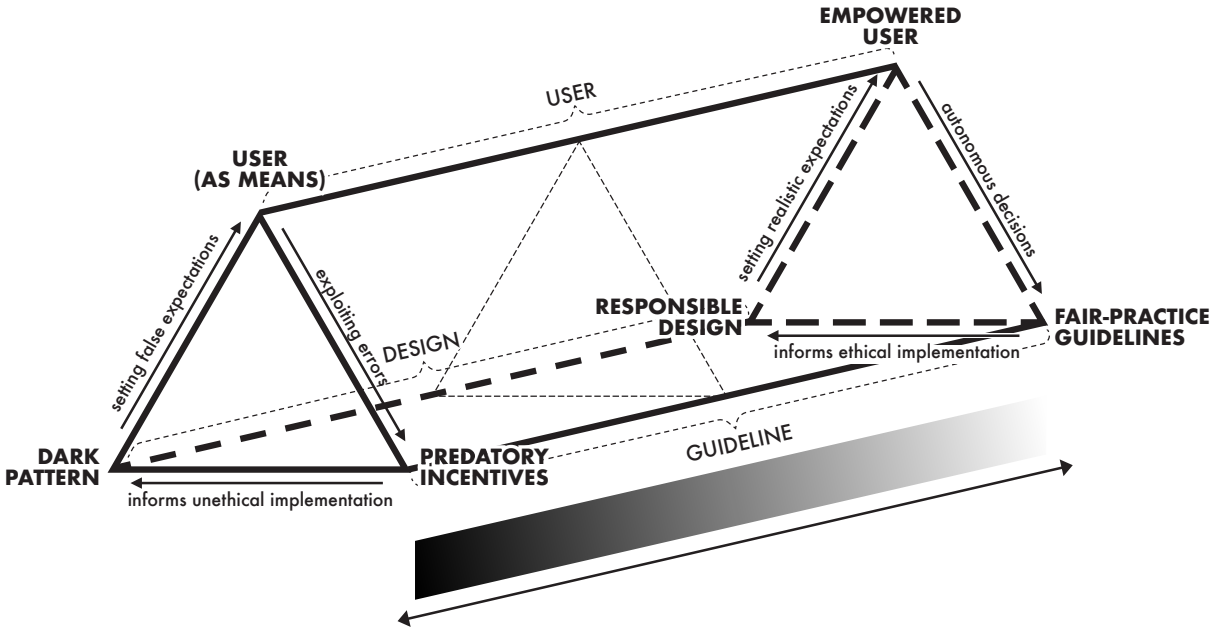


Fig. 1.3 This diagram visualises the *Responsible Design Triangle* within a three-dimensional space around the axes for *design*, *user*, and *guideline*. Moving forward, toward "darkness", the diagram illustrates how predatory incentives can override users' autonomy and lead to dark patterns, establishing false expectations. Consequently, these lead to erroneous interactions that service providers can exploit to maximise manipulative goals. Further back, on the other end, is the brighter alternative. Here, the user is in the focus as fair practices guide responsible design.

In line with Verbeek's theory of technology mediation (Verbeek, 2005), design is responsible for carrying realistic expectations about interactions. This entails a transparent communication of system capabilities and consequences, granting users autonomy to make informed decisions. In HCI, research has access to a variety of tools to investigate design in this regard. Qualitative interface analysis and quantitative user studies, for instance, can provide valuable insights into the presence of dark patterns that misguide users while revealing their experiences. Moreover, automatic processes can be described to detect problematic interfaces as more knowledge about dark patterns is collected.

Through sharing experiences with interactions and their consequences, users can help understand the resulting problems. However, they need to be able to articulate and report their experiences to allow any assessment of potential design issues, including the presence of dark patterns. These kinds of insights are precious for informing ethical design guidelines. Qualitative studies with developers, users, and experts, for instance, in focus groups or interview settings, can provide in-depth information about individual expectations toward interfaces. The findings could then be used for interface analysis of digital technologies to further detail the status quo of contained problematic design and inform design guidelines aiding practitioners' work. Returning design considerations to the design angle, these guidelines complete the three angles of the triangle.

Importantly, I do not view design through a binary lens as either a responsible design or a dark pattern. While deceptive and manipulative design attributes can be ascribed to dark patterns, nudges, as defined by Thaler and Sunstein (2008), should be evaluated depending on

their context. Work on persuasive design offers various examples in this regard (e.g. Alexandrovsky *et al.*, 2021). Furthermore, certain situations require fast actions and cannot afford extensive reflection. During a medical emergency that requires a cardiac massage without a medical expert present, it would be responsible for a public defibrillator to prioritise that its user quickly understands how to apply it on a person in need without them understanding all the consequences. Alternatively, *Nagging* dark patterns (see Figure 1.1) can remind users to engage in certain actions repeatedly, keeping them informed about (un)available choices, but take away their agency to permanently dismiss future prompts. Considering these examples, user empowerment is not always realised through autonomy but also through foresight that enables responsible actions. Nevertheless, this thesis mainly concerns everyday situations where users should be given time to reflect on their choices.

Situating the *Responsible Design Triangle* within the continuum between ethical and unethical design practices, Figure 1.3 offers examples on both ends. Drawing on the ethical mediators proposed by Gray and Chivukula (2019), the diagram exemplifies possibilities for dark patterns to occur when predatory incentives, or other bad intentions (e.g. unethical, commercial exploitations, blatant lies, or thievery) become main drivers to inform design instead of following fair, human-centred, or design guidelines that advocate design ethics. The negative and positive ends of this continuum function as examples. I chose the included terms not as immutable but as pointers to the possible root cause for unethical design, knowing that certain incentives can lead to good practice instead of dark patterns and well-meant fair practices can result in unwanted persuasion. Moreover, alternatives to unethical design are growing aiming to benefit service providers without disadvantages to its users. In Figure 1.3, the term *Fair Practice Guidelines* is meant to capture them all but could be replaced with any functioning guideline that upholds user autonomy and empowerment. The most relevant distinction between both ends of this continuum is the user angle, or better, how it is approached. In the sense of Immanuel Kant's categorical imperative (Kant, no date), a user should never be "used" to fulfill one's means to an end, but, instead, should be seen as the means itself. The purpose of design should be to aid its users in achieving their goals, solving their problems, and not those of a service provider. However, design is complex and full of constraints that require attention. As with most attempts to abstract complex topics, the diagram in Figure 1.3 is limited to considering only fragmented examples. In sum, if service providers disregard user autonomy to accomplish their goals, design can devolve into exploitative dark patterns that set false expectations, resulting in erroneous user interactions.

While I hope that our work aids regulatory bodies in considering establishing policies where design, users, and guidelines fall short of sufficiently protecting against dark patterns, my expertise is within the field of HCI, and so is this thesis. Therefore, the *Responsible Design Triangle* foremost considers guidelines where my competence is strongest. The model makes an appeal to the responsibility of designers to leverage expectations and create truthful interactions that users can assess effectively before carrying them out. Where this is not the case, and harm is imposed on the user, it problematises guidelines that incentivise ethical design. Consequently, as highlighted in Figre 1.2, the model considers usability as a basis for good design but places user empowerment above it to ensure informed decision-making.

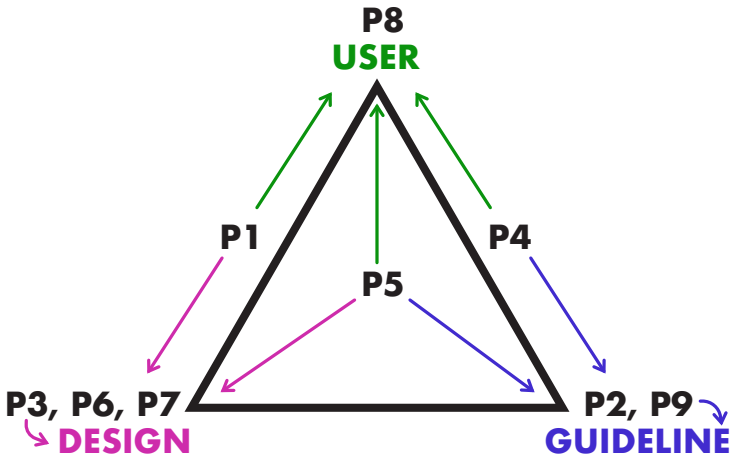


Fig. 1.4 The Figure positions the nine publications included in this thesis within the Responsible Design Triangle. Their location corresponds with the three angles to highlight their contributions to the model and this thesis.

1.4 Publications Included in this Thesis

As a cumulative thesis, this dissertation consists of individual publications in the field of HCI. In total, this thesis spans nine publications. While a large portion of included publications focus on dark patterns in SNS contexts (P1, P3, P4, and P8), this thesis expands this scope to answer its research questions. This is partly because the findings from these publications carry implications relevant to the general dark pattern discourse as well as responsible design. But also because additional contributions explored how dark patterns manifest in CUIs (P2 and P5) and revealed underlying mechanisms that enable dark patterns (P9). Also, I contributed to technologically ambiguous research that aims to enhance transdisciplinary work on dark patterns in the future, including regulatory efforts (P6 and P7). The *Responsible Design Triangle* synthesises the accumulated findings of the nine publications into a comprehensive model. The individual works thereby offer implications for individual or multiple angles of the model. To highlight individual contributions, Figure 1.4 positions the included publications within the triangle.

At the time of submitting this thesis, publications P1-P6 were published in relevant conference proceedings, whereas publications P7-P9 were either still under review or finished, but not yet submitted as submission dates were still upcoming. They were, however, published at *arXiv* as pre-prints. Because all publications are the result of collaborative efforts, I will state my personal contributions for each following the ICMJE guidelines for authorship criteria (International Committee of Medial Journal Editors, 2023). If I was the first author of a publication, I was responsible for submitting the final version of each paper. Where I was co-author, I approved the final version of the paper. I take full responsibility for all aspects of the included work and acknowledge that I followed good research practices and ethics appropriate to the conducted studies and their evaluation. The following publications contribute to this cumulative thesis and are included in Part II. I added labels to each publication to highlight their contributions to the angles of the *Responsbile Design Triangle*.

DESIGN USER

P1 **Mildner, T.** and Savino, G.-L. (2021). Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7). Association for Computing Machinery. https://doi.org/10.1145/3411763.3451659

My contribution to this paper was the study design, data collection, and analysis for both studies. I interpreted the results and wrote the manuscript, which I revised and submitted for the final publication.

GUIDELINE

P2 **Mildner, T.**, Doyle, P., Savino, G.-L. and Malaka, R. (2022). Rules Of Engagement: Levelling Up To Combat Unethical CUI Design. 4th Conference on Conversational User Interfaces, 1-5. https://doi.org/ 10.1145/3543829.3544528

My contribution to this paper was the analysis of related work, discussing their implications and the development of a process to assess unethical design in Conversational User Interfaces. I drafted the manuscript and revised it before the final publication. This publication received an honourable mention award (top 5% of papers).

DESIGN

P3 **Mildner, T.**, Savino, G.-L., Doyle, P. R., Cowan, B. R. and Malaka, R. (2023). About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1-15. https://doi.org/10.1145/3544548.3580695

My contribution to this paper was the design and supervision of the study, the qualitative analysis, and the interpretation of the data leading to the development of dark patterns specific to social networking services. Since qualitative analysis is best done between multiple coders, a co-author assisted in this process. I administered and structured the analysis. I drafted the manuscript and revised it before the final publication.

USER GUIDELINE

P4 **Mildner, T.**, Freye, M., Savino, G.-L., Doyle, P. R., Cowan, B. R. and Malaka, R. (2023). Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. Proceedings of the 2023 ACM Designing Interactive Systems Conference, 2362-2374. https://doi.org/10.1145/3563657.3595964

My contribution to this paper was the study design, data collection, and analysis of the results. Further, I interpreted the data to develop a process for assessing dark patterns in user interfaces. I drafted the manuscript and revised it before the final publication.

DESIGN USER GUIDELINE

P5 **Mildner, T.**, Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R. and Niess, J. (2024). Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1-18. https://doi.org/10.1145/3613904.3642542.

My contribution to this paper was the design and supervision of the study, the qualitative coding and conduction of reflexive thematic analysis on the transcripts as well as the creation of the CUI Expectation Cycle model. Best practice for qualitative research states that thematic analysis should be done between researchers, including discussions. I administered the analysis together with co-authors of this paper. I drafted the manuscript and revised it before the final publication.

DESIGN

P6 Gray, C. M., Santos, C., Bielova, N., and **Mildner, T.** (2024). An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1-22. https://doi.org/10.1145/3613904.3642436.

My contribution to this paper was in defining high-, meso-, and low-level types of dark patterns for the presented dark pattern ontology. Moreover, I contributed three case studies demonstrating the extension of the ontology with new dark patterns, the findings' interpretation, and the writing of the manuscript with an emphasis on the extension of the ontology. I approved the final draft before it was submitted by the first author.

DESIGN

P7 Gray, C. M., **Mildner, T.** and Bielova, N. (2023). Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow". http://arxiv.org/abs/2309.09635 – *At the time of submitting this thesis, this paper was finalised but in submission.*

My contribution to this paper was ideating dark patterns as temporal processes as well as the coding and analysis of the case study. I contributed to the findings' interpretation and the writing of the manuscript with an emphasis on the introduction, results, and discussion of this work. I approved the final draft before it was submitted for review by the first author.

USER

P8 **Mildner, T.**, Savino, G.-L., Putze, S., Malaka, R. (2024). Finding a Way Through the Social Media Labyrinth: Learning From User Perspectives. http://arxiv.org/abs/2405.07305 *At the time of submitting this thesis, this paper was finalised but in submission.*

My contribution to this paper was the design and conduction of the study, data collection, and analysis. I contributed to the interpretation of the data and wrote and revised the manuscript before submitting it for review.

GUIDELINE

P9 **Mildner, T.**, Inkoom, A., Malaka, R., and Niess, J. (2024). Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Deceptive Design Patterns. http://arxiv.org/abs/2405.07378 *At the time of submitting this thesis, this paper was finalised but in submission.*

My contribution to this paper was the design and conduction of the study, the qualitative coding of the data, and the interpretation of our codebook. Through discussions with a co-author, I developed the Relationship Model of Cognitive Biases and Dark Patterns. I drafted the manuscript and revised it before submitting it for review.

1.5 Outline of the Thesis

Following common practices for a cumulative thesis, this manuscript is split into two main parts. The first part presents an introduction, laying out the foundation of this work and bridging the contributions of the included publications. I divided it into seven chapters. This first chapter begins with my motivation to pursue this research and conceptualises the different angles from which the included publications are to be seen. More precisely, the *Responsible Design Triangle* connects these contributions, situating them within the relevant research agendas to respond to the research questions this thesis seeks to answer. Throughout the thesis, I will continuously reference the angles (design, users, and guidelines), by allocating research streams and my publications toward them. Chapter 2 paves the way to understanding the background of dark patterns and contemporary scholarship that has kept (and still keeps) me engaged in this research field. To this end, the chapter delves into research that laid the foundation for work on dark patterns, spotlighting relevant literature in the peripheral of my research. As the field of dark patterns research is in its relative infancy, with fundamental work sprouting just over a decade ago, the chapter follows academic efforts chronologically, consistent with the three main themes of this dissertation.Collectively, the first two chapters set the stage for works included in this thesis. The Chapters 3, 4, and 5 revisit the aforementioned research questions respectively. I decided to follow the *Responsible Design Triangle* in this order — first design, second users, and third guidelines — to mirror the course of adjacent literature as well as a certain chronology within my contributions to the field. Each of these

three chapters will answer one research question by drawing from included publications within contemporary research. After I offer answers to my research questions, Chapter 6 then ties the work together by discussing the high-level research question concerning the distribution of responsibility between actors regarding designs and their impacts to protect users from deceptive, unethical design and dark patterns. Moreover, the chapter presents important limitations of this thesis and offers an outlook into possible, future directions. Finally, concluding the first part of this thesis, Chapter 7 summarises all previous chapters.

As a cumulative thesis, the second part then presents the nine publications comprising this thesis, as listed in Section 1.4. Importantly, all manuscripts are unaltered and reflect the final publications as submitted to the respected conferences. The same is the case for papers that were still under review at the time of submitting this thesis but were published via *arXiv* as author versions. Each paper is introduced, including its full citation as well as digital object identifier (DOI), linking the work to the official online version of each publisher.

CHAPTER 2

Understanding Dark Patterns

More than a decade of transdisciplinary efforts have fostered a growing understanding of deployed design patterns that harm users in various scenarios and environments — in their midst are dark patterns. While a majority of the included publications in this thesis investigate design practices in SNSs, the main focus concerns general harmful strategies used by dark patterns, how they are deployed, and influence their users' decision-making. An important aspect thereby lies within the question of responsibility and how to develop interfaces that afford transparent and reflected interactions. In this regard, Figure 1.2 visualises the *Responsible Design Triangle*, where the three angles consider the perspectives of design, users, and guidelines.

After explaining the term "dark pattern", its origin, and current controversy in Section 2.1, this chapter follows the three angles to provide a thorough theoretical and practical background while drawing from related work for each section. Section 2.2 offers an overview of dark pattern scholarship and its presence in various digital interfaces. As most dark pattern research has been conducted for describing instances in different interfaces, it is the most substantial of this chapter. Section 2.3 then delves into user-centred studies, offering insights into users' ability to identify and avoid problematic interfaces and designs. Finally, Section 2.4 revisits ethically driven concepts, guidelines, and regulatory countermeasures that aim to restrict otherwise harmful effects of dark patterns. Each section is concluded by a sub-section that outlines implications for each angle, placing it within the general context of this thesis. Finally, Section 2.5 briefly summarises this chapter.

2.1 The Term

To understand dark patterns, it is pertinent to also understand its connection to Alexander *et al.'s* book "A Pattern Language" (Alexander *et al.*, 1977). Although the book mainly focuses on practised architecture and design paradigms, the usage of (design) patterns has transcended to describe digital interfaces as well. Summarised, the authors describe a pattern as a reliable solution to a frequent problem. Originating in architecture, Alexander *et al.* (1977) initialised the corpus with 253 patterns to formulate a language with the intention of guiding architects, urban planners, and designers in creating more humane, functional, and harmonious spaces. Patterns are often interconnected and can be combined to develop designs in various environments, such as buildings or office spaces. Furthermore, Alexander *et al.* (1977) emphasise the importance of addressing human needs and cultural context in design — quite similar to common HCI practices such as User-Centred Design (UCD).

From a development perspective, relying on patterns makes sense. It benefits efficient production through available options and common practices instead of running expensive user studies to find suitable interactions. In digital technologies, UI elements, such as the established close button represented by an "X" to exit programs and frames, are omnipresent on desktop devices across operating systems. Similarly, the two-finger pinch gesture has been established as the interaction to zoom in and out in touch-based devices — again across operating systems. From a user perspective, the reliance on patterns makes likewise sense. It allows them to only memorise single interactions across applications for the same or similar actions. Another advantage of patterns is the ability to address problems with adverse effects. In software programming, "anti-patterns" (Koenig, 1998) are used to describe the opposite of efficient solutions when pointing toward technical inefficiencies and severe problems that may result in vulnerable programs. Importantly, dark patterns should not be confused with anti-patterns as they describe distinct issues.

The term "dark pattern" was first introduced by Harry Brignull in 2010 and describes "*tricks used in websites and apps that make you do things that you didn't mean to, like buying or signing up for something*" (Brignull, 2010). In the same year, Conti and Sobiesk, 2010 studied UI elements that would fit Brignull's definition well. But instead of referring to dark patterns, the authors opted to describe "malicious interface design techniques" defined as practices where "[s]ome interface designers deliberately violate usable design best practices in order to manipulate, exploit, or attack the user" (Conti and Sobiesk, 2010). Yet, "dark pattern" as a term sustained and became canonical among researchers.

However, at the time of writing this thesis, however, a new discussion emerged questioning the term to describe coercive, deceptive, and manipulative design strategies. Although initiated by Brignull, he reconsidered his choice of words after voices emerged criticising "dark patterns" due to the potential posing racial bias the term poses, arguing that the term "dark" can be associated with "bad". In this vein, the Association for Computing Machinery (ACM) has administered a call to avoid using the term, placing it on its "Words Matter" list (Association for Computing Machinery, 2023). Instead, they offer possible alternatives, such as "deceptive/manipulative design" or "deceptive/manipulative pattern". Yet, the opposing side which supports the original term (Obi *et al.*, 2022) makes an effort to justify the retention of "dark pattern", stating that practitioners' intent is not the core issue but the harmful consequences of interface design, which is eventually experienced by end-users. Thereby, they distance themselves from any connotation to "bad" intentions and suggest that "dark", instead, refers to hidden features not obvious to the user, who, thus, falls victim to the unethical practices. Additionally, proposed alternatives limit the scope to deceptive or manipulative practices, whereas "dark pattern" describes a range of strategies limiting users' ability to make informed decisions.

The discourse has been echoed by the community (Gray *et al.*, 2023e) with the aim to find a fitting term that describes the range of strategies currently reflected by "dark patterns". "Damaging design" (Sinders, 2022) has been proposed as a suitable alternative by collecting strategies that harm users, regardless of any underlying mechanisms. However, only limited work (Monge Roffarello *et al.*, 2023) has adopted this alternative at the time of writing this thesis.

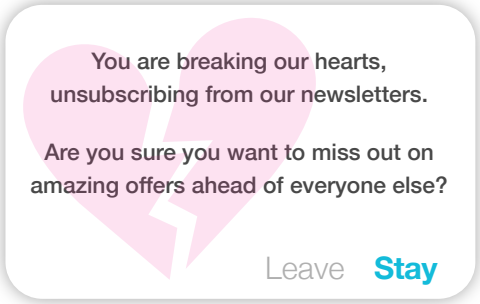


Fig. 2.1 Example of a *Confirmshaming* dark pattern. Emotionally manipulative language is used to shame users for controlling their decisions.

Fig. 2.2 Example of a *Playing by Appointment* dark pattern. Daily quests demand players to return regularly to progress in some games.

As this discourse remains ongoing, I will use the original term “dark patterns” in this thesis for continuity in line with past work and to avoid confusion arising from shifting terminology. I acknowledge and recognise the potential issues pertaining to the term. However, through my research, particularly the user studies, I understand that the harm is the result of obfuscated consequences users are often (kept) unaware of when engaging with interfaces. Thus, any incentives and intentions of practitioners become secondary, while the protection of users should remain the focus of this research and further regulation.

2.2 Dark Patterns in User Interfaces

Although the future term for dark patterns may be undecided, research has made a monumental effort to catalogue various unethical and harmful practices into a growing taxonomy of dark pattern types. In this section, I will follow these efforts chronologically and provide insights into relevant work extending the corpus of identified dark patterns. Table 2.1 offers an overview of the identified types based on empirical research, alongside references to the work in which they were first introduced. That is except for the first column. While not published in academic literature, I also included Brignull’s original twelve instances (Brignull, 2010) as the starting point for any following work. The underlying timeline mirrors the growth in interest as research contributions stretch over a growing landscape.

2.2.1 Expanding the Dark Pattern Taxonomy

The originator of the term “dark patterns”, Harry Brignull (Brignull, 2010), initiated this body of research in 2010 after introducing twelve types mostly found in online shopping sites and overall e-commerce. These initial types include patterns such as *Confirmshaming* (an example is given in Figure 2.1), linguistic attempts that emotionally pressure or shame users into specific actions, or the infamous *Roach Motel*[1], accounts or subscriptions that are easily created but unnecessarily difficult to cancel. Other patterns, such as *Sneak into Basket*, *Hidden Costs*, and

[1] The name originally described a product to capture cockroaches, but has since become an eponym to describe trap-like situations that are easy to get into but difficult to escape.

BRIGNULL* 2010	CONTI & SOBIESK 2010	ZAGAL ET AL. 2013	GREENBERG ET AL. 2014	BÖSCH ET AL. 2016	GRAY ET AL. 2018
· Trick Questions · Sneak Into Basket · Roach Motel · Privacy Zuckering · Confirmshaming · Disguised Ads · Price Comparison Prevention · Misdirection · Hidden Costs · Bait and Switch · Forced Continuity · Friend Spam	· Coercion · Distraction · Forced Work · Manipulating Navigation · Restricting Functionality · Trick · Confusion · Exploiting Errors · Interruption · Obfuscation · Shock	· Grinding · Impersonation · Monetized Rivalries · Pay to Skip · Playing by Appointment · Pre-Delivered Content · Social Pyramid Schemes	· Attention Grabber · Bait and Switch · The Social Network Of Proxemic Contracts Or Unintended Relationships · Captive Audience · We Never Forget · Disguised Data Collection · Making Personal Information Public · The Milk Factor	· Privacy Zuckering · Hidden Legalese Stipulations · Shadow User Profiles · Bad Defaults · Immortal Accounts · Information Milking · Forced Registration · Address Book Leeching	· Forced Action - *Gamification* - *Social Pyramid* · Interface Interference - *Aesthetic Manipulation* - *False Hierarchy* - *Hidden Information* - *Preselection* - *Toying With Emotions* · Nagging · Obsturction - *Intermediate Currency* · Sneaking

MATHUR ET AL. 2019	GRAY ET AL. 2020	GUNAWAN ET AL. 2021	HIDATA ET AL. 2023	MILDNER ET AL. 2023	MONGE ROFFARELLO ET AL. 2023
· Forced Action - *Forced Enrollment* · Misdirection - *Pressured Selling* - *Visual Interference* · Obstruction - *Hard To Cancel* · Scarcity - *High-Demand Messages* - *Low-Stock Messages* · Sneaking - *Hidden Subscriptions* · Social Proof - *Activity Notifications* - *Testimonials Of Uncertain Origins* · Urgency - *Countdown Timer* - *Limited-Time Messages*	· Automating the User Away · Two-Faced · Controlling · Entrapping · Nickling-And-Diming · Misrepresenting	· Account Deletion Roadblocks · Free Trials · Extraneous Badges · Missing Consent Notices, Consent Checkboxes, Or Settings Options · Needless Message Centers · No 'Bulk' Options For Settings · Paying For Ad-Free Experiences · Settings Do Not Save Properly	· Linguistic Dead-Ends - *Alphabet Soup* - *Untranslation*	· Engaging Strategies - *Interactive Hook* - *Social Brokering* · Governing Strategies - *Decision Uncertainty* - *Labyrinthine Navigation* - *Redirective Condition*	· Infinite Scroll · Casino Pull-to-refresh · Neverending Autoplay · Guilty Pleasure Recommendations · Disguised Ads and Recommendations · Recapture Notifications · Playing by Appointment · Grinding · Attentional Roach Motel · Time Fog · Fake Social Notifications

*These twelve dark patterns refer to Brignull's original set published on the website darkpatterns.org (Brignull, 2010). In 2023, the website got updated and now lists sixteen distinct types (Brignull, 2023). Moreover, in his book "Deceptive Design Patterns: Exposing the Tricks Tech Companies Use to Control You" (Brignull, 2023), Brignull names eight deceptive strategies in an attempt to group similar types.

Table 2.1 Overview of a total of 112 types of dark patterns from ten individual works in chronological order. Notably, some work took over previously defined types and extended their application through their research.

Price Comparison Prevention, operate by obscuring certain information from users of online shopping sites. The result of this obfuscation is that people using these sites cannot make a fully informed decision and can be misled into buying unwanted products. On the other hand, Brignull's dark patterns *Forced Continuity* or *Privacy Zuckering* are strategies that compromise the options and decisions available to people when using online services. This particular work marks the beginning of what is now more than a decade of dark pattern scholarship, describing dark patterns in various digital technologies. Independent from Brignull's initial work, however, released at the same time, the aforementioned taxonomy of malicious interface design techniques (Conti and Sobiesk, 2010) was based on findings from a twelve-month-long study and included eleven high-level behaviours associated with the application of dark patterns. For example, the *Coercion* technique describes interfaces that mandate users' decisions by restricting alternative options and enforcing compliance. Other techniques noted by the authors include *Interruptions* that interfere with a user's task flow and the *Obfuscation* of important information hindering informed decision-making.

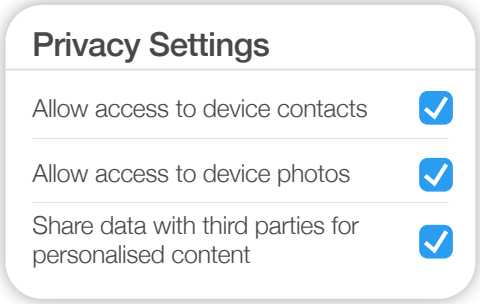


Fig. 2.3 Example of a *Bad Defaults* dark pattern. Privacy settings are automatically toggled to allow the sharing of sensitive data unless users change them.

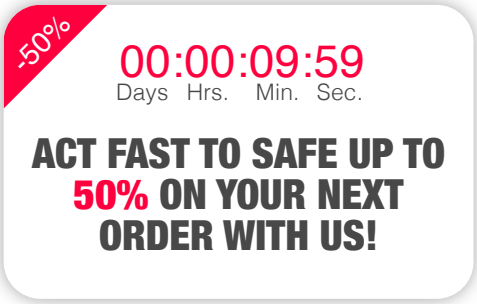


Fig. 2.4 Example of an *Urgency* dark pattern. An often false timer pushes potential customers into making quick and irrational decisions.

Looking at video games, Zagal *et al.* identify and describe seven related dark patterns (Zagal *et al.*, 2013). While certain patterns exploit a game's ecosystem of connected users, such as *Social Pyramid Schemes* and *Impersonation*, others impact game-play experiences like *Grinding* and *Playing by Appointment* (see Figure2.2). Although the latter can be harmful, the authors also discuss how these types of design can manifest as both dark and bright patterns. There are certain scenarios where they potentially cause harm and others where they can be beneficial to users. Elsewhere, Greenberg *et al.* (2014) consider dark patterns in conjunction with proxemics theory (Hall, 1966), leading them to offer warnings about the unethical use of existing technologies (Greenberg *et al.*, 2014). In their taxonomy of nine dark patterns, the authors discuss interactions with potentially abusive systems in spatial environments. Here, for example, they find the dark pattern *Attention Grabber* in the form of digital billboards, which change their appearance while pedestrians pass by. Such incidents create possibilities for brands to use people's proximity data to target them with personalised adverts as they draw close to digital advertising billboards. Inspired by the concept of *Privacy by Design* developed by a joint project of the Dutch Data Protection Authority and the Ontario Information Commissioner (Hustinx, 2010), Bösch *et al.* (2016) introduce privacy-related dark patterns. Comprising of seven underlying principles, the authors present what are effectively inverse strategies to the privacy strategies developed in the Privacy by Design project. The *Bad Defaults* dark pattern (as illustrated by Figure 2.3), for instance, describes privacy settings where every option is set to share personal data by default and has to be manually changed.

Spearheading dark pattern research by contributing what is likely the most impacting work since Harry Brignull coined the term, Gray *et al.* (2018), categorised a sample of 118 interface artefacts. For the first time, the authors introduce a hierarchy of lower and higher-level dark patterns, including five distinct dark pattern strategies: *Forced Action*; *Interface Interference*; *Nagging*; *Obstruction*; and *Sneaking*. Building on the first twelve dark pattern types, these strategies find wide application across technologies (Gray *et al.*, 2018). *Interface Interference* and *Nagging*, for example, represent interfaces that often manipulate visual elements to promote certain choices over others and constant reminders that have no option to be terminally dismissed.

Mathur *et al.* (2019), investigated e-commerce websites for instances of dark patterns users encounter when shopping online. Through an automated crawling approach, they collected a large corpus of text-based dark patterns and described seven dark patterns that incorporate previous work by Gray *et al.* (2018). Newly described instances include pressuring tactics like *Social Proof* and *Urgency* (as shown in Figure 2.4) that either expect users to follow social norms or make irrational decisions through limited time for reflection.

An equally important contribution of this work is the identification of common characterisations of dark patterns. Later extended to six such characterisations (Mathur *et al.*, 2021), the two works allow a better description of dark patterns within distinct dimensions. These characteristics span *Asymmetric*, *Covert*, *Deceptive*, *Disparate Treatment*, *Hiding Information*, and *Restrictive*. Importantly, these characteristics highlight the mutual "hidden" theme behind dark patterns, not necessarily the intent or malice of practitioners. Demonstrating the application of their work, the authors provide a preliminary assignment of previous dark patterns into these characteristics, showing that individual dark patterns can borrow from multiple characteristics. Publications in this thesis not only utilise these characteristics in multiple studies (P2, P3, P4, and P5), but the authors included a working definition for dark patterns that I have constantly been using since: "Dark patterns are user interface design choices that benefit an online service by coercing, steering, or deceiving users into making decisions that, if fully informed and capable of selecting alternatives, they might not make" (Mathur *et al.*, 2019). As with their characteristics, Mathur *et al.'s* definition does not require malintent as a condition while highlighting the contrast between benefits for service providers and tricked users that unknowingly tapped into interactions that they may regret later. While its limitation to "online service" makes sense in Mathur *et al.'s* work investigating online shopping sites, related work has since spotlighted dark patterns in various domains. Hence, I adapt the definition to consider design in general, thus also capturing dark pattern instances that are technology or interface-agnostic.

To understand how users perceive unethical design artefacts, Gray *et al.* (2020) conducted a content analysis based on user-generated content from the social network Reddit[2]. Defining "asshole designers", the work contributes practitioner properties for deploying design strategies such as dark patterns. In total, the work lists six such properties that, when considered from an interface perspective, can be used comparably to other dark pattern types as they follow similar sentiments. In this regard, *Two-Faced* is defined as "contradictory and conflicting information, confusing the user", which is in line with Conti and Sobiesk's *Confusion* dark pattern. *Controlling*, on the other hand, features elements from *Misdirection* (Brignull, 2010) and *Manipulating Navigation* (Conti and Sobiesk, 2010).

In this sense, Gunawan *et al.* (2021) provide a valuable overview of where dark pattern types occur across different screen modalities. Comparing mobile and browser UIs, the study highlights distinctions where certain interactions are prohibited depending on the modality a user uses to access a service. Parts of their results entail eight types of dark patterns linked to existing literature that restrict user choices through particular screen modalities. Privacy controls, for instance, were unavailable on some mobile applications (*Missing Consent*

[2] Based on the Reddit forum *assholedesign*, www.reddit.com/r/assholedesign. Accessed 08.05.2023.

Notices, Consent Checkboxes, or Settings Options) as users wanting to delete their accounts were only able to do so via specific modalities (*Account Deletion Roadblocks*). Restricting choices, particularly by disparity across modalities, strongly impedes informed decisions when users are left to believe that a service presents all possible options equally.

Looking into a different, Hidaka *et al.* (2023) make an important contribution when considering language-based dark patterns. Based on their analysis of 200 Japanese applications, the authors expand Gray *et al.* (2018) six dark pattern strategies with a seventh, which they coined *Linguistic Dead-Ends*. This class entails two subscribing dark patterns: *Alphabetic Soup* and *Untranslation* — both restricting users from accessing functionality or information by not providing sufficient information in a particular language. This is mainly the case for applications developed in a single language but which were then released into multi-language markets without properly ensuring good translations of original texts. This discrimination of certain speakers can yield grave problems when lack of translation leads to unawareness of consent and privacy controls. Hidaka *et al.* thus shine a light on many dark patterns that target marginalised users, pointing to a desperate need to explore dark patterns in this regard.

This section follows recent work on dark patterns which I summarised into a comprehensive taxonomy of dark patterns in chronological order (as Table 2.1 shows). I was also able to contribute to this corpus in our investigation of SNSs (P3). Thus, I want to expand this section with a brief summary of this work, although the publication is included in this thesis and will be more detailed in later chapters. Comparing four popular social media platforms, our study records types of dark patterns based on a subset of Table 2.1 until the work by Gray *et al.*, 2020. While the study surfaced instances of various dark patterns, we also noticed domain-specific interactions, fitting the dark pattern definition by Mathur *et al.* (2019), that were not previously described. This led to the definition of two strategies that include two and three low-level dark patterns respectively: Engaging strategies that entertain or embark users into interactions that they did not plan to execute and governing strategies that dictate users' decisions through the (un-)availability of certain choices.

Related to our engaging strategies and published simultaneously, Monge Roffarello *et al.* (2023) focused on interfaces that cause attention-based harms, often recorded in SNSs. In their effort, the authors conducted a systematic literature review to distill the surrounding discourse into a definition for "attention-capturing" dark patterns. This important work described eleven such dark patterns, including *Infinite Scroll* and *Neverending Autoplay* to hook users to content consumption. Their *Time Fog* dark pattern describes deflections about time spent using a service in this regard.

Collectively, these works on social media point toward dark patterns that follow alternative strategies when causing harm. Online shopping sites, as explored by Brignull (2010) and Mathur *et al.* (2019), feature dark patterns that can lead to immediate financial disadvantages. SNSs, on the other hand, have an incentive to keep users satisfied and return to their services. In following this incentive, these dark patterns have to balance user satisfaction against deceptive interactions restricting informed user choices.

The purpose of Table 2.1 is to give an overview of the widening dark pattern terrain. The twelve contributions mentioned above conclude the table's content. As an epilogue to this

collective effort, discussions and advances in research have not gone unnoticed by Brignull. In 2023, together with colleagues (Bringull *et al.*, 2023), Brignull updated his former website[3] through a change of the term "dark pattern" to "deceptive design" and, eventually, "deceptive patterns". Moreover, the dark pattern list now includes 16 types, featuring some from the original twelve but also taking on some proposed types from research (however, with slightly different names).

2.2.2 Implications for the Design Angle

The research journey to understand dark patterns began just over a decade ago. Surrounding work has grasped the many facets of dark patterns and, in doing so, identified various hosts where they are being deployed. Publications included in this thesis have contributed to advancing this body of research. However, previous work has often considered dark patterns as instances and interactions detached from contexts. To connect the various, often independent strands of work, I contributed to research establishing a dark pattern ontology (P6), included in this thesis, that aims to enable transdisciplinary work in the future. Further extending this knowledge outside traditional Graphical User Interface (GUI) technologies, particularly exploring CUIs, we investigated how unethical design emerges from design and broken expectations in P5. The granularity of dark patterns is important to understand the diversity of problematic designs. However, users are usually forced to navigate multiple interactions in their user journeys, leading to a demand to consider dark patterns in temporal sequences. Our studies on SNSs (P3) attempt to follow interaction flows as we explore this concept further in a later study included in P7.

2.3 Users' Ability to Identify and Recognise Dark Patterns

The definition of dark patterns implies difficulty for recipients to avoid them. But while studies have fostered a growing corpus of related interface techniques, less effort has gone into understanding the users' point of view. As one of the front liners interested in their perspectives, Di Geronimo *et al.* (2020) inspected popular mobile applications sampled from the Google Play Store. The authors used a cognitive walkthrough (Nielsen, 1994) to identify dark patterns in 240 applications sampled from popular categories in Google's Play Store. Each application was used for ten minutes following a task protocol, ensuring that any application was thoroughly explored. Their findings indicate that 95% of the tested applications contain dark patterns. In comparison to the work conducted by Mathur *et al.* (2019), who automated the capturing process, resulting in 11% of 11,000 shopping websites containing at least one dark pattern, the difference is quite apparent. Notably, the methodologies used in the two studies differ vastly. Limited to text-based elements, Mathur *et al.'s* approach cannot detect instances of *Interface Interference*, for example, as such dark patterns exploit visual tricks. Also, their scope is set on online shopping sites, whereas Di Geronimo *et al.* consider mobile applications. Though a direct comparison of the two studies should be taken tentatively, an interpretation could suggest that dark patterns perceive recipients on a visual level predominantly and

[3] Formerly www.darkpattern.org now redirects to www.deceptive.design. Accessed 08.05.2023.

occur differently between domains. Di Geronimo *et al.'s* methodology has inspired multiple studies to capture instances of dark patterns similarly (Gunawan *et al.*, 2021; Hidaka *et al.*, 2023) including publications part of this thesis (P3, P5). In a second study, Di Geronimo *et al.* evaluate users' ability to recognise dark patterns using an online survey, with most users exhibiting difficulty recognising dark patterns before falling victim to their schemes.

Understanding how users perceive dark patterns when encountered is an essential step toward identifying proper means for user safeguarding. Following a qualitative approach, Maier and Harr (2020) conducted a focus group and interviews with nine participants to collect data and show that their participants were mainly able to recognise problematic interfaces. Their findings brought forth three themes: Perception, Conduct, and Countermeasures. They describe how participants' awareness of dark patterns depends on the encountered types, which also causes differing perspectives regarding impact. While some instances were deemed highly problematic, others were rather seen as a nuisance. Interestingly, participants recognised responsibility in both actors relying on unethical practices as well as themselves when unable to avoid them or through their dependence on certain services. Although the study is limited by its sample size, it casts a prospect on users' awareness about themselves as well as service providers in an environment where deploying dark patterns has become an acceptable means to persuade their decisions.

Adopting a similar user-centred approach, Bongard-Blanchy *et al.* (2021) study peoples' awareness of manipulative interface designs and their recognition of dark patterns. While participants understood what dark patterns were and how they might manifest, they were still deceived by them during interactions. Studying the online video platform YouTube, Lukoff *et al.* (2021) analysed its interface to learn how internal mechanisms might be used to increase users' sense of agency. The researcher outlines design mechanisms that users feel more or less in control of. Specifically, users did not feel in control over YouTube's recommendation system, advertisements, or autoplay functionality. The study, which consisted of an online survey of 413 participants, found that users could recognise dark patterns, which aligns with previous results from Maier and Harr. However, being able to identify problematic designs did not reduce the likelihood of being influenced by them, which is in line with our findings within SNSs (P4).

Curious about how users perceive manipulative design, Gray *et al.* (2021a) collected qualitative response data from 169 participants when building on previous findings by Maier and Harr (2020). The authors noted a general awareness of problematic design among participants but further described their inability to communicate these perceptions. This is explained through dark patterns' subtle persuasion of user behaviour that often goes unnoticed. An interesting finding included differing perceptions between dark patterns; certain interfaces were perceived just obvious enough before engagement that users became partially aware of them (for example *Interface Interference* dark patterns). In other cases, dark patterns became apparent after longer usage — as is the case with *Nagging* dark patterns. The authors spotlight emotional components and felt manipulation of users when engaging with dark patterns, linking their work to a need for stricter policies to better protect users.

The persuasiveness of dark patterns is further demonstrated in work conducted by Graßl *et al.* (2021). Concerned with consent banners that trick users into sharing their data, the authors observe a culprit in the display of choices. By elevating a consent option through its design, while an alternative to decline is unnecessarily complicated to discover, the design preselects a choice for the user. A user study with 228 participants supports this implication showing that 94% of participants followed the design's suggestion. A second comparative study by the same authors underlines these implications further. The authors alternated the consent banner's design by highlighting the option to disagree instead of the one for consent and referred to this change as a "bright pattern". The study revealed that only 54% of participants would follow the now highlighted option. The authors explain this notable difference by discussing the long-term impacts dark pattern exposure has on users, conditioning them to comply without reflecting on decisions. This claim finds support in work by Habib *et al.* (2022), who point toward potential fatigue in users overriding otherwise informed and reflected decisions. The work, through its two studies, makes a strong case for the power of persuasive design and nudges.

2.3.1 Implications for the User Angle

Insights into the effects of dark patterns on users are essential for creating effective countermeasures. Although more work is required to understand all the implications in which users are affected by dark patterns, research outlines the manipulative techniques when demonstrating the difficulty to avoid falling victim to their effects (Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021; Maier and Harr, 2020). Replicating some of this work in the context of SNSs, publications included in this thesis (Mildner and Savino, 2021; Mildner *et al.*, 2023a) support these findings while promoting a possibility of assessing problematic design (Mildner *et al.*, 2023a). An important problem arises from users' unmet expectations when engaging with technologies, which we investigated in the context of CUIs (Mildner *et al.*, 2024a). Retracing these insights in SNSs, we compiled valuable design considerations from expectations users have when engaging with social media platforms (Mildner *et al.*, 2024c).

2.4 Countermeasures for Dark Patterns

By describing a broad dark pattern terrain and understanding how they persuade users into harmful interactions, research in HCI has laid the groundwork for better user protection through countering unethical designs. Academia is devoting an increasing amount of resources toward the development of ethics in design and technology; both in HCI research (Bruckman, 2014; Shilton, 2018; Shapiro *et al.*, 2020) and its education (Gray and Boling, 2016). However, more work is needed to address the harmful effects of dark patterns and ensure user safety. Meanwhile, regulations (for instance (Commission, 2016; California State Legislature, 2018; European Commission, 2022a)) have begun to take on ethical issues that arise from the development of new technologies to protect users where practitioners and service providers exploit unethical design practices.

2.4.1 How Dark Patterns Emerge

A crucial part of countering the harmful effects of unethical and problematic interactions is looking into their origin. Conceptualised as *ethical design complexity* (Gray and Chivukula, 2019), related work problematises outside pressures that influence the work of practitioners of technological design. In their work, Gray and Chivukula (2019) were not only able to describe the complex situations in which practitioners work but also shed light onto necessary ethical mediators that help practitioners to develop ethical, human-centred designs. In some instances, individual incentives contradict organisational goals, which again can break applied ethics practices. An interesting byproduct of this work is its implication of unintentional problematic and unethical design, which can be the result of environmental factors. These findings can further be interpreted as indicators for dark patterns resulting from environments where practitioners are unable to follow their moral compasses but have to submit to their organisation's aims. Instead, empowering workplace conditions should enable practitioners to create ethical and user-centred designs, rather than being restricted by organisational or corporate structures and incentives. A primary result of this work is described by the ethical design complexity model, demonstrating the tensions between actors.

Continuing this work to further understand practitioners' identity claims, Chivukula *et al.* (2021a) use thematic analysis of twelve semi-structured interviews to classify influences on ethical decision-making among professionals who work in technological fields. Overall, the work suggests eight identity claims among participants, including, but not limited to, educators and learners, but also policy-followers and activists. These sometimes connected but also contrasting identity claims mirror the complexity of working environments in which design emerges. Connecting these identity claims with the introduced ethical design complexity model, the work again emphasises the importance of establishing working conditions to limit the deployment of unethical designs such as dark patterns. Provoking this necessity further, Chivukula *et al.* (2023) discuss the responsibility of practitioners to take certain stances: Ethical dilemmas, ethical tensions, and ethical situations, each express identified practices of practitioners and their work environments and open space for active engagement in ethically difficult-to-navigate scenarios. Similarly, Sánchez Chamorro *et al.* (2023) apply the framework earlier proposed by Gray and Chivukula (2019) in their study and make a series of recommendations for practitioners to assess notions of unethical practice in their work. Based on value-sensitive suggestions, each recommendation includes a simple question for mitigating dark patterns. Drawing from prior work on ethical design complexity, Gray *et al.* (2024a) amplify the stream's notions that further support is needed for practitioners to engage with ethical and value-senstive methods. The study investigates how practitioners develop ethics-focused action plans and describe roles practitioners employ to realise their plans, process moves practitioners follow to operate within their problem space, and trajectories of practitioners regarding their action plans.

The work with practitioners offers explanations and insights into studies noticing a host of dark patterns in various online shopping sites (Mathur *et al.*, 2019; Moser *et al.*, 2019). To understand how businesses steer users into making impulsive purchases, Moser *et al.* (2019) compared and analysed 200 e-commerce websites. The work found that 84% of their

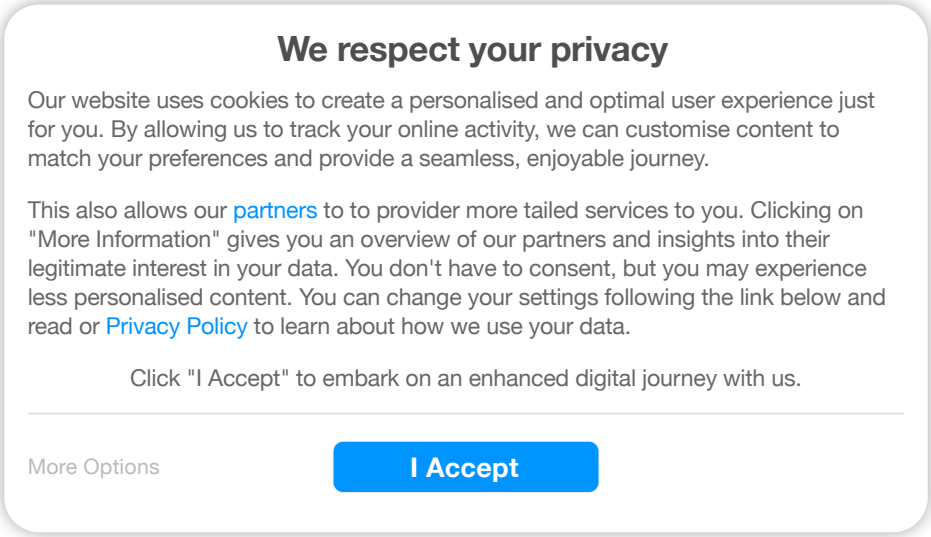


Fig. 2.5 By using dark patterns such as *Hidden Information* or *Interface Interference*, deceptive consent banners trick users into allowing services to track personal data.

sample used at least one design technique that was devised to encourage impulsive buying. A subsequent online survey with 151 participants revealed that potential customers would appreciate more salient information on pricing and generally desire less scarcity and urgency features that encourage impulsive buying. These findings are in line with work conducted by Mathur *et al.* (2019) (as described in Subsection 2.3).

Together, this strand of research demonstrates how the circumstances under which practitioners operate affect design ethics. Economic incentives, thereby, often dictate design processes according to their goals — leading to the deployment of unethical designs or dark patterns. Potentially causing conflicts and tensions between employer's goals and the individual beliefs of practitioners, there is an urgent need to change professional environments in order to ensure that practitioners are able to engage in ethical and human-centred design practices that afford user needs.

2.4.2 Recognising a Need for Regulatory Countermeasures

The previous Section 2.3 has detailed the various challenges users have to overcome when facing dark patterns. A segue from users' ability to recognise dark patterns to countering the effects can be drawn from work considering seemingly omnipresent consent banners (Figure 2.5 offers an example of a deceptive version). Exhibiting the dawning effects economic incentives can have on online interfaces, consent banners have repeatedly been observed to persuade users into making preselect choices (Maier and Harr, 2020; Gray *et al.*, 2021b). Although certain regulations, such as the GDPR (Commission, 2016), are in place to protect end-users, requiring service providers to gain informed consent before tracking any of their users' data, practitioners still try to find creative ways to overcome such obstacles through design (Matte *et al.*, 2020; Bielova *et al.*, 2024).

To understand the ramifications of particular design techniques used by consent banners, Utz *et al.* (2019) studied the influence of nudging design exploited by consent banners based on a large-scale study with 80,000 users. As the authors noticed, users were generally willing to interact with the consent banners. The study revealed specific design choices that influence users' behaviour: Position, available options, nudges, and wording of consent banners all had substantial effects on users' decision-making, which is in line with similar research conducted by Maier and Harr (2020). Comparing their data with requirements of the European Union (EU)'s GDPR (Commission, 2016), the authors indicate that consequential implementation of consent banners should result in less than 0.1% of users giving their consent. While adjacent work highlights the difficulty users have when navigating certain interfaces and dark patterns, Utz *et al.* (2019) take an important step toward highlighting the apparent misconduct of service providers. Similar work conducted by Bielova *et al.* (2024) supports these findings by illustrating how the design and availability of choices impact user preferences.

In a joint effort, Gray *et al.* (2021b) also investigated consent banners in a transdisciplinary effort, combining relevant disciplines. From the lenses of three perspectives — computer science, HCI, and law — the authors discuss the deceptive effects these interfaces have while considering their legal circumstances. In demonstrating the problematic strategies used by many consent banners and pointing to potential legal issues in many interfaces, the authors showcase the importance of synergies between disciplines for effectively addressing unethical design. Each field's unique perspective adds valuable insights into dark patterns and how they can be countered.

With work highlighting the manipulative and exploitative effects of dark patterns (Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021), this body of research exemplifies the need for further regulation and stricter enforcement of existing law within the single case of consent banners. As these studies mainly focus on consent banners within EU's GDPR, more work is needed to investigate and confirm similar legal issues in other domains and regulatory contexts. Nevertheless, the studies amplify the positive impact transdisciplinary collaborations can have on protecting users from dark patterns.

2.4.3 Regulatory Efforts Toward Countering Dark Patterns

A relevant piece within the intersection of HCI and law, Gunawan *et al.* (2022) reviews transdisciplinary literature from both contexts to study whether people, who experienced harm through dark patterns, should be given redress. As a basis for their work, the authors draw from GDPR consent requirements and developed a case study on informed consent. The authors exemplify the deployment and legal boundaries for dark patterns when discussing the tremendous implications dark patterns have on informed consent as defined by regulations of the EU. Spotlighting harms on the user's side that are indeed the results of dark patterns, the authors demonstrate how these design techniques can invalidate consent in terms of the GDPR: Thus, the work displays how damaged recipients should be given redress, inline with studies suggesting similar issues (Utz *et al.*, 2019; Bielova *et al.*, 2024).

With a background in law, Calo (2013) debated contemporary issues of service providers manipulating their customers' behaviour to their advantage. Businesses fully understand their

consumers' behaviour and are thus able to address them anywhere with personalised content. Noticing unethical implications, he calls for a change of this trend, encouraging positive alternatives for building healthier customer relationships. As of writing this thesis, Calo made this call for change over a decade ago. The growing body of work describing more and more dark patterns and other unethical practices across interfaces and technologies suggests that many service providers did not listen or find suitable alternatives. Instead, an increasing number of regulations and policies have been implemented internationally to protect users from deceptive practices. Similar to Di Geronimo *et al.* (2020) or Bongard-Blanchy *et al.* (2021), Luguri and Strahilevitz (2021) studied users' perception of dark patterns and noticed that it was predominantly dark patterns that mildly affected their choice architecture which successfully tricked users. Moreover, the results of their studies spotlighted that users did not find the exposure of all dark patterns necessarily problematic. Following up on these findings, the authors discussed the legal situation of dark patterns in the context of US law, particularly the Federal Trade Commission (FTC). While the context is complex and not always clear-cut, Luguri and Strahilevitz (2021) noticed precedents were harmful design was already successfully countered, making room for future legal work to connect.

In 2022, a series of major milestones in regulation happened. In the US, California's California Privacy Rights Act (CPRA) (California Privacy Protection Agency, 2022) as well as the FTC addressed certain dark patterns directly, regulating certain deceptive designs in online interfaces. Other regulatory bodies following this move include the UK's Competition and Market Authority Competition and Market Authority (CMA), the European Commission and the European Data Protection Board (EDPB) (Board, March, 2022), as well as the Organisation for Economic Co-operation and Development (OECD) (OECD, 2022). Also in 2022, the EU, for the first time, incorporated the term "dark patterns" into its Digital Service Act (DSA) (European Commission, 2022a), Digital Markets Act (DMA) (European Commission, 2022c), and the Data Act (DA) (European Commission, 2022b), introducing stricter protection of users from harmful interfaces. Shortly after, in 2023, the Department of Consumer Affairs in India followed up on these efforts by releasing guidelines against dark patterns (Ministry of Consumer Affairs, Food & Public Distribution, 2023). Importantly, these regulations provide guidelines for various dark pattern types formerly brought up in academic research. These efforts could be seen as a precedent for the benefits of HCI and law collaboration, following advice from Gray *et al.* (2021b) and Gray *et al.* (2023c).

2.4.4 Design Countermeasures

Considering the increasing regulatory efforts in place, even codifying the term "dark pattern" into legal statutes across nations, related HCI scholarship has done a tremendous job in problematising unethical design and its consequences. However, the question remains of alternative designs that offer simple-to-use interfaces while respecting user autonomy. Ethical design concepts, such as VSD (Friedman *et al.*, 2013), have been around for some time; nonetheless, environmental constraints in which practitioners work (Gray and Chivukula, 2019; Chivukula *et al.*, 2021a) often do not leave enough space for the implementation of more human-centred design. According to a study included in this thesis (P5), practitioners

echoed missing effective design guidelines and best practices as a go-to for designing without deceptions.

Within the particular context of dark patterns, the work by Graßl *et al.* (2021) introduced "bright patterns" as alternatives to their unethical counterparts. However, the technique relies on the same underlying strategies exploited by dark patterns to redirect user choices. Referring back to the definition provided by Mathur *et al.* (2019), dark patterns concern interactions with unexpected or hidden consequences to the user. While their effects are often reported as harmful consequences, arguably, it is the obfuscating and manipulating nature of dark patterns that deceive users into interactions they would not have engaged with if given full information about their consequences. To support this argument, work by Bielova *et al.* (2024) spotlights that 46% of surveyed participants feel comfortable sharing their personal data with service providers. If a design is utilised to create interactions governing decisions without enabling users to assess the outcomes of their actions, even if deployed in good faith, it becomes difficult to argue how that design differs from another dark pattern. Ultimately, design should be crafted with the intent to support users' decisions, fully respecting their autonomy.

In an attempt to do so, in the same popular context of consent banners, Leimstädtner *et al.* (2023) build on prior work on responsible nudging by Hansen and Jespersen (2013) and design friction to assist users in reflecting on their decisions before taking actions. In a study with 297 participants, the authors compared four interface designs following Hansen and Jespersen's framework of four nudges reflectively: The first included manipulation of choice, the second an influence of behaviour, the third was designed to include manipulation of choice, and the fourth included prompting reflective choice. Their study's findings imply that solely the reflection prompt helped users make choices in line with their preferences. Although the work does not directly link their work to usability issues noted as a result of design friction (Mejtoft *et al.*, 2023), Leimstädtner *et al.* (2023) illustrate the challenges arising when designing for informed decision-making. Design friction is commonly used to interrupt user-flows to aid users in acknowledging consequences before interactions and making them transparent (Wang *et al.*, 2013; Mejtoft *et al.*, 2019). As a consequence, the interruptions increase the cognitive load of users compared to nudges that deliver preselected choices (Mejtoft *et al.*, 2023).

2.4.5 Implications for the Guideline Angle

While regulatory bodies increased defensive measures against unethical and harmful online practices, currently, the landscape of design alternatives as countermeasures is relatively scarce. However, it may be our best line of defense to provide practitioners with alternatives as we promote the need to foster user autonomy. Addressing this important gap, this thesis entails frameworks to allow the assessment of dark patterns in interfaces (P2 P4). Other studies attempt to bridge the expectations between users and systems (P5) as well as identify elements of an intricate relationship between dark patterns and cognitive biases to deflect informed decision-making P9. By contributing theory that offers answers to the diverse challenges

behind developing ethical designs, I hope to inform fairer, human-centred interfaces in the future.

2.5 Chapter Summary

Although dark pattern research is still in its infancy, the transdisciplinary work of the past decade has established thorough understanding of underlying design mechanisms and concepts exploited by dark patterns. Research in HCI on ethical design and dark patterns offer a substantial contribution to landscaping the dark pattern terrain in a range of technologies. The subsequent vocabulary describing 112 dark patterns (see Table 2.1) is one of many important gains of this research. In response to these academic efforts, first regulatory bodies deployed various counteracting design practices that harm users on various levels. With most efforts gone into identifying dark patterns, surfacing the need for better user protection, practitioners currently lack sufficient and practical alternatives while the user perspectives have yet to be fully explored. The three angles, this thesis is structured around, aim to offer further insights and responses for this otherwise complex and intertwined field. The publications included in this thesis address these gaps and, collectively, answer the research questions as proposed in Chapter 1. With this chapter setting the stage and providing a dense background of dark pattern scholarship, the following three chapters will each tackle on answering the research questions by following the *Responsible Design Triangle's* angles — design, user, and guidelines — respectively.

CHAPTER 3

The Design Angle

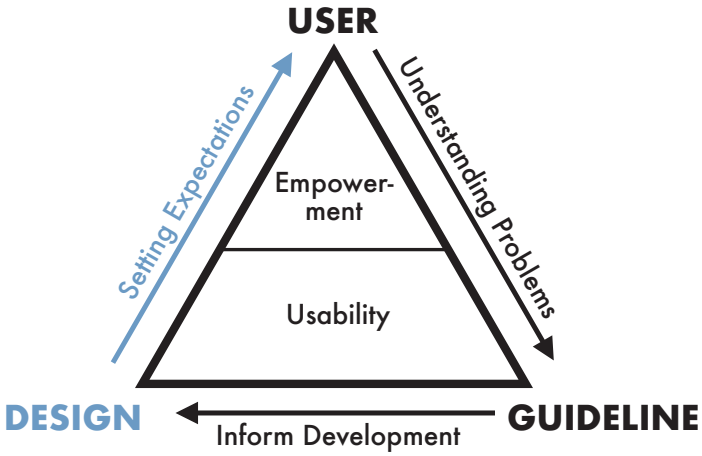


Fig. 3.1 Responsible Design Triangle highlighting the design path covered in this chapter.

Expectations are essential when designing systems or interactions (P5) as they enable users to develop plans and estimate consequences before engaging with a system. Affording realistic expectations should thus be a distinct design goal. Nevertheless, many designs create false beliefs, misleading users into unaccounted or undesired interactions. Importantly, expectations are often set before engaging with a system or an interaction for the first time (Luger and Sellen, 2016). Advertisement for a product and its overall appeal strongly impact users' expectations (Oliver, 1977; Oliver, 1980).

To this end, practitioners can devise certain design elements to capture and keep their audience's attention (O'Brien and Toms, 2008). Aesthetic appeal and easy access, for instance, can be used to draw potential users toward a system (Bron *et al.*, 2017). Convoluting emotional, experiential, and enjoyable design dimensions, outside of sole pragmatic requirements, User Experience (UX), in its complexity, provides further angles through which design can become engaging (Hassenzahl and Tractinsky, 2006). Based on a literature review, O'Brien and Toms (2008) developed a model to explain user engagement further, segmented into three phases: An initial engaging point, an engagement period, and, lastly, disengagement. For each phase, the authors provide attributes that impact the phases. These attributes indicate particular influences practitioners can utilise to trigger and prolong user engagement and re-engage them if they should disengage. In line with the disengagement attributes proposed by O'Brien and Toms (2008), Frøkjær *et al.* (2000) shed light on particular design considerations to keep users engaged. Users may stop using a system or service if individual aspects, such as effi-

ciency, effectiveness, or satisfaction, are neglected. Conclusively, various mechanisms exist for practitioners to create and control engagement and influence expectations.

In times of surveillance capitalism, where free-to-use services generate revenue by selling data to advertisers and other third parties (Zuboff, 2023), user engagement is essential for many service providers to secure a steady income. However, the task of designing long-lasting engagement with services is all but an easy one. Perhaps to avoid otherwise difficult design challenges, the body of work describing unethical and problematic dark patterns illustrates the alternative routes some service providers choose to take (Gray *et al.*, 2018; Gray *et al.*, 2020). Instead of fostering satisfaction, trust, and user autonomy, services have become deceptive and manipulative through greedy and profit-maximising incentives (Gray and Chivukula, 2019; Chivukula *et al.*, 2023).

As highlighted in Figure 3.1, this chapter concerns the design angle and how applications set expectations for the users — and break them. While the *Responsible Design Triangle* incentivises user empowerment, it is built around research demonstrating harmful interactions throughout interfaces. In this vein, this chapter seeks to answer the research question: "**RQ1:** ***How can design be used to create and break expectations that lead to dark patterns?***" Understanding how applications manipulate expectations and deploy dark patterns opens avenues for countermeasures as well as design strategies to align users' intentions with system capabilities. This chapter is based on contributions from the following publications:

> **P1** Mildner, T. and Savino, G.-L., "Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7, ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3451659
>
> **P3** Mildner, T., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, ISBN: 9781450394215. DOI: 10.1145/3544548.3580695
>
> **P5** Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R., and Niess, J., "Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642542
>
> **P6** Gray, C. M., Santos, C. T., Bielova, N., and Mildner, T., "An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642436
>
> **P7** Gray, C. M., Mildner, T., and Bielova, N., "Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow"," arXiv:2309.09635, 2023

As the first author of P1, P3, and P5, my contributions were to provide the initial idea behind the papers, the drafting and submission of the manuscripts, design and conduction of the studies, data collection, as well as the analysis and interpretation of results or findings. For P6, I contributed to the analysis of dark pattern types from various sources and their organisation into the final ontology, including their definitions. Moreover, I contributed three case studies to demonstrate the ontology's application and possible expansion in the future, while writing parts of the paper and approving the final draft before its submission. For P7, Colin Gray and I designed the concept of Temporal Analysis of Dark Patterns (TADP) together, while I was responsible for the content analysis of Amazon's "Illiad Flow". As with P6, I contributed to the writing of the paper and approved the final manuscript before its submission for review.

3.1 Designing Deceptions

My interest in investigating dark patterns first ignited after reading studies that reported how SNSs were responsible for their users' decreased mental health and well-being in longitudinal study contexts (Shakya and Christakis, 2017; Twenge *et al.*, 2018). Meanwhile, HCI-related research that took a design or user-centred perspective to consider harms resulting from SNS interfaces was scarce. This lack motivated me to capture deceptive interface elements in SNSs, with the first results published in P1. Beginning with Facebook, an almost historic yet relevant platform in its field, we collected imagery material of its ever-changing UI between the years 2004, when Facebook first went public, and 2020, when the last big update was released before we conducted the study. Using this material, we conducted an interface analysis to better understand the specific changes made by the platform. Figure 3.2 follows these changes and illustrates the displacement of account and privacy-related features, obfuscating the discoverability of critical settings. Notably, the "Account Settings" and "Privacy Settings" were moved from a first-level view, where they were quickly accessible, to a nested menu, decreasing their discoverability. Similarly, the "Logout-Button" and, later, the "Privacy Shortcuts" were nested deeper within Facebook's UI. The study revealed a host for dark patterns in Facebook but required further research to explore how and where precisely dark patterns manifest in various SNSs platforms.

To this end, I designed a comparative study to consider mobile applications of four popular SNSs. Next to Facebook, P3 included the platforms Instagram, TikTok, and Twitter. Following a more rigorous procedure that draws on work done by Di Geronimo *et al.* (2020), we conducted a cognitive walkthrough of each application designed to explore features of SNSs to their full extent. For the study, we recruited six participants, all of whom were PhD students in HCI at the time. To prepare them for the study and ensure an equal understanding of the topic, each received an introduction to dark patterns prior to executing their walkthroughs. Equipped with a definition for dark patterns by Mathur *et al.* (2021) and a comprehensive taxonomy of dark patterns, comprised of over 80 individual types from eight works, participants were tasked to identify instances of dark pattern types throughout the SNSs. To gain deeper insights into their decisions, we asked participants to comment on their decisions and what they perceived in a think-aloud fashion (Jaspers *et al.*, 2004).

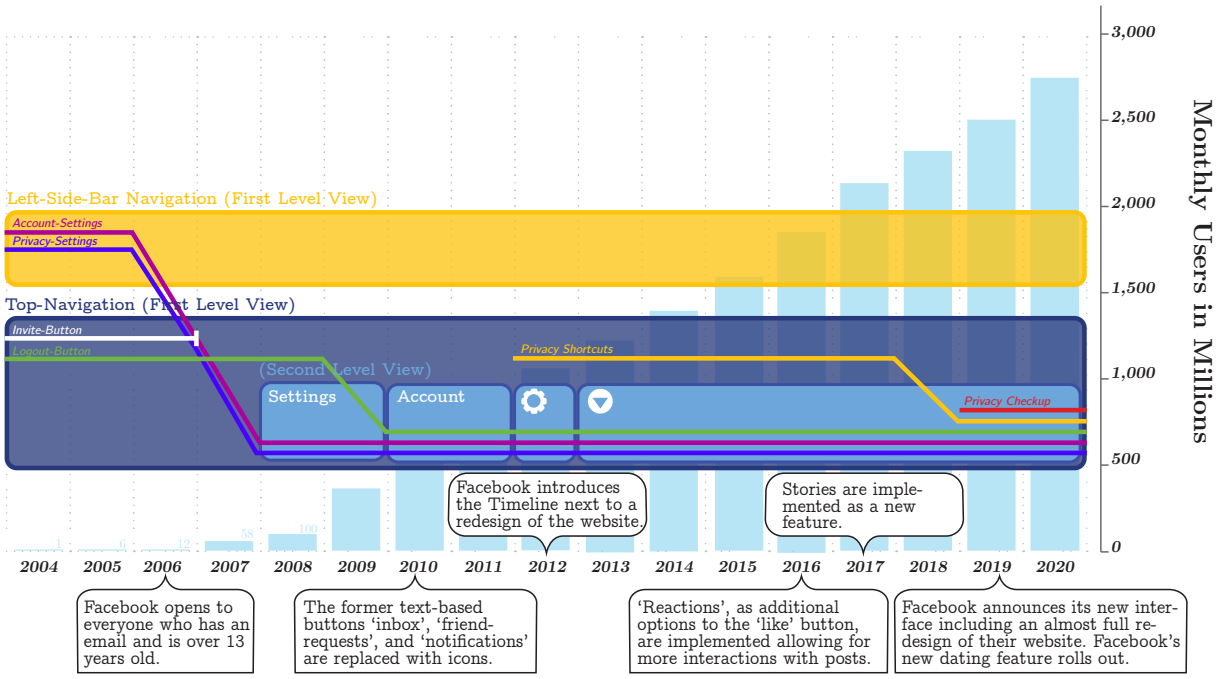


Fig. 3.2 This diagram describes the changes within Facebook's interface between the years 2004 and 2020. The figure contains multiple levels of data. Following the years, it shows the nesting of existing, introduction of new, and exclusion of relevant account and privacy-related features across Facebook's UI complemented by critical junctions in the design of speech bubbles. The diagram also contains a bar chart reflecting the increasing user count per year. This figure was published in Mildner and Savino (2021) (P1).

The choices for this study design were threefold: Firstly, it allowed us to observe previously described dark patterns outside their original scopes to gain a dense overview of their prevalence and variety in alternative interfaces. Secondly, we were curious about the similarities and differences between individual dark pattern types. We noticed instances where some work carried over and refined dark patterns first described by others. For example, Brignull's initial *Privacy Zuckering* (Brignull, 2010) was later adopted, but slightly changed, by Bösch *et al.* (2016); Brignull described his version of the dark pattern as a trick to mislead users into sharing more personal information than intended, Bösch *et al.'s* adaptation concerns restricted access to related settings and the ease with which users can control them. Thirdly, to our knowledge, no other study has considered using a similarly rich corpus to investigate interfaces for the presence of dark patterns.

After completing the study, we analysed the collected data in a combination of inductive and deductive coding (Mayring, 2020) following a reflexive thematic analysis approach (Braun and Clarke, 2006). The inductive codes consisted of the same dark pattern types previously handed to our six participants for their walkthroughs. The deductive codebook was generated based on open coding of the initial data and iterative discussions between two coders to resolve possible disagreements (Blandford *et al.*, 2016). Through this comprehensive research, we garnered a unique experience using dark pattern taxonomies as tools to recognise problematic and unethical designs while adding dark patterns — that have not yet been captured — meaningfully.

3.1.1 Findings of the Deductive Coding

Author	Dark Pattern	F	I	Ti	Tw
Brignull (2010)	Bait And Switch	●	○	○	●
	Confirmshaming	●	●	●	●
	Disguised Ads	○	●	○	●
	Forced Continuity	○	○	○	○
	Friend Spam	○	○	○	○
	Hidden Costs	○	○	○	○
	Misdirection	●	●	●	○
	Price Comparison Prevention	○	○	○	○
	Privacy Zuckering	●	●	●	●
	Roach Motel	●	●	●	●
	Sneak Into Basket	○	○	○	○
	Trick Question	○	○	○	○
Conti and Sobiesk (2010)	Coercion	○	○	○	○
	Confusion	●	○	○	●
	Distraction	●	●	●	●
	Exploiting Errors	○	○	○	○
	Forced Work	●	●	●	●
	Interruption	●	●	●	●
	Manipulating Navigation	●	●	●	●
	Obfuscation	●	●	●	●
	Restricting Functionalities	●	●	○	○
	Shock	○	●	○	○
	Trick	○	○	○	○
Zagal *et al.* (2013)	Grinding	○	○	○	○
	Impersonation	○	○	○	○
	Monetized Rivalries	○	○	○	○
	Pay To Skip	○	○	○	○
	Playing By Appointment	○	○	○	○
	Pre-Defined Content	○	○	○	○
	Social Pyramid Schemes	○	○	○	○
Greenberg *et al.* (2014)	Attention Grabber	●	●	●	○
	Bait And Switch	○	○	○	○
	Captive Audience	○	○	○	○
	Disguised Data Collection	○	○	○	○
	Making Personal Info. Public	○	○	○	○
	The Milk Factor	○	○	○	●
	Unintended Relationships	○	○	○	○
	We Never Forget	○	○	○	○
Bösch *et al.* (2016)	Address Book Leeching	●	●	●	●
	Bad Defaults	●	●	●	●
	Forced Registration	○	○	○	○
	Hidden Legalese Stipulations	●	●	●	●
	Immortal Accounts	○	○	○	○
	Information Milking	●	○	○	○
	Privacy Zuckering	●	○	●	○
	Shadow User Profiles	○	○	○	○
Gray *et al.* (2018)	Forced Action	●	●	●	●
	Gamification	●	●	●	●
	Social Pyramid	●	●	●	●
	Interface Interference	●	●	●	●
	Aesthetic Manipulation	●	●	●	●
	False Hierarchy	●	●	●	●
	Hidden Information	●	●	●	●
	Preselection	●	●	●	●
	Toying With Emotions	●	●	●	●
	Nagging	●	●	●	●
	Obstruction	●	●	●	●
	Intermediate Currency	●	●	●	●
	Sneaking	●	●	○	○
Gray et al. (2020)	Automating The User	●	●	○	●
	Controlling	●	●	●	●
	Entrapping	○	○	○	○
	Misrepresenting	●	●	●	●
	Nickling-And-Diming	○	○	○	○
	Two Faced	○	○	○	○
Mathur *et al.* (2019)	Forced Acrtion (see Gray *et al.* (2018))				
	Forced Enrollment	○	○	○	○
	Misdirection	●	●	●	●
	Pressured Selling	●	●	●	●
	Visual Interference	●	●	●	●
	Obstruction (see Gray *et al.* (2018))				
	Hard To Cancel	●	●	●	●
	Scarcity	●	●	●	●
	High-Demand Messages	○	○	○	○
	Low-Stock Messages	○	○	○	○
	Sneaking (see Gray *et al.* (2018))				
	Hidden Subscriptions	○	○	○	○
	Social Proof	●	●	●	●
	Activity Notifications	○	○	○	○
	Testimonials	○	○	○	○
	Urgency	●	●	●	●
	Countdown Timer	○	○	○	○
	Limited-Time Messages	○	○	○	○

Legend: **F** - Facebook
I - Instagram
Ti - TikTok
Tw - Twitter

Table 3.1 Based on 80 deductive codes created from related research on dark patterns, this table visualises which social media platforms deploy particular dark patterns. The deductive codes were applied in four SNSs (**F** - Facebook, **I** - Instagram, **Ti** - TikTok, and **Tw** - Twitter). The presence of a particular dark pattern is highlighted with a "●" whereas "○" indicates its absence based on our analysis. This table was first published in Mildner *et al.* (2023b) (P3).

The deductive approach resulted in a thorough overview of the prevalence of dark patterns based on previous dark pattern typologies. Table 3.1 presents an overview in this regard. In total, we noticed 44 types of dark patterns across the SNSs. However, we found certain typologies more applicable than others. For instance, the more abstractly described dark patterns by Gray *et al.* (2018) proved to be easily applicable, while others were not identified

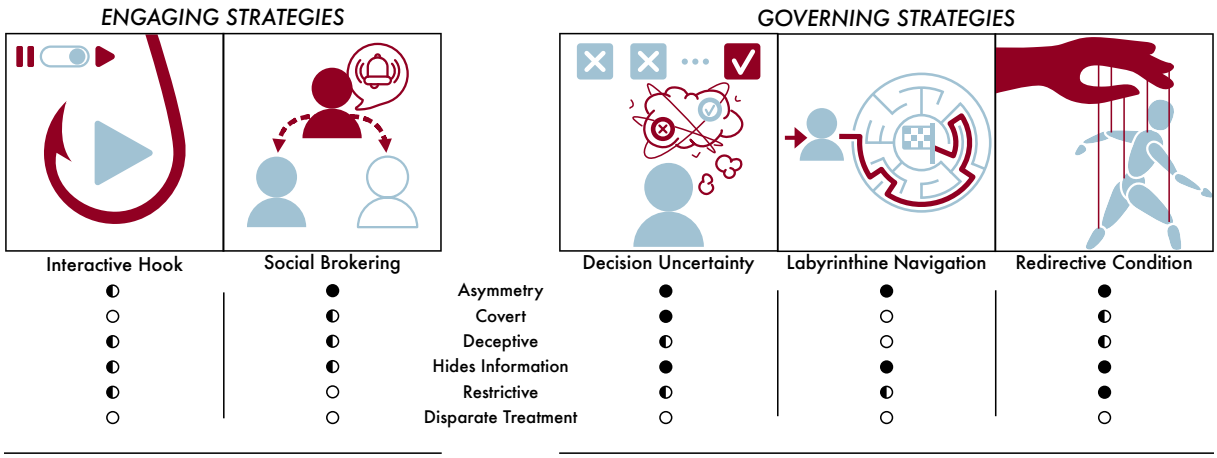


Fig. 3.3 The results of the inductive coding, the engaging and governing strategies contain five types of SNS-specific dark patterns. To situate the five dark patterns' in existing research, we categorised each based on Mathur *et al.* (2021) six dark pattern charactersitics (asymmetry, convert, deceptive, hides information, restrictive, and disparate treatment). Following their originator's application, "●" highlights the presence of a characteristic, "◐" indicates optional presence, and "○" the absence of a dark pattern characteristic. This figure was published in Mildner *et al.* (2023b) (P3).

at all, as is the case with the game-related dark patterns described by Zagal *et al.* (2013). The generalisability of dark patterns from Gray *et al.* (2018) allowed a diverse, almost interface agnostic, application, whereas dark patterns from Zagal *et al.* (2013) were precisely described in gaming contexts. Our decision to stay as close to original definitions as possible, to test the utility of a taxonomy for revealing dark patterns, restricted us from simply transferring domain-specific types to SNSs.

3.1.2 Findings of the Inductive Coding

Aside from existing dark pattern types, we recorded any instance where we noticed a problematic interface design fitting the dark pattern definition by Mathur *et al.* (2021). After ensuring that any prior descriptions did not already cover such instances, through the deductive codebook, we analysed them by conducting a thematic analysis. Finally, this resulted in two overarching strategies: Those that *engaged* users in interactions they did not plan to engage in, and those that *governed* their decisions through presenting choices in a way that restricted users from perceiving all available options. Figure 3.3 presents an overview of these two strategies with five identified SNS-specific dark patterns subscribing to them. Learning from the advantages of more abstractly defined dark patterns, we established our findings within this hierarchy to enable future work to recognise them in alternative contexts as well. Moreover, we grounded each dark pattern within the six dark pattern characteristics (asymmetry, convert, deceptive, hides information, restrictive, and disparate treatment) introduced by Mathur *et al.* (2019) and later expanded by Mathur *et al.* (2021), indicating in which dimensions they manifest.

We noticed two dark pattern types that belong to the engaging strategies. First, *Interactive Hooks* that, for example, use gamification elements for provoking unsolicited interactions to get users to disclose personal information as seen in Figure 3.4. Second, *Social Brokering*, as

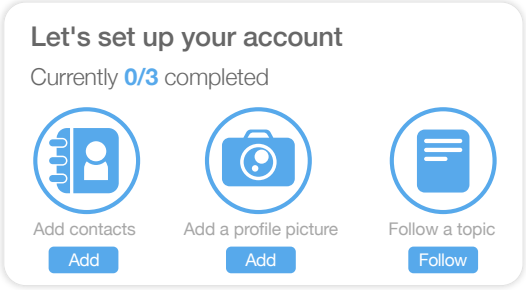


Fig. 3.4 Example of a *Interactive Hooks* dark pattern.

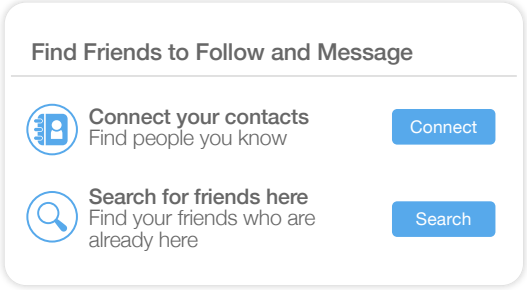


Fig. 3.5 Example of a *Social Brokering* dark pattern.

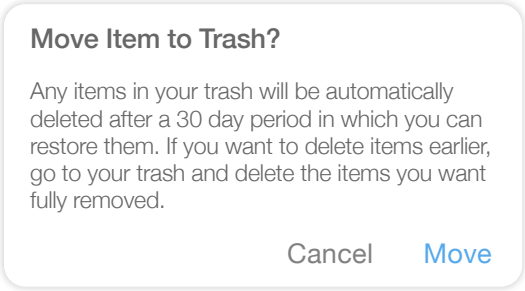


Fig. 3.6 Example of a *Redirective Condition* dark pattern.

displayed in Figure 3.5 manifests in attempts to create artificial connections between SNS users or ask them to share other peoples' contact information. The latter is especially problematic when non-users are not given the chance to withhold their consent before users share their data with a service, similar to *Addressbook Leeching* (Bösch *et al.*, 2016).

With regard to the governing strategies, we described three types of dark patterns. Unfortunately, two describe instances that cannot be illustrated in still images. The first, *Decision Uncertainty* uses overwhelming interface design to deflect or confuse users, denying informed decision-making. Particularly when creating a new account on TikTok, our participants felt distracted by video and music media playing. At the same time, the SNSs asked them to accept their terms and conditions and allow the platform to use their data for advertising purposes. Our participants described the interaction as problematic as it kept them from focusing and making an informed decision. Second, *Labyrinthine Navigation* describes maze-like, obstructive interface structures commonly found in settings menus in the SNSs. The ease with which users get lost and forget their initial goals, such as maintaining privacy settings, highlighted the problems arising from inefficient interface design. Third, *Redirective Conditions* intercept user interactions and redirect them toward service providers' goals. Figure 3.6 demonstrates an example where a user's goal to delete personal data was disrupted, forcing them to wait for 30 days before their data is eventually deleted. We observed similar behaviours when trying to delete an SNS account, which got unnecessarily postponed by 30 days before finalising the process.

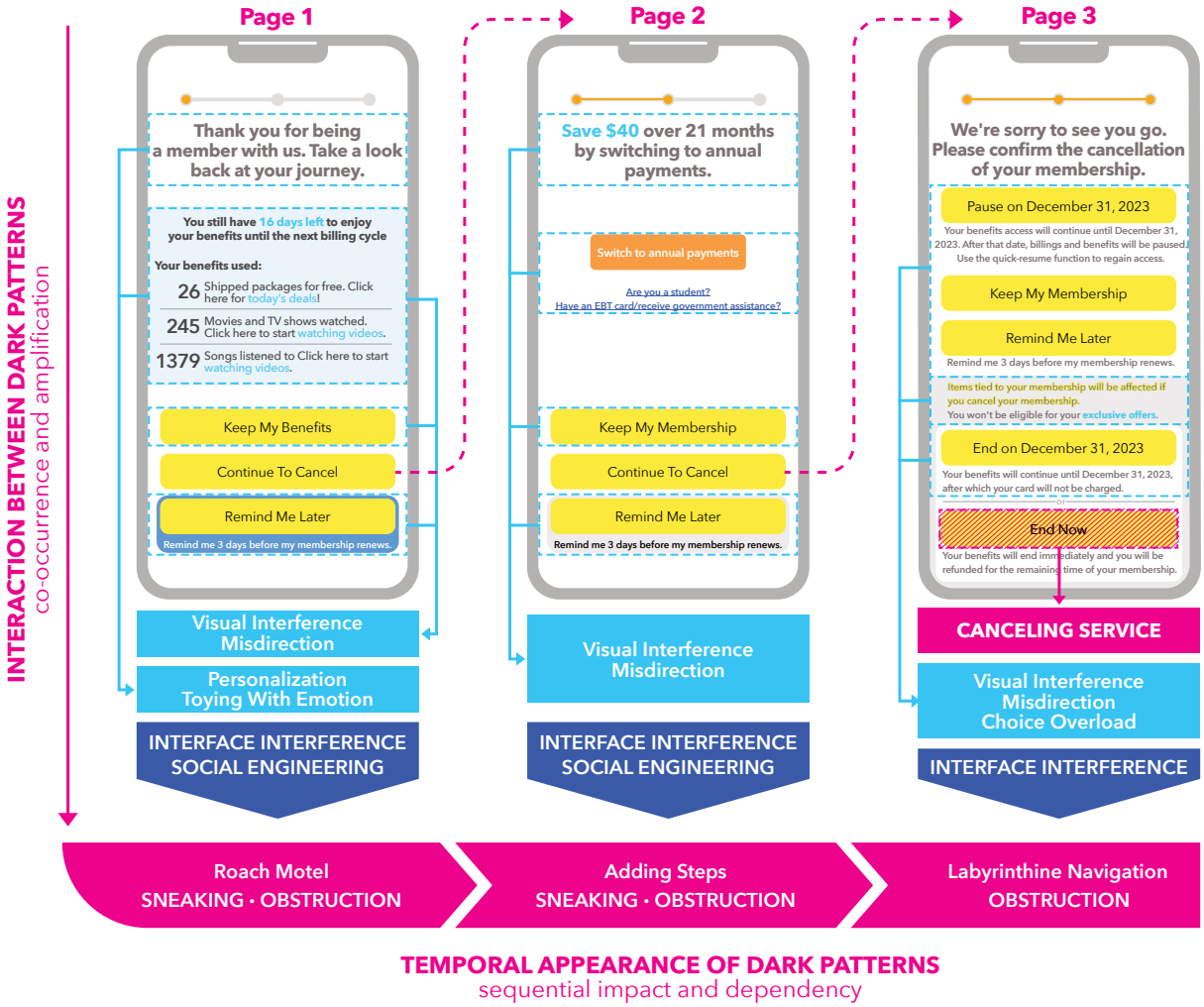


Fig. 3.7 A summary of our temporal analysis of dark patterns in Amazon's "Iliad Flow." For brevity, we simplified the interface complexity but maintained key options, including three screens users have to navigate to be able to cancel their membership. Vertically underneath each page, we summarized co-occurring and amplifying dark patterns. Horizontally, we follow the sequential impacts and dependency of dark patterns. Dark patterns in small caps refer to high-level types, while lower-case dark patterns refer to meso- and low-level instances from P6. This figure was published in Gray *et al.* (2023b) (P7).

3.2 Temporal Analysis of Dark Patterns

While contributing novel insights to dark pattern research, my investigations of SNSs included the limitation that we evaluated interfaces mainly based on dark patterns as individual design elements. While this is a common approach (see for example Bongard-Blanchy *et al.*, 2021, or Gray *et al.*, 2018), it restricted us from assessing the accumulation of dark patterns throughout user journeys. Yet, the analysis made it clear that dark patterns rarely come solely. In most cases, users face a range of dark patterns both simultaneously and chronologically throughout interaction sequences. This led to the conceptualisation of a novel, alternative approach to evaluating the presence of dark patterns in UIs. Importantly, this approach could also be relevant for regulatory bodies to show the different dimensions in which service providers manipulate or deceive their users. To this end, we decided to not focus on SNSs, where P3 and P4 partially discussed this phenomenon. Instead, the shift to other contexts where users are

harmed allowed me to expand my research focus. To demonstrate the relevance of our TADP for future dark pattern research as well as regulators, a complaint by the FTC (Federal Trade Commission, 2023) against Amazon proved to be a perfect and timely case study.

In their complaint, the FTC criticised Amazon for its overly obstructive and complicated account-deletion process, shedding light on the arising problems when users face highly deceptive interaction sequences. While in 2022, EU citizens benefited from new EU regulations (European Comission, 2022) requiring Amazon to simplify this process into a two-click option, other users still had to complete the trial laid out by the online service — who internally referred to their deletion process as "Iliad Flow"[1] (Federal Trade Commission, 2023). Between the attention it received and the obvious use of dark patterns, the "Iliad Flow" presented a suitable scenario for a case study to explore our TADP (P7).

Before reaching the "Iliad Flow", in order to delete an Amazon account, users had to progress through various steps; the actual sequence then, however, *only* features three pages. The diagram in Figure 3.7 illustrates the procedure in a simplified visualistion and highlights the various dark patterns users encountered per interaction and across the entire scheme. Briefly, Amazon designed the process to require users to click specific buttons to progress, as any other option would terminate the process and force them to start over. These interactions are aggravated by targeting a user's memories to recall positive emotions, through social engineering, as well as interface manipulations that would place the option to continue into the background.

We described the methodology for TADP within three steps: First is the identification of dark pattern type(s) deployed, including combination and sequence thereof. Second is the identification of UI elements that are affected by dark patterns and the assessment of how they amount to impact users. Third is the description of interactions between dark patterns in terms of co-occurrence between types and amplifying effects. Using these steps, we identified 70 instances of dark patterns across the three pages of the "Iliad Flow".

3.3 Organising Dark Patterns into an Ontology

The TADP continued some aspects of our SNS-based walkthrough study (P3) to identify dark patterns in different kinds of contexts, derived into a methodological approach. However, it relies on solid dark pattern typologies to be effective. Consequently, I idealised the benefits of a thorough dark pattern ontology. In 2023, Gray *et al.* (2023d) have begun to synthesise the existing dark pattern discourse into a preliminary ontology. Having started to pursue a similar goal to simplify future studies through a uniform and consistent catalogue, I was happy to join their endeavor and contribute to the development of the finalsied ontology (P6). At the time of writing this thesis, the ontology offers the most comprehensive collection of dark patterns from both academic and regulatory sources. Its hierarchical structure will support future (transdiscipliary) work through a common basis and language.

[1] *Iliad* is an ancient name for a city before it was replaced with its more commonly known name of Troy. The "Iliad Flow" likely refers to the Greek epic by Homer centering around Achilles and the Trojan War (Britannica, The Editors of Encyclopaedia., 2023).

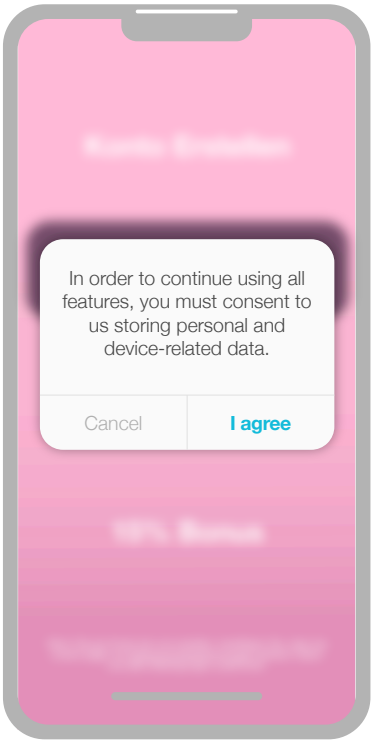


Fig. 3.8 Example of a *Forced Action* dark pattern. Translated and adopted from from Mildner *et al.* (2024d).

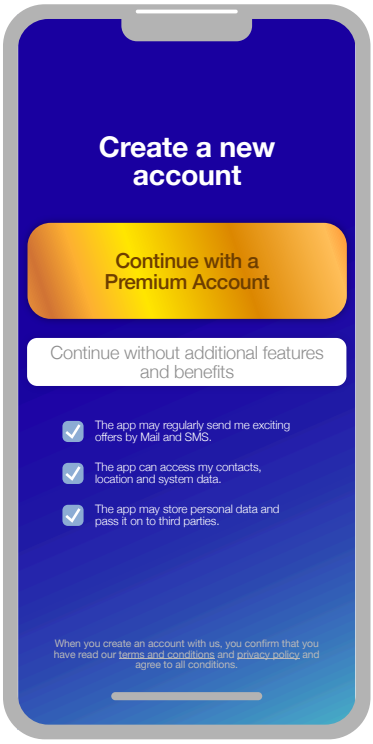


Fig. 3.9 Example of a *Interface Interference* dark pattern. Translated and adopted from from Mildner *et al.* (2024d).

In line with the broad applicability of dark patterns by Gray *et al.* (2018) in our previous SNS study, the five high-level dark patterns contained *Forced Action* (Figure 3.8), *Interface Interference* (Figure 3.9), *Obstruction* (Figure 3.10), *Sneaking* (Figure 3.11), and, newly added, *Social Engineering* (Figure 3.12). Unlike the preliminary ontology published a year in advance, we decided to remove *Nagging* as a high-level dark pattern and, instead, embed it as a new meso-level inside *Forced Action*. In the following, I will introduce each high-level strategy by providing our original definitions from the ontology (Gray *et al.*, 2024b) as well as examples[2] to elaborate on how they manifest.

Forced Action. "Forced Action is a strategy which requires users to perform an additional and/or tangential action or information to access (or continue to access) specific functionality, preventing them from continuing their interaction with a system without performing that action" (P6).Briefly, dark patterns under the *Forced Action* strategy restrict user choices, prohibiting interactions unless users meet certain conditions, such as granting consent. Figure 3.8 shows an example for *Forced Action* in the form of an interface that requires the user to disclose personal information before allowing them to use an otherwise free-to-use service. Especially, if services are not dependent on those data to offer their users better functionality and may sell it to third parties.

[2] The examples are based on illustrations I created for a German article in the Bundesgesundheitsblatt (Mildner *et al.*, 2024d). Here, I translated them into English and made minor edits. As is the article, these images were published under the Creative-Commons license, granting everyone permission to (re-)use them.

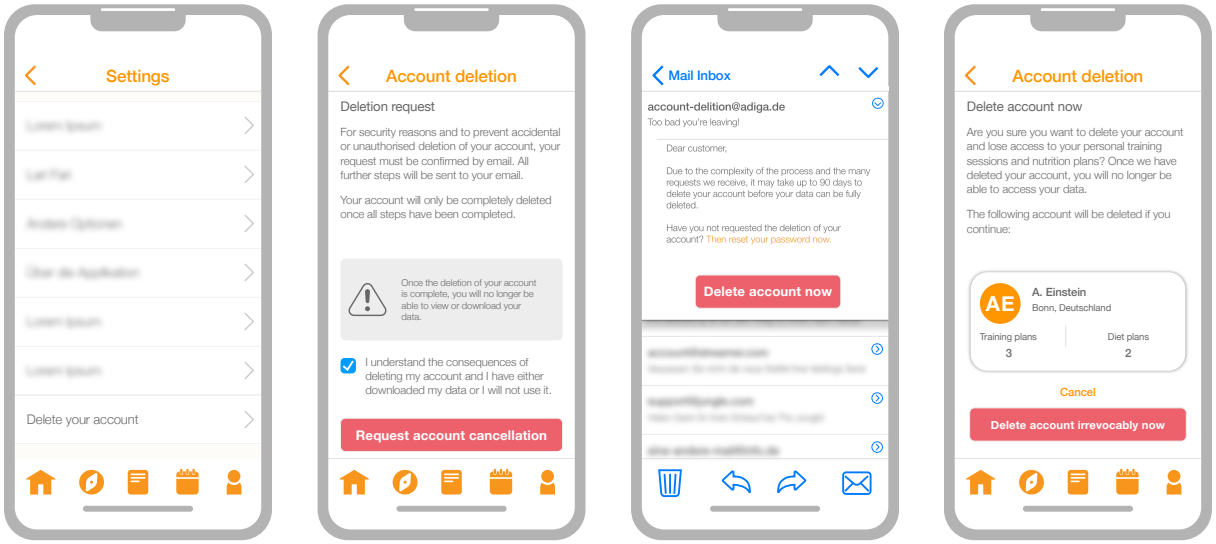


Fig. 3.10 Example of an *Obstruction* dark pattern throughout a user journey. Translated and adopted from from Mildner *et al.* (2024d).

Interface Interference. "Interface Interference is a strategy which privileges specific actions over others through manipulation of the user interface, thereby confusing the user or limiting discoverability of relevant action possibilities" (P6). Dark patterns exploiting *Interface Interference* steer users' perception to specific UI elements and thereby away from other choices. In this sense, Figure 3.9 illustrates how some interfaces highlight premium options over other, cheaper alternatives. The service provider prioritises their business incentives over users' ability to make an unbiased decision.

Obstruction. "Obstruction is a strategy which impedes a user's task flow, making an interaction more difficult than it inherently needs to be, dissuading a user from taking an action" (P6). This strategy manifests in the form of obstacles that disrupt interaction flows or user journeys. Consequently, they impede users' ability to achieve their goals if it goes against a service provider's interests. Figure 3.10 presents a userjourney across different interactions, throughout which a user experiences a series of obstacles when trying to delete their account with a service.

Sneaking. "Sneaking is a strategy which hides, disguises, or delays the disclosure of important information that, if made available to users, would cause a user to unintentionally take an action they would likely object to" (P6). An example can be found in sales funnels that often navigate users through multiple *Bait and Switch* instances, each contributing to small price increases that amount to unwanted surprises at their end. In this vein, Figure 3.11 illustrates how services try to trap users into expensive subscriptions, for instance, through conditionally "free" trials (that are only free if a user subscribes to the service).

Social Engineering. "Social Engineering is a strategy which presents options or information that causes a user to be more likely to perform a specific action based on their individual and/or social cognitive biases, thereby leveraging a user's desire to follow expected or imposed social

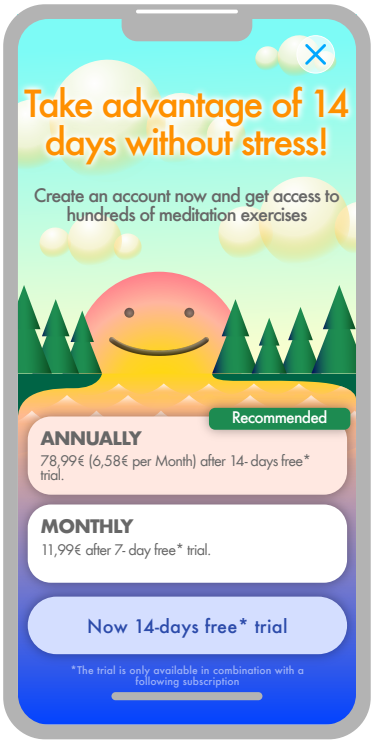


Fig. 3.11 Example of a *Sneaking* dark pattern. Translated and adopted from Mildner *et al.* (2024d).

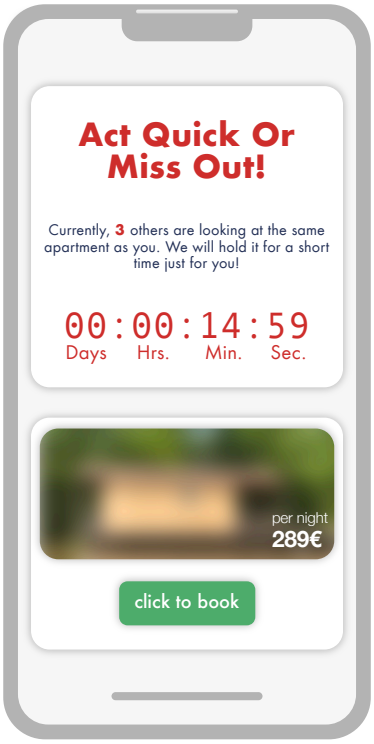


Fig. 3.12 Example of a *Social Engineering* dark pattern. Inspired but changed from Mildner *et al.* (2024d).

norms" (P6). Related dark patterns often rely on emotional pressure to gain users' attention and engage them in actions. To this end, Figure 3.12 includes text that aims to pressure users through positive reinforcement while displaying a ticking clock to create urgency in users to misguide them into quick instead of informed decisions.

3.3.1 Levels of Granulartiy

As outlined in Chapter 2, where I followed prior dark pattern research and its recognition of a plethora of dark patterns, more than a decade of related research has fostered a thorough understanding of types in various technologies and contexts. The scale of this discourse, including the varying terminologies used to describe similar concepts, has led to the ontology respecting a hierarchy among dark patterns.

To this end, the ontology organises dark patterns within a three-level hierarchy, listing five high-level dark patterns at the top which describe general strategies as previously presented. Low-level dark patterns, on the other end, describe domain or context-sepcfic means of execution. Between high and low-level, we further describe meso-level dark patterns, which describe a certain angle of attack through which high-level dark patterns manifest. This granularity not only helps to place existing work within the same context but also aids future work by offering a common language and a framework that can be expanded. We informed the ontology by analysing 245 dark patterns from various academic and regulatory sources. In a five-step methodology, we (1) aggregated various types of dark patterns, (2) traced their provenance, and (3) clustered the sampled corpus based on direct citations, identical language,

or inferred similarity. We (4) introduced "meso-level" dark patterns as an addition to already used categories for high and low-level types and, eventually, (5) finalsed the ontology after iterative and thorough discussions around definitions. The final ontology contains 65 dark patterns (5 high-level, 25 meso-level, and 35 low-level). However, it is meant to become a living and growing document to foster transdisciplinary work.

One of our primary aims was to recognise dark patterns as domain and context agnostic while respecting instances that are domain and context-specific. The low-level *Address Book Leeching* (Bösch *et al.*, 2016), for example, describes a dark pattern where services ask users to upload their contacts, often by promising to quickly connect them with existing contacts and providing tailored services. This dark pattern is usually context-specific to SNSs and similar applications, which profit from the collected data. On the meso-level, the dark pattern subscribes to *Forced Communication or Disclosure*. As an angle of attack, it exploits users' false expectations of gaining benefits or access to functionalities only if they share personal data with a service provider. This meso-level dark pattern, again, belongs to the previously outlined *Forced Action* strategy. By incorporating a hierarchy, the ontology caters to diverse usage scenarios for transdisciplinary work and purposes. The full ontology and its definitions are included in P6.

3.4 Expectations in Design

The aforementioned work outlines my contributions to understanding dark patterns as design artefacts, occurring in a range of digital interfaces. Starting with a focus on SNSs, our work captured particular, domain-specific strategies. The work motivated further research and sparked the formalisation of a methodology (TADP) as well as a robust ontology for future work to build upon. However, all these works share a common limitation: A certain neglect of users' expectations when using digital services, as we first noticed early in our investigation of Facebook's interface (P1). Before the next Chapter 4 dives into the user's perspective, I want to elaborate on what I mean when writing about user expectations — based on our findings from interviewing researchers, practitioners, and users of CUI systems (see P5). For this chapter, however, I will translate these findings into the scope of digital interfaces to better answer this chapter's research question, since the tensions between design and its users make their expectations relevant here as well.

From the user's point of view, expectations are complex and individual, drawing from past experiences and projecting these onto something — possibly unfamiliar. When interacting with a digital interface, we may be led by assumptions and anticipations, which decide how we formulate plans and goals, followed by beliefs about possible outcomes. Additionally, we may be guided by certain needs that motivate our plans and goals in the first place. Consequently, expectations play a crucial role during decision-making processes. In Publication P5, we described these factors as *intrinsic*, carried by users inside and shaping their expectations.

Following a design perspective, as is the focus of this chapter, expectations can be accounted for — or, in the words of Peter-Paul Verbeek, can be inscribed into technologies by the designer (Verbeek, 2006). In the terminology of HCI, affordances and signifiers, previously described in Chapter 1, play an important role here, as practitioners design a system to

communicate its states and capabilities to its users, assisting them in navigating an interface following their needs. If done well, their audience develops realistic expectations and goals that they are able to carry out. This description for expectations is generally in line with usability and UCD principles (Norman, 2013; Nielsen and Molich, 1990; Nielsen, 1994). In Publication P5, we described these factors as *extrinsic*, influencing user expectations from the outside.

Both intrinsic and extrinsic factors yield implications that can negatively impact the user's expectations in the form of two delimiters: distrust and deception. Intrinsic factors draw from experiences and biases that result in attitudes toward technologies that, for instance, stem from prejudice or disappointment with the companies behind them. Extrinsic factors, on the other hand, are the result of practitioners' (bad) design choices trying to maneuver the user to a specific goal without consolidating their interests or intentions enough. Understanding that expectations can be specifically addressed through design has critical implications for exploitative and manipulative design. This contribution of P5is important to answer this chapter's research question.

3.5 Breaking Expectations — Answers for Research Question 1

This chapter comprises a selection of work collectively exploring underlying design aspects of dark patterns. The work identifies different sources in different digital interfaces but also considers design aspects that lead to harmful interactions. As shown in Publications P1 and P5, users engage with technologies with individual expectations that dark patterns break by exploiting their assumptions or restricting the available choice architecture. The publications as part of this chapter also show the varying contexts where dark patterns occur (P6). Offering answers to the first research question of this thesis, these works describe the various dark pattern strategies used to deflect expectations and coerce actions. Importantly, these strategies differ depending on the context and environment.

Online shopping sites, for example, could show deflective price tags that do not reflect the total cost (Mathur *et al.*, 2019), which becomes apparent only when checking out. Alternatively, services could sneak in additional items or lure customers into buying unneeded products. In either case, the user sets out to do one thing, having specific expectations regarding a product and its cost, but may experience problematic financial surprises. In many cases, the harm is immediately noticeable, even if there is nothing the user can do to revert it.

Dark patterns in SNS-related contexts often have a different approach to breaking users' expectations. Generating most of their revenue not through paid memberships but their users' data and advertisement implies incentives to keep users engaged and satisfied for them to return (P1 and P3). Consequently, engaging strategies keep users entertained while governing strategies ensure that users stay within an SNS's best interest by obstructing their choices if users' and platform's interests are not aligned (P3). In this context, the resulting harm of dark patterns may appear only after longer periods as users will not be able to notice it early (P4).

That time can be an issue is also shown in the TADP method proposed in P7. Even if a single dark pattern might be unsuccessful in exploiting users' actions, multiple instances appearing both simultaneously and prolonged over multiple interactions are less likely to have no effect on a user's behaviour. Unlike many prior studies investigating dark patterns in the form of

isolated artefacts (e.g. Gray *et al.*, 2018; Bongard-Blanchy *et al.*, 2021; Mildner *et al.*, 2023a), the additional temporal dimension adds relevance to the future of dark pattern research as design is continuously leveraged to distort users' expectations.

Practitioners, specifically designers, are equipped to shape and consolidate user expectations through design — and also break them. Consequently, to design ethical user interfaces, they are also responsible for doing this adequately and with the users' intentions and best interest in mind. In the words of the uncle of a certain "friendly spider from the neighbourhood": *"With great power comes great responsibility"* (Lee *et al.*, 1962). With this in mind, our analysis in Publication P5 opened a discussion for ethical caveats of CUI technologies (further outlined in Chapter 5). Not intending to imply malice, our discussion further included a warning of potential dark patterns that may be exploited if not cared for in this regard. With reference to the intrinsic and extrinsic factors as well as delimiters in the form of deception and trust, designers ought to be wary of any design choices that create tensions between users' expectations and system capabilities. Importantly, not all dark patterns break users' expectations, as is shown in the work on SNSs. Instead, bending expectations can be all that is needed to achieve similar results without the same risks of losing users who feel betrayed. On the contrary, users may remain satisfied if their goals are now aligned with the service provider's incentives, as findings from P1 demonstrate. In cases where users follow their personal intentions when using a service, the service provider may attempt to redirect them through omissions of relevant information. In other cases, where users may have no initial intentions, service providers may see an opportunity for exploitation and persuasion to steer their users towards their goals, regardless of what is in the best interest of the user.

CHAPTER 4

The User Angle

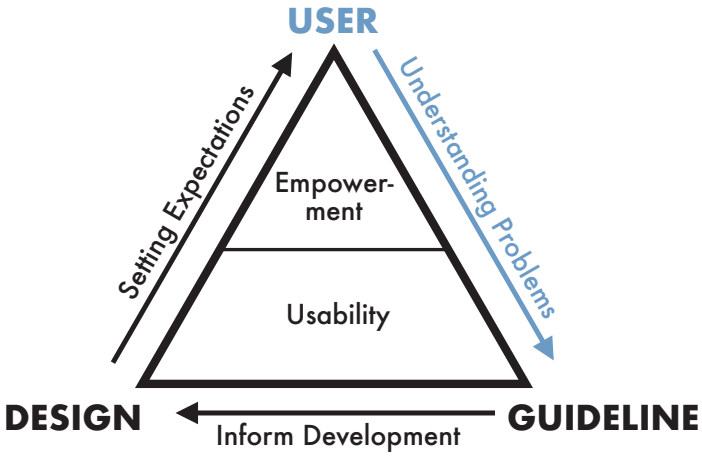


Fig. 4.1 Responsible Design Triangle highlighting the user path covered in this chapter.

Users' expectations arise from a mix of their perceptions, their intentions, and the design of an interface. After I discussed how design is responsible for setting expectations in the first place — or, in the event of dark patterns, exploiting users' actions to their disadvantage, this chapter is based on the user angle. To understand this angle and, thus, the user's perspective better, I first want to outline how users approach digital interfaces, develop goals, and attempt to execute them. Drawing from traditional HCI literature, a commonly used model is the seven stages of action by Norman (2013): It begins with the user (1) perceiving the world's state, followed by (2) their interpretation of what they perceive. Afterwards, they can (3) evaluate these interpretations, leading to (4) setting a goal. Based on their goals, the user (5) creates intentions for actions, which they (6) develop into a sequence. Last but not least, the user (7) executes their sequenced actions, which cause a reaction with the world, changing its state. From there, the model can restart with the first step where the user can reevaluate their actions and develop new and adapted goals, turning the seven stages of action into a cycle.

These seven stages find support in several other works explaining how human perception impacts people's goals as well as their development (for example, Davis, 1993; Hilton and Darley, 1991). For example, the cognitive walkthrough methodology for studying interface design by Polson *et al.* (1992), which I used to study HCI experts' ability to identify dark patterns (P3, P4), relies on people's perceptual abilities to construct plans and derive actions in order to understand their interactions with (digital) interfaces. How this process can be intercepted through design is shown in the previous chapter. In order to develop realistic goals, design has to foster realistic expectations for users to perceive. Although Norman does not really mention

the term "expectation" when describing the seven stages, I would argue that users develop their expectations during the first three, between perceiving the world's state and evaluating their interpretation thereof. To their aid, the design's purpose is to communicate an interface's capabilities effectively. This can happen, for instance, through affordances, feedback, and signifiers (Norman, 2013). However, dark patterns set a strong example of how easily users can be misguided or manipulated by diverting them from following their individual goals or persuading them into developing pre-set ones that are rather beneficial for the service provider (Gray *et al.*, 2021b; Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021).

To highlight the user's importance and point of view, the *Responsible Design Triangle* purposefully locates the user angle at its top. However, compared to the invaluable efforts that have gone into capturing and describing dark patterns in various interfaces, we lack a similarly dense understanding here as well. And, to that end, their ability to fend for themselves, with few general studies investigating the issue (Di Geronimo *et al.*, 2020; Maier and Harr, 2020; Bongard-Blanchy *et al.*, 2021). To better understand their challenges when faced with dark patterns, this chapter offers answers to the research question: **RQ2:** ***To what degree are users able to identify dark patterns in interfaces to safeguard themselves?*** Recognising vulnerabilities among users elevates the importance of responsible design strategies and stricter countermeasures against dark patterns to ensure users are not harmed. This chapter is based on contributions from the following publications:

> **P1** Mildner, T. and Savino, G.-L., "Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7, ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3451659
>
> **P4** Mildner, T., Freye, M., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "Defending Against the Dark Arts: Recognising Dark Patterns in Social Media," in *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA*, 2023. DOI: 1010.1145/3563657.3595964
>
> **P5** Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R., and Niess, J., "Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642542
>
> **P8** Mildner, T., Savino, G.-L., Putze, S., and Malaka, R., "Finding a Way Through the Social Media Labyrinth: Guiding Design Through User Expectations," arXiv:2405.07305 [cs], 2024

As the first author of the publications P1, P4, and P5, my contributions were to provide the initial idea behind the papers, the drafting of the manuscripts, the design and conduction of the studies, data collection, as well as the analysis and interpretation of results or findings. For

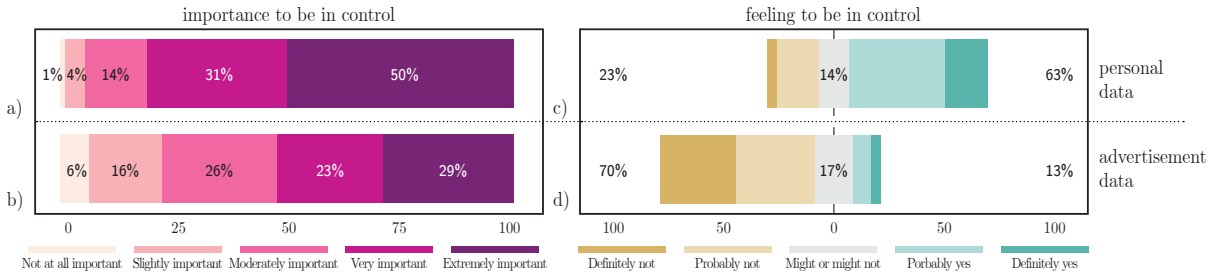


Fig. 4.2 This graphic shows the incongruity between SNS users' priorities to control their data versus their actual feeling of being in control. The questions asked were *a) How important is it for you to be in control over the information other people can see about you on Facebook, b) How important is it for you to be in control over the information about you which Facebook uses for targeted advertisement?, c) Do you feel in control over the information other people can see about you on Facebook?* and *d) Do you feel in control over the information about you which Facebook uses for targeted advertising.* This figure is published in Mildner and Savino (2021) (P1).

P3 and P5 specifically, I drafted the theoretical contributions before discussing them among co-authors. With the helpful assistance of my co-authors, I was responsible for the writing of the papers and submitting them. P8 is the result of equal efforts between Gian-Luca Savino and myself. The idea to view SNS UIs from a navigable network perspective is the result of many fruitful discussions. The resulting study design, analysis, and writing of the manuscript were also equally distributed among us.

4.1 Users' Perception on SNS

Users' behaviour, including their interactions with digital media, depends on their perception (Hilton and Darley, 1991; Davis, 1993). However, research outside dark pattern scholarship has shown for some time now that users do not always act in line with their beliefs or preferences. Particularly on SNS platforms, users regularly engage with features, but later regret their activities (Wang *et al.*, 2011). In a similar vein, users show misaligned values and behaviour when it comes to their privacy. In this regard, the privacy paradox (Barth and Jong, 2017) is a well-studied and understood phenomenon. The paradox concerns users' awareness regarding their online data, including their desire to protect themselves. However, users frequently disregard these values if actions taken to secure their data become overly cumbersome or an obstacle to other desires — like to keep using SNS applications. Depending on the circumstances, conflicts between values and goals (privacy preferences versus using SNSs) often result in neglecting the values. As a consequence, the user quite knowingly disregards their beliefs to accomplish their goals.

P1 is not only chronologically the first but also reflects my initial interest to better grasp people's behaviour on SNS. Ultimately, its results sparked my motivation to pursue research on dark patterns in this domain. Based on an online survey, we asked 94 Facebook users about their usage behaviour, including their values regarding control of their data and their feelings regarding their ability to control how it is seen or used by Facebook. Furthermore, we prompted them to reply both for personal data and data used for advertising purposes. Figure 4.2 demonstrates their perceptions in this regard. The responses show that a majority of our participants generally desire control of their data. However, while 63% actually replied

that they feel in control of personal data, only 13% were similarly confident when it comes to data used for advertising.

In addition to these value-focused questions, we were interested in their motivation and satisfaction when using Facebook. In line with other work on this matter (Lin and Lu, 2011), our participants reported the strongest motivations in staying in touch with family and friends, participating in groups of interest, retrieving news, or using the messaging feature. In terms of satisfaction, we found that, collectively, 63% use Facebook sometimes or generally more than wanted. This is reflected by another 80% who would not want to use the platform more than they currently do. When asked whether they think they overuse the SNSs, however, 68% see themselves not spending too much time there. Hinting at another incongruence, we learned that the majority of our participants were either extremely satisfied (29%) or somewhat satisfied (39%) with their time spent on Facebook. Another 20% stated neutrality in this regard, leaving only 12% reporting some dissatisfaction. Interestingly, no participant answered with extreme dissatisfaction. Consequently, Facebook's users generally seem satisfied with their time on the platform.

4.2 Understanding Users' Expectations in SNS

The incongruencies in users' beliefs and actions, as shown in P1, open dangerous opportunities for exploitation through unethical design. Interactions can (purposefully) hinder users from following through on their preferences but still offer easy access to fulfill their current intentions. The dark pattern we described as *Labyrinthine Navigation* presents a fitting example here: A difficult-to-navigate interface obstructs users from finding certain UI features. At the same time, engaging elements offer users alternatives that draw them away from their goals but keep them satisfied nonetheless.

Eager to understand the *Labyrinthine Navigation* dark pattern better through investigating SNS users' general expectations about relevant features, I designed a card sorting study based on Facebook's mobile interface (see Publication P8, Mildner *et al.* (2024c)). The study included 58 common features that we sampled from a total of 102 UI elements. Extending the traditional card sorting method of grouping related features together, we asked participants to also rate each feature in terms of its "importance" to be included in SNSs as well as the "frequency" with which they use it. These additional insights allowed us to consider each feature within additional dimensions, both individually as well as within their respective groups. Based on hierarchical clustering, our findings capture six sensible UI feature groups: User support; legal and policy compliance; data security and privacy; profile and account management; visibility control; and user experience customisation. Visualised in Figure 4.3, each group is situated within four quadrants that spanned between our participants' importance and frequency ratings. In the figure, each dot represents a single UI feature. Convex hulls further highlight the groups' locations within this space.

Generally, our findings support the values of user-centred design principles to align digital interfaces with their users' expectations. As a dark pattern, *Labyrinthine Navigation* obstructs users' from customising settings in line with their preferences (Mildner *et al.*, 2023b) or leaves them in the dark regarding *Bad Defaults* (Bösch *et al.*, 2016) when privacy settings default to

Fig. 4.3 This figure contains a scatterplot and their convex hulls for each of the six UI feature groups. Each subfigure describes one group and its distribution across the importance and frequency dimensions. This figure was first published in Mildner *et al.* (2024c) (P8).

meet service providers' interests instead of their users'. Moreover, the interface analysis conducted in P1, previously described in Section 3.1 suggests that many features our participants deemed important are not readily accessible, such as UI features allowing users to log out or delete their accounts. By not aligning UI features with user preferences, labyrinthine interfaces enhance the effects of the privacy paradox and similar incongruent behaviour between users' values and behaviour.

4.3 Recognising Dark Patterns

Guided by the findings of P1 and continuing prior efforts from P3, I wanted to investigate the ability of users to identify dark patterns in the context of SNSs. In fact, the studies in P4 were conducted simultaneously with those in P3, aiming to detail the presence of dark patterns in SNSs and their impact on users. Thus, P4 also concerns the four SNSs platforms: Facebook, Instagram, TikTok, and Twitter. As described in Section 2.3, a general problematic among users to recognise dark patterns has previously been studied, with critical work spearheaded by Di Geronimo *et al.* (2020) and Bongard-Blanchy *et al.* (2021). However, at the time, no particular study has investigated SNSs to the same degree we aimed to do. The relevance for an SNS-specific study received additional support after learning about certain conceptual differences between dark patterns deployed in SNSs and elsewhere.

In the first study, six expert participants demonstrated that schooled eyes were able to recognise dark patterns. However, each participant would miss certain instances, with only the accumulated data allowing a sense of completion. Still, the hidden nature of dark patterns leaves doubts about whether the experts were able to capture all dark patterns deployed by

Mathur *et al.* (2019)

Dark Pattern Characteristics

Characteristic	Question
Asymmetric	Does the user interface design impose unequal weights or burdens on the available choices presented to the user in the interface?
Covert	Is the effect of the user interface design choice hidden from the user?
Deceptive	Does the user interface design induce false beliefs either through affirmative misstatements, misleading statements, or omissions?
Hides Information	Does the user interface obscure or delay the presentation of necessary information to the user?
Restrictive	Does the user interface restrict the set of choices available to users?

Table 4.1 This table lists the introductory questions Mathur *et al.* (2019) gave for each dark pattern characteristic. We used these questions in P4 to detail SNS users' ability to recognise dark patterns from screenshots.

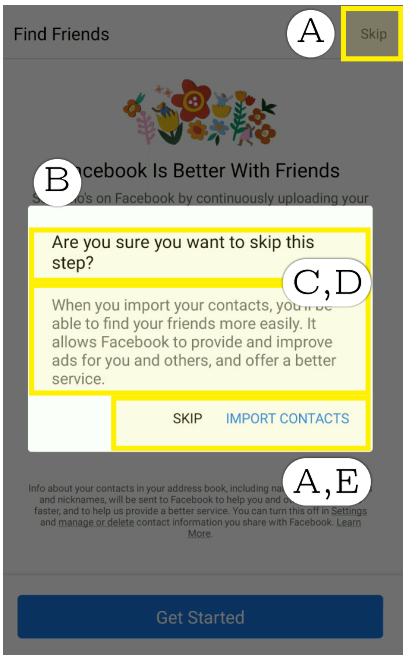
the four SNSs. This expert analysis, based on a cognitive walkthrough (Polson *et al.*, 1992) and taking inspiration from prior work conducted by Di Geronimo *et al.* (2020), formed a baseline for this work.

In a second study, we proceeded to research SNS users' ability in this respect. Instead of conducting further cognitive walkthroughs, I designed an online study by sampling screenshots from the video material recorded during the expert analysis and asking 193 participants to rate various SNS UI screenshots during an online survey. Each participant rated a total of sixteen images, eight of which contained dark patterns based on the experts' reviews and another eight that did not contain any dark patterns. In all instances, we revised the screenshots to ensure the presence or absence of dark patterns. Figure 4.4 contains a subset of four screenshots used in the study. The highlights around dark patterns served for clarity in the context of the publication and this thesis and were not included in the actual survey.

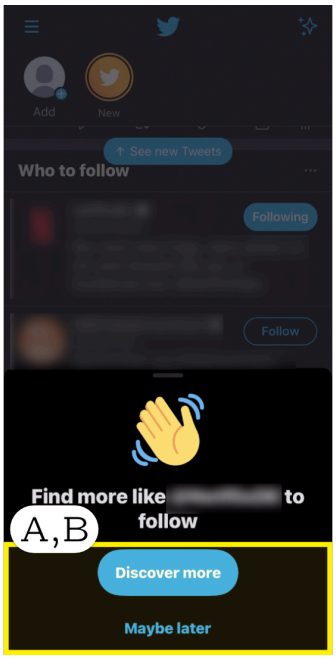
Participants rated each screenshot based on a custom questionnaire that drew from the five dark pattern characteristics as described by Mathur *et al.* (2019). The original authors introduced each characteristic through a brief introductory question that allowed an easy application for responses in a Likert-scale format (see Table 4.1). Moreover, the survey included a definition of dark patterns (as described in Mathur *et al.* (2021)) and a general question asking about any presence of dark patterns per screenshot[1].

While the generalised question opened a binary scope into participants' abilities to recognise dark patterns, the five characteristic-based questions allowed detailed observations across five dimensions, as proposed by Mathur *et al.* (2019): (1) Asymmetry; (2) covert; (3) restrictive; (4) deceptive; and (5) hides information. In both cases, we identified significant differences in our participants' ratings between the groups of screenshots. Consequently, SNS users are generally able to recognise dark patterns. However, the results spotlight the vagueness with

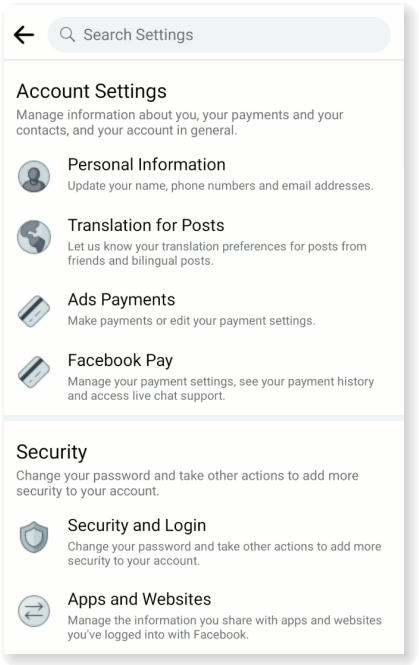
[1] In the survey, we replaced the term "dark pattern" with "malicious design" as we suspected most participants being unfamiliar with this subject's terminology.



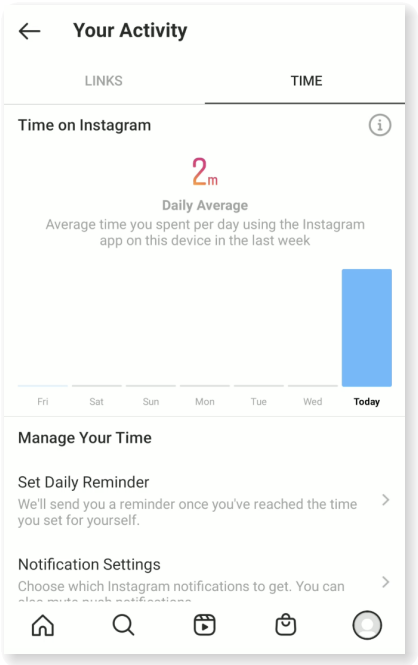
(a) Facebook Screenshot With Dark Patterns



(b) Twitter Screenshot With Dark Patterns



(c) Facebook Screenshot Without Dark Patterns



(d) Instagram Screenshot Without Dark Patterns

Fig. 4.4 Among others, these four screenshots were used in Study 2 and sampled from Study 1 of P4. **Figure 4.4a** contains the dark patterns *Interface Interference (A), Confirmshaming (B), Address-Book Leeching (C), Privacy Zuckering (D),* and *Visual Interference (E)*. **Figure 4.4b** contains the dark patterns *Interface Interference (A),* and *Visual Interference (B)*. Importantly, **Figure 4.4a** and **Figure 4.4b** were presented to participants without annotations. Neither **Figure 4.4c** nor **Figure 4.4d** contain any dark patterns. In total, sixteen screenshots were used in Study 2 of P4 — eight containing dark patterns and eight that do not. The figure was published in Mildner *et al.* (2023a) (P4).

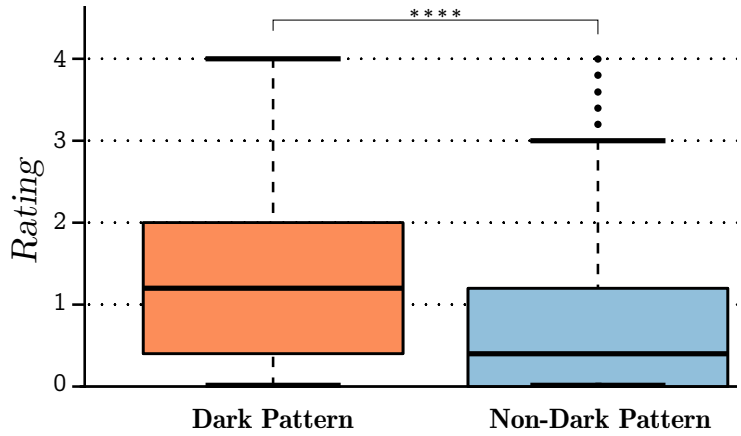


Fig. 4.5 This boxplot graphic compares users' ability to identify problematic design in SNSs screenshots containing dark patterns and screenshots that do not. Ratings on the y-axis were determined by accumulating five individual scores that were based on the dark pattern characteristics proposed by Mathur *et al.* (2019). Per characteristic, participants were asked a question in a Likert-scale fashion from 0 ("Not at all") to 4 ("Extremely"). This boxplot is published in Mildner *et al.* (2023a) (P4).

which our participants were able to do so. Visualising the small difference, Figure 4.5 compares the accumulated ratings of the characteristic questions. Although the ratings show significant differences, both mean ratings remain close together with standard deviations spreading widely. This finding is in line with Bongard-Blanchy *et al.* (2021), who also observed users' ability to identify dark patterns across multiple digital interfaces but noted similar challenges. Together, the studies investigating users' inability to effectively safeguard themselves from dark patterns foreshadow a necessity for other safeguarding measures to protect them instead. The limitation across these studies to mainly focus on still images further informed the idea of temporal analysis of dark patterns (TADP) to concentrate on more realistic user scenarios where dark patterns appear together and in sequential as discussed in P7.

These results already present answers to the second research question of this thesis. But at the same time, I missed some of the reasons explaining the struggles among users. While Publication P4 spotlighted the challenges to avoid dark patterns, P1 and P8 surfaced certain limitations regarding users' expectations with digital interfaces as well as incongruent behaviour. To this end, I wanted to gain additional qualitative insights into the perspective of users that may provide explanations for the misalignment of design and users' expectations.

4.4 Listening to Users

The apparent gap between the design and user angles, with respect to establishing and breaking expectations as well as harming users through dark patterns, has motivated me to devise a qualitative interview study inviting researchers, practitioners, and users to share their views, published as P5. Shifting the focus away from SNSs towards CUIs systems, I saw the opportunity to get a headstart on identifying potential ethical design problems as the technology was still in its infancy, as compared to widespread GUI-based dark patterns. In Section 3.4, I previously delved into the study, particularly into the role design plays in establishing expectations. Here, I want to focus on the users' perception and setting of expectations.

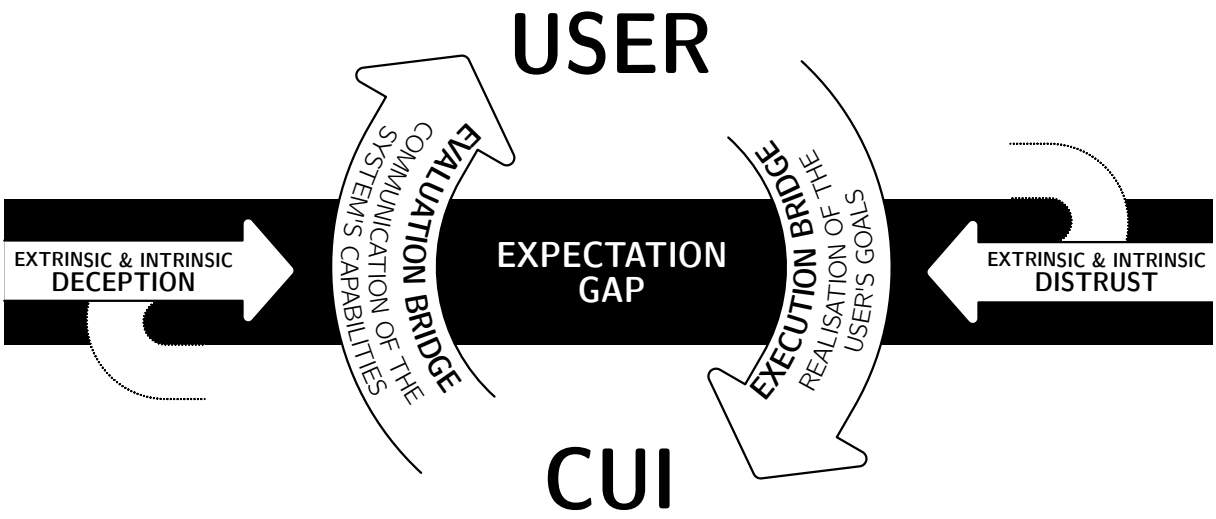


Fig. 4.6 Created to be agnostic of specific domains, the CUI Expectation Cycle (CEC) comprises the gap between user and CUI expectations. The model adopts Norman (2013) gulf of execution and evaluation to bridge the *Expectation Gap*. Either bridge includes a delimiter, both intrinsic and extrinsic in nature, impairing the bridges' purposes: *Deception* delimits the *Evaluation Bridge* and *Distrust* the *Execution Bridge*. The CEC was first published in Mildner *et al.* (2024a) (see Publication P5).

While practitioners iterated over the diverse capabilities contemporary CUIs were able to perform, users faced frustration as their devices did not react to prompts the way they would imagine. The responses from interviewed users brought forth a series of design issues, some of which were based on misconceptions; other issues, however, manifested as dark patterns. We noticed particular challenges for vulnerable users, including elderly or technologically illiterate users who do not always understand the consequences of their actions, leaving further room for exploitation. Collectively, the interviews across the three cohorts — researchers, practitioners, and users — led to constructing the CUI Expectation Cycle (CEC), a model bridging the gap between CUI design and user expectations (see Figure 4.6). Thereby, the model adopts established concepts from Norman's Action Cycle (Norman, 2013) and Oliver's Expectation Confirmation Theory (Oliver, 1977; Oliver, 1980). As a model that guides the development of CUIs technologies, the CEC is part of the guideline angle in Chapter 5. However, just as the *Responsible Design Triangle* describes a loop with no single starting point to set it off, the CEC is ideal for looping into each of its angles, offering particular insights into the users' angle.

The CEC is the result of a qualitative analysis of interview data involving researchers, practitioners, and users of CUI technology and was inspired by contemporary work on unethical design at the time. After analysing the transcripts, we noticed different perspectives across cohorts regarding interactions with CUIs. To accommodate these individual points of view, the CEC considers the system on one side and its users on the other, divided by the *Expectation Gap*. The model provides two bridges to overcome this gap: The *Evaluation Bridge* and *Execution Bridge*. These bridges adopt Norman's gulfs of evaluation and expectation (Norman, 2013), demonstrating how a system should communicate its capabilities to allow users to shape realistic goals. However, two delimiters negatively impact these bridges. *Deception* endangers the *Evaluation Bridge* when design choices are not truthful to a system's capabilities, obfuscating a user's expectations. On the other side, *Distrust* can disable the *Execution*

Bridge, infringing user goals through opaque consequences. Both of these delimiters can be extrinsic and intrinsic in nature — as previously described Section 3.4. This means that past experience with companies and devices (intrinsic) as well as design choices (extrinsic) impact users' perception when engaging with CUIs.

4.5 Avoiding Dark Patterns — Answers for Research Question 2

In this chapter, I followed four studies illuminating a recurrent detachment of the design of digital interfaces and users' expectations. Mirroring certain design implications of breaking user expectations, made in Chapter 3, the Publications P1 and P8 showcase users' behaviours and preferences in the context on SNS interfaces. The work further elaborates how incongruent user goals and actions add design challenges while leaving space for exploitation in the form of dark patterns. Taking a user-centric view throughout these works, I identified roots partially in the negligence of user needs and their perceptions of system capabilities (P5), as well as the implementation of unethical design and dark patterns that exploit errors P4. Collectively, the publications included in this chapter provide answers to reflect on the second research question about users' ability to safeguard themselves.

To this end, the work done in P4 is central in providing an answer to this research question. Reproducing similar studies (Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021) within the context of SNSs, this body of research observed users' inability to avoid dark patterns well enough to not fall victim to their trapping practices in several instances. Although the study showed that users generally noticed differences between the screenshots containing dark patterns and those that did not, the relatively low scores with which they were rated suggest that most users would not be able to avoid harm from dark patterns. Whether intentionally deployed or by mistake, the harm remains with the users. As a consequence, the burden to protect themselves must not fall solely on the users' end.

However, P4 presents an interesting opportunity to investigate possible reasons. While the incongruent user behaviour in P1 misses insights that would explain their willingness to spend more time on SNSs than planned for, the domain-specific dark patterns described in P3 and the inability to recognise those in P4 suggest certain design mechanisms that coerce SNS users' behaviour while keeping their satisfaction with the platforms high enough to keep using them. Returning to this phenomenon, P8 revealed the important role SNS users' expectations play when designing ethical user interfaces. Obfuscating interfaces dissuade users from reaching their goals by amplifying their struggles to navigate complex digital interfaces successfully. The *Labyrinthine Navigation* dark pattern (P3), further demonstrated in the study included in P8 presents a clear example in this regard. With hindsight to P1, the study showed that many features were not necessarily placed to be easily accessible.

The role of user expectations and their relationship with dark patterns when misguided or being broken recurred throughout my work and led to the development of the CEC, part of P5. Although expectations are critical for users in order to develop goals from their perceptions (Norman, 2013), dark pattern scholarship leaves no doubt about how easily users can get manipulated or deceived. Addressing ethical caveats for the design of CUIs while attempting to be domain agnostic, the overall purpose of this model is to serve as a guiding tool for

practitioners, uncover design problems, and avoid unethical design in their systems. As such, it will also be relevant in the following Chapter 5. Here, its purpose is to offer additional insights explaining the cause for unrealistic expectations through deceptive design practices or trust issues from a user's point of view. The findings of P5, including the CEC, help to fill the gaps left by P1 and P8 by elevating user expectations, shaped through their perceptions of a design and therefore what it offers them. A constant reminder in this thesis is the pertinence of good design to allow users to make realistic assumptions of capabilities, informing realisable goals. However, as a tool, the model is limited to only concern design problems not resulting from malintent. While the CUI can serve to protect users when used alongside design development, it is up to practitioners to pick it up and utilise it willingly.

Although more work is needed to fully grasp how users are affected by dark patterns, specifically vulnerable populations, this chapter has shown that users are not equipped to safeguard themselves from dark patterns and their harmful effects. It further dug up a root issue in misaligned user expectations and surfaced a gap between users and practitioners, particularly designers, with room left for better, user-centred dialogues.

CHAPTER 5

The Guideline Angle

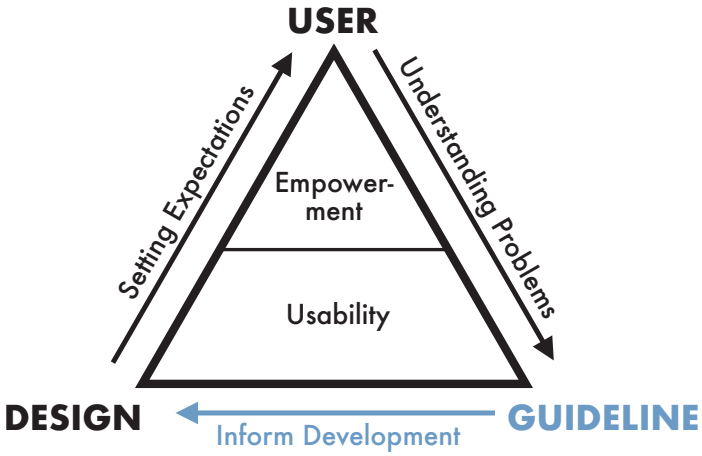


Fig. 5.1 Responsible Design Triangle highlighting the guideline path covered in this chapter.

Some dark patterns cause immediate harm and frustration (Gray *et al.*, 2018), while others maintain user satisfaction over periods of time (P3, P6). Nonetheless, by collectively restricting users' choice architecture (P3), breaking their expectations (P5), and being difficult to avoid (P4), the topic demands special requirements in order to protect users. Therefore, I included the guideline angle in the *Responsible Design Triagnle* to instantiate a third perspective whenever design harms its users who are incapable of protecting themselves. In such cases, guidelines can function as mediators that inform ethical and user-centred design through consulting the needs and expectations of users. Because this thesis' scope is in HCI, I want to reiterate that my main concern is guidelines for designers and other practitioners to consider during development phases. Still, recent movements from regulatory bodies and policymakers across the globe[1] reflect shortcomings of guidelines when greed and malintent dictate service providers' incentives to neglect their users' well-being. Thus, I see an opportunity for the guideline angle to include legal statutes when design guidelines are not sufficient in safeguarding users.

Any regulatory intervention yields certain risks in our Western economic systems (Majone, 1999). Thus, the focus on promoting ethical and responsible design guidelines, that cater both to service providers and users by aligning their incentives, is a crucial step for safeguarding. Before diving deep into the publications part of this chapter, I want to take a step back and look at certain milestones for responsible, sustainable, and ethical design in HCI.

[1] For example, European Commission (2022a), European Commission (2022b), California State Legislature (2018), California Privacy Protection Agency (2022), Ministry of Consumer Affairs, Food & Public Distribution (2023)

A centrepiece to every following work on ethical design in the field of HCI, work by Friedman and Nissenbaum (1997) set the foundation for VSD by pointing toward the relevance of systems to increase their users' autonomy. Continuing this strand of work, Friedman later collaborated with colleagues to establish the VSD principles (Friedman and Kahn, 2002). More than two decades later, this discipline can look back at flourishing work interested in increasing people's mental and physical health (Alqahtani *et al.*, 2021; Wagener *et al.*, 2023), well-being (Fleck and Fitzpatrick, 2010), and happiness (Desmet and Hassenzahl, 2012). Meanwhile, the term 'positive computing' (Calvo and Peters, 2014) has been fostered to describe supportive technologies whose aim is to impact people's lives positively. Following this trend, designers and developers have branched out to make technology more accessible for, but not limited to, marginalised demographics (Harrington *et al.*, 2022) or people with disabilities (Pradhan *et al.*, 2018). Fortunately, great work has been done to improve situations for many people through technology. Unfortunately, mentioning all noteworthy contributions of this kind would exceed the bounds of this dissertation.

On the other side of these positive ambitions are current efforts to fight unfair technologies and unethical design practices. Important work — that keeps motivating my research — stems from the non-profit organisation Algorithmic Justice League (Buolamwini *et al.*, 2016), founded by Joy Buolamwini in 2016. The team behind this fantastic project noticed potential discrimination and harm caused by technologies that neglected fundamental principles of inclusiveness and fairness. Surveillance technologies, for instance, were developed including strong racial biases (Gray, 2020; Benjamin, 2023) that present disadvantages to many populations who are not represented adequately and do not fall under the categories "white", "cis", "male", and "western" (Harrington *et al.*, 2022). In a similar vein, work by (Henrich *et al.*, 2010) explores diversity in research — and the lack thereof. With the acronym W.E.I.R.D. ("western", "educated", "industrialised", "rich", "democratic"), the authors describe sampling issues of psychology and behaviour-related research in cross-cultural contexts. Even though the acronym does not include equally important factors, such as race, it shines a light on the often ignored gap between technologies and their users, which aligns well with findings in P5.

On the shoulders of these important works, this chapter looks forward to providing ethically aligned tools for practitioners to develop designs that meet users' expectations and afford their needs. In doing so, it offers answers to the third research question of this thesis: **RQ3: *Which ethical design considerations are necessary to avoid the implementation of dark patterns?***. Although my work is largely situated in SNS or CUI contexts, these works also carry relevant implications for the design of any ethical UI. While staying truthful to their origins, this chapter discusses underlying implications in a more general scope. Understanding the challenges raised in the design and user angle, this chapter aids the identification of dark patterns while surfacing root mechanisms that enable their exploitative nature. This chapter is based on contributions from the following publications:

P2 Mildner, T., Doyle, P., Savino, G.-L., and Malaka, R., "Rules Of Engagement: Levelling Up To Combat Unethical CUI Design," in *Proceedings of the 4th Conference on Conversational User Interfaces*, 2022, ISBN: 9781450397391. DOI: 10.1145/3543829.3544528

P4 Mildner, T., Freye, M., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "Defending Against the Dark Arts: Recognising Dark Patterns in Social Media," in *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA*, 2023. DOI: 1010.1145/3563657.3595964

P5 Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R., and Niess, J., "Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642542

P9 Mildner, T., Inkoom, A., Malaka, R., and Niess, J., "Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Dark Patterns," arXiv:2405.07378 [cs], 2024

As the first author of the publications P2, P4, P5, and P9, my contributions were to provide the initial idea behind the papers, the drafting of the manuscripts, design and conduction of the studies, data collection, as well as the analysis and interpretation of results or findings. I was further responsible for submitting the final manuscripts.

5.1 Ethical Caveats for CUI Design

Unlike dark patterns in SNSs, or other GUI contexts where their manifestation has been relatively well understood, other domains lack the same depth of research. Exploring ethical implications of language or speech-based interactions as well as the mitigation of dark patterns in underlying technologies, P2 opens the discussion for potential dark patterns in CUI contexts. As the technology is still in its infancy, P2 conceptualises how interactions can be designed to exploit users in order to raise awareness about unethical practices. Moreover, with this work, we wanted to get a head start to understand the ethical caveats for CUI design and warn about potential design exploitations.

Picking up the provocations made in P2, the interview study in P5 contributed to each angle of the *Responsbile Design Triangle*.As a key contribution to answering RQ1, this publication describes the importance of user expectations and their implications on ethical caveats for CUIs. On one hand, the work highlights the relevance of maintaining expectations while staying truthful to a system's capabilities. On the other, it further outlines the implications of exploiting, breaking, or bending expectations, leading to dark patterns. Relevant to RQ2, we built on these implications to construct our CEC model. The CEC offers guidance for designers to acknowledge users' expectations as discussed in Section 4.4.

Finally, the work makes considerable contributions to answering RQ3. To this end, I want to focus on the findings the CEC is based on, namely, what we coined as five ethical caveats

Theme	Ethical Caveat	Guiding Question
Building Trust and Guarding Privacy	Users feel vulnerable to use CUIs, posing a need for CUI developers to prioritise transparency and control over data handling.	Does the system/interaction provide accessible and transparent information about personal data with easy control thereof?
Guiding Through Interactions	Guidelines and frameworks need to educate developers to develop accessible CUIs that empower users with diverse technological literacy to confidently interact with available features.	Does the system/interaction adequately inform users about its technical capabilities to enable full utilisation of its features?
Human-like Harmony	Anthropomorphic features should be implemented with care and in line with a CUI's capabilities to support intuitive and authentic interactions, preventing unrealistic expectations.	Does the system/interaction clarify the presence of anthropomorphic features to avoid misconceptions and unrealistic expectations?
Inclusivity and Diversity	The development and design of CUI interactions need to consider individual needs and characteristics of users, especially marginalised groups, ensuring equitable CUI interactions.	Does the system/interaction cater towards users with diverse needs, potentially through alternative interactions where otherwise inaccessible?
Setting Expectations	CUI capabilities should avoid deceptive interactions and, instead, be transparent to users to prevent frustration and mistrust.	Does the system/interaction handle user prompts truthfully, clarifying the scope of its capabilities to provide realistic expectations?

Table 5.1 This table summarises the five identified themes and design questions per ethical caveat. It was published in Mildner *et al.* (2024a) (P5).

for CUI design. These ethical caveats are the result of a thematic analysis of all 27 interviews conducted between researchers, practitioners, and users of CUI systems. Table 5.1 lists each ethical caveat per theme alongside a guiding question that we propose for practitioners to reflect upon during the development stages of their systems. Below, I introduce each of the five themes and their ethical caveats based on their introduction in P5:

Building Trust and Guarding Privacy: Operating Extrinsic and Intrinsic Factors This theme spotlights extrinsic and intrinsic challenges that result in trust and privacy deficits in CUI interactions. Across cohorts, a reoccurring theme echoed fears of untrustworthy handling of personal data, kindled by prior negative experiences and the reputation of companies and practitioners. A call from researchers to increase transparency to bridge users' concerns was underlined by users' mention of their own safeguarding strategies, which limit interactions to basic functionalities and do not require them to disclose private information. Practitioners

acknowledged these problems but noted counterintuitive regulations and design limitations obscuring access to settings rooted in speech-based interactions.

Guiding Through Interactions: Overcoming Knowledge Gaps This theme illuminates the importance of providing users with informed guidance about possible CUI interactions, as a lack of technological literacy and limited experience results in difficulties in accessing available features. While researchers noted that research findings need to be better aligned with industry development, practitioners echoed a desire for further guidelines.

Human-like Harmony: Providing Authentic Anthropomorphism This theme describes the importance of implementing an appropriate amount of human-like features so that a system can support intuitive interaction without eliciting false beliefs that may leave users feeling deceived. Finding the right balance between humanness and technological limitations is challenged by the phenomena of anthropomorphism. That is the tendency to attribute human characteristics to non-human objects that most people engage into some degree (Waytz *et al.*, 2010) from early childhood (Piaget, 1997). Anthropomorphous behaviour appears significantly heightened in dialogue with technological devices endowed with gendered voices and names (Gong and Nass, 2007). Other influences that might encourage anthropomorphous behaviour in this context include: expectations for social affordances implied by representations of speech interfaces in media and advertising (Murad and Munteanu, 2020); the fact that these systems conduct tasks typically carried out by humans using human language (Nass *et al.*, 1994); and that language use itself might be inherently social and agentic (Fausey *et al.*, 2010; Jia *et al.*, 2013). When developing CUIs, practitioners should conscientiously navigate the amount of human-like features to avoid inadvertently manipulating our bias for anthropomorphic characteristics.

Inclusivity and Diversity: CUIs in the Wild This theme delves into the unexpected design and interaction challenges emerging when CUIs are introduced to diverse user groups with distinct characteristics and needs. Participants spotlighted several groups susceptible to poor design choices or technical limitations within CUI interactions. They emphasised a regression towards the mean, acknowledging how CUIs are designed with an "average" user in mind, leading to the "othering" (Mengesha *et al.*, 2021) — the marginalisation or exclusion of certain people — of individuals and groups that do not fit this profile. This encompasses users with stronger accents, colloquial dialects, second language users, people with deficits in speech or cognition, or users with lower technical literacy.

Setting Expectations: Transparency to Mitigate Frustration This theme emphasises the pivotal role of transparency when designing CUI interactions to set realistic expectations. Unintended device reactions can result in disappointment and frustration in the user when a device appears more capable than it actually is. To deliver users an optimal experience, commercial incentives should be aligned with human-centred practices of HCI to avoid ethical concerns tied to deceptive and manipulative interactions. Instead, transparent and evident CUI capabilities should empower users to use the system autonomously and easily.

Setting the foundation for the CEC, these five ethical caveats provide practitioners with grounding resources and create awareness for avoiding unethical CUI design. Thereby, each ethical caveat includes topics that draw from previous related work and align with the core principles of VSD. Conclusively, they promote the development of truthful and accessible (CUI) technologies that empower users to use them on their own terms. Moreover, they can be useful for the design of any UI. I phrased the guiding questions (see Table 5.1) deliberately to be agnostic of specific technologies to support designers of different systems or interfaces. Users' trust and privacy, for instance, are just as relevant for SNSs or any online application, as it is for CUIs. The importance of guidelines to empower users with diverse technological literacy can also be extracted from CUI contexts, as users should be kept aware of any system's capabilities and features. Anthropomorphism is part of the innate nature of every human, making the third theme important for any UI where its users may allocate human traits. Diverse user needs also require special catering in non-CUI interactions. Here, screenreaders or haptic feedback are excellent examples of supporting blind or visually impaired users to access GUI elements. Finally, representing capabilities truthfully, as prompted by the fifth guiding question, is an important consideration for any GUI as well. While all contributions of P5 are the result of CUI-based research, the work adds to the overall understanding of design and how dark patterns manifest. Nevertheless, any such implications should be taken tentatively when considered outside their original contexts.

5.2 Identifying Dark Patterns

While P2 and P5 provide insights into ethical considerations to avoid dark patterns, particularly for CUIs, P4 explores how to identify the presence of dark patterns. Although we conducted the work across SNS platforms, it promotes an approach to assess dark patterns in any UI. While effective, the process used in P3 to identify existing and uncaptured dark patterns in SNSs, by utilising a comprehensive dark pattern taxonomy, was lengthy and would likely be inefficient in any other scope. After reflecting on the implications for identifying dark patterns, I found that this approach may not be feasible in practice, especially in regulatory procedures, and may introduce a cumbersome task instead. While P6 addresses the issues of differing terminologies by offering a shared vocabulary in the form of an ontology, P4 considers a streamlined and simplified process for their identification.

Building on the proposition made in P2, we tested the feasibility of dark pattern characteristics by Mathur *et al.* (2019)[2] as dimensions in this regard. The results from our online survey, described in Section 4.3, indicated that these dark pattern characteristics can be used to distinguish between interfaces that do and those that do not contain dark patterns. To this end, Figure 5.2 envisions a possible implementation to assess dark patterns across the five dimensions in UIs. The included diagram follows each of the five dimensions based on

[2] The five characteristics include asymmetry, covert, deception, information hiding, and restriction. For their full descriptions, Table 4.1 in Chapter 4 presents each characteristic individually. In 2021, Mathur *et al.* (2021) introduced a sixth characteristic with disparate treatment, mainly based on the gaming dark patterns from Zagal *et al.* (2013).

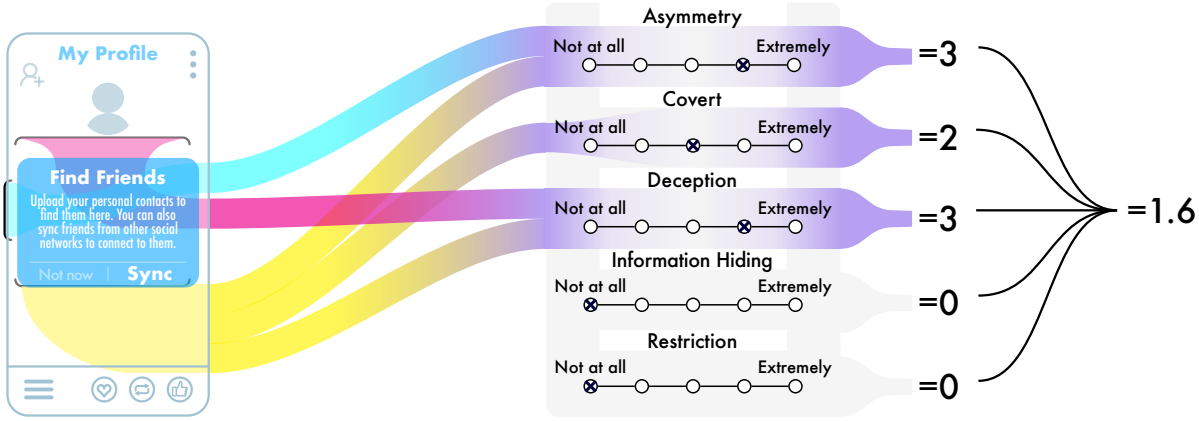


Fig. 5.2 This diagram demonstrates how the dark pattern characteristics by Mathur *et al.* (2019) can be used to determine a dark pattern score, using five questions with Likert-scale ratings. This Figure was published in Mildner *et al.* (2023a) (P4).

the same sample screenshot. For each dimension, a subsequent question is answered in a Likert-scale fashion from "Not at all" to "Extremely", resulting in a score.

In our analysis in P4, we considered the multiple dimensions individually as well as an accumulated score. The differentiation of dimensions could be useful to identify specific causes of unethical design, whereas a mean would allow a more general estimate of interfaces. Finally, the resulting score(s) allow assessments of interfaces in the complex continuum between dark patterns and responsible design, further explored by the expanded *Responsible Design Triangle* in Figure 1.3, Chapter 1. Thereby, this process offers a flexible determination of acceptable interfaces and what should be deemed unethical and problematic. In regulatory contexts, the threshold between dark patterns and responsible design could be set depending on the context and constraints.

Although the result of studying e-commerce websites, a benefit of the characteristics by Mathur *et al.* (2019) lies in their domain-agnostic nature. Thus, they allow for our proposed procedure to work outside GUIs as well as in TADP contexts, when dark patterns become emergent after sequential interactions. Moreover, the procedure could be extended by new characteristics in the future. Since conducting our study, Mathur *et al.* (2021) published a sixth characteristic to introduce a dimension for "disparate treatment" of users. While not included in our diagram, additional dimensions could offer more granulated insights into dark patterns' presence. Furthermore, our dark pattern ontology (P6) could serve in a similar way.

5.3 Relationship between Cognitive Biases and Dark Patterns

While analysing digital interfaces for the presence of dark patterns offers one way to safeguard users, it is just as important to understand the underlying mechanisms through which dark patterns affect the choice architecture of their users. A considerable amount of effort has gone into research that studies persuasive techniques that influence people's choice architecture — most prominently surrounding the concept of nudges (Thaler and Sunstein, 2008). As potential enablers for dark patterns, as shown in Publication P9, I want to dive into the underlying concepts before continuing to describe our findings.

5.3.1 The Discourse Surrounding Nudges

In the words of their originators, a nudge is "is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not" (Thaler and Sunstein, 2008). As this quote makes apparent, a core concept of nudges lies in their avoidability. A prominent positive and frequently referenced example stems from public health strategies of various nations regarding organ donorships as opt-in or opt-out choices (Sharif and Moorlock, 2018; Molina-Pérez *et al.*, 2019). Germany, for instance, asks its citizens to actively decide to become organ donors. Other countries, such as Austria or France, consider all citizens donors unless they actively decide against it. Here, anchoring and pre-selection biases positively impact public health aspects of nations when strategies deploy an opt-out approach.

Despite such positive notions, exploitation of nudges has been shown to negatively impact people's choice architecture with adverse effects on Thaler and Sunstein's original aims. Foreshadowing harmful implications, Munson *et al.* (2015) demonstrated how certain nudges can hinder people from achieving their goals. After publically sharing their personal aims, people's commitments decreased in order to avoid being criticised. These two examples hint at the responsibility necessary when choosing to use design tools as powerful as nudges.

Thaler and Sunstein's (Thaler and Sunstein, 2008) nudge theory illustrates the effectiveness of design decisions by altering people's choice architecture. Later work of the original authors reflects on critics against paternalistic and ethical implications on agency or exploitation of cognitive biases (e.g., (Rizzo, 2009; Hausman and Welch, 2010; McCrudden and King, 2015)). Appreciative of the concerns, Sunstein (2015) responded that a good nudge does not affect people's agency at all and, instead, provides healthy or good defaults. Thaler (2018) later explained how nudges were meant to help people make choices that they themselves feel good about. Instead, he proposed the term 'sludges' for negative and harmful exploitation. Following up on the dual-process model (Tversky and Kahneman, 1974), Hansen and Jespersen (2013) entered the discourse considerate of both streams and proposed two types of nudges in the form of an "epistemic distinction between transparent and non-transparent nudges" that accounts for unethical, manipulative exploitation. Recent work by Leimstädtner *et al.* (2023) built on this concept to study responsible nudges with results demonstrating the effectiveness of design friction to positively affect participants in making informed decisions.

Generally implicating the power and responsibility designers have, the Fogg Behavior Model (FBM) (Fogg, 2009) demonstrates how users' decisions can be steered toward desired goals using persuasive design techniques. While the definition of nudges provides the important criteria that they should be "easy to avoid" (Thaler and Sunstein, 2008), upholding users' autonomy, persuasive design includes a recipe to address motivation and ability to afford a desired action. Although the aforementioned work includes often recited fundamentals of HCI scholarship, recent studies highlight their consistent impact on today's research (e.g., (Caraban *et al.*, 2019; Kornfield *et al.*, 2022)) and industry impact (e.g., (Souza-Neto *et al.*, 2023)). While studies utilise the FBM and derived methods to assist people in, for instance, health-related

contexts (Diethei *et al.*, 2021; Alexandrovsky *et al.*, 2021; Agapie *et al.*, 2022), it is important to recognise that participants made an autonomous decision and consented before giving up full autonomy of assistive health-care interventions. However, the same strategies can be exploited unknowingly by a user, posing ethical questions about intent and consequences.

The work on dark patterns introduced a plethora of examples of exploitation of trust and non-transparent display of options. While HCI research has predominantly concerned design artefacts to study the effects of dark patterns (Gray *et al.*, 2018; Gray *et al.*, 2024b), from a human cognition standpoint, it is often users' perception and decision-making that is of interest. In this vein, Chen *et al.* (2022a) investigated the role dual-process theory (Kahneman, 2003) plays in motivating impulsive purchases without giving customers much opportunity to rationally evaluate their decision. While effective for (online) sellers, the exploitation of cognitive biases and manipulation of users' decision-making poses serious ethical problems (Chen *et al.*, 2022b), which are only further enhanced through the misuse of personal data, as described by Nabbosa and Kaar (2021).

5.3.2 The Relationship Model of Cognitive Biases and Dark Patterns

The ethical tensions surrounding nudges and other persuasive techniques led me to design a study investigating a possible relationship with dark patterns. To this end, we conducted a focus group study involving dark pattern scholars and experts in psychology and cognitive science to explore the relationship dark patterns have with cognitive biases. The study concerns cognitive biases specifically, as nudges rely on them as well. Also, prior works have discussed that dark patterns exploit cognitive biases (Mathur *et al.*, 2019; Waldman, 2020). Building on these works, we were the first, to our knowledge, to investigate the particularities of the two domains at the time of conducting the study. In total, we conducted four focus groups with 15 participants, eight with backgrounds in dark pattern research and seven with backgrounds in psychology or cognitive science. Except for one, each focus group paired two experts from each domain to discuss similarities, differences, and possible facilitators between the two subjects.

We conducted a thematic analysis based on the focus groups' transcripts which resulted in the "Relationship Model of Cognitive Biases and Dark Patterns". Before finalising the model, we validated it through collected feedback gained from participants of the focus groups. Outlined in Figure 5.3, the final model comprises three stages encompassing five phases while following a process from design to its real-world impacts. Drawing inspiration from Verbeek's theory of technology mediation (Verbeek, 2005; Verbeek, 2006), the three stages of our model include: (1) Inscription/delegation from a designer's perspective, (2) mediation of the technology or particular interface, and (3) users' interpretation thereof. The model demonstrates how dark patterns emerge from exploiting cognitive biases.

To this end, our model breaks the three stages down into five phases. The first stage describes the designer's perspective to inscribe or delegate functionalities. Happening in phases one and two, design addresses particular cognitive biases determining the balance between autonomy versus coercion. The second stage — mediation — contains the third phase: the exploitation of cognitive biases, which can ultimately lead to deceptive practices and harm.

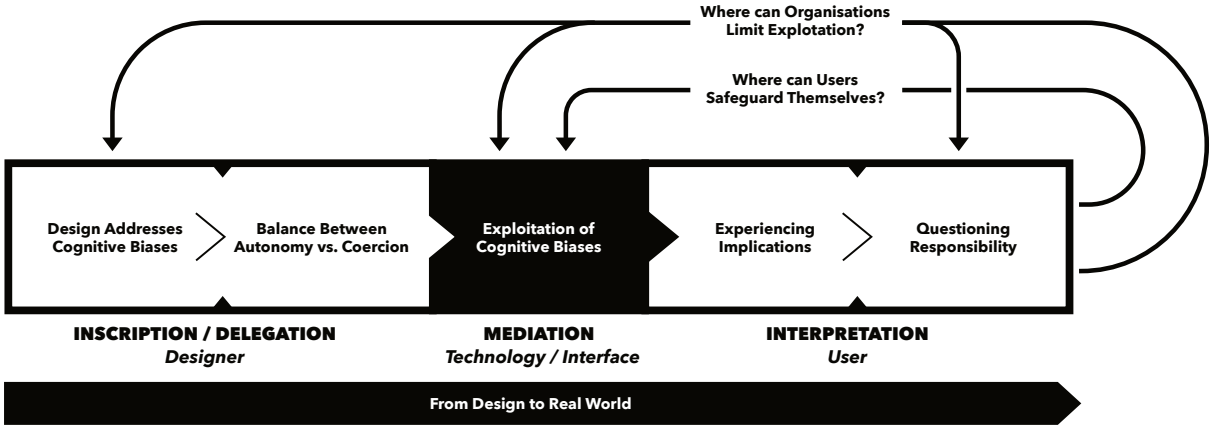


Fig. 5.3 This Figure presents the Relationship Model of Cognitive Biases and Dark Patterns. Following a continuum from (potentially unethical) design to real-world applications, the model comprises three stages spanning five phases. Adopting Verbeek's theory of technology mediation (Verbeek, 2005; Verbeek, 2006), the model follows designers' inscription of functionalities into technology to users' interpretation, leading to the questioning of responsibilities. Depending on the impact and implications of the (unethical) design, end-users may need safeguarding measures, while policy and regulation may be required for their protection. This Figure was first published in Mildner *et al.* (2024b).

The third stage focuses on the users' point of view, interpreting the design throughout the fourth and fifth phases. First, users experience the design's implications, leading them to question the responsibility of its cause. Furthermore, we identified crossroads for safeguarding strategies of end-users as well as opportunities for organisations to limit harmful design through exploiting cognitive biases. In the following subsections, we outline each phase in detail. We support the descriptions for each phase through quotes from our participants and connect individual phases to related work where applicable. For improved readability, we slightly altered some statements, ensuring words and sentiment were maintained.

We intended the Relationship Model of Cognitive Biases and Dark Patterns to support both researchers and practitioners by providing them with reflective phases that can support them in considering the ethical caveats and impacts of their designs. It is not meant to be prescriptive or paternalistic but provides a roadmap that highlights ethical implications throughout the design's lifespan and the interplay of cognitive biases and dark patterns. While the aim of our model cannot change any malicious objectives of practitioners, it can serve as a tool to reveal implied consequences of utilising cognitive biases in design and can guide toward potential countermeasures. The model can, therefore, be applied in situations where dark patterns are observed. Especially in the early development stages of designs, the Relationship Model of Cognitive Biases and Dark Patterns can complement decisions made alongside existing, traditional design paradigms that may not always prioritize user agency and autonomy. As such, it complements the CEC that takes a more user-centred approach.

5.4 Understanding Dark Patterns — Answers for Research Question 3

Covering the third angle of the *Responsible Design Triangle*, this chapter followed publications that produced tools to guide the design of ethical interfaces. When design fails to meet users'

needs or expectations and users cannot avoid exploitative techniques of digital interfaces, design guidelines can offer the necessary support and pave the way for safer interactions. Looking through a design lens in these studies, I presented means for practitioners to understand the ethical caveats of their work (P5) and identified underlying problems (P2, P4) before their designs enter the real world (P9). Thus, these publications present answers to the third research question of this thesis, asking about the ethical design considerations necessary to avoid implementing dark patterns.

In a first line of defense, P2 conceptualised a process to derive a dark pattern score based on the dark pattern characteristics by Gray and Chivukula (2019). Soon after, I continued developing this idea in P4, where I described it more systematically. It supports anyone in evaluating digital interfaces for the presence of dark patterns. This includes users curious about the design of systems they are using, practitioners who may want to confirm that their design does not expose their users to unethical design, or potentially regulatory bodies who require an efficient tool to estimate the harms caused by problematic interfaces. Overall, the possibility to asses digital interfaces through dark pattern characteristics informs about the unethical dimensions included in their design.

To avoid the deployment of unethical design from the start, P5 promotes five ethical caveats and respective guiding questions. Drawing on a growing body of value-centred research (e.g., Friedman *et al.*, 2013; Luger and Sellen, 2016; Harrington *et al.*, 2022), the findings in P5 amplified voices that called for increased alignment of design with users' expectations. Collectively, each ethical caveat provides a partial response to this chapter's research question. Developing systems that are truthful to their capabilities and offer inclusive and accessible interactions cater to diverse user needs while affording realistic user goals. Leading to the creation of the CEC, this publication demonstrates how to align users' expectations with system capabilities to support design processes with ethical considerations.

Fostering our conceptions of the foundational mechanisms that enable dark patterns, P9 further explores the intricate relationship between dark patterns and cognitive biases. The work contributes a model that presents a roadmap following design from its development to its implications when entering the real world. Acknowledging that all design relies on perceptual factors and human cognition, including cognitive biases and heuristics, the Relationship Model of Cognitive Biases and Dark Patterns further draws practitioners' attention towards the (im)balance between user autonomy and coercion as the latter leads to harmful interactions through exploitation of cognitive biases. Adopting the theory of technology mediation by Verbeek (2006), the model further reminds that once a designed interface is released into the real world, it is out of the control of its designer to mitigate negative consequences or inevitable harm. If respecting the phases of this model, practitioners will be reminded of their responsibilities to avoid unethical design practices that exploit cognitive biases.

Based on the ethical considerations provided by these works, I hope to offer practitioners sufficient guidance to reflect on the implications of their designs and avoid utilising dark patterns. Unfortunately, all these and other design countermeasures are without effect if practitioners cannot or will not use them. Aside from this, the ecosystem in which practitioners operate does not always allow the production of ethical designs (Gray and Chivukula, 2019;

Chivukula *et al.*, 2023). With hindsight to the limitations of this angle to remain within an HCI context, I do believe that it requires legal statutes if users' safety cannot otherwise be ensured.

CHAPTER 6

Discussing the Responsibilites Of Digital Interfaces

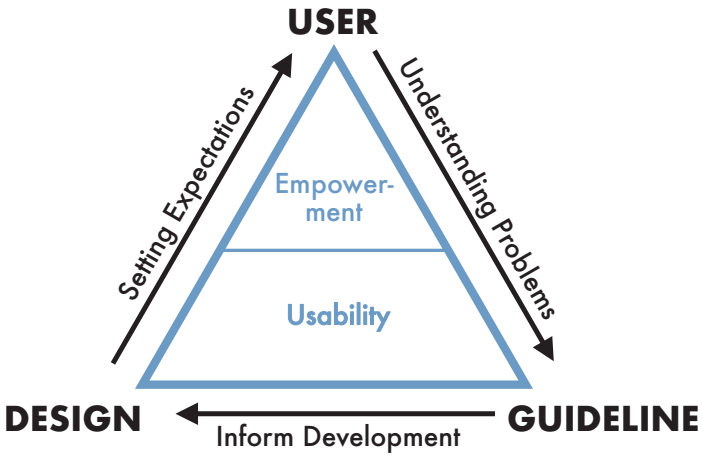


Fig. 6.1 Responsible Design Triangle highlighting the centre showing empowerment above usability covered in this chapter.

Everything artificial is designed (Ulrich, 2011; Norman, 2023). However, if everything artificial is designed, why does the design process seem so difficult? At least this thesis may make it seem so and leave the impression that the design of "good" UIs is rather rare. Fortunately, the discipline of HCI has blossomed throughout the past decades widening its original scope of human factors and user-centred aims, including traditional UCD goals, towards considering the impacts technologies have on societal scales (Dombrowski *et al.*, 2016) as well as individual and social well-being (Friedman *et al.*, 2013; Niess and Woźniak, 2020).

I would like to think that the publications included in this thesis do their part in making digital interfaces safer for their users. Nonetheless, I want to emphasise that technologies, such as SNSs, are not inherently "bad" or "evil" (Allen *et al.*, 2014; Given *et al.*, 2017). Whether unintentional harm of well-meant persuasion (Brynjarsdottir *et al.*, 2012) or greed-driven intentions in a surveillance capitalistic system (Zuboff, 2023), it is not necessarily the technology but its design that exposes unethical consequences on its users. Picking up critical voices (e.g., Gray *et al.*, 2018; Mathur *et al.*, 2019; Beyens *et al.*, 2020; Velthoven *et al.*, 2018) that echo negative consequences as a result of problematic designs, this thesis focused on the responsibility of designing interfaces that, rather than diminish user agency, empower them to use technologies for their individual means. Following the *Responsible Design Triangle* from my motivation in Chapter 1, I posed the research question: ***RQ: How can the responsibility of designs and their impacts be distributed between actors to protect users from deceptive,***

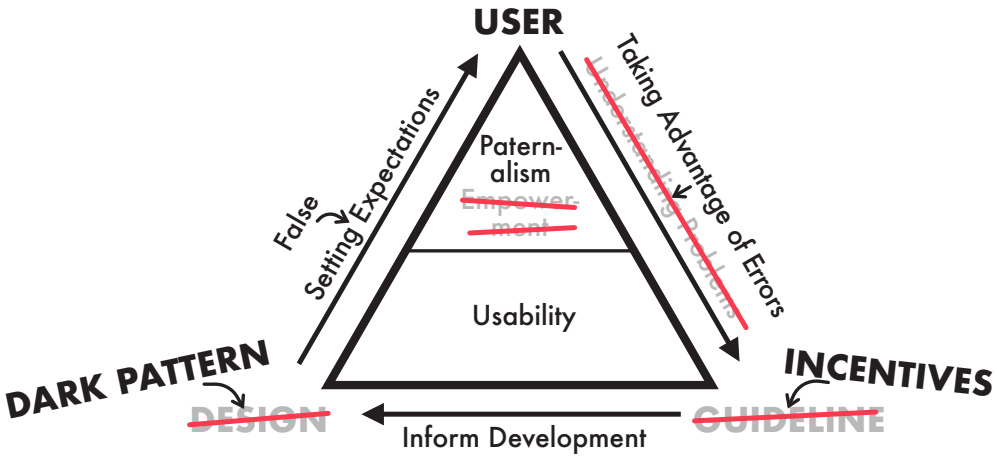


Fig. 6.2 The *Irresponsible Design Triangle* illustrates how incentive-oriented design may override users' autonomy by deploying dark patterns to establish false expectations. These lead to erroneous interactions that service providers can exploit.

unethical design practices and dark patterns? To allow for a detailed and thorough answer to this research question, I divided it into three questions and structured their answers around respective chapters, each tackling one angle of the *Responsible Design Triangle*.

Each angle takes a particular perspective relevant to describing the shared responsibility between actors. In doing so, they outline the centre of the triangle, which will be part of the focus of this chapter: the extension of usability through empowerment. Traditionally, to evaluate the design of systems or interfaces, usability is an often-used metric for commenting on the ease with which users can access and utilise functionalities. Today, usability has become a widely used application for understanding systems both in research and industry terms (Brooke, 2013). However, usability does not directly account for experience or emotional attachment to a system. Moreover, it fails to provide any ethical implications of a design. And while a plethora of metrics exist to evaluate UX (Laugwitz *et al.*, 2008) or, to a degree, emotions (Watson and Clark, 1988), ethics is a seemingly tricky topic to approach in this prospect (Gray and Chivukula, 2019; Gray *et al.*, 2023a; Sánchez Chamorro *et al.*, 2023), although a range of work provides numerous methods which include ethical considerations (Chivukula *et al.*, 2021b).

6.1 Empowering Users — Answers for the Meta Research Question

While raising the importance of including ethical considerations in design, the *Responsible Design Triangle* can be viewed as an ecosystem between its three actors. Any change needs careful reflection as — with any ecosystem — critical change may result in severe repercussions and cause a stable system to fall out of balance. In this vein, the current presence of dark patterns from the design angle poses a serious risk to the users' angle and requires compensation from the guideline angle. For demonstration, Figure 6.2 (showing the *Irresponsible Design Triangle* as a nod to incentive-oriented design restricting user autonomy) draws on the ethical mediators proposed by Gray and Chivukula (2019), exemplifying possibilities for dark patterns to occur when service providers' incentives become main drivers to inform design instead of

following human-centred or UCD practices or design guidelines that advocate design ethics. While the *Responsible Design Triangle* elevates user empowerment, dark patterns paternalise user choices and restrict autonomy. If service providers disregard user autonomy to accomplish their goals, design can devolve into exploitative dark patterns that set false expectations, resulting in erroneous user interactions. Reiterating this thesis' structure, I want to emphasise the responsibilities of each angle of the *Responsible Design Triangle*.

The Design Angle's Responsibility. As the emergent environment for digital interfaces, the design angle attributes all capabilities into interfaces, or in terms of Verbeek (2006), designers inscribe functionalities onto technologies. As any implications are out of the designers' hands, once their products leave their desks and fall into the hands of users, every precaution needs to be taken beforehand. Chapter 3 demonstrates the array of dark patterns in digital interfaces. The line between design aiding users through systems that present, at best, satisfying and enjoyable experiences and dictating persuasive technologies that govern their actions by restricting autonomy is often a fine one. As the degree to which users require assistance varies between contexts and situations, it remains within the design angle's responsibility to ensure no harm is done through developed technologies. This can be achieved through rigorous user studies and consultation of ethical design guidelines.

The User Angle's Responsibility. Along others (Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021), the publications mentioned in Chapter 4 imply that users' are not sufficiently equipped to recognise and avoid dark patterns before getting harmed. Consequently, these findings remove any responsibility from users to ensure their own safety, as they cannot be burdened to protect themselves. However, the user angle is responsible for giving users opportunities to voice their expectations, especially when digital interfaces fail to transparently communicate capabilities or deceive through deceptive design strategies. Echoing broken expectations and experienced harm is exceptionally important for vulnerable and too often overseen populations, especially with regard to design produced for the masses. Through understanding the roots of unethical design practices and caused problems, additional means can safeguard users where they are incapable of protecting themselves.

The Guideline Angle's Responsibility. The necessity of additional guidelines becomes evident considering the current dark pattern landscape, spreading throughout technologies. As long as the design angle fails to ensure user safety and continues to deploy unethical practices unavoidable to users, a mediating entity needs to ensure users' safety instead. This mediating role can be taken by design guidelines and aiding tools, such as those presented in Chapter 5. Alongside user-centred design processes, ethical design guidelines can remind practitioners about ramifications and inform the development of empowering interfaces that do not restrict users' choices. It is thus within the responsibility of the guideline angle to understand users' righteous concerns and provide the design angle with all necessary tools to avoid unethical designs, like dark patterns.

Collectively, each angle encompasses aspects of the shared distribution of responsibility and, thus, respective answers to my meta research question. This leaves the centre of the *Responsible Design Triangle* for me to discuss. The results and findings of the included publications spread across the three angles share a common theme of user-centred approaches for aligning technologies with users' values and intentions. Traditionally, usability is commonly attributed to five core dimensions: efficient, effective, error tolerant, engaging, and easy to learn (Quesenbery, 2003) which are sometimes slightly altered (Nielsen, 2012; Komninos, 2020). Together, these attributes translate to interfaces users should find accessible, easy to learn and to use, and appealing. Although error handling includes requirements for users to revert choices, the key aspects of usability do not necessarily empower users or facilitate their well-being.

By using the term empowerment, I am referring to the autonomy to make reflected and informed decisions and the enablement of self-determination that allows people to achieve personal goals. In terms of digital technologies and interfaces, they can empower users by retaining their autonomy and affording their goals. The concept is thereby very close to VSD (Friedman *et al.*, 2013). I acknowledge that not all technologies can be empowering in every sense their users may desire. Nonetheless, they can subscribe to empowering elements within their capabilities and avoid deception that may otherwise obfuscate users' perceptions.

Importantly, I do not intend to extend the concept of usability by another term. It conveys perfectly well what it attempts to do. Instead, I want to elevate user empowerment above usability as a superior goal for responsible and ethical design. While digital interfaces should maintain usability goals, arguably, dark patterns are very usable except for the error-handling dimension. Hence, additional ethical considerations for user empowerment could introduce the necessary constraints for mitigating dark patterns.

6.2 The Responsible Design Triangle Falling Out of Balance

Earlier, I described the *Responsible Design Triangle* as an ecosystem where each actor carries responsibilities that secure its overall stability. From a critical design perspective (Bardzell and Bardzell, 2013; Dunne and Raby, 2013), I want to discuss the ramifications of the model more reflectively when it falls out of balance. Not in the sense where malintent and incentive-oriented design leads to dark patterns (as shown in Figure 6.2), but where one angle fails to support the others.

Considering the responsibilities I ascribed to each angle, the design angle creates and releases design into the world, often following organisational goals or commercial imperatives (Chivukula *et al.*, 2023). If practitioners are, for whatever reason, unable to ensure that their designs foster ethical practices that empower their users, particularly vulnerable populations, before introducing them to the public, it is out of their control to stop immediate consequences. Revisiting Verbeek's theory of technology mediation (Verbeek, 2005; Verbeek, 2006), this lack of control highlights the necessity for reflexivity within the design angle, where designers critically examine underlying motives, values, and impacts in the real world.

Designs that are not necessarily dark pattern-related may still pose risks to users. These risks could leave them vulnerable to, for example, privacy issues without their or the practi-

tioner's awareness. Here, the user angle becomes critical. Users must be empowered to identify and articulate any issues they encounter. To this end, they would have to have the agency and knowledge to recognise the problems and communicate their concerns.

For smaller issues, designers could try to directly fix their design by understanding their users' perspective. In cases of more severe issues, it may make sense for the guideline angle to translate users' concerns and experiences into actionable design guidelines. This could ensure that similar issues do not reoccur in the future, if the practitioners in the design angle incorporate them into their future developments. By incorporating users' concerns and feedback into design practices, practitioners can build more responsible and sustainable solutions.

Depending on the magnitude of a problematic design, the *Responsbile Design Triangle* can be brought back into balance through the other angles, as long as only one fails. However, if multiple angles are unable to inform the others about a design's issues, the overall balance would be terminally endangered. When design causes problems to a user without their or the practitioner's awareness, even existing guidelines would be unable to rectify the causes.

6.3 Limitations

As with most academic work, this thesis and the included publications are limited in various ways. Most importantly, the proposed *Responsible Design Triangle* is an abstraction of concepts and relationships between its angles. It is an attempt to illustrate opportune moments to avoid dark patterns and utilise ethical design principles. As an abstraction, it may, however, oversimplify otherwise complex backgrounds. The development of design is constrained by various factors, such as organisational stipulations (Chivukula *et al.*, 2023). At the same time, users may not perceive their surroundings, including UIs, equally, changing the way dark patterns may impact individuals. Moreover, the thesis is the result of work done in the field of HCI. While it draws from relevant work in human psychology and cognition as well as regulatory and policy work, it is limited by my expertise as the author. Furthermore, Chapter 2, providing a background into dark pattern scholarship, is limited by focusing mostly on academic research in the context of HCI, neglecting equally important pushes from the hands of policy-makers and regulators that were not in the direct scope of this thesis. Also, additional surrounding work continues to foster ethical design alternatives by investigating technological problems that impair users' well-being, security, or privacy. For example, contributions to general design theory but also more specific research fields like usable security. Unfortunately, taking these works into account would have exceeded the scope of the chapter, whose aim is to give an introduction to the topic of dark patterns. As each publication contains a detailed discussion of its individual limitations, this thesis naturally adopts them as it constantly draws from their findings. To avoid unnecessary limitations, all study designs conducted across the publications included specific precautions. Yet, certain limitations remain stemming from the context in which studies took place, available resources, and the COVID-19 pandemic as a steady companion during most of these efforts.

6.4 Implications for Future Technologies

While this thesis makes a series of theoretical contributions to our understanding of design and its ethical considerations to mitigate dark patterns in different technological contexts, based on several quantitative and qualitative user-centred studies, it lacks foundational technological contributions. A reason for this is the relatively young research area surrounding dark patterns. Starting in 2010, the first years were spent exploring design phenomena that harmed users with the aim of understanding deployed strategies, mainly in the context of online GUIs. This research has since expanded to describing dark patterns in alternative technologies, including CUIs, with our ontology published in (P6) representing the currently most comprehensive corpus of dark patterns since the beginning of this body of research.

Now, with the theoretical groundwork laid, future dark pattern research is equipped to advance the field and build more fair and ethical technologies. To this end, my contributions, such as the CEC, the *Relationship Model of Cognitive Biases and Dark Patterns*, and finally the *Responsible Design Triangle* can help pave the way in this regard. Specific to SNSs, included work could spark the development of UIs that gives users autonomy over their data, allowing them to enhance their social connectedness and online well-being. Despite (and at the same time because of) my research, I believe that if done responsibly, SNSs have the potential to offer its users media to express and share ideas, engage in communities, and entertain meaningful relationships across the world.

I also found that CUIs are currently not meeting the high expectations of their users. However, as a technology that enables hands-off and verbal interactions, it has immense potential as an accessible and inclusive technology. If done in line with the theoretical contributions and ethical considerations shared in this thesis, future iterations could offer users context-aware and timely access to otherwise unavailable information and interactions.

But also outside this thesis' scope, its contents can have positive impacts on future technologies. As AI technologies become more powerful, just recently through the vast availability of generative AI and Large Language Models (LLMs), ethical and responsible considerations are crucial to ensure the empowerment of users through transparent and fair implementations, for example by informing them about the origins of content. Although these technologies have several useful implications for assistive interactions, instances of misuse for deception have already been shown (Zhou *et al.*, 2023; Hua *et al.*, 2024). These risks make responsible implementation and handling of future AI-based technologies a necessity to ensure people's safety and well-being. But we could also utilise AI to our advantage for combatting dark patterns, as we discuss in P6. The established ontology can be the foundation for building automated systems that can identify dark patterns in various digital interfaces. This would not only support future research but also aid regulators in their work to protect users.

6.5 Outlook

The background provided in Chapter 2 shows how far the dark pattern scholarship has come in just over a decade of research. Still, there is much to be done with respect to these related works and the publications in this thesis. The ontology included in Publication P6 consists of a dense

catalogue of various dark patterns sorted in a granulated structure. As such, it presents various opportunities for future research to build on it and extend it through novel considerations of underrepresented technologies. Moreover, it opens avenues for transdisciplinary work by offering a common baseline that different works can utilise. Specifically, in the crossroads between regulation and HCI, impactful work can actively ensure users' safety.

A large portion of this thesis inspects SNS, whose users require those additional protections because SNSs constantly confront them with engaging and governing design strategies. As these UI choices overshadow otherwise existing benefits, it would be interesting to see if the findings included in this thesis can aid the development of an alternative, ethics-driven SNSs. A careful design could empower people to maintain relationships across the globe and provide them with a shared platform to voice opinions or access news. Without being exposed to the ethical caveats of today's SNSs, platforms could foster people's well-being — a goal that could be expanded to all technologies (Calvo and Peters, 2014).

Some of the publications, specifically P4, P5, and P9, contain guidelines or suggestions to avoid dark patterns and other unethical designs. It will be interesting to see future work building on the CEC or the Relationship Model of Cognitive Biases and Dark Patterns to foster ethical design with an impact outside of research. Overall, the nine publications included in this thesis, alongside the plethora of research contributing to the dark pattern scholarship, have paved the way for exciting research. The theoretical contributions prepare future studies to develop holistic or specific frameworks for detecting dark patterns or scales that allow quantifiable measurements. They also prepare effective regulatory protection for people by codifying this knowledge into legal statutes.

Conclusively, the *Responsible Design Triangle* situates the individual responsibilities shared between design, user, and guideline angles. It describes each angle's influence on another to raise awareness about concerning design practices while offering considerations to mitigate deceptive design and dark patterns. Supported by nine publications that detail backgrounds for the respective angles, it makes its contribution to the discipline of HCI by reminding about the importance of user-centred design, elevating empowerment to enable users make informed and reflected choices.

CHAPTER 7

Conclusion

More than a decade ago, Brignull (2010) first set of twelve dark patterns marked the beginning of research studying unethical design practices across digital interfaces. The continuing body of work investigates unethical design resulting from a neglect of user needs and interests as well as profit maximising greed of service providers. Inspired by this work and driven to understand underlying mechanisms as the root cause of users' harm, this thesis focuses on responsible design. As a concept to mitigate dark patterns by elevating user empowerment, the thesis follows nine publications investigating dark patterns, particularly in Social Networking Services (SNSs) and Conversational User Interfaces (CUIs). Divided into three perspectives, this thesis considers design, user, and guideline angles to construct the *Responsible Design Triangle*. This framework situates the included publications alongside its angles while highlighting opportunities to focus on user-centred approaches that empower users of digital interfaces.

The thesis includes a comprehensive summary of contemporary dark pattern scholarship that is extended by findings in SNS and CUI technologies. Thus, the thesis builds up a thorough understanding of where dark patterns emerge and provides detailed information about their exploiting strategies. Following up on past work (Di Geronimo *et al.*, 2020; Bongard-Blanchy *et al.*, 2021), the studies included in this thesis describe the inability of users to recognise and avoid harm when facing dark patterns. To this end, the thesis discusses how design can be utilised to raise expectations in their users, which dark patterns exploit through misguidance or break to cause frustrations. In this vein, this thesis emphasises the importance of design that is truthful to its capabilities and, thus, allows realistic goals for users to achieve. Offering guidance for practitioners to consult, the thesis contains publications that propose additional frameworks and models that can be used to expose ethical implications of design. After highlighting the intertwined interactions between the design, user, and guideline angles of the *Responsible Design Triangle*, the thesis discusses the individual responsibilities respectively.

Nonetheless, this thesis makes a contribution to uncovering and understanding dark patterns and developing countermeasures to safeguard users. Although included guidelines remain propositions for ethical design, future work has yet to consult them during design stages to demonstrate their effectiveness. The variety of dark pattern types shows how widestretched mischievous strategies in online domains can be. Still, they all have one thing in common: They harm users. Regulators and legislation already have powerful tools to ensure the protection of end-users. However, not all regulations are equally effective. To support this, findings from HCI research on dark patterns can aid existing approaches to protect peoples' privacy on problematic designs. With a focus on Human-Computer Interaction (HCI), the thesis could pave the way for regulators and policy-makers through its insights leading to the *Responsible*

Design Triangle. Recent pushes within the European Union (EU)[1], for instance, foreshadow the positive impact this kind of work can have to protect users in online interfaces in the future.

Enhanced by growth hacking, particularly in contexts of surveillance capitalism, dark patterns have spread throughout domains like a pandemic of manipulative, coercive, and deceptive design strategies. As a countermeasure, responsible design must reflect on the implications design has in real-world scenarios. It should empower users to engage with technologies based on the truthful representation of functional capabilities and consequences, prioritising humans' well-being. If design, however, falls short of delivering systems that users can navigate without experiencing harm, additional guidelines, and ultimately regulations, need to ensure users' safeguarding instead. This thesis includes empirical observations into the root causes of dark patterns and offers guidelines based on user-centred principles to help avoid them in digital interfaces.

[1] For example European Comission (2022), European Parliament (2024), and European Commission (2024).

Literature References

[1] Agapie, E., Areán, P. A., Hsieh, G., and Munson, S. A., "A Longitudinal Goal Setting Model for Addressing Complex Personal Problems in Mental Health," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–28, 2022, ISSN: 2573-0142. DOI: 10.1145/3555160.

[2] Ahn, D. and Shin, D.-H., "Is the social use of media for seeking connectedness or for avoiding social isolation? mechanisms underlying media use and subjective well-being," *Computers in Human Behavior*, vol. 29, no. 6, pp. 2453–2462, 2013.

[3] Alexander, C., Ishikawa, S., and Silverstein, M., *A Pattern Language: Towns, Buildings, Construction*. OUP USA, 1977, ISBN: 978-0-19-501919-3.

[4] Alexandrovsky, D., Friehs, M. A., Grittner, J., Putze, S., Birk, M. V., Malaka, R., and Mandryk, R. L., "Serious Snacking: A Survival Analysis of how Snacking Mechanics Affect Attrition in a Mobile Serious Game," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 113, 2021, pp. 1–18, ISBN: 978-1-4503-8096-6.

[5] Allen, K. A., Ryan, T., Gray, D. L., McInerney, D. M., and Waters, L., "Social Media Use and Social Connectedness in Adolescents: The Positives and the Potential Pitfalls," *The Educational and Developmental Psychologist*, vol. 31, no. 1, pp. 18–31, 2014, ISSN: 0816-5122, 1839-2504. DOI: 10.1017/edp.2014.2.

[6] Alqahtani, F., Winn, A., and Orji, R., "Co-designing a mobile app to improve mental health and well-being: Focus group study," *JMIR Formative Research*, vol. 5, no. 2, e18172, 26, 2021. DOI: 10.2196/18172.

[8] Bardzell, J. and Bardzell, S., "What is "critical" about critical design?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 3297–3306, ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2466451.

[9] Barth, S. and Jong, M. D. T. de, "The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review," *Telematics and Informatics*, vol. 34, no. 7, pp. 1038–1058, 2017, ISSN: 0736-5853. DOI: 10.1016/j.tele.2017.04.013.

[10] Benjamin, R., "Race after technology," in *Social Theory Re-Wired*, 3rd ed., 2023, ISBN: 978-1-00-332060-9.

[11] Beyens, I., Pouwels, J. L., Driel, I. I. van, Keijsers, L., and Valkenburg, P. M., "The effect of social media on well-being differs from adolescent to adolescent," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[12] Bielova, N., Litvine, L., Nguyen, A., Chammat, M., Toubiana, V., and Hary, E., "The Effect of Design Patterns on (Present and Future) Cookie Consent Decisions," in *33rd USENIX Security Symposium*, vol. 33, 2024, pp. 1–18.

[13] Blandford, A., Furniss, D., and Makri, S., *Qualitative HCI Research: Going Behind the Scenes*. Morgan & Claypool, 2016, vol. 9, ISBN: 978-1-62705-759-2.

[15] Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., and Lenzini, G., "I am Definitely Manipulated, Even When I am Aware of it. It s Ridiculous! – Dark Patterns from the End-User Perspective," *Designing Interactive Systems Conference 2021*, pp. 763–776, 2021. DOI: 10.1145/3461778.3462086.

[16] Bösch, C., Erb, B., Kargl, F., Kopp, H., and Pfattheicher, S., "Tales from the dark side: Privacy dark strategies and privacy dark patterns.," *Proc. Priv. Enhancing Technol.*, vol. 2016, no. 4, pp. 237–254, 2016.

[17] Braun, V. and Clarke, V., "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. DOI: 10.1191/1478088706qp063oa.

[19] Brignull, H., *Deceptive Patterns: Exposing the Tricks Tech Companies Use to Control You*, 1st. Testimonium Ltd (30 July 2023), 2023, p. 272, ISBN: 978-1739454401.

[22] Bron, M., Redi, M., Lalmas, M., Silvestri, F., Evans, H., and Chute, M., "Friendly, appealing or both? characterising user experience in sponsored search landing pages," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 699–707, ISBN: 9781450349147. DOI: 10.1145/3041021.3054193.

[23] Brooke, J., "Sus: A retrospective," *J. Usability Studies*, vol. 8, no. 2, pp. 29–40, 2013.

[24] Bruckman, A., "Research Ethics and HCI," in *Ways of Knowing in HCI*, 2014, pp. 449–468, ISBN: 978-1-4939-0378-8. DOI: 10.1007/978-1-4939-0378-8_18.

[25] Brynjarsdottir, H., Håkansson, M., Pierce, J., Baumer, E., DiSalvo, C., and Sengers, P., "Sustainably unpersuaded: How persuasion narrows our vision of sustainability," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 947–956, ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2208539.

[28] Calo, R., "Digital Market Manipulation," *George Washington Law Review*, vol. 82, no. 4, pp. 995–1051, 2013.

[29] Calvo, R. A. and Peters, D., *Positive Computing: Technology for Wellbeing and Human Potential*. The MIT Press, 2014, ISBN: 978-0-262-32568-4. DOI: 10.7551/mitpress/9764.001.0001.

[30] Caraban, A., Karapanos, E., Gonçalves, D., and Campos, P., "23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15, ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300733.

[31] Chen, H., Chen, H., and Tian, X., "The dual-process model of product information and habit in influencing consumers' purchase intention: The role of live streaming features," *Electronic Commerce Research and Applications*, vol. 53, p. 101 150, 2022, ISSN: 1567-4223. DOI: 10.1016/j.elerap.2022.101150.

[32] Chen, Z., Piao, J., Lan, X., Cao, H., Gao, C., Lu, Z., and Li, Y., "Practitioners versus users: A value-sensitive evaluation of current industrial recommender system design," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, 2022. DOI: 10.1145/3555646.

[33] Chivukula, S. S., Hasib, A., Li, Z., Chen, J., and Gray, C. M., "Identity Claims that Underlie Ethical Awareness and Action," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13, ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445375.

[34] Chivukula, S. S., Li, Z., Pivonka, A. C., Chen, J., and Gray, C. M., "Surveying the Landscape of Ethics-Focused Design Methods," *arXiv:2102.08909 [cs]*, 17, 2021.

[35] Chivukula, S. S., Obi, I., Carlock, T. V., and Gray, C. M., "Wrangling ethical design complexity: Dilemmas, tensions, and situations," in *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 179–183, ISBN: 9781450398985. DOI: 10.1145/3563703.3596632.

[37] Conti, G. and Sobiesk, E., "Malicious interface design: Exploiting the user," in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, p. 271, ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772719.

[39] d'Aquin, M., Troullinou, P., O'Connor, N. E., Cullen, A., Faller, G., and Holden, L., "Towards an "ethics by design" methodology for ai research projects," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 54–59, ISBN: 9781450360128. DOI: 10.1145/3278721.3278765.

[40] Davis, F. D., "User acceptance of information technology: System characteristics, user perceptions and behavioral impacts," *International Journal of Man-Machine Studies*, vol. 38, no. 3, pp. 475–487, 1993.

[41] Desmet, P. and Hassenzahl, M., "Towards happiness: Possibility-driven design," in *Human-Computer Interaction: The Agency Perspective*, 2012, pp. 3–27, ISBN: 978-3-642-25691-2. DOI: 10.1007/978-3-642-25691-2_1.

[42] Devon, R. and Van de Poel, I., "Design ethics: The social ethics paradigm," *International Journal of Engineering Education*, vol. 20, no. 3, pp. 461–469, 2004.

[43] Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., and Bacchelli, A., "UI Dark Patterns and Where to Find Them," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14, ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376600.

[44] Diethei, D., Niess, J., Stellmacher, C., Stefanidi, E., and Schöning, J., "Sharing Heartbeats: Motivations of Citizen Scientists in Times of Crises," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15, ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445665.

[45] Dombrowski, L., Harmon, E., and Fox, S., "Social justice-oriented interaction design: Outlining key design strategies and commitments," in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 2016, pp. 656–671, ISBN: 9781450340311. DOI: 10.1145/2901790.2901861.

[46] Dunne, A. and Raby, F., *Speculative Everything: Design, Fiction, and Social Dreaming*. The MIT Press, 2013, ISBN: 0262019841.

[47] Enders, A., Hungenberg, H., Denker, H.-P., and Mauch, S., "The long tail of social networking.: Revenue models of social networking sites," *European Management Journal*, vol. 26, no. 3, pp. 199–211, 2008, ISSN: 0263-2373. DOI: 10.1016/j.emj.2008.02.002.

[56] Fausey, C., Long, B., Inamori, A., and Boroditsky, L., "Constructing agency: The role of language," *Frontiers in Psychology*, vol. 1, 2010, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2010.00162.

[58] Findeli, A., "Ethics, aesthetics, and design," *Design Issues*, vol. 10, no. 2, pp. 49–68, 1994, ISSN: 07479360, 15314790.

[59] Fleck, R. and Fitzpatrick, G., "Reflecting on reflection: Framing a design landscape," in *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, 2010, pp. 216–223, ISBN: 9781450305020. DOI: 10.1145/1952222.1952269.

[60] Fogg, B., "A behavior model for persuasive design," in *Proceedings of the 4th International Conference on Persuasive Technology*, 2009, ISBN: 9781605583761. DOI: 10.1145/1541948.1541999.

[61] Friedman, B. and Kahn, P. H., "Human values, ethics, and design," in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. 2002, pp. 1177–1201, ISBN: 0805838384.

[62] Friedman, B., Kahn, P. H., Borning, A., and Huldtgren, A., "Value Sensitive Design and Information Systems," in *Early engagement and new technologies: Opening up the laboratory*, vol. 16, 2013, pp. 55–95, ISBN: 978-94-007-7843-6 978-94-007-7844-3. DOI: 10.1007/978-94-007-7844-3_4.

[63] Friedman, B. and Nissenbaum, H., "Software agents and user autonomy," in *Proceedings of the first international conference on Autonomous agents - AGENTS '97*, 1997, pp. 466–469, ISBN: 978-0-89791-877-0. DOI: 10.1145/267658.267772.

[64] Frøkjær, E., Hertzum, M., and Hornbæk, K., "Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2000, pp. 345–352, ISBN: 1581132166. DOI: 10.1145/332040.332455.

[65] Given, L. M., Winkler, D. C., and Hopps-Wallis, K., "Social Media for Social Good: A Study of Experiences and Opportunities in Rural Australia," in *Proceedings of the 8th International Conference on Social Media & Society*, 2017, pp. 1–7, ISBN: 978-1-4503-4847-8. DOI: 10.1145/3097286.3097293.

[66] Gong, L. and Nass, C., "When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference," *Human communication research*, vol. 33, no. 2, pp. 163–193, 2007.

[67] Graßl, P., Schraffenberger, H., Zuiderveen Borgesius, F., and Buijzen, M., "Dark and Bright Patterns in Cookie Consent Requests," *Journal of Digital Social Research*, vol. 3, no. 1, pp. 1–38, 2021. DOI: 10.33621/jdsr.v3i1.54.

[68] Gray, C. M. and Boling, E., "Inscribing ethics and values in designs for learning: A problematic," *Educational Technology Research and Development*, vol. 64, no. 5, pp. 969–1001, 2016, ISSN: 1556-6501. DOI: 10.1007/s11423-016-9478-x.

[69] Gray, C. M., Chen, J., Chivukula, S. S., and Qu, L., "End User Accounts of Dark Patterns as Felt Manipulation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–25, 2021, ISSN: 2573-0142. DOI: 10.1145/3479516.

[70] Gray, C. M. and Chivukula, S. S., "Ethical Mediation in UX Practice," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11, ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300408.

[71] Gray, C. M., Chivukula, S. S., Carlock, T. V., Li, Z., and Duane, J.-N., "Scaffolding Ethics-Focused Methods for Practice Resonance," in *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 10, 2023, pp. 2375–2391, ISBN: 978-1-4503-9893-0. DOI: 10.1145/3563657.3596111.

[72] Gray, C. M., Chivukula, S. S., and Lee, A., "What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit," in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2020, pp. 61–73, ISBN: 978-1-4503-6974-9. DOI: 10.1145/3357236.3395486.

[73] Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A. L., "The dark (patterns) side of ux design," pp. 1–14, 2018. DOI: 10.1145/3173574.3174108.

[74] Gray, C. M., Mildner, T., and Bielova, N., "Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow"," arXiv:2309.09635, 2023.

[75] Gray, C. M., Obi, I., Chivukula, S. S., Li, Z., Carlock, T., Will, M., Pivonka, A. C., Johns, J., Rigsbee, B., Menon, A. R., and Bharadwaj, A., "Building an Ethics-Focused Action Plan: Roles, Process Moves, and Trajectories," in *roceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, 2024, pp. 1–18, ISBN: 979-8-4007-0330-0/24/05. DOI: 10.1145/3613904.3642302.

[76] Gray, C. M., Sanchez Chamorro, L., Obi, I., and Duane, J.-N., "Mapping the landscape of dark patterns scholarship: A systematic literature review," in *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 188–193, ISBN: 9781450398985. DOI: 10.1145/3563703.3596635.

[77] Gray, C. M., Santos, C., and Bielova, N., "Towards a Preliminary Ontology of Dark Patterns Knowledge," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, ISBN: 978-1-4503-9422-2. DOI: 10.1145/3544549.3585676.

[78] Gray, C. M., Santos, C., Bielova, N., Toth, M., and Clifford, D., "Dark patterns and the legal requirements of consent banners: An interaction criticism perspective," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–18, ISBN: 9781450380966. DOI: 10.1145/3411764.3445779.

[79] Gray, C. M., Santos, C., Tong, N., Mildner, T., Rossi, A., Gunawan, J. T., and Sinders, C., "Dark Patterns and the Emerging Threats of Deceptive Design Practices," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–4, ISBN: 978-1-4503-9422-2. DOI: 10.1145/3544549.3583173.

[80] Gray, C. M., Santos, C. T., Bielova, N., and Mildner, T., "An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642436.

[81] Gray, K. L., *Intersectional tech: Black users in digital gaming*. LSU Press, 2020, p. 201, ISBN: 978-0807174555.

[82] Greenberg, S., Boring, S., Vermeulen, J., and Dostal, J., "Dark patterns in proxemic interactions: A critical perspective," 2014, pp. 523–532, ISBN: 9781450329026. DOI: 10.1145/2598510.2598541.

[83] Gunawan, J., Pradeep, A., Choffnes, D., Hartzog, W., and Wilson, C., "A Comparative Study of Dark Patterns Across Web and Mobile Modalities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–29, 2021, ISSN: 2573-0142. DOI: 10.1145/3479521.

[84] Gunawan, J., Santos, C., and Kamara, I., "Redress for Dark Patterns Privacy Harms? A Case Study on Consent Interactions," in *Proceedings of the 2022 Symposium on Computer Science and Law*, 2022, pp. 181–194, ISBN: 978-1-4503-9234-1. DOI: 10.1145/3511265.3550448.

[85] Habib, H., Pearman, S., Young, E., Saxena, I., Zhang, R., and Cranor, L. F., "Identifying user needs for advertising controls on facebook," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 1–42, CSCW1 30, 2022, ISSN: 2573-0142. DOI: 10.1145/3512906.

[86] Hall, E. T., *The hidden dimension*. Doubleday, 1966, ISBN: 0-385-08476-5 978-0-385-08476-5.

[87] Hansen, P. G. and Jespersen, A. M., "Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy," *European Journal of Risk Regulation*, vol. 4, no. 1, pp. 3–28, 2013, ISSN: 1867-299X, 2190-8249. DOI: 10.1017/S1867299X00002762.

[88] Harrington, C. N., Garg, R., Woodward, A., and Williams, D., ""It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–15, ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3501995.

[89] Hassenzahl, M. and Tractinsky, N., "User experience - a research agenda," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, 2006. DOI: 10.1080/01449290500330331.

[90] Hausman, D. M. and Welch, B., "Debate: To Nudge or Not to Nudge*," *Journal of Political Philosophy*, vol. 18, no. 1, pp. 123–136, 2010, ISSN: 09638016, 14679760. DOI: 10.1111/j.1467-9760.2009.00351.x.

[91] Henrich, J., Heine, S. J., and Norenzayan, A., "The weirdest people in the world?" *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 61–83, 2010, ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X0999152X.

[92] Hidaka, S., Kobuki, S., Watanabe, M., and Seaborn, K., "Linguistic Dead-Ends and Alphabet Soup: Finding Dark Patterns in Japanese Apps," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–13, ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580942.

[93] Hilton, J. L. and Darley, J. M., "The effects of interaction goals on person perception," in vol. 24, 1991, pp. 235–267. DOI: https://doi.org/10.1016/S0065-2601(08)60331-7.

[94] Hua, Y., Niu, S., Cai, J., Chilton, L. B., Heuer, H., and Wohn, D. Y., "Generative AI in User-Generated Content," 2024. DOI: 10.1145/3613905.3636315.

[95] Hustinx, P., "Privacy by design: Delivering the promises," *Identity in the Information Society*, vol. 3, no. 2, pp. 253–255, 2010, ISSN: 1876-0678. DOI: 10.1007/s12394-010-0061-z.

[97] Jaspers, M. W., Steen, T., Bos, C. van den, and Geenen, M., "The think aloud method: A guide to user interface design," *International Journal of Medical Informatics*, vol. 73, no. 11, pp. 781–795, 2004, ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2004.08.003.

[98] Jia, H., Wu, M., Jung, E., Shapiro, A., and Sundar, S. S., "When the tissue box says "bless you": Using speech to build socially interactive objects," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 1635–1640, ISBN: 9781450319522. DOI: 10.1145/2468356.2468649.

[99] Kahneman, D., "A perspective on judgment and choice: Mapping bounded rationality.," *American Psychologist*, vol. 58, no. 9, pp. 697–720, 2003, ISSN: 1935-990X, 0003-066X. DOI: 10.1037/0003-066X.58.9.697.

[100] Kant, I., *Gesammelte Schriften*. Bd. 1-22 Preussische Akademie der Wissenschaften, Bd 23 von der Deutschen Akademie der Wissenschaften zu Berlin, ab Bd 24 von der Akademie der Wissenschaften zu Göttingen, no date, 1900ff.

[101] Koenig, A., "Patterns and antipatterns," in *The Patterns Handbooks: Techniques, Strategies, and Applications*. 1998, pp. 383–389, ISBN: 0521648181.

[103] Kornfield, R., Meyerhoff, J., Studd, H., Bhattacharjee, A., Williams, J. J., Reddy, M., and Mohr, D. C., "Meeting users where they are: User-centered design of an automated text messaging tool to support the mental health of young adults," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, ISBN: 9781450391573. DOI: 10.1145/3491102.3502046.

[104] Laugwitz, B., Held, T., and Schrepp, M., "Construction and Evaluation of a User Experience Questionnaire," in *HCI and Usability for Education and Work*, 2008, pp. 63–76, ISBN: 978-3-540-89350-9.

[105] Lee, S., Ditko, S., and Kirby, J., *Amazing Adult Fantasy*. Atlas Comics, 3, 1962, vol. 1, p. 36, ISBN: 9788891207388.

[106] Leimstädtner, D., Sörries, P., and Müller-Birn, C., "Investigating Responsible Nudge Design for Informed Decision-Making Enabling Transparent and Reflective Decision-Making," in *Mensch und Computer 2023*, 2023, pp. 220–236. DOI: 10.1145/3603555.3603567.

[107] Lin, K. Y. and Lu, H. P., "Why people use social networking sites: An empirical study integrating network externalities and motivation theory," *Computers in Human Behavior*, vol. 27, no. 3, pp. 1152–1161, 2011. DOI: 10.1016/j.chb.2010.12.009.

[108] Luger, E. and Sellen, A., ""Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5286–5297, ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858288.

[109] Luguri, J. and Strahilevitz, L. J., "Shining a Light on Dark Patterns," *Journal of Legal Analysis*, vol. 13, no. 1, pp. 43–109, 2021, ISSN: 2161-7201, 1946-5319. DOI: 10.1093/jla/laaa006.

[110] Lukoff, K., Liao, J. V., Choi, J., Fan, K., Munson, S. A., and Hiniker, A., "How the Design of YouTube Influences User Sense of Agency," in *CHI'21*, 2021, ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445467.

[111] Maier, M. and Harr, R., "Dark design patterns: An end-user perspective," *Human Technology*, vol. 16, no. 2, pp. 170–199, 2020.

[112] Majone, G., "The regulatory state and its legitimacy problems," *West European Politics*, vol. 22, no. 1, pp. 1–24, 1, 1999, ISSN: 0140-2382. DOI: 10.1080/01402389908425284.

[113] Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., and Narayanan, A., "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–32, 2019, ISSN: 2573-0142. DOI: 10.1145/3359183.

[114] Mathur, A., Mayer, J., and Kshirsagar, M., "What Makes a Dark Pattern ... Dark ? Design Attributes, Normative Considerations, and Measurement Methods," in *CHI'21*, 2021, p. 18, ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445610.

[115] Matte, C., Bielova, N., and Santos, C., "Do Cookie Banners Respect my Choice? : Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework," in *2020 IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 791–809, ISBN: 978-1-72813-497-0. DOI: 10.1109/SP40000.2020.00076.

[116] Mayring, P., "Qualitative Inhaltsanalyse," in *Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren*, 2020, pp. 495–511, ISBN: 978-3-658-26887-9. DOI: 10.1007/978-3-658-26887-9_52.

[117] McCrudden, C. and King, J., "The dark side of nudging: The ethics, political economy, and law of libertarian paternalism," *Choice Architecture in Democracies, Exploring the Legitimacy of Nudging (Oxford/Baden-Baden: Hart and Nomos, 2015), Forthcoming, U of Michigan Public Law Research Paper*, no. 485, 2015.

[118] Mejtoft, T., Hale, S., and Söderström, U., "Design Friction," in *Proceedings of the 31st European Conference on Cognitive Ergonomics*, 2019, pp. 41–44, ISBN: 978-1-4503-7166-7. DOI: 10.1145/3335082.3335106.

[119] Mejtoft, T., Parsjö, E., Norberg, O., and Söderström, U., "Design friction and digital nudging: Impact on the human decision-making process," in *Proceedings of the 2023 5th International Conference on Image, Video and Signal Processing*, 2023, pp. 183–190, ISBN: 9781450398381. DOI: 10.1145/3591156.3591183.

[120] Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., and Tuennerman, E., ""i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans," *Frontiers in Artificial Intelligence*, vol. 4, p. 169, 2021.

[121] Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R., and Niess, J., "Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642542.

[122] Mildner, T., Doyle, P., Savino, G.-L., and Malaka, R., "Rules Of Engagement: Levelling Up To Combat Unethical CUI Design," in *Proceedings of the 4th Conference on Conversational User Interfaces*, 2022, ISBN: 9781450397391. DOI: 10.1145/3543829.3544528.

[123] Mildner, T., Freye, M., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "Defending Against the Dark Arts: Recognising Dark Patterns in Social Media," in *Designing*

Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA, 2023. DOI: 1010.1145/3563657.3595964.

[124] Mildner, T., Inkoom, A., Malaka, R., and Niess, J., "Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Dark Patterns," arXiv:2405.07378 [cs], 2024.

[125] Mildner, T. and Savino, G.-L., "Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7, ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3451659.

[126] Mildner, T., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, ISBN: 9781450394215. DOI: 10.1145/3544548.3580695.

[127] Mildner, T., Savino, G.-L., Putze, S., and Malaka, R., "Finding a Way Through the Social Media Labyrinth: Guiding Design Through User Expectations," arXiv:2405.07305 [cs], 2024.

[128] Mildner, T., Savino, G.-L., Schöning, J., and Malaka, R., "Dark Patterns: Manipulative Designstrategien in digitalen Gesundheitsanwendungen," 2024. DOI: 10.1007/s00103-024-03840-6.

[130] Mittelstadt, B., "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0114-4.

[131] Molina-Pérez, A., Rodríguez-Arias, D., Delgado-Rodríguez, J., Morgan, M., Frunza, M., Randhawa, G., Wijdeven, J. R.-V. d., Schiks, E., Wöhlke, S., and Schicktanz, S., "Public knowledge and attitudes towards consent policies for organ donation in Europe. A systematic review," *Transplantation Reviews*, vol. 33, no. 1, pp. 1–8, 2019, ISSN: 0955-470X. DOI: https://doi.org/10.1016/j.trre.2018.09.001.

[132] Monge Roffarello, A., Lukoff, K., and De Russis, L., "Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–19, ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580729.

[133] Moser, C., Schoenebeck, S. Y., and Resnick, P., "Impulse Buying: Design Practices and Consumer Needs," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15, ISBN: 978-1-4503-5970-2.

[134] Munson, S. A., Krupka, E., Richardson, C., and Resnick, P., "Effects of public commitments and accountability in a technology-supported physical activity intervention," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1135–1144, ISBN: 9781450331456. DOI: 10.1145/2702123.2702524.

[135] Murad, C. and Munteanu, C., "Designing voice interfaces: Back to the (curriculum) basics," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12, ISBN: 9781450367080. DOI: 10.1145/3313831.3376522.

[136] Nabbosa, V. and Kaar, C., "Societal and ethical issues of digitalization," in *Proceedings of the 2020 International Conference on Big Data in Management*, 2021, pp. 118–124, ISBN: 9781450375061. DOI: 10.1145/3437075.3437093.

[137] Nass, C., Steuer, J., and Tauber, E. R., "Computers are social actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994, pp. 72–78, ISBN: 0897916506. DOI: 10.1145/191666.191703.

[139] Nielsen, J., "Usability inspection methods," in *Conference Companion on Human Factors in Computing Systems*, 1994, pp. 413–414, ISBN: 978-0-89791-651-6. DOI: 10.1145/259963.260531.

[140] Nielsen, J. and Molich, R., "Heuristic Evaluation Of Userinterfaces," *CHI'90*, vol. 01, pp. 249–256, 1990.

[141] Niess, J. and Woźniak, P. W., "Embracing Companion Technologies," in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 25, 2020, pp. 1–11, ISBN: 978-1-4503-7579-5. DOI: 10.1145/3419249.3420134.

[142] Norman, D. A., *Design for a Better World: Meaningful, Sustainable, Humanity Centered*. The MIT Press, 2023, ISBN: 9780262047951.

[143] Norman, D. A., *The design of everyday things*, Revised and expanded edition. Basic Books, 2013, ISBN: 978-0-465-05065-9.

[144] O'Brien, H. L. and Toms, E. G., "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008, ISSN: 1532-2882, 1532-2890. DOI: 10.1002/asi.20801.

[145] Obi, I., Gray, C. M., Chivukula, S. S., Duane, J.-N., Johns, J., Will, M., Li, Z., and Carlock, T., "Let's Talk About Socio-Technical Angst: Tracing the History and Evolution of Dark Patterns on Twitter from 2010-2021," 2022.

[147] Oliver, R. L., "A cognitive model of the antecedents and consequences of satisfaction decisions," *Journal of Marketing Research*, vol. 17, no. 4, pp. 460–469, 1980. DOI: 10.1177/002224378001700405.

[148] Oliver, R. L., "Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation," *Journal of Applied Psychology*, vol. 62, no. 4, pp. 480–486, 1977. DOI: doi.org/10.1037/0021-9010.62.4.480.

[149] Piaget, J., *The Child's Conception of the World*. Routledge, 1997, ISBN: 978-0-415-16887-8.

[150] Polson, P. G., Lewis, C., Rieman, J., and Wharton, C., "Cognitive walkthroughs: A method for theory-based evaluation of user interfaces," *International Journal of Man-Machine Studies*, vol. 36, no. 5, pp. 741–773, 1992, ISSN: 00207373. DOI: 10.1016/0020-7373(92)90039-N.

[151] Pradhan, A., Mehta, K., and Findlater, L., ""accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities," in *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 102–112, ISBN: 978-1-4503-5620-6/18/04. DOI: 10.1145/3173574.3174033.

[152] Quesenbery, W., "The five dimensions of usability," in *Content and complexity*, 1st ed., 2003, p. 22.

[153] Rizzo Mario J. Whitman, D. G., "Little brother is watching you: New paternalism on the slippery slopes themed issue: Perspectives on the new regulatory era," *Arizona Law Review*, vol. 51, p. 685, 2009.

[154] Sánchez Chamorro, L., Bongard-Blanchy, K., and Koenig, V., "Ethical tensions in UX design practice: Exploring the fine line between persuasion and manipulation in online interfaces," in *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 10, 2023, pp. 2408–2422, ISBN: 978-1-4503-9893-0. DOI: 10.1145/3563657.3596013.

[156] Shakya, H. B. and Christakis, N. A., "Association of Facebook Use With Compromised Well-Being: A Longitudinal Study," *American Journal of Epidemiology*, vol. 185, no. 3, pp. 203–211, 2017, ISSN: 0002-9262. DOI: 10.1093/aje/kww189.

[157] Shapiro, B. R., Meng, A., O'Donnell, C., Lou, C., Zhao, E., Dankwa, B., and Hostetler, A., "Re-Shape: A Method to Teach Data Ethics for Data Science Education," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13, ISBN: 978-1-4503-6708-0.

[158] Sharif, A. and Moorlock, G., "Influencing relatives to respect donor autonomy: Should we nudge families to consent to organ donation?" *Bioethics*, vol. 32, no. 3, pp. 155–163, 2018. DOI: https://doi.org/10.1111/bioe.12420.

[159] Shilton, K., "Values and Ethics in Human-Computer Interaction," *Foundations and Trends® in Human–Computer Interaction*, vol. 12, no. 2, pp. 107–171, 2018, ISSN: 1551-3955, 1551-3963. DOI: 10.1561/1100000073.

[160] Shilton, K., "Values Levers: Building Ethics into Design," *Science, Technology, & Human Values*, vol. 38, no. 3, pp. 374–397, 2013, ISSN: 0162-2439. DOI: 10.1177/0162243912436985.

[161] Sinclair, T. J. and Grieve, R., "Facebook as a source of social connectedness in older adults," *Computers in Human Behavior*, vol. 66, pp. 363–369, 2017.

[162] Sindermann, C., Löchner, N., Heinzelmann, R., Montag, C., and Scholz, R. W., "The Revenue Model of Mainstream Online Social Networks and Potential Alternatives: A Scenario-Based Evaluation by German Adolescents and Adults," *Technology in Society*, p. 102 569, 2024, ISSN: 0160-791X. DOI: 10.1016/j.techsoc.2024.102569.

[164] Souza-Neto, V., Marques, O., Mayer, V. F., and Lohmann, G., "Lowering the harm of tourist activities: A systematic literature review on nudges," *Journal of Sustainable Tourism*, vol. 31, no. 9, pp. 2173–2194, 2023. DOI: 10.1080/09669582.2022.2036170.

[165] Sunstein, C. R., "Nudges, Agency, and Abstraction: A Reply to Critics," *Review of Philosophy and Psychology*, vol. 6, no. 3, pp. 511–529, 2015, ISSN: 1878-5166. DOI: 10.1007/s13164-015-0266-z.

[166] Thaler, R. H., "Nudge, not sludge," *Science*, vol. 361, no. 6401, pp. 431–431, 2018. DOI: 10.1126/science.aau9241.

[167] Thaler, R. H. and Sunstein, C. R., *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, 2008, ISBN: 978-0-300-12223-7.

[168] Tversky, A. and Kahneman, D., "Judgment under uncertainty: Heuristics and biases," vol. 185, pp. 1124–1131, 1974. DOI: 10.1126/science.185.4157.1124.

[169] Twenge, J. M., Joiner, T. E., Rogers, M. L., and Martin, G. N., "Increases in depressive symptoms, suicide-related outcomes, and suicide rates among u.s. adolescents after 2010 and links to increased new media screen time," *Clinical Psychological Science*, vol. 6, no. 1, pp. 3–17, 2018. DOI: 10.1177/2167702617723376.

[170] Ulrich, K. T., "Design is everything?" *Journal of product innovation management*, vol. 28, no. 3, pp. 394–398, 2011.

[171] Utz, C., Degeling, M., Fahl, S., Schaub, F., and Holz, T., "(Un)informed Consent: Studying GDPR Consent Notices in the Field," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 973–990, ISBN: 978-1-4503-6747-9. DOI: 10.1145/3319535.3354212.

[172] Velthoven, M. H. v., Powell, J., and Powell, G., "Problematic smartphone use: Digital approaches to an emerging public health problem," *DIGITAL HEALTH*, vol. 4, p. 2 055 207 618 759 167, 2018. DOI: 10.1177/2055207618759167.

[173] Verbeek, P.-P., "Materializing Morality: Design Ethics and Technological Mediation," *Science, Technology, & Human Values*, vol. 31, no. 3, pp. 361–380, 2006, ISSN: 0162-2439. DOI: 10.1177/0162243905285847.

[174] Verbeek, P.-P., *What things do: philosophical reflections on technology, agency, and design*, 1. paperback print, trans. by R. P. Crease. Pennsylvania State Univ. Press, 2005, ISBN: 978-0-271-02539-1 978-0-271-02540-7.

[175] Wagener, N., Reicherts, L., Zargham, N., Bartłomiejczyk, N., Scott, A. E., Wang, K., Bentvelzen, M., Stefanidi, E., Mildner, T., Rogers, Y., and Niess, J., "SelVReflect: A guided VR experience fostering reflection on personal challenges," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 19, 2023, pp. 1–17, ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580763.

[176] Waldman, A. E., "Cognitive biases, dark patterns, and the 'privacy paradox'," *Current Opinion in Psychology*, vol. 31, pp. 105–109, 2020, ISSN: 2352-250X. DOI: 10.1016/j.copsyc.2019.08.025.

[177] Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F., "Privacy nudges for social media: An exploratory facebook study," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 763–770, ISBN: 9781450320382. DOI: 10.1145/2487788.2488038.

[178] Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F., ""i regretted the minute i pressed share": A qualitative study of regrets on facebook," in *Proceedings of the Seventh Symposium on Usable Privacy and Security*, 2011, ISBN: 9781450309110. DOI: 10.1145/2078827.2078841.

[179] Watson, D. and Clark, L. A., "Development and validation of brief measures of positive and negative affect : The PANAS scales," *Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1988.

[180] Waytz, A., Cacioppo, J., and Epley, N., "Who sees human?: The stability and importance of individual differences in anthropomorphism," *Perspectives on Psychological Science*, vol. 5, no. 3, pp. 219–232, 2010. DOI: 10.1177/1745691610369336.

[181] Zagal, J. P., Björk, S., and Lewis, C., "Dark patterns in the design of games," in *Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013)*, 2013, pp. 39–46, ISBN: 978-0-9913982-0-1.

[182] Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and De Choudhury, M., "Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–20, ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581318.

[183] Zuboff, S., "The age of surveillance capitalism," in *Social Theory Re-Wired*, 2023, pp. 203–213.

Miscellaneous and Online References

[7] Association for Computing Machinery. “Words matter: Alternatives for charged terminology in the computing profession.” en. (2023), [Online]. Available: https://www.acm.org/diversity-inclusion/words-matter (visited on 07/31/2023).

[14] Board, E. D. P., *Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them | European Data Protection Board*, March, 2022.

[18] Brignull, H. “Deceptive design - types of deceptive design.” en. (2010), [Online]. Available: http://web.archive.org/web/20220525230009/https://www.deceptive.design/types (visited on 07/27/2023).

[20] Bringull, H., Leiser, M., Santos, C., and Doshi, K. “Deceptive Patterns - Types of Deceptive Pattern.” en. (2023), [Online]. Available: https://www.deceptive.design/types (visited on 07/27/2023).

[21] Britannica, The Editors of Encyclopaedia. “Iliad.” T. E. o. E. Britannica, Ed. (2023), [Online]. Available: https://www.britannica.com/topic/Iliad-epic-poem-by-Homer (visited on 01/26/2024).

[26] Buolamwini, J., Chock, S. C., Petty, T., Lopez, M. A., Lizárraga, J. R., Taye, B., and Benhamin, R. “Mission, Team and Story - The Algorithmic Justice League.” (2016), [Online]. Available: https://www.ajl.org (visited on 08/01/2023).

[27] California State Legislature, *CCPA-18 2018. California Consumer Privacy Act of 2018 [1798.100 - 1798.199] (CCPA)*. California State Legislature, 28, 2018.

[36] Commission, E., *Gdpr-16 2016. regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance)*, European Commission, 4, 2016.

[38] California Privacy Protection Agency, *California privacy rights act*, 2022.

[48] European Comission, *Consumer protection*, Text, 2022.

[49] European Commission. “Commission opens formal proceedings against x under the DSA,” European Commission - European Commission. (), [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_23_6709 (visited on 02/26/2024).

[50] European Commission, *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*, 2021.

[51] European Commission, *Proposal for a regulation of the european parliament and of the council on a single market for digital services (digital services act) and amending directive 2000/31/ec*, 2022.

[52] European Commission, *Proposal for a regulation of the European Parliament and of the Council on harmonized rules on fair access to and use of data (Data Act)*, 2022.

[53] European Commission, *Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives)*, 2022.

[54] European Parliament, *Digital servicess act***i. european parliament [a9-0356/2021]*, 2022.

[55] European Parliament. “TA-9-2023-0459_en.pdf,” Addictive design of online services and consumer protection in the EU single market. (), [Online]. Available: https://

www.europarl.europa.eu/doceo/document/TA-9-2023-0459_EN.pdf (visited on 02/26/2024).

[57] Federal Trade Commission, *Complaint for Permanent Injunction, Civil Penalties, Monetary Relief, and Other Equitable Relief*, 2023.

[96] International Committee of Medial Journal Editors. "ICMJE | recommendations | defining the role of authors and contributors." (2023), [Online]. Available: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html (visited on 09/22/2023).

[102] Komninos, A. "An introduction to usability," The Interaction Design Foundation. (Jul. 22, 2020), [Online]. Available: https://www.interaction-design.org/literature/article/an-introduction-to-usability (visited on 02/25/2024).

[129] Ministry of Consumer Affairs, Food & Public Distribution, *Central consumer protection authority issues 'guidelines for prevention and regulation of dark patterns, 2023' for prevention and regulation of dark patterns listing 13 specified dark patterns*, 2023.

[138] Nielsen, J. "Usability 101: Introduction to usability," Nielsen Norman Group. (2012), [Online]. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (visited on 02/25/2024).

[146] OECD, *Dark commercial patterns*, 2022.

[163] Sinders, C. "What's In a Name?" en. (Jun. 2022), [Online]. Available: https://medium.com/@carolinesinders/whats-in-a-name-unpacking-dark-patterns-versus-deceptive-design-e96068627ec4 (visited on 09/20/2023).

PART II

PUBLICATIONS

PUBLICATION P1

Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook

Authors:
Thomas Mildner & Gian-Luca Savino

The publication contributes to the following angles:
DESIGN USER

This publication analysises the changes in Facebook's User Interface (UI) during the years from 2004-2020 and notices obfuscations where later UI iterations limit the discoverability of specific features and privacy settings. Based on a second user survey with 116 Facebook users, we studied users' perception and usage behaviour and learned about incongruencies regarding their feelings of being in control, their usage of the Social Networking Service (SNS), and misalignment with actual behaviour. Interestingly, our participants reported satisfaction with Facebook despite these incongruencies.

Its contribution to the thesis is twofold. By following Facebook's historic UI changes, it identifies roots for design strategies that can become hosts to dark patterns in the design angle. Further studying users' perspectives surfaces the complexity with which users engage with SNS and demonstrate opportunities for dark patterns to exploit their behaviours relevant to the user angle.

My contribution to this paper was the study design, data collection, and analysis for both studies. I interpreted the results and wrote the manuscript, which I revised and submitted for the final publication.

The contents of this chapter originally appeared in: Mildner, T. and Savino, G.-L., "Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7, ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3451659

Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook

Thomas Mildner
University of Bremen
Germany

Gian-Luca Savino
University of Bremen
Germany

ABSTRACT

Many researchers have been concerned with whether social media has a negative impact on the well-being of their audience. With the popularity of social networking sites (SNS) steadily increasing, psychological and social sciences have shown great interest in their effects and consequences on humans. In this work, we investigate Facebook using the tools of HCI to find connections between interface features and the concerns raised by these domains. Using an empirical design analysis, we identify interface interferences impacting users' online privacy. Through a subsequent survey (n = 116), we find usage behaviour changes due to increased privacy concerns and report individual cases of addiction and mental health issues. These observations are the results of a rapidly changing SNS creating a gap of understanding between users' interactions with the platform and future consequences. We explore how HCI can help close this gap and work towards more ethical user interfaces in the future.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; HCI theory, concepts and models; *Empirical studies in interaction design*; Interaction design theory, concepts and paradigms; • **Security and privacy** → *Usability in security and privacy*.

KEYWORDS

SNS, social media, Facebook, interface design, dark patterns, well-being, ethical interfaces

ACM Reference Format:
Thomas Mildner and Gian-Luca Savino. 2021. Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451659

1 INTRODUCTION & MOTIVATION

During the rapid rise of social networking sites (SNS) and the number of people using them in a little more than a decade, an increase of mental health issues among their audiences has been recorded and published [23, 27, 29, 33], while further studies looked at their impacts on social capital and connectedness [1, 2, 25]. Although participants of longitude studies self-reported a negative impact on physical and mental health when using Facebook [30], they still kept using the application regularly. A possible explanation can be derived from existing research on internet addiction [5, 19, 26]. Christakis et al. has further linked an experienced decrease of future well-being in Facebook users to the disease in earlier works. [7, 8]. Other researchers, however, did not find a reason to believe that SNS cause their audience to experience higher risks of mental health problems, such as depression [17]. Moreover, significant findings show that audiences of SNS may perceive higher levels of social connectedness [1, 15, 31] and improved overall well-being with regards to reduced stress and social support [21].

As people perceive and interact with SNS on the level of user interfaces, we believe that the perspective of the HCI community can add important insights to describe the relationship between a person and an interface with a focus on well-being. Recent research in HCI has already shown some interest in social media and their features. The 'liking' behaviour of people on SNS, for example, has been investigated, noticing specific elements that influence the count of likes received [16]. With a focus on the social network Facebook, Wang et al. found that people often feel regret about certain content that they have shared which ultimately caused them disadvantages in social and professional areas [35]. In a follow-up study, Wang et al. implemented additional interface features that gave people the chance to review and correct content while displaying all recipients before publishing [34]. They found that most participants approved their features and thus demonstrated possibilities in which interface design can respond to people's concerns. A recent study from Andalibi et al. describes anxiety among interviewees concerning emotion recognition technologies that is expected to become a big part in SNS perhaps giving new possibilities to show fitting content [3].

As a result of the fast changes and growth in technology, including SNS, society and research have fallen behind to provide ethical guidelines creating a 'cultural lag' [14, 22]. This lag is especially problematic when people are being misguided into doing something that they either did not expect or even tricked into actions with harmful results. With regards to Brignull's et al. work defining Dark Patterns [6], Grey et al. have showcased such occurrences in various digital interfaces. Often, Dark Patterns utilise knowledge about human psychology in combination with usable design to create deceiving design practices which do not have the user's interests in mind [14]. Because SNS want their audience to be recurring and promoting their service, it is unlikely that they want them to be associated with any form of harm. Thus, traditional Dark Patterns as defined by Brignull et al. [6] mostly seem unfit for SNS.

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3451659

Prominent SNS, such as Facebook, Twitter, and Instagram, offer their services for free and generate their revenue through advertisement [11, 18]. The more time people spend on them, the more income can thus be generated. This motivates our research to take a look at design and interface strategies of SNS that may be linked to the decreased well-being through Dark Patterns or features similar to them. We, therefore, investigate Facebook, the currently largest social networking site, using common HCI methods like interface and user behaviour analysis. In this work, we present two studies: (1) In an empirical design analysis we identify and describe Facebook's key interface features and their changes in the past 16 years. (2) Through a survey with 116 participants, we analyse people's usage behaviour of Facebook with a particular focus on motivation, satisfaction, and privacy concerns. Based on these studies, we highlight scopes in which Facebook implements Dark Patterns, especially interface interferences. We further noticed a change in people's motivation to use Facebook due to privacy concerns and that those increase with age. Thus, especially younger generations might be exposed to harm through loose privacy settings and harmful features like 'endless scrolling'. Eventually, we noticed a few general behavioural patterns but found worrying behaviour especially in individual cases, which we urge to further investigate in future research. With this paper we hope to spark the interest of the scientific community and contribute (1) an analysis of Dark Patterns in Facebook's desktop interface and (2) a discussion on how HCI can help to identify potential impacts of user interfaces to peoples' well-being based on the responses of 116 survey participants.

2 METHODOLOGY

Since their launch date in 2004, Facebook's interface received multiple updates. Meanwhile, additional features brought new options for people to engage with others and the application itself. In our research, we take a close look at Facebook's interface, its changes, and their effects on Facebook's audience while considering previous research in HCI, psychology, and social sciences. The presented study is split into two parts: (1) An empirical design analysis of Facebook's user interface and (2) a survey asking about people's Facebook usage.

2.1 Empirical Design Study

Two HCI researchers analysed Facebook's desktop user interface on an empirical level. To collect the necessary images of past UI iterations we used manual web crawling and collected screenshots of Facebook's desktop interface from a wide range of years and were crosschecking screenshots from different sources to ensure their accuracy. Further analysis focused mainly on Dark Patterns and privacy concerns. It included tracking certain UI-elements like the logout button and its positioning in the navigation bar as well as several iterations of the 'account' menu. The understandings gained through this empirical approach led to the design of the Facebook user survey.

2.2 Facebook Usage Survey

To understand people's motivation, satisfaction, and privacy concerns when using Facebook, we created a questionnaire comprising of 30 questions including attention checks. The questionnaire consisted of 'yes/no/maybe', 5-point Likert scale, and open-ended questions. We structured it in four sections: (1) demographics, (2) current/past motivation to use Facebook and most used features, (3) satisfaction of the time spent on Facebook and specific features, and (4) the participants' attitudes towards privacy concerns. We recruited participants through *SurveySwap* and *Reddit*, and hosted the surveys on Qualtrics [24]. We received 126 responses from which we excluded 10 due to failed attention checks or double entries. In the following results section, we analyse the data of these 116 survey respondents.

3 RESULTS

3.1 Empirical Design Analysis

Developed as a student network at Harvard University in 2004, Facebook received their most recent visual update in 2019. In the following paragraphs, we recorded notable changes to Facebook's desktop interface made within these years. For readability, the most prominent findings are visualised in Figure 1. The following paragraphs present an overview of relevant changes to the logout button, the privacy settings, and the Privacy Checkup.

3.1.1 Interface Interference. Although being moved around the applications' interface, most of Facebook's features remained to be constant elements of the application. The logout button, for example, was part of the top-navigation until 2010. With an update, the button was moved into the 'Account' menu, limiting discoverability. The privacy settings have experienced similar changes. Shortly after Facebook was opened to a wider audience, the privacy settings were moved from the navigation bar into the site's settings drop-down menu in 2008. In 2012, Facebook offered users an alternative on the first-level view, the Privacy Shortcuts. This feature allowed them to quickly access selected privacy settings and tools and was removed from the first-level interface with Facebook's latest redesign in 2019. Even though these changes could be natural consequences of responsive interface design, it is important to note that Facebook does benefit from users not logging out or sharing more information due to lighter privacy settings. This way Facebook is able to track their users across the web and use the information to create content and targeted advertisement. Intentionally moving these buttons to limit their discoverability and prevent certain user actions (i.e. logouts) would be considered interface interference in the Dark Pattern terminology [14].

3.1.2 Novel Dark Patterns. During their latest redesign, Facebook introduced a new feature called Privacy Checkup. It allows users to edit privacy settings for pre-selected categories during a guided step-by-step process. Changes made in this feature directly affect the privacy settings, albeit with incomplete coverage. In addition to the privacy settings, users are further able to control ad-related settings. These also do not include the full spectrum of settings that are available in the general settings. Especially ad-related settings can interfere with Facebook's advertisement strategies. By offering a guided settings feature Facebook is able to curate which settings users will manage. If this is used intentionally to keep users from certain settings, it could be seen as a novel way of interface interference. While placing all privacy settings behind several interface

layers, Facebook actively offers a well designed but incomplete alternative to handle them.

3.1.3 Summary. The goal of the empirical study was to find indications for Dark Patterns within Facebook's interface. Moving the logout button and privacy settings into drop-down menus can be classified as interface interference [14]. Given that Facebook directly benefits from users not logging out (by being able to track them across the web as long as they are logged in), this can be a conscious choice to limit discoverability and thus prevent certain user actions (i.e. logouts). Traditionally, Dark Patterns cause some degree of harm to their audience. As SNS aim to keep theirs entertained and satisfied, this creates a contradiction to what is typically described by Dark Patterns. We did, however, find a novel way in which Facebook guides their users' actions in their interest. Past research has shown that managing privacy settings on Facebook is not an easy task for their users and most often does not result in the level of security that they expect [20]. Facebook now offers a feature which makes this process easier, but potentially still leaves the user with wrong expectations. This design choice leaves the impression of traditional Dark Patterns but is not yet covered by their terminology.

3.2 Facebook Survey

The following results were extracted from the survey and present people's usage of Facebook, as well as our three main focal points on their motivation, satisfaction, and privacy concerns when using Facebook. Since our survey is exploratory, we primarily used descriptive statistics.

3.2.1 Demographics. Of the 116 respondents 46 (40%) identify as male, 67 (58%) as female, one identified as other, and two did not prefer to say. Their age ranged from 16 to 52 with a mean age of 26 ($SD = 7.4$). Respondents were situated in 29 different countries with roughly half coming from the UK ($n = 19$), the US ($n = 23$) and the Netherlands ($n = 27$). Of all participants, 81% ($n = 94$) actually use Facebook currently and 19% ($n = 22$) do not. Of those who use Facebook, 74% joined the network between 2005 and 2012. The majority (65%) of respondents use Facebook on mobile more than on desktop, for 17% it is the other way round. 18% use it the same amount on both.

3.2.2 Motivation. To get an understanding of people's general motivation to use Facebook, we asked participants about their current and original motivation (why they initially signed up) as well as their most-used and most-regrettable features. The 94 active Facebook users were allowed to give multiple replies and named a total of 133 current motivations. We identified six main categories: (1) 'Keeping in Touch with Family and Friends' (31%); (2) 'Groups of Interest' (14%); (3) 'News/Information' (13%); (4) 'Messenger/Chatting Feature' (12%); (5) 'Events' (8%); and (6) 'Entertainment' (7%). 21 replies (15%) did not fit those categories and comprise of individual motivations. Regarding the original motivation (when first joining Facebook) respondents named a total of 88 which were sorted in the same six main categories: (1) 'Keeping in Touch with Family and Friends' (41%); (2) 'Groups of Interest' (3%); (3) 'News/Information' (0%); (4) 'Messenger/Chatting Feature' (6%); (5) 'Events' (2%); and (6) 'Entertainment' (3%). Within the remaining 39 replies we identified two other large categories: (1) 'Like/Share' (20%); (2) 'Peer Pressure' (11%). 'Keeping in Touch with Family and Friends' remains an important feature of Facebook from back when people original started using it to today. While Facebook's 'Groups of Interest' feature is mentioned more often as a current motivation compared to the original one, 'Like/Share' seems to have declined being mentioned only once as a current motivation. The participants were further asked to state their most used and most regrettable Facebook features. The two most used features of a total of 181 features mentioned, were (1) 'Messenger' (29%) and (2) 'Groups of Interest' (21%). Other notable features comprised of: (3) 'Like/Share' (8%); (4) 'Newsfeed' (8%); and (5) 'Entertainment' (7%). Additionally, participants mentioned 29 features they regret spending too much time on. People mostly regret spending time on 'Entertainment' (34%) features on Facebook, followed by the 'News Feed' feature (24%). Together 'Entertainment' and 'News Feed' makeup 15% of people's most used features. Thus, they are among the lesser-used features compared to others. Still, some people use them while, at the same time, regret spending too much time on them.

3.2.3 Satisfaction. We further asked participants about their satisfaction with the time they spent on Facebook and whether they would like to change something about it. Most users were extremely (29%) and somewhat (39%) satisfied with the time they spent on Facebook. 20% answered that they were neither satisfied nor dissatisfied and the remaining 12% were somewhat dissatisfied. No participant was extremely dissatisfied which was mostly due to them wanting to spend less time on Facebook rather than more time. 49% of respondents reported to sometimes find themselves longer on Facebook than they had planned to and 13% about half the time or more often. Only 37% never find themselves using it longer than they planned to. In contrast, 80% of respondents never want to spend more time on Facebook than they already have. When asked directly, 14% think they spend too much time on Facebook, 18% think they maybe do, and 68% do not think they spend too much time on Facebook.

The 32% ($n = 30$) who think that they do or maybe do spend too much time on Facebook listed the following reasons for why they think this is the case: 7 participants answered that they use Facebook to procrastinate and distract themselves from tasks they should be doing; 4 answered that it is addictive or has addictive features and that they would like to stop using it but still use it every day; 4 mentioned that they often mindlessly scroll and that the infinite scrolling behaviour of the timeline supports this habit. Other responses included that it is just entertaining, boring, or a waste of time. Finally, we asked all 94 participants if they would like to actually spend less time on Facebook. To this, 50% answered that they do not want to spend less time, 21% replied with maybe, and 29% do would like to spend less time.

3.2.4 Privacy Concerns. In order to learn more about people's privacy concerns, our questionnaire comprised four directed question items. We distinguished between social interaction and business related incentives from Facebook by asking targeted questions in a 5-point Likert-Scheme. A majority of the participants generally want to be in control over the information that other people can see about them. For 50% of them, this is an extremely important desire while 1% do not find this control important at all. When

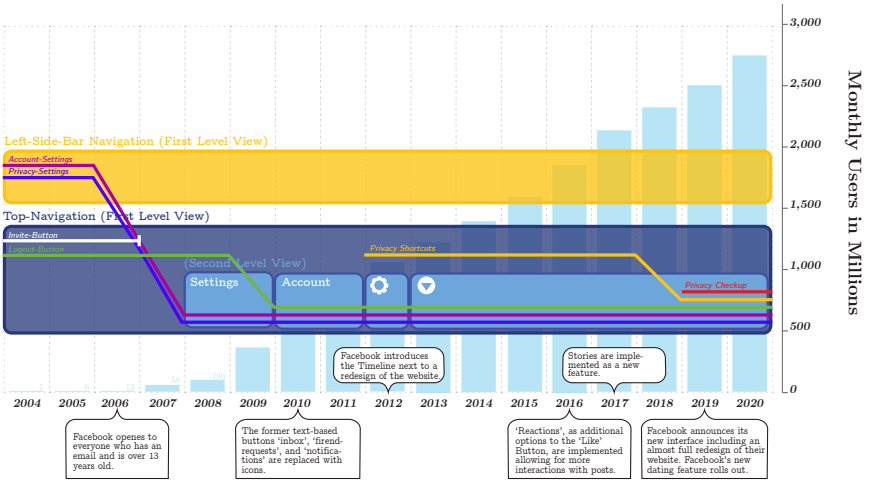


Figure 1: This diagram visualises Facebook's interface changes between 2004-2020. With increasing popularity the interface was extended in multiple iterations. Because no official and complete data-set was present at the time of writing this work, the data of monthly Facebook users was collected from two sources. Firstly, the data for the years between 2004-2008 were retrieved from Facebook Newsroom [13, 28]. As Facebook Newsroom no longer features their original data, an article from The Guardian that relies on the same source is cross-referenced for support. However, comparisons need to be tentative and taken with caution. Secondly, the data representing the years from 2008-2020 was gathered from Statista.com [9].

asked about whether they actually feel in control over such information, 63% agree while 23% rather do not. When instead asking about the control over the information that Facebook uses for advertisement, 29% find it extremely important whereas 6% find it not important at all. We again asked about whether they feel in control over Facebook's use of their data for advertisement. While 13% actually feel in control over the information that Facebook uses for advertisement, (definitely yes - 4%), 70% of our participants do not feel in control (definitely no - 34%). This is further visualised in Figure 2. Eventually, we asked participants to state their logout behaviour as well as their general concern about internet security. Most of the participants (79%) never log out of Facebook, 7% log out regularly, and 5% do so always. When it comes to internet security, 73% are generally concerned whereas 11% are not.

4 DISCUSSION

Research is divided on what effects social media and SNS have on people's mental health. While some results show small positive effects when spending time on social media, others yield small negative effects, or did not identify any effect at all. On one hand, various studies have found connections between increased depressive symptoms, that in occasions tragically resulted in suicides, and increased screen-time, new media, and SNS [32]. On the other hand, researchers did not identify any such links [10] or instead observed overall positive effects on people's well-being [1, 15, 31].

As so often in technology, SNS change quickly and thus require research to regularly re-evaluate their impact. HCI has the utilities to continuously analyse SNS and their effects on users with a particular focus on human-centred design. It can describe the relationship a person has with an interface and actively propose design guidelines to prevent possibly harmful design choices such as Dark Patterns.

In this work, we focused on this relationship between people and interfaces. We highlighted Facebook's continuous interface changes and noticed that design choices have led to potential Dark Patterns. In the presented empirical design study, we have identified two interface interferences. As can be seen in Figure 1, Facebook has changed the location of the logout button as well as the link to the privacy settings multiple times while offering alternative but incomplete features. Although the cause for these changes may be reasoned through aesthetics, they result in unnecessary obstacles that users have to overcome. In the following discussion, we aim to create possible links between the presented studies. Our interface analysis only focuses on Facebook's desktop application while most participants of our survey preferred the mobile application. Thus, the conclusions drawn from these comparison should not be generalised as they reflect only on a smaller portion of our sample.

Even though they might not be intentionally developed, Dark Patterns can be used to evaluate the ethical principles of interfaces by assessing the difficulty of, for example, basic user actions. In the case of Facebook's logout button, it could explain why a majority of the participants never log out of Facebook, while also showing concern about internet security. Since Facebook uses cookies to track personal data within and outside their website [12], our findings demonstrate an incongruity between people's privacy concerns and their actual activities when using SNS. As can be seen in Figure 2, our results show that participants want control over the data that they share with others. This preference also exists for the data which Facebook uses for advertisement, although to a lesser extent. During their updates, Facebook chose to implement their Privacy Shortcuts and Privacy Checkup as alternatives to regular settings.

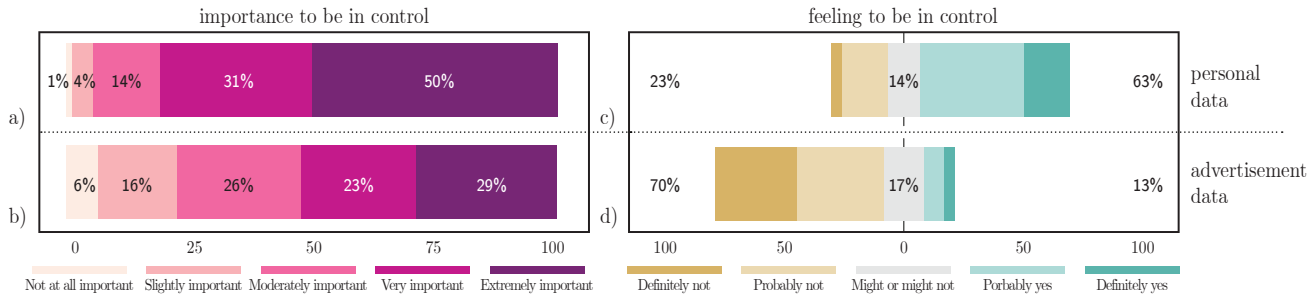


Figure 2: This graphic shows the results of the questions: a) 'How important is it for you to be in control over the information other people can see about you on Facebook?'; b) 'How important is it for you to be in control over the information about you which Facebook uses for targeted advertising?'; c) 'Do you feel in control over the information other people can see about you on Facebook?'; d) 'Do you feel in control over the information about you which Facebook uses for targeted advertising?' of the survey.

With both, they provide their audience with controls over the general privacy settings. However, by limiting the number of options Facebook governs over the settings that their audience will adjust. Although such alternative interfaces are not covered by traditional Dark Patterns, they create interferences that allow Facebook to navigate their audience's decision-making without causing immediate harm.

Governing, but never harming, users makes it possible to keep their general satisfaction with Facebook high. Even though the majority of participants are generally satisfied with their time spent on Facebook, most of them spend more time than they actually plan to. Those who actually think they do spend too much time mention reasons like procrastination, distraction, mindless scrolling and even addiction. This shows a worrying development in which individual participants even realise their bad habits: "No time at all would be the ideal time, but the fact that I still open the app once a day shows a bad habit." P(102). Facebook seems to keep general satisfaction up, but it is the individual person who develops problematic usage behaviours through features like endless scrolling (*"infinite scroll makes me stay longer"* (P113)). Investigation of such features offer insights into how user interfaces may affect people's well-being.

An important part of this process is to understand people's incentives and motivations when using a service. By understanding what features people use or do not use or how their usage behaviour has changed over time, we are able to identify user groups and can connect them to behaviours and habits. Individual comments of the participants describe a perceptional change in their motivation to use Facebook. Sharing content, for example, made up 20% of people's original motivation to create a Facebook account: *"To [...] post about my social life"* (P44). Interestingly, it was only mentioned by one participant as their current motivation. A potential reason for this change could be the increased privacy concern as people grew up: *"[I] used to be more active to share my life with others, but I don't anymore due to the privacy concern"* (P52). Our results show that with age participants' privacy concern also increased ($r = 0.23$, p-value = 0.029). Unfortunately, we were not able to show statistically significant relationships between features or motivations and users' satisfaction. Still, this analysis helped us to understand what people do on Facebook and why they might quit the service or never even use it in the first place. We asked the 22 participants (19% of our participants), who do not use Facebook actively, why they either stopped using the SNS or never even registered. Although their replies are highly individual, they foreshadow a link to the concerns of psychological and social science scholars, namely to well-being. P11 stated, that *"it was bad for [their] mental health"* while P13 felt *"judged by likes and comments"*. Although still a Facebook user, P24 wrote: *"I felt bad about myself and my life"*, as a reason for why their motivation has changed.

In this work, we explored how standard HCI methods like interface and user behaviour analysis can be used to identify potentially harmful aspects of Facebook's user interface and discuss how those could relate to effects on people's well-being. We do not find large numbers indicating general problems with SNS, but individual users show worrying usage behaviour and even specifically state mental health issues. As Beyens et al. [4] show, there can be vast individual differences, especially between adolescent users, in the way SNS affect their well-being. We, therefore, argue that researching the individual is just as important as the large scale studies which are already available. Through our approaches, we propose a first guideline for ethical interface design, namely considering users' motivation and expectations while testing user interfaces for Dark Patterns. Practitioners should further understand that high user satisfaction and usability is not necessary a result of ethical user interfaces.

5 LIMITATIONS & FUTURE WORK

The empirical design analysis focuses only on Facebook's desktop interface, as it proved to be difficult to find enough authentic screenshots for all mobile versions due to various screen layouts based on different device sizes and operating systems (iOS and Android). An in-depth analysis of the mobile interface was thus out of scope for our work. Yet, the results of the survey show that most participants prefer Facebook's mobile application. While Facebook's latest interface update makes comparison between both interfaces easier, since both layouts show little to no differences, caution should be excersised when drawing conclusions. Future research should distinguish between all interface variations.

As with most studies utilising qualitative surveys, participants assessed their experience with Facebook. Comparing users' current usage behaviour with Facebook to their original motivation is necessary to understand how design changes may alter people's handling of a system. Nevertheless, the presented study asked participants to recall their experiences making these replies susceptible to recall bias. Moreover, chosen terminology, especially regarding one's well-being, are subjective. This subjective assessment allowed us to conduct in-depth analyses. We do, however, acknowledge that our sample was too small to find significant results in other parts of the evaluation. Especially for the results on participants' motivation larger sample sizes are necessary to fully understand their relevance.

The presented research only focuses on Facebook since it is still the most widely used SNS. While more thorough research should continue to study Facebook, similar studies should investigate other SNS as the range of available SNS has grown in recent years. The research on the protection of human's well-being on SNS is an interdisciplinary effort. HCI adds a range of utilities that help to understand the interaction between people and SNS on the level of interface and technology. This community can thus support psychological and social sciences by actively impacting future design development and providing ethical guidelines while advancing its interdisciplinary efforts.

6 CONCLUSION

This paper presents two studies investigating the social medium Facebook. The first describes an empirical study analysing Facebook's continuous interface changes throughout recent years. We find cases of interface interference that Facebook has implemented with respect to the logout button and the privacy settings. In contrast to traditional Dark Patterns, these do not cause direct consequences but indirectly govern how they interact with the interface. Through a survey with 116 participants, the second study finds that users' motivation changed over time, making them share less personal data than they used to, due to increased privacy concerns. While being concerned about their private and ad-related data, users only partly feel in control about the amount they share. This feeling could be a result of the interface interference through which Facebook actively limits users' choices and guides their behaviour. This negatively impacts users (e.g. sharing information by not logging out), demonstrating an example for Dark Patterns on SNS. With this contribution, we explore different methods to show how HCI can contribute to investigating the well-being of social media users and hope to spark the interest of the scientific community to engage in this research.

REFERENCES

[1] Dohyun Ahn and Dong-Hee Shin. 2013. Is the social use of media for seeking connectedness or for avoiding social isolation? Mechanisms underlying media use and subjective well-being. *Computers in Human Behavior* 29, 6 (2013), 2453–2462.

[2] Kelly A. Allen, Tracii Ryan, DeLeon L. Gray, Dennis M. McInerney, and Lea Waters. 2014. Social Media Use and Social Connectedness in Adolescents: The Positives and the Potential Pitfalls. *The Educational and Developmental Psychologist* 31, 1 (2014), 18–31. https://doi.org/10.1017/edp.2014.2 arXiv:https://www.tandfonline.com/doi/pdf/10.1017/edp.2014.2

[3] Nazanin Andalibi and Justin Buss. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376680

[4] Ine Beyens, J Loes Pouwels, Irene I van Driel, Loes Keijsers, and Patti M Valkenburg. 2020. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports* 10, 1 (2020), 1–11.

[5] Roberta Biolcati, Giacomo Mancini, Virginia Pupi, and Valeria Mugheddu. 2018. Facebook Addiction: Onset Predictors. *Journal of Clinical Medicine* 7, 6 (2018), 1–12. https://doi.org/10.3390/jcm7060118

[6] Harry Brignull, Marc Miquel, Jeremy Rosenberg, and James Offer. 2015. Dark Patterns-User Interfaces Designed to Trick People. http://darkpatterns.org/ (visited on 01/05/2021).

[7] Dimitri A Christakis. 2010. Internet addiction: a 21 st century epidemic? *BMC medicine* 8, 1 (2010), 1–3.

[8] Dimitri A Christakis and Megan A Moreno. 2009. Trapped in the net: will internet addiction become a 21st-century epidemic? *Archives of pediatrics & adolescent medicine* 163, 10 (2009), 959–960.

[9] J. Clement. 2020. Facebook MAU worldwide 2020. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (visited on 01/05/2021).

[10] Sarah M. Coyne, Adam A. Rogers, Jessica D. Zurcher, Laura Stockdale, and McCall Booth. 2020. Does time spent using social media impact mental health?: An eight year longitudinal study. *Computers in Human Behavior* 104 (2020), 106160. https://doi.org/10.1016/j.chb.2019.106160

[11] Lisette de Vries, Sonja Gensler, and Peter S.H. Leeflang. 2012. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *Journal of Interactive Marketing* 26, 2 (2012), 83 – 91. https://doi.org/10.1016/j.intmar.2012.01.003

[12] Facebook, Inc. 2020. Facebook Cookies. https://www.facebook.com/policies/cookies/ (visited on 01/07/2021).

[13] Facebook, Inc. 2021. Facebook Newsroom. https://about.fb.com/news/ (visited on 01/05/2021).

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

[15] Rachel Grieve, Michaelle Indian, Kate Witteveen, G Anne Tolan, and Jessica Marrington. 2013. Face-to-face or Facebook: Can social connectedness be derived online? *Computers in human behavior* 29, 3 (2013), 604–609.

[16] Jin Yea Jang, Kyungsik Han, and Dongwon Lee. 2015. No Reciprocity in "Liking" Photos: Analyzing Like Activities in Instagram. In *Proceedings of the 26th ACM Conference on Hypertext; Social Media* (Guzelyurt, Northern Cyprus) *(HT '15)*. Association for Computing Machinery, New York, NY, USA, 273–282. https://doi.org/10.1145/2700171.2791043

[17] Lauren Jelenchick, J. Eickhoff, and M. Moreno. 2013. "Facebook depression?" social networking site use and depression in older adolescents. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 52 1 (2013), 128–30.

[18] Dong Hoo Kim, Yoon Hi Sung, So Young Lee, Dongwon Choi, and Yongjun Sung. 2016. Are you on Timeline or News Feed? The roles of Facebook pages and construal level in increasing ad effectiveness. *Computers in Human Behavior* 57 (2016), 312 – 320. https://doi.org/10.1016/j.chb.2015.12.031

[19] Daria J. Kuss and Mark D. Griffiths. 2011. Online Social Networking and Addiction—A Review of the Psychological Literature. *International Journal of Environmental Research and Public Health* 8, 9 (2011), 3528–3552. https://doi.org/10.3390/ijerph8093528

[20] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing Facebook Privacy Settings: User Expectations vs. Reality. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (Berlin, Germany) *(IMC '11)*. Association for Computing Machinery, New York, NY, USA, 61–70. https://doi.org/10.1145/2068816.2068823

[21] Robin L Nabi, Abby Prestin, and Jiyeon So. 2013. Facebook friends with (health) benefits? Exploring social network site use and perceptions of social support, stress, and well-being. *Cyberpsychology, Behavior, and Social Networking* 16, 10 (2013), 721–727.

[22] William F Ogburn. 1957. Cultural lag as theory. *Sociology & Social Research* 41 (1957), 167–174.

[23] Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, et al. 2011. The impact of social media on children, adolescents, and families. *Pediatrics* 127, 4 (2011), 800–804.

[24] Qualtrics. 2021. Qualtrics. https://www.qualtrics.com (visited on 01/07/2021).

[25] Tracii Ryan, Kelly A. Allen, DeLeon L. Gray, and Dennis M. McInerney. 2017. How Social Are Social Media? A Review of Online Social Behaviour and Connectedness. *Journal of Relationships Research* 8 (2017), e8. https://doi.org/10.1017/jrr.2017.13

[26] Tracii Ryan, Andrea Chester, John Reece, and Sophia Xenos. 01 Sep. 2014. The uses and abuses of Facebook: A review of Facebook addiction. *Journal of Behavioral Addictions JBA* 3, 3 (01 Sep. 2014), 133 – 148. https://doi.org/10.1556/jba.3.2014.016

[27] Christina Sagioglou and Tobias Greitemeyer. 2014. Facebook's emotional consequences: Why Facebook causes a decrease in mood and why people still use it. *Computers in Human Behavior* 35 (2014), 359–363.

[28] Ami Sedghi. 2014. Facebook: 10 years of social networking, in numbers. https://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics (visited on 01/05/2021).

[29] Maarten HW Selfhout, Susan JT Branje, M Delsing, Tom FM ter Bogt, and Wim HJ Meeus. 2009. Different types of Internet use, depression, and social anxiety: The role of perceived friendship quality. *Journal of adolescence* 32, 4 (2009), 819–833.

[30] Holly B. Shakya and Nicholas A. Christakis. 2017. Association of Facebook Use With Compromised Well-Being: A Longitudinal Study. *American Journal of Epidemiology* 185, 3 (02 2017), 203–211. https://doi.org/10.1093/aje/kww189 arXiv:https://academic.oup.com/aje/article-pdf/185/3/203/24329441/kww189.pdf

[31] Tara J Sinclair and Rachel Grieve. 2017. Facebook as a source of social connectedness in older adults. *Computers in Human Behavior* 66 (2017), 363–369.

[32] Jean M. Twenge, Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. 2018. Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science* 6, 1 (2018), 3–17. https://doi.org/10.1177/2167702617723376

[33] Regina JJM Van den Eijnden, Gert-Jan Meerkerk, Ad A Vermulst, Renske Spijkerman, and Rutger CME Engels. 2008. Online communication, compulsive Internet use, and psychosocial well-being among adolescents: A longitudinal study. *Developmental psychology* 44, 3 (2008), 655.

[34] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. Association for Computing Machinery, New York, NY, USA, 763–770. https://doi.org/10.1145/2487788.2488038

[35] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania) *(SOUPS '11)*. Association for Computing Machinery, New York, NY, USA, Article 10, 16 pages. https://doi.org/10.1145/2078827.2078841

PUBLICATION P2

Rules Of Engagement: Levelling Up To Combat Unethical CUI Design

Authors:
Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, & Rainer Malaka

The publication contributes to the following angle:
GUIDELINE

This publication considers ethical implications of the design of Conversational User Interface (CUI) technologies. It thereby translates Graphical User Interface (GUI)-related dark pattern scholarship for CUI-related contexts and provokes the need for further attention. By following related work in CUI and broader Human-Computer Interaction (HCI) disciplines, this publication illustrates potential roots for dark patterns in speech and voice design, specifically with regard to emotional dimensions that can be exploited to manipulate users. In order to counter deceptive design and dark patterns, the publication demonstrates a process to assess dark patterns in novel interfaces, based on dark pattern characteristics (Mathur *et al.*, 2019).

Its contribution to the thesis is to the guideline angle by provoking the presence of dark patterns in CUIs and a process allowing simple assessments of dark pattern across previously identified dimensions.

My contribution to this paper was the analysis of related work, discussing their implications and the development of a process to assess unethical design in CUIs. I drafted the manuscript and revised it before the final publication.

The contents of this chapter originally appeared in: Mildner, T., Doyle, P., Savino, G.-L., and Malaka, R., "Rules Of Engagement: Levelling Up To Combat Unethical CUI Design," in *Proceedings of the 4th Conference on Conversational User Interfaces*, 2022, ISBN: 9781450397391. DOI: 10.1145/3543829.3544528

Rules Of Engagement: Levelling Up To Combat Unethical CUI Design

Thomas Mildner
University of Bremen
Bremen, Germany
mildner@uni-bremen.de

Gian-Luca Savino
University of St.Gallen
St.Gallen, Switzerland
gian-luca.savino@unisg.ch

Philip Doyle
University College Dublin
Dublin, Ireland
philip.doyle1@ucdconnect.ie

Rainer Malaka
University of Bremen
Bremen, Gremany
malaka@tzi.de

ABSTRACT

While a central goal of HCI has always been to create and develop interfaces that are easy to use, a deeper focus has been set more recently on designing interfaces more ethically. However, the exact meaning and measurement of ethical design has yet to be established both within the CUI community and among HCI researchers more broadly. In this provocation paper we propose a simplified methodology to assess interfaces based on five dimensions taken from prior research on so-called dark patterns. As a result, our approach offers a numeric score to its users representing the manipulative nature of evaluated interfaces. It is hoped that the approach - which draws a distinction between persuasion and manipulative design, and focuses on how the latter functions rather than how it manifests - will provide a viable way for quantifying instances of unethical interface design that will prove useful to researchers, regulators and potentially even users.

CCS CONCEPTS

• **Human-centered computing** → **User interface design**; **User centered design**; *HCI theory, concepts and models.*

KEYWORDS

ethical design, dark patterns, conversational user interfaces, evaluation methods and techniques

ACM Reference Format:
Thomas Mildner, Gian-Luca Savino, Philip Doyle, and Rainer Malaka. 2022. Rules Of Engagement: Levelling Up To Combat Unethical CUI Design. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543829.3544528

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

ACM ISBN 978-1-4503-9739-1/22/07…$15.00
https://doi.org/10.1145/3543829.3544528

1 INTRODUCTION

Although ethical interface design has been a concern in HCI for some time, researchers are only beginning to formulate ways to understand, define and measure occurrences of unethical interface design. In terms of attempts to build knowledge around this issue, most progress can be found in literature on dark patterns, which began just over a decade ago. In 2010, Brignull introduced the term of dark patterns as *"tricks used in websites and apps that make you do things that you didn't mean to"* [3]. Throughout the past decade, many examples of dark patterns have been identified, generating a rich taxonomy of specific dark pattern types and dark pattern strategies [2, 3, 7, 10, 11, 13, 19, 31]. While this is a positive move in terms of better understanding unethical interface design practises, the body of work suffers somewhat from a lack of consensus, continuity, and clear definition about what exactly is being examined across various different studies. This creates a significant problem when trying to devise ways to protect users against their potentially harmful effects, including efforts to help users identify unethical design practises more easily, and attempts to regulate them.

This lack of consensus and continuity in how concepts are defined also echoes a general problem in human-machine dialogue (HMD) research, which has been highlighted in a number of recent reviews of research in the field [4, 5, 28, 29]. The aim of this provocation paper is to draw attention to the potential for learning from the problems within these bodies of work and to further encourage efforts to bring clarity to how we define and measure examples of unethical design in CUIs. As a research community, we might then be able to establish more effective ways of identifying and measuring unethical design practises in the context of CUI development before they become as common and problematic as they already in graphical interface interaction.

2 FUNDAMENTALS: PERSUASION & MANIPULATION

To a certain degree persuasion is a fundamental feature of design. Ideally, the aim is to design objects that signify their potential uses and constraints so people can interact with them as intuitively as possible. That is, we look to create objects that gently nudge people toward using them in a certain way. This is a common general understanding of design in HCI research, and echoes widely known basic principles for ensuring alignment between a user's mental

model and a system forwarded by Norman in '*The Design of Everyday Things*' [22]. Yet, in recent years the notion of persuasive design has begun to acquire negative connotations, largely due to a rise in design techniques aimed at governing and exploiting peoples' decision-making. Thaler and Sunstein provide an example of this type of design technique when introducing the term of *Nudges* [17] to describe interventions that alternate peoples' decision making process in a predictable way, allowing design to be used to navigate a users' focus into a predefined direction or goal. However, approaches of this nature, that take advantage of our cognitive biases, can be used equally efficiently for benevolent or malevolent ends; as is evidenced by the growth in research on dark patternsacross numerous domains.

This raises questions about how we define persuasive design, and indeed, how we draw distinctions between designs regarded as persuasive versus manipulative. Nudges have certainly been used in beneficial ways, including improving peoples' eating [27], healthcare [32], fitness [21], sleep [15] and relaxation practices [30]. It would also be inaccurate to assume people are naive to how these systems work, with the implicit nature of persuasive cognitive approaches being both part of the appeal and part of the reason they are effective [18]. However, they also share fundamental similarities with designs used to encourage users to make potentially harmful decisions; so-called dark patterns. They disarm a person of their autonomy, albeit temporarily, by encouraging them to make choices they might not have made otherwise. Therein lies our ethical obligation and challenges as HCI researchers: to ensure we combat unethical acts of manipulative interface design, whilst ensuring we continue to deliver persuasive interface designs that improve the lives of people who use them. We argue that establishing clear conceptual definitions is the first step toward consistent identification and measurement of unethical interface design in CUIs.

Based on established psychological definitions, we suggest drawing a distinction between persuasive design and manipulative design. Here, persuasion is defined as, *"an active attempt by one person to change another person's attitudes, beliefs, or emotions associated with some issue, person, concept, or object"* [23]. To bring this into context with HCI research, we suggest a slight adaptation, defining persuasive design as, *'an active attempt to influence a person's behaviours, attitudes, beliefs, or emotions associated through interface design'*. On the other hand, manipulation is described in psychology as, *'behaviour designed to exploit, control or otherwise influence others to one's own advantage'* [24]; or in the context of HCI, *'designs aimed at exploiting, controlling or otherwise influencing users to one's own advantage'*. Our challenge in bringing clarity to this distinction lies in developing ways to identify and measure occurrences of unethical manipulative design, which will aid us in defending the benefits of using psychological knowledge to improve people's lives through technology also.

3 UNDERSTANDING AND MEASURING UNETHICAL DESIGN: LESSONS FROM DARK PATTERNS

Research into unethical interface design practises began with examinations of e-commerce websites by Brignull [3]. The work identified twelve specific examples of unethical design aimed at inhibiting people's ability to make informed choices. These include, *Sneak into Basket*, *Hidden Costs*, and *Price Comparison Prevention*, which all operate on the premise of obscuring information and potentially misleading users into buying unwanted or unnecessarily expensive products. Other dark patterns defined by the author include, *Forced Continuity*, *Privacy Zuckering*, and *Roach Motel*, which are all tactics aimed at forcing people to sign-up, or stay signed up for accounts and services they might not require anymore. These examples are also used by service providers to gain access to private data without fully informing users why it is being collected, who has access to it, or even what it might be used for. Since then, researchers have described and defined a plethora of other examples, with a recent review of the literature from Mathur et al. [20] identifying 62 specific types of dark patterns. For readability, Table 1 offers a complete overview. Understanding how these existing dark patterns were established and how they are used to facilitate unethical design in graphical user interfaces (GUIs) could allow us to prevent similar developments in CUIs.

While it may seem like unethical GUI design has been a perennial problem for HCI researchers, literature on dark patterns shows that these tactics and the form they take develop and change over time through conscious efforts made by interface design practitioners. Grey et al. [9, 11] identified multiple dark patterns as well as design constraints of practitioners which lead to their creation. In a first study, the authors analysed an image-based corpus to define five types of dark patterns that practitioners engage in when developing manipulative designs [11]. Most of Gray at al.'s dark patterns include descriptions from prior works. For example, the *Obstruction* dark pattern, used to make processes unnecessarily difficult, incorporates Brignull's *Roach Motel*, *Price Comparison Prevention*, and *Intermediate Currency* [3]. In a follow up to this work, Gray et al. also analysed 4775 user-generated posts of the Reddit sub-forum *r/assholedesign* [9]. Analysis resulted in identifying six properties of "asshole designers". The work is particularly useful for understanding the origins of dark patterns and how they emerge from constraints under which practitioners work.

While many of the previously described dark patterns are applicable to different domains, research has also found domain specific examples. For example, Zagal et al. identified seven dark patterns that related to video game mechanics [31]. While certain patterns exploit a game's ecosystem of connected users, such as *Social Pyramid Schemes* and *Impersonation*, others impact game-play experience like *Grinding* and *Playing by Appointment*. Elsewhere, Greenberg et al. [13] consider dark patterns in conjunction with proxemics theory [14]. Identifying nine types of dark pattern in total, the authors discuss interactions with manipulative design in spatial environments. The *Attention Grabber* and *Disguised Data Collection* dark patterns, for instance, could be used in the design of digital billboards that exploit people's proximity and personal data to deliver personalised advertising.

Understanding the creation of dark patterns and analysing their occurrences helps to close the "cultural lag" [25] where the creation of ethical guidelines inevitably lags behind the release of novel dark patterns and even novel technologies. However, recent attempts by the authors to apply the aforementioned corpus of dark patterns in the domain of social media highlight a central problem in this body of work: it is hallmarked by a high degree of overlap between

Brignull 2010 [3]	**Conti & Sobiesk** 2010 [7]	**Zagal et al.** 2013 [31]	**Greenberg et al.** 2014 [13]	**Bösch et al.** 2016 [2]	**Gray et al.** 2018 [11]	**Gray et al.** 2020 [10]	**Mathur et al.** 2019 [19]
· *Trick Questions*	· *Coercion*	· *Grinding*	· *Attention Grabber*	· *Privacy Zuckering*	· *Nagging*	· *Automating the User*	· *Countdown Timers*
· *Sneak Into Basket*	· *Distraction*	· *Impersonation*	· *Bait and Switch*	· *Hidden Legalese Stipulations*	· *Obstruction*	· *Two-Faced*	· *Limited-time Messages*
· *Roach Motel*	· *Forced Work*	· *Monetized Rivalries*	· *The Social Network Of Proxemic Contracts Or Unintended Relationships*	· *Shadow User Profiles*	· *Sneaking*	· *Controlling*	· *High-demand Messages*
· *Privacy Zuckering*	· *Manipulating Navigation*	· *Pay to Skip*	· *Captive Audience*	· *Bad Defaults*	· *Interface Interference*	· *Entrapping*	· *Activity Notifications*
· *Confirmshaming*	· *Restricting Functionality*	· *Playing by Appointment*	· *We Never Forget*	· *Immortal Accounts*	· *Forced Action*	· *Nickling-And-Diming*	· *Confirmshaming*
· *Disguised Ads*	· *Trick*	· *Pre-Delivered Content*	· *Disguised Data Collection*	· *Information Milking*		· *Misrepresenting*	· *Testimonials of Uncertain Origins*
· *Price Comparison Prevention*	· *Confusion*	· *Social Pyramid Schemes*	· *Making Personal Information Public*	· *Forced Registration*			· *Hard to Cancel*
· *Misdirection*	· *Exploiting Errors*		· *The Milk Factor*	· *Address Book Leeching*			· *Visual Interference*
· *Hidden Costs*	· *Interruption*						· *Low-stock Messages*
· *Bait and Switch*	· *Obfuscation*						· *Hidden Subscriptions*
· *Forced Continuity*	· *Shock*						· *Pressured Selling*
· *Friend Spam*							· *Forced Enrollment*

Table 1: This table lists 62 types of dark pattern described in prior research.

definitions, inconsistent terminology, and descriptions that operate on different levels often without explicit acknowledgement. That is, the difference between specific dark pattern designs and broader dark pattern strategies is not always recognised. Attempts to apply taxonomies of dark patterns also shows that while some established dark patterns types are applicable to other domains, it is also likely that novel CUI specific dark patterns types and strategies will need to be described and defined. Further, while this is particularly problematic from a research perspective, we see an even more urgent need to begin this work so we might also better protect users.

4 WAYS TO COMBAT UNETHICAL INTERFACE DESIGN

With regards to technology, particularly online interfaces, studies have shown that the burden to counteract dark patterns often falls on users [1, 8]. Although regulations, such as the GDPR [6] or the CCPA [16], aim to protect users in online environments, the previously mentioned cultural lag [25] means these efforts are struggling to counter all problematic designs described under the umbrella of manipulative design. Indeed, one could argue they have led to the creation of new dark patterns. The design of cookie consent banners that favour 'accept all' options over offering users greater control, shows how design can be used to easily negate efforts to combat dark patterns, and how current regulatory efforts fail to address the fundamental nature of manipulative design. This has lead to designs exploiting cognitive biases not covered by the GDPR, such as anchoring effects, to steer users into giving consent that they might not have given were they provided with a neutral choice [12, 19].

In January 2022, the European Union proposed a new article 13a, as part of the Digital Service Act [26], to address previous concerns. Offering a generalised definition of problematic design that is closely worded to Mathur et al's. [20] definition of dark patterns, article 13a contains an extendable list of specific interface designs to be regulated. However, we argue that this approach may lead to a similar problem seen in attempts to apply dark pattern taxonomies across different interfaces and domains. That is, the taxonomies quickly become very large, difficult to maintain, and not always appropriate across interfaces types and use cases. Further, creating an ever growing list of examples may also deepen this problem over-complicating crucial efforts to combat unethical design.

Mathur et al. 2019 [19]

Dark Pattern Characteristics

Characteristic	Question
Asymmetric	Does the user interface design impose unequal weights or burdens on the available choices presented to the user in the interface?
Covert	Is the effect of the user interface design choice hidden from the user?
Deceptive	Does the user interface design induce false beliefs either through affirmative misstatements, misleading statements, or omissions?
Hides Information	Does the user interface obscure or delay the presentation of necessary information to the user?
Restrictive	Does the user interface restrict the set of choices available to users?

Table 2: This table lists the introductory questions Mathur et al. (2019) [19] gave for each dark pattern characteristic.

5 LEVELLING UP TO UNETHICAL INTERFACE DESIGN

By categorising specific dark pattern types into five broad characteristics, Mathur et al. [19] offer an alternative and potential useful avenue for combating manipulative designs. This higher level categorisation is based on the cognitive biases specific designs are developed to exploit. By relying on more fundamental concepts, which are much less interchangeable than specific interface layouts, the approach offers a way to understand manipulative design that focuses on their impact on users, in a broadly applicable fashion, whilst remaining agnostic to the interface or domain in question. By

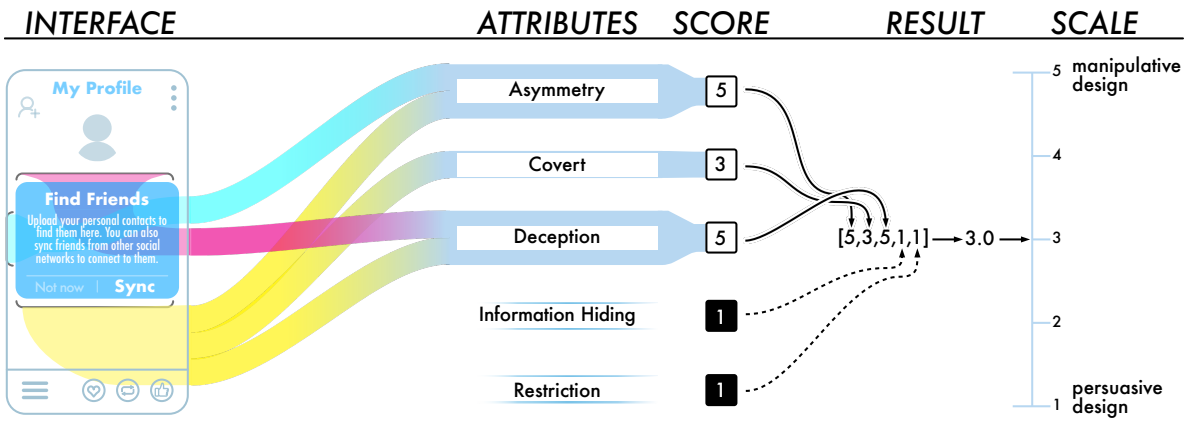


Figure 1: This figure demonstrates the derivation of a five-dimensional dark pattern vector, which can further be reduced into a single digit value, inspired by Mathur et al.'s five dark pattern characteristics and attributes [19]. Although the example is based on a screenshot of a mobile interface, it could be easily adapted for any conversational user interface.

being based on cognitive biases, instead of mere dark pattern definitions that stem from GUI domains, this model allows enhanced evaluation outside its original scope, such as CUIs.

In an attempt to advance their model and differentiate between manipulative and persuasive designs (i.e, between dark patterns and bright patterns), we developed a technique that allows us to evaluate individual examples based on characteristics that stem from the cognitive biases they target. We therefore consider each of the five characteristic by asking a specific question to assess impact on each of these dimensions, as seen in Table 2. By assigning each a value from 1 (not at all) to 5 (extremely), we are able to evaluate persuasive designs across the five dimensions, indicating the degree to which they might be regarded as persuasive or manipulative. The benefit of this approach is that we gain an overall score that can be used to determine the degree to which a specific design is either persuasive or manipulative, whilst identifying which dimensions of persuasion a manipulative design targets. Depending on the context and situation in which a design is evaluated, the score determining the degree of persuasive versus manipulative design can be adapted by alternating the threshold set to identify what is acceptable design depending on the score. Figure 1 visualises the steps of this process. By providing this clearly defined and measurable conceptualisation of persuasive and manipulative design, this model yields a certain duality. On one hand, we might help regulators combat unethical interface design practices in a consistent and broadly applicable fashion, whilst protecting the benefits of persuasion. On the other, this approach could allow practitioners to evaluate their own designs through user studies.

6 CURRENT CAVEATS TO KEEP IN MIND

This provocation paper addresses unethical design in CUIs but discusses the topic with the means of manipulation and exploitation of cognitive biases. We understand a distinction between ethical design and measurement of unethical practices. Yet, we argue that by learning which unethical practices are at play, we are able to compare and understand practitioners' strategies better while promoting more conscious handling of design techniques that, in the wrong context, may exploit cognitive biases harming the user. The currently available amount of tools to assess the good or bad in design is limited while a growing demand to evaluate interfaces ethically can be seen across disciplines. By utilising knowledge about cognitive biases, and exploitation thereof, we are able to understand malicious interface strategies better and can further classify them to share new knowledge between research communities. As a basis for this research, cognitive biases describe basic behaviour traits, certain strategies and heuristics under which decisions are formed, shared among all humans. Building on this existing knowledge, we aim to establish a robust measurement that also allows for comparison of interfaces.

We acknowledge the early stage of this endeavour and are aware of current limitations. Arguably, a numeric and finite score as a determiner for how ethical an interface is may be appealing to different cohorts, whether in the context of regulation or user interface design. As all ordinal scales, however, the proposed approach can only represent a limited abstraction of manipulative dimensions in an interface. Moreover, the questionnaire has not yet been verified and thus it is uncertain how effective it will prove to be in action. In this early stage, we rely on the five characteristics proposed by Mathur et al. [19]. By rooting their characteristics in cognitive bias research, they gain the advantage of being similarly effective across domains. Still, it is questionable whether these exact five dimensions are as effective in the context of CUIs when compared to their origin in GUIs. This could easily be addressed in future research by investigating the differences and unanimity between cognitive biases exploited throughout different kinds of interface modalities. Further studies could then look at the variety of described cognitive biases to identify exploitation across domains to offer alternative sets of questions and target each modality precisely. For example, where colourisation of certain buttons may promote some choices over others in GUI contexts, in voice based CUIs colour of voice, emotional speech, and pitch could be misused for similar deceptions.

7 CONCLUSION

In this provocation piece, we argue that current attempts to account for manipulative interface design are curtailed by a lack of clarity and continuity around how concepts are defined; hampering efforts to measure and combat their use. We also argue that there are valuable lessons to learn from previous work on GUI dark patterns that might help us head off these problems in the realm of CUIs. Indeed, the approach we suggest may prove useful across multiple domains and interface contexts, and stands to benefit researchers, regulators and potentially users also. Further, by aligning the conceptualisation with how manipulative designs function, rather than how they manifest, we hope to make it much more difficult for designers to circumnavigate. Manipulative design that targets our cognitive biases represents a real and pertinent danger to people, and difficulties faced, even by knowledgeable users and regulators, highlight an urgent need to develop quantifiable approaches that can be easily understood by a range of stakeholders involved in the fight against unethical interface design practices.

ACKNOWLEDGMENTS

The research of this project was partially supported by the Klaus Tschira Stiftung gGmbH.

REFERENCES

[1] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I Am Definitely Manipulated, Even When I Am Aware of It. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[2] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proc. Priv. Enhancing Technol.* 2016, 4 (2016), 237–254.

[3] Harry Brignull, Marc Miquel, Jeremy Rosenberg, and James Offer. 2010. Dark Patterns-User Interfaces Designed to Trick People. http://darkpatterns.org/ visited on 18-02-2022.

[4] Birgit Brüggemeier, Michael Breiter, Miriam Kurz, and Johanna Schiwy. 2020. User Experience of Alexa when controlling music: comparison of face and construct validity of four questionnaires. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9.

[5] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers* 31, 4 (2019), 349–371.

[6] European Commission. 2016. GDPR-16 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

[7] Gregory Conti and Edward Sobiesk. 2010. Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, Raleigh, North Carolina, USA, 271. https://doi.org/10.1145/1772690.1772719

[8] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[9] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300408

[10] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. *What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit*. Association for Computing Machinery, New York, NY, USA, 61–73. https://doi.org/10.1145/3357236.3395486

[11] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. *The Dark (Patterns) Side of UX Design*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[12] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. *Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective*. Association for Computing Machinery, New York, NY, USA, pp. 1–18. https://doi.org/10.1145/3411764.3445779

[13] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. *Dark Patterns in Proxemic Interactions: A Critical Perspective*. Association for Computing Machinery, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[14] Edward T. Hall. 1966. *The hidden dimension*. Doubleday, Garden City, N.Y.

[15] Alexandra Hosszu, Daniel Rosner, and Michael Flaherty. 2019. Sleep Tracking Apps' Design Choices: A Review. In *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*. IEEE, 426–431.

[16] California State Legislature. 2018. CCPA-18 2018. California Consumer Privacy Act of 2018 [1798.100 - 1798.199] (CCPA). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[17] Thomas C. Leonard. 2008. Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy* 19, 4 (Dec. 2008), 356–360. https://doi.org/10.1007/s10602-008-9056-2

[18] BGDA Madhusanka and Sureswaran Ramadass. 2021. Implicit intention communication for activities of daily living of elder/disabled people to improve well-being. In *IoT in Healthcare and Ambient Assisted Living*. Springer, 325–342.

[19] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–32. https://doi.org/10.1145/3359183

[20] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. *What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods*. Association for Computing Machinery, New York, NY, USA, pp. 1–18. https://doi.org/10.1145/3411764.3445610

[21] EDIZIONI MINERVA MEDICA. 2018. Fitness mobile apps positively affect attitudes, perceived behavioral control and physical activities. *The Journal of sports medicine and physical fitness* (2018).

[22] Donald A. Norman. 2002. *The design of everyday things*. Basic Books, [New York]. http://www.amazon.de/The-Design-Everyday-Things-Norman/dp/0465067107/ref=wl_it_dp_o_pC_S_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I262V9ZRW8HR2C

[23] APA Dictionary of Psychology. 2022. Manipulation. https://dictionary.apa.org/persuasion visited on 18-02-2022.

[24] APA Dictionary of Psychology. 2022. Manipulation. https://dictionary.apa.org/manipulation visited on 18-02-2022.

[25] W.F. Ogburn. 1922. *Social Change with Respect to Culture and Original Nature*. B.W. Huebsch, Incorporated, University of California.

[26] European Parliament. 2022. Digital Servicess Act***I. European Parliament [A9-0356/2021]. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.html

[27] Alessandra Sarcona, Laura Kovacs, Josephine Wright, and Christine Williams. 2017. Differences in eating behavior, physical activity, and health-related lifestyle choices between users and nonusers of mobile health apps. *American Journal of Health Education* 48, 5 (2017), 298–305.

[28] Katie Seaborn and Jacqueline Urakami. 2021. *Measuring Voice UX Quantitatively: A Rapid Review*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451712

[29] Katie Seaborn and Jacqueline Urakami. 2021. Measuring voice UX quantitatively: A rapid review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

[30] Taelyr Weekly, Nicole Walker, Jill Beck, Sean Akers, and Meaghann Weaver. 2018. A review of apps for calming, relaxation, and mindfulness interventions for pediatric palliative care patients. *Children* 5, 2 (2018), 16.

[31] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013)* (May 14-17). Society for the Advancement of the Science of Digital Games, Chania, Crete, Greece, 39–46. http://www.fdg2013.org/program/papers.html

[32] Jing Zhao, Becky Freeman, Mu Li, et al. 2016. Can mobile phone apps influence people's health behavior change? An evidence review. *Journal of medical Internet research* 18, 11 (2016), e5692.

PUBLICATION P3

About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services

Authors:
Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, & Rainer Malaka

The publication contributes to the following angle:
DESIGN

This publication explores the four Social Networking Service (SNS) platforms: Facebook, Instagram, TikTok, and Twitter, based on thematic analysis. By utilising a taxonomy of 80 dark patterns drawn from related research, the study demonstrates the presence of various dark pattern types across the considered SNS. Furthermore, it captures five previously not identified and domain-specific instances within engaging and governing strategies.

Its contribution to the thesis is to the design angle by analysiing several SNS interfaces and landscaping the presence of dark patterns. It further extends this body of work through previously unidentified types and strategies.

My contribution to this paper was the design and supervision of the study, the qualitative analysis, and the interpretation of the data leading to the development of dark patterns specific to the domains of social networking services. Since qualitative analysis is best done between multiple coders, a second co-author helped with this process, which I administered and structured. I drafted the manuscript and revised it before the final publication.

The contents of this chapter originally appeared in: Mildner, T., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, ISBN: 9781450394215. DOI: 10.1145/3544548.3580695

About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services

Thomas Mildner
University of Bremen
Bremen, Germany
mildner@uni-bremen.de

Gian-Luca Savino
University of St. Gallen
St.Gallen, Switzerland
gian-luca.savino@unisg.ch

Philip R. Doyle
University College Dublin
Dublin, Ireland
philip.doyle1@ucdconnect.ie

Benjamin R. Cowan
University College Dublin
Dublin, Ireland
benjamin.cowan@ucd.ie

Rainer Malaka
University of Bremen
Bremen, Germany
malaka@tzi.de

ABSTRACT

Research in HCI has shown a growing interest in unethical design practices across numerous domains, often referred to as "dark patterns". There is, however, a gap in related literature regarding social networking services (SNSs). In this context, studies emphasise a lack of users' self-determination regarding control over personal data and time spent on SNSs. We collected over 16 hours of screen recordings from Facebook's, Instagram's, TikTok's, and Twitter's mobile applications to understand how dark patterns manifest in these SNSs. For this task, we turned towards HCI experts to mitigate possible difficulties of non-expert participants in recognising dark patterns, as prior studies have noticed. Supported by the recordings, two authors of this paper conducted a thematic analysis based on previously described taxonomies, manually classifying the recorded material while delivering two key findings: We observed which instances occur in SNSs and identified two strategies — engaging and governing — with five dark patterns undiscovered before.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; HCI theory, concepts and models; *Empirical studies in interaction design*; Interaction design theory, concepts and paradigms; • **Security and privacy** → *Usability in security and privacy*.

KEYWORDS

SNS, social media, social networking services, interface design, dark patterns, well-being, ethical interfaces

ACM Reference Format:
Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3580695

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580695

1 INTRODUCTION

"Reading this paper makes you a better person!" Emotionally pressuring language is just one example of a growing body of unethical and malicious design practices, often referred to as "dark patterns". Brignull [7] first introduced the term over a decade ago, describing dark patterns as interface design strategies that coerce or steer people into actions they would not necessarily engage in if fully informed [31]. Since then, significant effort has been expended on identifying, capturing, and describing examples of dark patterns, most notably in the domain of e-commerce websites where a large corpus of dark patterns has been catalogued [31]. There is, however, a need to extend this work further to capture dark patterns across other domains, such as social networking services (SNSs). Although SNSs are used extensively across the globe, a clear view as to the types of patterns used on these sites and how these vary between major SNSs platforms has yet to be defined. Existing literature does not focus exclusively on SNSs [4, 12, 15, 21] or focuses on specific elements only, for instance advertising dark patterns only [22] and SNSs' deletion processes [39]. The work does, however, emphasise an urgent need to deepen our understanding of how SNSs harm their users through unethical design practices [12, 21, 22, 39].

Gaining this understanding is a critical step in informing current efforts to legislate against such practices. Current legislation often limits regulation to data and privacy protection, in many cases neglecting potentially harmful consequences interface designs can have on individuals. For instance, both the GDPR [8] and CCPA [27] require website providers to make their reasons for collecting data transparently whilst also offering users the option to decline any storage of personal data. However, interface designs that encourage excessive use of social media or designs that obfuscate account deletion are currently unregulated. To support future regulatory endeavours, such as a proposed draft for guidelines by the European Data Protection Board (EDPB) [3], we first need to understand how SNSs use dark patterns and identify what domain-specific dark patterns might be used on these platforms.
Our work contributes to this understanding by answering two key research questions:

RQ1 What types of dark patterns are currently used in the four SNSs Facebook, Instagram, TikTok, and Twitter?
RQ2 Do SNSs contain dark patterns currently unique to their domain?

To address these questions, we had six HCI researchers conduct reviews across four widely used SNS platforms, based on their mobile applications: Facebook, Instagram, TikTok, and Twitter. Reviewers were assigned to identify instances of dark patterns whilst completing a range of tasks commonly carried out by users on these platforms. The decision fell on HCI researchers as the ability to not only recognise dark patterns but also to reflect and react to them was deemed a crucial requirement for reviewers to generate best possible results. Prior studies researching non-expert users' ability to recognise dark patterns demonstrate a significant difficulty for such tasks [4, 12] which we aim to mitigate by relying on experts instead. These sessions were recorded, producing 16 hours of interaction data, which was then evaluated using thematic analysis. Based on screen recordings created during the study, instances of 44 out of 80 previously established dark patterns were identified. Thematic analysis of these identified five consistent themes the describe SNS-specific dark patterns: (1) *interactive hooks*; (2) *social brokering*; (3) *decision uncertainty*; (4) *labyrinthine navigation*; and (5) *redirective conditions*. These themes were then further organised into two overarching strategies that cover more high-level incentives allowing for broader application: engaging strategies and governing strategies. These emerge from SNS-specific incentives that differ from how dark patterns are used in other domains. The strategies include two and three types of the SNS-specific dark patterns. Falling under the umbrella of engaging strategies are the *interactive hooks* and *social brokering* dark patterns, which are designed to keep users occupied and entertained with SNS for as long as possible. Whereas, SNS dark patterns that can be considered governing strategies include *decision uncertainty*, *labyrinthine navigation*, and *redirective conditions*, which are designed to navigate users' decision-making ability on these platforms. We contribute to the current dark pattern discourse and literature by considering the impact dark patterns have on four popular SNSs, and by extending current taxonomies with instances specific to this domain.

2 RELATED WORK

In this section, we will review the relationship between research on dark patterns and work on the persuasive design in social media. We highlight the necessity to carry the dark pattern discourse over to SNSs, bridging the current gap between these two strands of research. The first two subsections present a brief overview of current dark pattern taxonomies, which were used to guide the thematic analysis conducted in this study. The overview also highlights approaches and methodologies used in previous work aimed at identifying dark patterns. Some of these are adopted in this study which builds on previous work to better understand dark patterns, specifically within the context of SNS interface design. We then continue with work that dealt with users' perception of dark patterns and their ability to recognise them in different environments. Lastly, we establish the importance of considering dark patterns in social media while following the discourse of SNS interface strategies that lead to problematic or even harmful usage behaviour.

2.1 Early Research On Dark Patterns

The past decade of research into dark patterns has defined and described a comprehensive taxonomy of different types across several different domains. In recent work, Mathur et al. [32] offer a summary of the current dark pattern landscape resulting in a dense taxonomy comprising relevant works. This taxonomy is the result of their attempt to characterise dark patterns based on the cognitive biases that they exploit. As this corpus depicts a thorough overview of past dark pattern collections, we decided to use this corpus as the basis for our study. However, we decided only to include work based on empirical research, which means that we excluded reports, such as the NCC [10] and the CNIL [14] that promote dark patterns to the public while recommending guidelines to make informed decisions.

One of the earliest pieces of research on dark patterns was conducted by Brignull [7], who captured and described interface strategies used to harm people through interface tricks. Producing the first taxonomy of dark pattern types, Brignull described twelve interface tricks designed to misguide users. Among these were dark patterns, such as *sneak into basket*, *hidden costs*, and *Price Comparison Prevention*, which all operate by obscuring certain information from users of online shopping sites. The intention here is to inhibit customers' ability to make informed decisions and potentially mislead them into buying unwanted products. Alternately, the dark patterns *forced continuity*, *privacy zuckering*, and *roach motel* use strategies that limit the options and decisions available to people when using online services. In the same year as, Brignull, Conti and Sobiesk introduced their own taxonomy based on findings from a twelve-month-long study aimed at cataloguing a wide range of malicious interface practices. This second taxonomy includes eleven types of dark patterns, such as *coercion*, which describes interfaces that mandate users' decisions by restricting alternative options and enforcing compliance. Other techniques noted by the authors include *interruptions* that interfere with a user's task flow and the *obfuscation* of important information, both of which operate by hindering informed decision-making. Elsewhere, Zagal et al. examined dark patterns in video games, identifying seven types of dark patterns that specifically focus on game mechanics [47]. The research shows that while certain patterns exploit a game's ecosystem of connected users, such as *social pyramid schemes* and *impersonation*, others impact game-play experience like *grinding* and *playing by appointment*.

Elsewhere, Greenberg et al. [20] consider dark patterns in conjunction with proxemics theory [23]. Identifying nine types of dark patterns in total, the authors discuss interactions with potentially abusive systems in spatial environments. For example, the *attention grabber* and *disguised data collection* dark patterns could be used in the design of digital billboards and involves brands exploiting people's proximity and personal data to deliver personalised advertising to specific pedestrians as they pass by. In a similar vein – inspired by the concept of *Privacy by Design* project [24] – Bösch et al. introduce nine further types of dark patterns, which are effectively inverse strategies to the privacy strategies developed in the Privacy by Design project. The work also highlights the role design strategies can play in manipulating users both for good and for nefarious reasons. Collectively, these early studies show that dark patterns can appear in a variety of contexts and situations, highlighting the importance of establishing a broad understanding of their origins.

2.2 Understanding the Origins of Dark Patterns

Reflecting on Brignull's original work [7], Gray et al. looked to investigate how dark patterns are created in the first place. Here, researchers adopted a qualitative approach, using established taxonomies to analyse an image-based corpus of potential types of dark patterns [19]. The work defines five high-level strategies that practitioners engage in when developing manipulative designs. For instance, the *obstruction* dark pattern was used to make processes unnecessarily difficult, and incorporates Brignull's *roach motel*, *price comparison prevention*, and *intermediate currency* [7]. In later work that adopts a similar approach, Gray et al. analysed 4775 user-generated posts of the Reddit sub-forum *r/assholedesign* [17]. Following multiple iterations of content analysis, the authors describe a set of six properties of "asshole designers" that portray malicious motivations of designers. The *two-faced* property, for instance, describes designers who offer conflicting information that limits users' ability to make an informed decision. Understanding the origins of dark patterns as constraints under which practitioners work offers relevant insights into where to look when trying to recognise dark patterns anywhere.

Focusing more centrally on the frequency with which dark patterns are embedded in online interfaces, Mathur et al. [31] built a web-crawler application to collect data from over 11K shopping websites. The work identifies instances of dark patterns in more than 11% of their samples. Although this percentile already presents a significant number of occurrences, the authors limited the breadth of their analysis by not analysing imagery material and suggest that many more dark patterns could be identified had other factors been taken into account. Nonetheless, guided by prior works from Gray et al. [19] and Brignull [7], the authors were able to compose a taxonomy of fifteen types of dark patterns. Extending these works further, Mathur et al. [32] more recently looked to identify clusters or relationships among established dark patterns, setting the basis for this work's thematic analysis. Taking the existing five high-level characteristics from their previous work [31], the authors added a sixth characteristic that incorporates Zagal et al.'s online gaming dark pattern *disparate treatment* [47]. Their proposed model further categorises these six dark pattern characteristics into two choice architectures that distinctly affect users: (1) modification of decision space and (2) manipulation of information flow. The authors thus propose an interesting three-tiered hierarchical framework (choice architectures > high-level dark pattern characteristics > specific manifestations of dark pattern designs) under which discovered and yet-to-be-defined dark patterns can fit into. Overall, the above-mentioned work collectively describe 81 specific types of dark patterns from various domains, mostly unique in how they operate. However, we noticed an omission in regard to SNSs. Filling this gap, we build on previous work by developing a deductive codebook containing this taxonomy, which was then used during the execution of the thematic analysis.

2.3 Recognising and Identifying Dark Patterns

With a more central focus on end-users' perspectives of dark patterns, Di Geronimo et al. [12] inspected popular mobile applications sampled from the Google Play Store. The authors use a cognitive walkthrough [37] to identify dark patterns across a total of 240 apps, each used for ten minutes. Findings showed that 95% of the tested applications contained dark patterns. In a second study, the authors evaluate users' ability to recognise dark patterns. Using an online survey, the work suggests most users had problems recognising dark patterns. Adopting a similar user-centred approach, Bongard-Blanchy et al. study peoples' awareness of manipulative interface designs and their recognition of dark patterns [4]. The study, which consisted of an online survey of 413 participants, also found that 59% of their participants were able to identify "interface elements that can influence users' choices" more than half of the time, suggesting that they were somewhat able to recognise dark patterns. This aligns with previous results from Di Geronimo et al. [12] and Maier and Harr [30]. However, the work also showed that while participants understood what dark patterns were and how they might manifest, they were still deceived by them during interactions. Elsewhere, Gunawan et al. [21] use thematic analysis based on video recordings of online services to study differences between the various web modalities and resulting dark patterns. A contribution of their analysis is the addition of a further twelve specific types of dark patterns to the body of work. Their *account deletion roadblocks* dark pattern, for instance, describes the insufficient communication between the service provider and user when the latter is trying to delete their account. Together, these works demonstrate processes for assessing and evaluating dark patterns and how they manifest. In this work, we consider a wider corpus of dark patterns on a relatively small set of applications. This allows us to offer profound insights in SNS-specific dark patterns while building on established work and their methodologies.

2.4 Design Strategies on SNSs

Social media plays a major role in the daily lives of millions of people worldwide. Although prior dark pattern work has considered SNS in their research [12, 21, 22], we still lack a thorough understanding of social media-specific dark patterns and the different kinds of harm they might present compared to malicious practices observed in other domains. In this context, early work by Roffarello and De Russis [35] proposes five dark patterns in Facebook and YouTube which aim to capture their users' attention. Although offering insights into SNS specific strategies, dark patterns that exploit alternative strategies than attention-capturing have yet to be described. Aside from work with an explicit focus on dark patterns, a growing body of work suggests certain uses and users of SNS may be prone to negative consequences in terms of mental health and well-being [1, 41, 43, 45]. A well-documented reason for this may be found in research comparing users' self-reported time spent on SNS to actual times, suggesting a lack of self-control and self-determination. Indeed, numerous studies have demonstrated a lack in users' ability to self-report accurately the time they spent on any SNS [13, 26, 40]. By showing that users generally spend less time on Facebook than they think while opening the application more often than realised, both Junco's and Ernala et al.'s works highlight a problem around self-awareness when it comes to frequency and length of social media use. A similar disparity was described by Mildner and Savino [34], where a contradiction in people's perceived and actual usage behaviour was observed among 116 participants. Most Facebook users who participated in

the study admitted that they spent more time on Facebook than planned, though most also declared they had no desire to spend less time on the platform. Another interesting result of their survey is noted in participants' perceived feeling of low control over ad-related data. These findings are in line with prior research noting a general dissatisfaction in users seems to arise from limited options to protect their privacy combined with an urge to have more control over their personal data [11, 44]. Investigating advertising controls on Facebook, Habib et al. [22] consider the impacts of dark patterns when conducting an online survey to support a thematic analysis identifying users' desired advertisement controls. The authors described users' difficulty in finding Facebook's Ad Preference section, to begin with. These findings affirm a prior suggestion by Gunawan et al. [21], who argue that the granularity of interfaces may discourage users from making desired changes to their preferences. Both users' difficulty in self-reporting the amount of time spent on SNSs and their lack of agency to control personal settings outline unethical practices that may fall under the umbrella of dark patterns. Consequently, we see an urgent need to better understand dark patterns in SNSs considering domain-specific strategies. This necessity is further highlighted by Schaffner et al. [39], who demonstrate how difficult account deletion is across 20 popular SNSs. Not only does the possibility to entirely delete an account vary depending on the modality of a particular service, but the authors further notice a difficulty among users to follow through with a deletion process. In our research, we further investigate this problem by recording and analysing reviewers' usage of four SNSs, a deletion process, allowing for detailed insights.

3 EXPERT REVIEW & DATA COLLECTION

To collect necessary data of SNS usage for the thematic analysis, we asked six HCI experts to record their usage of four mobile SNS applications. By expanding the scope, we gain a general understanding of common practices across SNSs. The decision fell on Facebook, Instagram, TikTok, and Twitter, as they present some of the most popular SNSs while satisfying comparable user needs. To assist their review in targeting potential unethical practices, each expert reviewer was provided ten tasks that afforded them to use the mobile applications of each SNS intensively. To get a better understanding of their actions, we also asked them to narrate their decisions to retrieve data in the form of a think-aloud protocol [25]. As we estimated that 30 minutes were required to complete all tasks, we asked each reviewer to record their usage of two of the four SNS. Thus, three independent recordings per SNS were collected. As the experiment was conducted during the COVID-19 pandemic, reviewers completed the study without supervision.

3.1 Reviewers

For reviewers, we reached out to HCI researchers, choosing six to investigate the four SNSs (3 female, 3 male, mean age = 28.33 years, SD = 1.63). All have multiple years of experience in HCI research, with backgrounds in cognitive science, computer science, and media science. At the time of the study, their years of experience varied from two to six years, with a mean of 3.83 years (SD = 1.47). At the time of conducting this research, all reviewers were employed as researchers at academic research faculties based in Germany, focusing on HCI related topics. Five reviewers are German citizens, while one is of Russian nationality. Except for one reviewer, participants did not have prior experience conducting dark pattern related research although all shared a general conceptualisation of the field. Participation in this study was voluntary and without compensation. The decision fell on HCI researchers to conduct this study as regular users have been repeatedly shown to have difficulties in detecting dark patterns [4, 12]. In contrast, the researcher's strong expertise of usability best practices and design heuristics makes them more sensitive towards interface strategies, enabling them to uncover and discuss persuasive techniques better. Hence, their expertise allows them to identify a wider variety of dark patterns than regular users would be able to. Nonetheless, we acknowledge that this qualitative research was conducted entirely by people with strong HCI backgrounds, possibly influencing our findings and interpretations.

3.2 Preparation

To counter any systematic issues caused by a particular operating system, each reviewer was provided with two smartphone devices: an iPhone X with iOS and an Android running Google Pixel 2 device. Reviewers were asked to use both devices simultaneously and to pay attention to device-specific differences when solving the tasks. Before giving them the devices, each phone was factory reset and only contained fresh installations of the SNS[1] and some media content containing photos with creative-commons licenses. Each participant was also provided with a new phone number and email address, allowing them to create a new social media account while protecting their privacy. For the same reason, the media content was provided, as some of the tasks required the participants to create and post content.

3.3 Tasks

Each evaluator was asked to execute ten tasks during their recording sessions. We decided to implement five tasks similar to Di Geronimo et al.'s [12] tasks as they were shown to be effective for this kind of study. To ensure coverage of SNS-specific interface problems, we also added five additional tasks tailored to this domain. Before employing the tasks to the reviewers, they were tested to ensure their accuracy and ability to cover a wide range of SNS-specific functionalities. After minor revisions following piloting, the ten tasks comprised the following exercises (whereas tasks that were taken from or worded closely to Di Geronimo et al.'s experiment are highlighted by an asterisk):

1. *Turn on screen recording on each device.*
*2. *Open the app and create an account to log in and then out.*
*3. *Close and reopen the app.*
4. *Create any kind of content, post it, and delete it.*
5. *Follow and unfollow other accounts.*
*6. *Visit the personal settings.*
*7. *Visit the ad-related settings.*
*8. *Use the application for its intended use (minimum of five minutes):*

[1]Installed versions consistent throughout the study: Facebook (iOS: 321.0.0.53.119; Android: 321.0.0.37.119); Instagram: (iOS: 191.0.0.25.122; Android: 191.1.0.41.124); TikTok (iOS:19.3.0; Android: 19.3.4); Twitter (iOS: 8.69.2; Android: 8.95.0-release.00).

I *Describe the natural flow of the app – what did you use it for?*
II *Could you use the app as you wanted, or did some features 'guide' your interactions?*
III *how easy was it to get distracted, and if so, what distracted you?*

9. *Delete your account.*
10. *Turn off screen recording and save the recording.*

3.4 Procedure

Before completing their walkthrough of the SNSs, reviewers were prepared for the study via a one-hour introduction session, outlining recent research on dark patterns and established taxonomies. This took the form of an online meeting. The introduction involved instructions around the 81 dark pattern types provided in the related work section, which was followed by an open discussion aimed at answering any unresolved questions reviewers may have. Afterwards, each reviewer was provided with a copy of the information, including a detailed list of dark patterns described in the introduction session. This material is included in the supplementary material of this work. Even though all participants have backgrounds in the field, this preparation ensured a common understanding of current conceptualizations of dark patterns. Information provided to reviewers also included a sheet explaining the cognitive biases particular dark patterns were designed to exploit [31]. This was included to help them identify dark patterns that were not already described in taxonomies. Each reviewer was then randomly assigned two of the four SNSs, ensuring that each SNS was examined independently three times by three different HCI researchers. The SNS applications that reviewers used were randomly assigned while ensuring that collectively reviewers did every possible sequential permutation.

Reviewers then conducted their reviews without supervision, following a detailed step-by-step guide they were provided with. The steps can be seen in Figure 1, and explain each task along with when to start and end the screen and audio recording. The manual further reminded reviewers to openly discuss their actions allowing us to retrieve data in the form of a think-aloud protocol, which we later evaluated to understand their decisions. The quality of the reviews are twofold: (1) The tasks were designed to get detailed insights into SNSs, affording reviewers to engage with the available features of each platform. (2) By recording the reviewers' commentary, we also gained information about their perception and judgement of potentially malicious artifacts. All recordings were automatically stored on each device and retrieved when reviewers returned both smartphones.

4 FINDINGS

As a result of this data collection, the recordings provided over 16 hours of video and audio material, including the think-aloud commentary from reviewers. On average, each session lasted 41 minutes ($SD = 19$) for each device. Between the six participants, each SNS was thus used for an average of 247 minutes. Looking at different approaches to identifying dark patterns, we noticed varying methodologies in prior work. Of the eight taxonomies considered here, two relied on crowd-sourced or user-generated imagery material to apply qualitative methodologies, such as the constant comparative method [17, 19], whereas another used a web crawler to generate a large enough corpus to apply hierarchical clustering [31]. Although highlighting important interface issues that need to be addressed, the remaining five works provide limited information to recreate their results. We decided to follow prior works and conducted a thematic analysis based on the 81 dark pattern types contained in this literature while further relying on screen recordings and audio commentary of participants. Furthermore, by limiting the scope of this research to four applications only, we were able to study each service more extensively.

4.1 Analysis of the Data

To answer our research questions, two researchers conducted a thematic analysis supported by the resulting data from the screen and audio recordings [6]. While the screen recordings contained the visual captures of dark patterns and interactions thereof, the audio served as complementary material, which allowed the coders to get a detailed impression of reviewers' perception and judgement of a scene. This aided the coding process in a supportive role. However, all labelling was conducted by the two coders. The thematic analysis was carried out through a combination of deductive and inductive coding [33] using the software ATLAS.ti [16]. In line with works from Di Geronimo et al. [12] and Gunawan et al. [21], we applied codes whenever an interface was perceived problematic, rather than being driven by a concern with designer intent. This allows for the identification of dark patterns that emerge from both intentionally manipulative and unintended yet potentially harmful design choices. The thematic analysis was conducted by two researchers who designed and administered the experiment and aimed to identify indications of dark patterns present in the four SNSs. Although the recorded sessions were coded independently, both coders met for discussions after each session to align their interpretations of dark pattern definitions, rendering testing for inter-rater reliability unnecessary [2]. Additionally, all instances where coders diverged from reviewers' comments were specifically marked and later discussed among both coders. Once all sessions were analysed, they met for a final thorough discussion to establish full agreement over all 12 sessions.

4.1.1 Creating The Deductive and Inductive Codebook. The deductive codebook was derived from the descriptions of existing dark patterns as seen in Table 1. Each of the 81 codes included in the codebook denotes a specific type of dark pattern. That is 81 specific interface design elements that - by accident or by design - impinge on the users' autonomy. As the 12 dark patterns by Gunawan et al. were not described by the time we conducted the thematic analysis, we were not able to include them in this study. In an attempt to reduce the number of codes, we collapsed dark pattern types that shared the same name. This resulted in combining Brignull and Mathur et al.'s *confirmshaming* dark patterns, which share a name and very similar descriptions, which reduced the codebook to 80 codes from an initial 81. Other candidates were also considered (*privacy zuckering* [5, 7] and *bait and switch* [7, 20]), but were not collapsed as their descriptions were deemed too distinct. We thus concluded the deductive codebook with 80 codes.

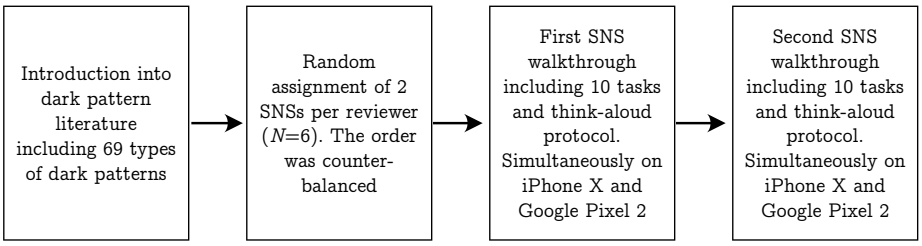


Figure 1: This flowchart describes the four steps in which our data was collected. First, reviewers received an in-depth introduction into the dark pattern literature. Each reviewer was then randomly assigned two of the four SNSs Facebook, Instagram, TikTok, and Twitter. Per SNS, the reviewers conducted a walkthrough based on 10 tasks on two devices (iOS and Android).

Once the deductive codebook had been generated, a random session was chosen and then independently coded by both coders. Deductive codes were applied wherever a type of dark pattern from the taxonomy sufficed to describe a recognised problem within the interface. In those cases where an aspect of the interface was deemed a potential dark pattern but was not present within the deductive codebook, a new code was generated and added to a separate inductive codebook. Descriptions for inductive codes focused on describing the design or interaction extracted from the recordings. Once the first session was completed, the two annotating researchers met to discuss the adequacy of established inductive codes and to form a common agreement. Afterwards, overlapping codes were eventually merged to compile a single inductive codebook comprising 22 codings by using the affinity diagramming technique [2].

4.1.2 Coding Remaining Data. Both coders then proceeded to code the remaining sessions independently, following the same procedures, only interrupted to discuss and resolve inconsistencies after each session, based on Blandford et al. [2]. By relying on screen recordings of the interfaces, instead of sampled screenshots, we gained an important advantage that allowed us to thoroughly investigate not only dark patterns on particular frames but observe sequenced interactions that would be invisible in still images. This was further affirmed by the recordings of the think-aloud protocol. If necessary, we could stop, repeat, and compare interface situations to get a deep understanding of the interactions the reviewers performed. Because of this decision, we also noticed dark patterns that only surfaced after sequential interactions were taken. This allowed us to observe dark patterns that were not previously described and were potentially unique to SNSs.

4.2 Deductive Codebook Analysis: Findings

Answering our first research question, we applied 44 out of the 80 deductive codes across our dataset, highlighting how many distinct dark patterns occurred across the four different SNSs. Table 1 displays all dark patterns and shows for which SNS they were applied during the analysis. Of these 44 that were applied, we observed 32 instances on each of the four SNSs. Regarding the original scope of individual dark pattern taxonomies, we noticed that the dark patterns described by Conti and Sobiesk [9] and Gray et al. [19] were noticed most commonly in the context of SNS, suggesting easier applicability of these taxonomies. Interestingly, these sets of dark patterns were both originally created by analysing a wide range of interfaces rather than focusing on a specific domain. Further, those described by Gray et al. [19] even subsumed dark pattern types formerly described by Brignull [7], generalising them even further. We also investigated which SNSs feature the largest variety of dark pattern types. Facebook, which contained 41 different types of dark patterns, exhibited the most variety, followed by Instagram, featuring 39, Twitter, where 35 different types of dark patterns were observed, and finally Tiktok, with 37 different types identified.

While this result highlights that SNS seem to make use of a wide variety of dark pattern types, 36 dark pattern codes were still left unused, implying that these were not appropriate to describe dark patterns in SNSs. Three groups of dark patterns from the initial 80 were not or only rarely applied: the proxemic dark patterns by Greenberg et al. [20], the gaming dark patterns by Zagal et al. [47], and the e-commerce dark patterns by Mathur et al. [31]. Many of these dark patterns were generated by describing design decisions in a specific domain or context, making it almost impossible to apply them elsewhere. One example would be Greenberg's *captive audience* dark pattern, which requires a "person [to] [enter] a particular area" [20] relying heavily on the actual proximity, hence why they are less applicable to a domain as ubiquitous as SNSs.

Beyond overly-specific dark patterns, others share a name, albeit with varying definitions. As a result, two dark patterns with the same name do not necessarily apply identically to SNS. For example, *privacy zuckering* was described by Brignull [7] and Bösch et al. [5]. While Bösch et al. focuses specifically on sharing more data than a user intends to give based on privacy settings, Brignull does not make this specification and thus leaves more room for interpretation and application. Hence, Brignull's version of the dark pattern was applied across all four SNS during our analysis, while the version by Bösch et al. was only applied explicitly on Facebook and TikTok. In other cases, some names of dark pattern promise applicability in SNSs but were actually never applied as their definition hindered accurate coding. An example is illustrated by the dark pattern *immortal accounts* by Bösch et al. [5]. While the name of the dark pattern suggests that user accounts are (almost) impossible to delete, its definition refers to service providers requiring new users to sign up for accounts to use their service.

4.3 Inductive Codebook Analysis: Findings

Although 44 codes from the deductive codebook could be applied in the SNSs, often, new codes were required to describe interface issues

Author	Dark Pattern	F	I	Ti	Tw
Brignull [7]	Bait And Switch	●	○	○	●
	Confirmshaming	●	●	●	●
	Disguised Ads	○	●	○	●
	Forced Continuity	○	○	○	○
	Friend Spam	○	○	○	○
	Hidden Costs	○	○	○	○
	Misdirection	●	●	●	○
	Price Comparison Prevention	○	○	○	○
	Privacy Zuckering	●	●	●	●
	Roach Motel	●	●	●	●
	Sneak Into Basket	○	○	○	○
	Trick Question	○	○	○	○
Conti & Sobiesk [9]	Coercion	○	○	○	○
	Confusion	●	○	○	●
	Distraction	●	●	●	●
	Exploiting Errors	○	○	○	○
	Forced Work	●	●	●	●
	Interruption	●	●	●	●
	Manipulating Navigation	●	●	●	●
	Obfuscation	●	●	●	●
	Restricting Functionalities	●	●	○	○
	Shock	○	●	○	○
	Trick	○	○	○	○
Zagal et al. [47]	Grinding	○	○	○	○
	Impersonation	○	○	○	○
	Monetized Rivalries	○	○	○	○
	Pay To Skip	○	○	○	○
	Playing By Appointment	○	○	○	○
	Pre-Defined Content	○	○	○	○
	Social Pyramid Schemes	○	○	○	○
Greenberg et al. [20]	Attention Grabber	●	●	●	○
	Bait And Switch	○	○	○	○
	Captive Audience	○	○	○	○
	Disguised Data Collection	○	○	○	○
	Making Personal Info. Public	○	○	○	○
	The Milk Factor	○	○	○	●
	Unintended Relationships	○	○	○	○
	We Never Forget	○	○	○	○
Bösch et al. [5]	Address Book Leeching	●	●	●	●
	Bad Defaults	●	●	●	●
	Forced Registration	○	○	○	○
	Hidden Legalese Stipulations	●	●	●	●
	Immortal Accounts	○	○	○	○
	Information Milking	●	○	○	○
	Privacy Zuckering	●	○	●	○
	Shadow User Profiles	○	○	○	○
Gray et al. [19]	Forced Action	●	●	●	●
	Gamification	●	●	●	●
	Social Pyramid	●	●	●	●
	Interface Interference	●	●	●	●
	Aesthetic Manipulation	●	●	●	●
	False Hierarchy	●	●	●	●
	Hidden Information	●	●	●	●
	Preselection	●	●	●	●
	Toying With Emotions	●	●	●	●
	Nagging	●	●	●	●
	Obstruction	●	●	●	●
	Intermediate Currency	●	●	●	●
	Sneaking	●	●	○	○
Gray et al. [18]	Automating The User	●	●	○	●
	Controlling	●	●	●	●
	Entrapping	○	○	○	○
	Misrepresenting	●	●	●	●
	Nickling-And-Diming	○	○	○	○
	Two Faced	○	○	○	○
Mathur et al. [31]	Forced Acrtion (see Gray et al. [19])				
	Forced Enrollment	○	○	○	○
	Misdirection	●	●	●	●
	Pressured Selling	●	●	●	●
	Visual Interference	●	●	●	●
	Obstruction (see Gray et al. [19])				
	Hard To Cancel	●	●	●	●
	Scarcity	●	●	●	●
	High-Demand Messages	○	○	○	○
	Low-Stock Messages	○	○	○	○
	Sneaking (see Gray et al. [19])				
	Hidden Subscriptions	○	○	○	○
	Social Proof	●	●	●	●
	Activity Notifications	○	○	○	○
	Testimonials	○	○	○	○
	Urgency	●	●	●	●
	Countdown Timer	○	○	○	○
	Limited-Time Messages	○	○	○	○

Legend: **F** - Facebook
I - Instagram
Ti - TikTok
Tw - Twitter

Table 1: This table offers an overview of all 80 deductive codes used in the thematic analysis, sorted by authors. We only considered original sources for dark patterns to rely on unique codes and avoiding redundancy. As both Gray et al. [19] and Mathur et al. [31] include both high-level and low-level definitions, we used indentation to highlight categorical differences for these types of dark patterns. For transparency, we added three of Mathur et al.'s [31] high-level dark pattern categories, which were carried over from Gray et al. [19] and labeled accordingly. The codes were applied in four SNSs (F - Facebook, I - Instagram, Ti - TikTok, and Tw - Twitter). Codes that were identified by either of the two coders conducting the thematic analysis are indicated with "●" whereas "○" indicates that a code for a dark pattern was not found in the SNSs.

not yet covered. We decided early on to stick rigidly to established definitions of the types of dark patterns when applying deductive codes to avoid ambiguity among coders. During the coding of a first sample, we noticed issues were close in nature to certain deductive codes, but narrow wording hindered precise usage in different contexts. For instance, this was the case with the *confirmshaming* code. The underlying dark pattern, first coined by Brignull [7], describes texts that guilt users into certain actions based on shameful language. Applying an inverse strategy, we noticed various cases in which language was used in the form of positive encouragement to steer users towards a certain direction. We, therefore, created an additional code, *persuasive langugage*, that we defined bidirectionally describing any instance where language is used to push decisions. In total, we defined 22 unique codes describing problematic interactions or otherwise questionable interfaces that could not be coded using deductive codes (the entire inductive codebook, including code descriptions, are provided in Appendix A). As seen in Table 2, the 22 inductive codes did not occur across each of the four SNSs equally. Yet, 16 codes were applied to recording sessions on Facebook, 15 to TikTok and Twitter, while 14 were applied on Instagram sessions, showing a similar propagation of dark patterns across SNSs. Once all twelve sessions were analysed, we applied axial coding to the 22 codes. This resulted in the identification of five themes that describe the various types of dark patterns that are specific to SNSs. Answering our second research question, these themes encompass: (1) *interactive hooks*; (2) *social brokering*; (3) *decision uncertainty*; (4) *labyrinthine navigation*; and (5) *redirective conditions*. Moreover, we were able to assign these five themes to two broader strategies, to describe practitioners' and SNSs' intentions to navigate users' decision-making: (1) Engaging strategies and (2) Governing strategies. Engaging strategies envelope the themes *interactive hoooks* and *social-broking*, whilst governing strategies incorporate *decision uncertainty*, *labyrinthine navigation*, and *redirective conditions*. Table 2 provides a complete summary of these findings, including the 22 inductive codes, resulting in five themes and overarching strategies.

4.4 Engaging Strategies

In the context of SNSs, engaging strategies cover dark patterns where the goal is to keep users occupied and entertained for as long as possible. Prior dark patterns described by Roffarello and Russis [35] fall under this strategy alongside Lukoff et al.'s [28] discussion of potentially existing *attention-capture* dark patterns, which the OECD recently integrated [38]. As Table 2 indicates, the overall count of engaging strategies and subordinated codes is higher compared to those assigned to governing strategies, which is in line with research on people's motivation to use SNSs [46]. Emerging from the five themes, we identified two SNS-specific dark patterns - *interactive hooks* and *social brokering* - that subscribe to the engaging strategies (see Table 2).

4.4.1 Interactive Hooks. We define *interactive hooks* as design mechanisms that use rewarding schemes to keep users entertained and spend more time on a service. Throughout our recordings, we found multiple cases where such mechanisms were implemented. For example, some form of *gamification* elements [42] were coded within each SNS, galvanising users to share more information about themselves or connect to new people (see Figure 2). In another example, we found that many artefacts utilise *addictive mechanisms* [36] that often provide seemingly infinite content, further coded with *pull-to-refresh*, *infinite scrolling*, and *auto playing* media. With this finding, we confirm previous results of Lukoff et al. [29] who looked at auto-playing mechanisms on YouTube. We extend their work by also considering four alternative SNSs. By relying on a sequence of interactions, these addictive mechanisms can only be described with difficulty from still images, demonstrating the complexity in which dark patterns manifest in SNSs.

4.4.2 Social Brokering. We define *social brokering* as design mechanisms that nudge users to create multiple connections with people (e.g. based on similar characteristics) while suggesting new people to connect to, leading users to share more than they may want to a wider public. The name is inspired by agents whose aim is to facilitate connections between potential (transaction-) partners - here in the context of social networks. Although gamification strategies of the *interactive hooks* pattern already encourage users to increase their social connections, we found a range of artefacts specifically designed for this purpose codes as *social connector*. Moreover, we noticed that each SNS customised content presented based on the reviewers' usage behaviour. While in some instances the news feed content appeared a poor fit for the reviewers' preferences (coded as *false content customisation*), we also found very popular content reappearing across the recordings, captured in the code *regression toward the mean*. Visualised in Figure 3, Brignull's bad defaults pattern [7] is exploited to promote one's account to other users also outside the platform itself. Figure 4 shows how Instagram nudges its users to upload their device's local contacts to connect to more accounts quickly. However, by giving away details from private contacts, SNS providers can also store information from people who do not use their services without getting their consent first. In a similar case, Facebook uses persuasive language by telling users to add friends in order to see more content (see Figure 5).

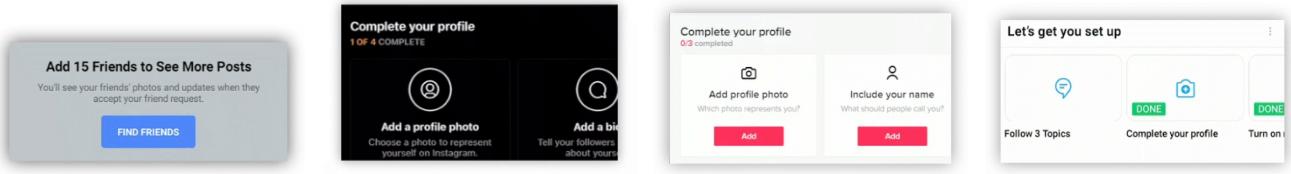
4.5 Governing Strategies

Governing strategies describe interface designs that navigate users' decision-making towards the designers' and/or platform providers' goals. Essentially, these are designed to control or govern user behaviour. While existing dark patterns, such as *interface interference* [17], fit this strategy's scope, we shine a light on not yet discussed dark pattern types: (1) *decision uncertainty*; (2) *labyrinthine navigation*; (3) and *redirective conditions*. All these strategies share a limitation of users' agency as SNS providers override users' goals with their own incentives.

4.5.1 Decision Uncertainty. We define *decision uncertainty* as strategies that are confusing to users by diminishing their ability to assess situations, leaving the user clueless as to what is expected of them or what options are available. Most similar to this dark pattern is the *confusing* dark pattern by Conti and Sobiesk [9]. However, in SNS, the strategy is not necessarily limited to incomprehensible questions or information but includes other elements, like *distraction* [9], that overwhelm users. Interface elements of the *decision uncertainty* code were so striking that we decided to promote it to a theme.

Strategy	Theme	No.	Inductive Code	F	I	Ti	Tw
Engaging Strategies	Interactive Hook	1.	Addictive Design	○	●	●	●
		2.	Autoplay Content	●	●	●	●
		3.	Fear Of Missing Out	○	○	○	●
		4.	Gamification	●	●	●	●
		5.	Infinite Scrolling	●	●	●	●
		6.	Pull To Refresh	●	●	●	●
		7.	Reduced Friction	●	●	●	●
	Social Brokering	8.	False Content Customisation	●	○	○	○
		9.	Regression Toward The Mean	●	○	●	●
		10.	Social Connector	●	●	●	○
Governing Strategies	Decision Uncertainty	11.	Decision Uncertainty	●	○	○	●
		12.	Clinging To Accounts	●	●	●	●
		13.	Persuasive Language	●	●	●	●
	Labyrinthine Navigation	14.	External Solution Search	○	●	○	○
		15.	Labyrinth	●	●	○	●
		16.	Hidden In Plain Sight	●	●	●	●
	Redirective Condition	17.	Auto Accept Third Party Terms	○	○	○	●
		18.	Decision Governing	●	●	○	○
		19.	Forced Access Granting	○	○	●	○
		20.	Forced Dialogue Interaction	●	○	●	○
		21.	Forced Grace Period	●	●	●	●
		22.	Plain Evil	○	○	●	○
			Total	16	14	15	15

Table 2: This table lists inductive codes and their presence within the four SNSs Facebook (F), Instagram (I), TikTok (Ti), and Twitter (Tw). Inductive codes observed by either of the two coders in a particular SNS session are visualised with "●" while "○" implies that a code was not found. This table further shows the themes each inductive code was assigned to based on axial coding of the data and the high-level strategies they subscribe to - engaging strategies and governing strategies. To align with prior research of this field, the developed themes are later referred to as types of dark patterns.



(a) Screenshot from Facebook (b) Screenshot from Instagram (c) Screenshot from TikTok (d) Screenshot from Twitter

Figure 2: Four examples of *interactive hooks*. Each sub-figure contains gamification elements that galvanise users to widen their social networks or publish more information about themselves.

Other elements obscured decision-making further. Some interfaces obfuscated the account deletion process (coded with *clinging to accounts*) while others used *persuasive language* to confuse users, similar to the *confirmshaming* dark pattern [7]. We found a quite unique design choice TikTok users faced when first logging on to the platform: When first opening the app after logging in, users are prompted with an interface asking them to choose preferred ad-related settings. While making their decision, both video and audio media is running in the background, contributing to cognitive overload. During the recordings, we noticed that reviewers of the data collection quickly complied with the platform's preference that utilises interface interference [17] and visual interference [31] dark patterns (see Figure 6). In a second example, Twitter users trying to delete their accounts will only find an option to deactivate it. Although it is possible to fully delete their account by following this path, the wording obfuscates this possibility.

4.5.2 Labyrinthine Navigation. We define *labyrinthine navigation* as nested interfaces that are easy to get lost in, disabling users from choosing preferred settings. This pattern is often seen in SNS settings menus. Related to *manipulating navigation* [9], but not necessarily steering users towards a designer's objective. Instead, this dark pattern describes interface architectures, such as menus, that users will easily get lost in, leaving them unsuccessful in achieving their goals. While recording the data, especially tasks six (visit the personal settings) and seven (visit the ad-related settings), surfaced

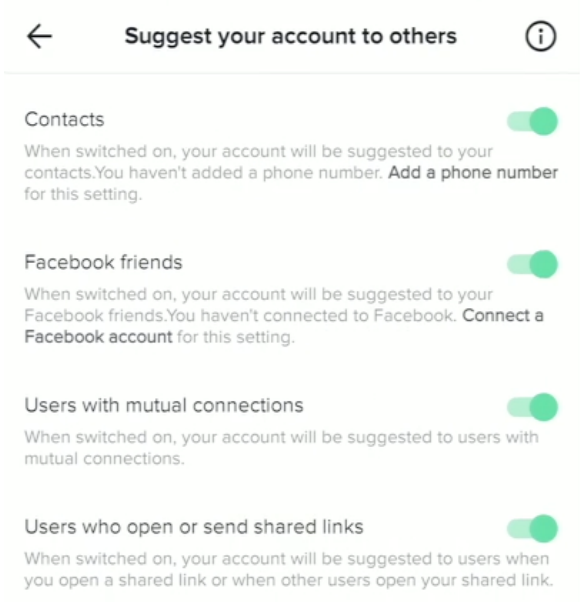


Figure 3: Example of *social brokering* from TikTok where settings are pre-set to suggest an account within and outside the SNS. The interface includes *bad defaults* and *privacy zuckering* dark patterns

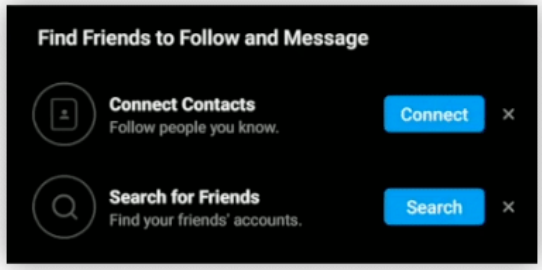


Figure 4: Example of *social brokering* from TikTok where settings are pre-set to suggest an account within and outside the SNS.

difficulties for reviewers to find specific settings, some of them coded *hidden in plain sight* camouflaged between a wide collection of other options. We noticed this issue across all four SNSs. In a particular case, we noticed one participant using Instagram to switch applications to an online search engine to look up the solution for how to find a specific setting.

4.5.3 Redirective Conditions. We define *redirective condition* as choice limitations that force users to overcome unnecessary obstacles before being able to achieve their goals. Redirective conditions usually favour the objectives of the SNS. The *forced action* dark pattern [17] plays an important role here, as users are required to first comply with certain demands before being able to do what they want. In the context of SNS, this dark pattern includes passive functionalities, like the restriction of services that are only lifted

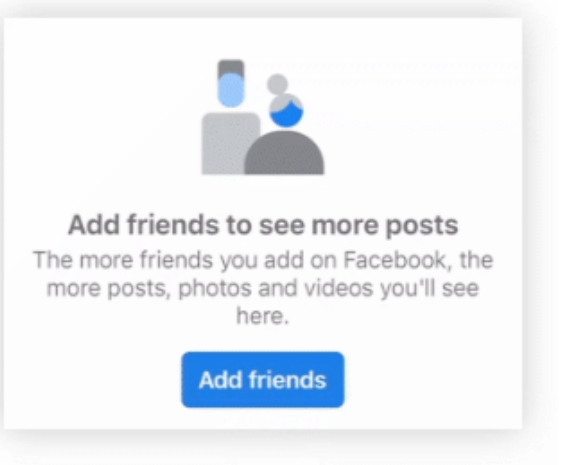


Figure 5: Example of *social brokering* from Facebook. Once users reach the bottom of their timeline content, this message appears, nudging them to add friends to see new posts.

after users give permissions unrelated to the functionality (coded as *forced access granting*). In related situations, we noticed *forced dialogue interactions* that required users to engage with text elements otherwise occupying desired functionalities. More specific yet similar to the *roach motel* [7], *hard to cancel* [31], or *account deletion roadblocks* [21] dark patterns, a grace period of up to 30 days was announced by all SNSs before certain deletion processes were accepted. The *Forced Grace Period* code was used in such instances. When deleting content, Facebook keeps the targeted item, as seen in Figure 7. Trying to bypass this rule, participants found it difficult to find the particular settings that would delete their content immediately. Interestingly, all four SNS denied users an immediate account deletion, as demonstrated by TikTok in Figure 8. Each had a 30 days return option that would automatically reactivate accounts, even if users accidentally logged in to the SNS.

5 DISCUSSION & IMPLICATIONS

Our examination of four SNSs — Facebook, Instagram, TikTok, and Twitter — identified instances of 44 dark patterns from a taxonomy of 80 codified in prior dark pattern literature. The work also identified instances of 22 inductive codes capturing malicious design artefacts that were not outlined in earlier research. Findings stem from expert reviews of the aforementioned SNSs carried out by six trained HCI researchers. Each researcher reviewed two SNSs and executed ten tasks designed to thoroughly understand the applications. The 22 inductive codes were then analysed, supported by axial coding. This resulted in the generation of five common themes describing SNS-specific dark patterns and two high-level strategies: Engaging strategies and governing strategies. Figure 9 summarises these findings while further allocating Mathur et al.'s [32] dark pattern characteristics.

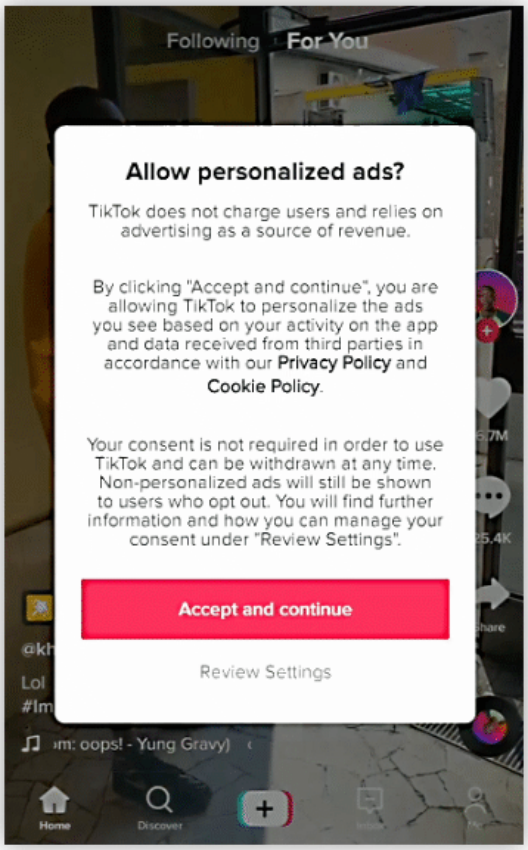


Figure 6: Example of *decision uncertainty* from TikTok. After first logging in, users of the SNS are asked to choose preferences for their ad-related settings. Meanwhile, video and audio media are playing in the background, complicating the interaction. The interface further contains *interface interference, visual interference* and hidden-legalese stipulation dark patterns.

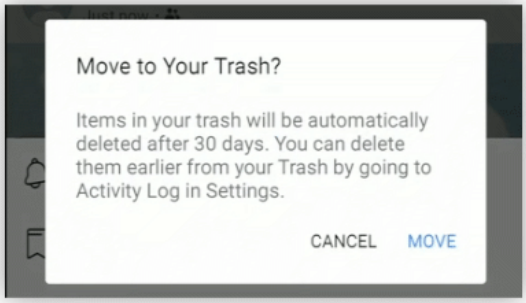


Figure 7: Example of *redirective conditions* from Facebook. When trying to delete a post, Facebook will instead keep the item stored for another 30 days.

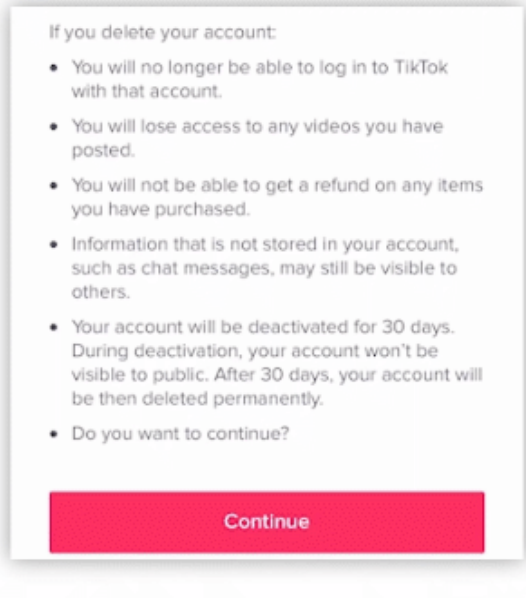


Figure 8: Example of *redirective conditions* from TikTok. Accounts will be kept and able to be restored for 30 days without giving users the choice to bypass this decision.

5.1 Dark Patterns In SNSs

To explore the potential existence of dark patterns in SNSs, we began the thematic analysis by initialising a deductive dark pattern codebook comprising 80 dark pattern types from prior works [5, 7, 9, 18–20, 31, 47]. Answering our first research question regarding the types of dark patterns that are used across the four SNSs, we identified instances of 44 types. This might suggest that the other 36 types of dark patterns are, as our experience conducting thematic analysis attests, highly domain-specific, hindering their potential to be applied elsewhere.

Indeed, despite the relative success of our attempts to apply these dark patterns to SNSs, different levels of generalisability and specification caused ambiguity decreasing confidence of coders when applying certain dark patterns. Some works provide alternating abstractions for their taxonomies. Conti and Sobiesk [9], for instance, call their patterns malicious interface design *techniques*, as the term dark pattern was not widely established at the time of their publication. In another work, Gray et al. [19] speak of dark pattern *strategies*, placing their findings on a more abstract level as, for example, Zagal et al.' [47] dark patterns, who speak of game dark patterns without further labels.

Because we agree with the necessity to understand dark patterns in different hierarchical contexts, we chose to include all taxonomies in our study to learn about differences when they are applied in situations outside their original field of application. Further guided by our decision to remain close to provided definitions,

we learned that dark patterns with abstract definitions were applied more often throughout the SNSs. On the contrary, 36 dark patterns were left unused during our study. Certain dark patterns shared a name with varying definitions leading to ambiguity when used. Some dark patterns were derived from highly specific contexts, such as e-commerce [7, 31] or games [47], which hindered their usage elsewhere. Others contained overly precise descriptions descriptions [5] but could find usage elsewhere if phrased more generically. This implies that general and unspecific dark patterns gain the utility to describe unethical practices and, thus, better help to identify interface problems.

This implication is in line with prior research by Mathur et al. [31]. Aside from the identification of 12 dark patterns, the authors found an alternative approach to abstract the basic operations of dark patterns by mapping them onto cognitive biases that they exploit [31, 32]. The authors developed six characteristics distinguishing overall approaches to dark patterns: (1) asymmetric; (2) covert; (3) deceptive; (4) hides information; (5) restrictive; and (6) disparate treatment. Through these characteristics, the authors were able to characterise a comprehensive dark pattern taxonomy in which we notice an accessible and extendable framework for future works.

5.2 SNS-Specific Dark Patterns & Strategies

Answering the second research question, asking about dark patterns unique to SNSs, our study revealed five types of dark patterns not contained in previous work. When defining the dark patterns, one goal was to provide enough abstraction to enable applicability outside SNSs. For the same reason, we placed them in Mathur et al.'s [31, 32] six characteristics to allow future works to consider these dark patterns under their lens as well. As can be seen in Figure 9, not all of Mathur et al.'s characteristics could be applied with the same effectiveness or explicitness. Interestingly, we were not able to apply the *disparate treatment* characteristic at all because it solely contains Zagal et al.'s [47] game dark patterns which describe mechanics where games benefit players who, for example, buy advantages. We found none of the four considered SNSs to offer distinct benefits to certain users while posing disadvantages to others.

Another reason why Mathur et al.'s characteristics show lower effectiveness in SNSs may be described in how the two strategies operate, distinguishing them from prior dark pattern settings. As an incentive to increase time spent on their platforms, SNS need satisfied users. Since the implementation of harmful designs could act as an antagonist to this goal, jeopardising potential advertisement revenue, we describe engaging and governing strategies that navigate users' decision-making while possibly keeping their satisfaction with the SNS high [34]. Although we cannot draw an easy causal connection between how dark patterns affect users and users' well-being, assessing identified SNS dark patterns further resulted in the description of two design strategies, *engaging strategies* and *governing strategies* (see Figure 9). While one group aims to increase the time users spend on SNS, the other governs their decision-making by nudging users into desired directions or keeping them from other options by, for example, designing for increased friction.

With the intent to enable future work to better build on our findings and inspired by the types of dark patterns that showed more pertinence in our study, our aim was to describe our dark patterns to allow application outside SNS-specific contexts. As a different environment, many games feature achievement tracks requiring players to regularly play the game to progress. Similar to Zagal et al.'s *playing by appointment*, it is imaginable that these games feature *interactive hooks* to keep players engaged or contain distracting elements obfuscating users' decisions and, thus, deploying *decision uncertainty*.

Although this work mainly contributes to the academic community, this strand of research has gained traction outside, including regulation and guidelines. In a first line of defence against violations of GDPR [8] requirements in SNS contexts, the EDPB [3] developed a guideline for designers and users to recognise and avoid dark patterns. However, the included dark pattern categories (overloading, skipping, stirring, hindering, fickle, and left in the dark) lack alignment with taxonomies developed in the academic community that relies on empirical evidence. In this regard, we acknowledge a potential for better cooperation between law and HCI research to work on the protection of users in a joint effort. To offer some aid in these efforts, our goal was to create definitions that are broadly applicable. The five dark patterns expand current taxonomies by the scope of SNSs, whereas the two strategies provide high-level categories to describe alternate burdens placed on users not yet covered.

6 LIMITATIONS & FUTURE WORK

As previously stated in Section 3, we faced certain limitations in the data collection phase of this work. Firstly, we decided to focus on only four SNSs that we deemed comparable after prior considerations looking at a wider range of applications for the user count and similar functionalities. This limitation extends to the fact that we limited the review to only looking at their mobile applications. Although this limitation allowed us to study each SNS in a single modality on a deeper level, our findings do not necessarily represent all SNSs. None of the considered SNSs indicates any usage of the *disparate treatment* characteristic, which describes discrepancies in the treatment of paid and free customers, respectively. This might be more relevant in a study of LinkedIn, for example, as it offers both free and paid subscriptions, the latter of which affords certain advantages like hiding one's identity when looking at another user's profile. Future work could extend our findings by considering other SNSs. Secondly, we decided early on to only focus on mobile applications of considered SNSs, following users' preferences. However, users may be faced with alternative dark patterns based on SNSs' modalities, as the research by Guanawa et al. [21] or Schaffner et al. [39] suggest. Future studies could consider our results when expanding on other SNSs and modalities. Thirdly, this research was conducted during the COVID-19 pandemic. The data collection was thus conducted without supervision after providing necessary information, including a comprehensive manual. While neither screen nor voice recordings indicated confusion about the tasks, room for misunderstanding remained. Thirdly, we relied on HCI researchers as expert reviewers. This decision is

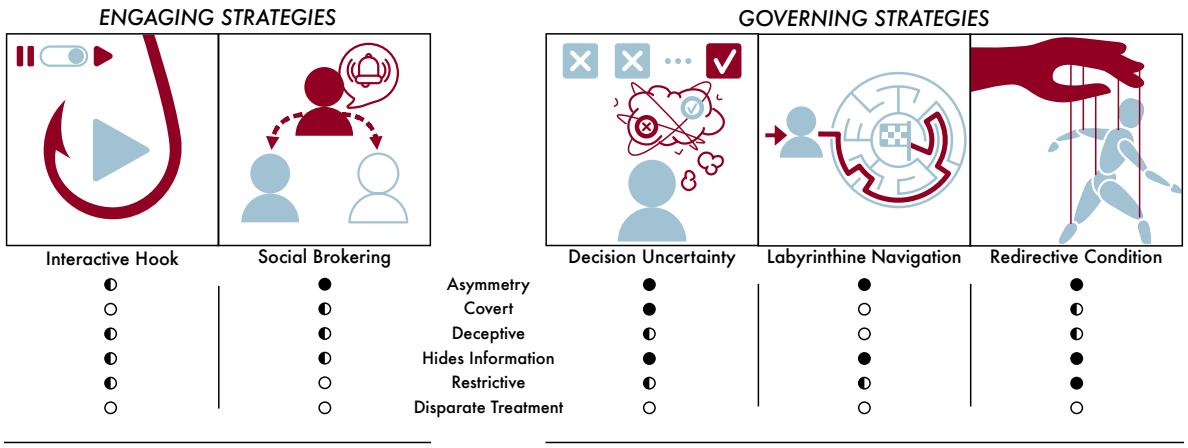


Figure 9: Summary of engaging and governing strategies with five SNS dark patterns in two strategies - engaging and governing. For future work, each dark pattern was assigned corresponding attributes following Mathur et al. [31] six dark pattern characteristics.

justified by their competence to understand and recognise state-of-the-art design practices, unlike regular users would be able to. However, by limiting reviewers' expertise to a single domain, we may have missed alternative expertise that may surface additional findings. Future work could consider recruiting experts with backgrounds in cognitive science and psychology, including a thorough understanding of cognitive biases. Their expertise could be particularly interesting in establishing connections between our current understanding of dark patterns and of cognitive biases.

During the thematic analysis, we also faced some limitations. Firstly, we limited the deductive codebook to eight works comprising 81 dark patterns. We relied on Mathur [32] dark pattern review to get a comprehensive taxonomy. However, we decided to neglect dark patterns that were described outside the academic community, including guidelines such as those published by the NCC [10], CNIL [14], or the EDPB [3]. Moreover, dark patterns by Gunawan [21] were not included, as they were not mapped onto Mathur et al. [31, 32] dark pattern characteristics, making comparisons more difficult. Future work could generate a complete mapping of a comprehensive dark pattern taxonomy onto Mathur et al.'s characteristics. Connecting dark patterns to cognitive biases, as Mathur et al. propose, seems to be a natural next step for the dark pattern discourse. To prepare this work for this direction, we assigned our findings to appropriate characteristics. In this process, we noticed that the current scope of characteristics does not cover all problematic design strategies contained in SNS. However, a thorough analysis of potential cognitive biases active in SNS would have been outside the scope of this research. Future work could address this gap and extend this work by bringing together the studies surrounding cognitive biases and dark patterns to generate a resourceful and sustainable framework.

7 CONCLUSION

In recent years, the dark pattern landscape has expanded into various different domains. In this paper, we contribute to this body of research through the application and expansion of a comprehensive taxonomy of dark patterns in the context of SNSs. Supported by thematic analysis, we investigate Facebook, Instagram, TikTok, and Twitter and confirm that the platforms all deployed a variety of dark patterns and design strategies aimed at limiting users' agency, steering their decision-making. Findings suggest that dark patterns with higher grades of abstraction are easier to be applied in multiple contexts compared to those given narrower definitions. Lastly, our results yield evidence for two high-level strategies - engaging and governing - containing five types of dark patterns previously not described.

ACKNOWLEDGMENTS

The research of this work was partially supported by the Klaus Tschira Stiftung gGmbH.

REFERENCES

[1] Ine Beyens, J Loes Pouwels, Irene I van Driel, Loes Keijsers, and Patti M Valkenburg. 2020. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports* 10, 1 (2020), 1–11.

[2] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. *Synthesis Lectures on Human-Centered Informatics* 9, 1 (April 2016), 1–115. https://doi.org/10.2200/S00706ED1V01Y201602HCI034 Publisher: Morgan & Claypool Publishers.

[3] European Data Protection Board. March, 2022. Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them | European Data Protection Board. https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-32022-dark-patterns-social-media_en Visited on 2022-03-29.

[4] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I Am Definitely Manipulated, Even When I Am Aware of It. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[5] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proc. Priv. Enhancing Technol.* 2016, 4 (2016), 237–254.

[6] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, Pranee Liamputtong (Ed.). Springer Singapore, Singapore. https://doi.org/10.1007/978-981-10-5251-4_103

[7] Harry Brignull. 2010. Deceptive Design – formerly darkpatterns.org. https://www.deceptive.design/ Visited on 2022-03-29.

[8] European Commission. 2016. GDPR-16 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement

of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

[9] Gregory Conti and Edward Sobiesk. 2010. Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, Raleigh, North Carolina, USA, 271. https://doi.org/10.1145/1772690.1772719

[10] Frobrukerrådet (Norwegian Consumer Council). 2019. Deceived by design: How tech companies use dark patterns to discourage us from exercising our rights to privacy. https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf5

[11] Ratan Dey, Zubin Jelveh, and Keith Ross. 2012. Facebook users have become much more private: A large-scale study, In 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2012. *2012 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2012* 01, 346–352. https://doi.org/10.1109/PerComW.2012.6197508

[12] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[13] Sindhu Kiranmai Ernala, Moira Burke, Alex Leavitt, and Nicole B. Ellison. 2020. How Well Do People Report Time Spent on Facebook? An Evaluation of Established Survey Questions with Recommendations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376435

[14] Commission Nationale Informatique et Libertés. 2019. On the practical procedures for collecting the consent provided for in article 82 of the french data protection act, concerning operations of storing or gaining access to information in the terminal equipment of a user (recommendation "cookies and other trackers"). https://www.cnil.fr/sites/default/files/atoms/files/draft_recommendation_cookies_and_other_trackers_en.pdf

[15] Lothar Fritsch. 2017. Privacy dark patterns in identity management. In *Open Identity Summit 2017 : Proceedings (Lecture Notes in Informatics, 277)*. Gesellschaft für Informatik, Bonn, Germany, 93–104.

[16] ATLAS.ti Scientific Software Development GmbH. 2021. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[17] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300408

[18] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. *What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit*. Association for Computing Machinery, New York, NY, USA, 61–73. https://doi.org/10.1145/3357236.3395486

[19] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. *The Dark (Patterns) Side of UX Design*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[20] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. *Dark Patterns in Proxemic Interactions: A Critical Perspective*. Association for Computing Machinery, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[21] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2021. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1–29. Issue CSCW2. https://doi.org/10.1145/3479521

[22] Hana Habib, Sarah Pearman, Ellie Young, Ishika Saxena, Robert Zhang, and Lorrie Faith Cranor. 2022. Identifying User Needs for Advertising Controls on Facebook. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1–42. Issue CSCW1. https://doi.org/10.1145/3512906

[23] Edward T. Hall. 1966. *The hidden dimension*. Doubleday, Garden City, N.Y.

[24] Peter Hustinx. 2010. Privacy by design: delivering the promises. *Identity in the Information Society* 3, 2 (Aug. 2010), 253–255. https://doi.org/10.1007/s12394-010-0061-z

[25] Monique W.M. Jaspers, Thiemo Steen, Cor van den Bos, and Maud Geenen. 2004. The think aloud method: a guide to user interface design. *International Journal of Medical Informatics* 73, 11 (2004), 781–795. https://doi.org/10.1016/j.ijmedinf.2004.08.003

[26] Reynol Junco. 2013. Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior* 29, 3 (May 2013), 626–631. https://doi.org/10.1016/j.chb.2012.11.007

[27] California State Legislature. 2018. CCPA-18 2018. California Consumer Privacy Act of 2018 [1798.100 - 1798.199] (CCPA). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[28] Kai Lukoff, Cissy Yu, Julie Kientz, and Alexis Hiniker. 2018. What Makes Smartphone Use Meaningful or Meaningless? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 22 (March 2018), 26 pages. https://doi.org/10.1145/3191754

[29] Kai Lukoff, Cissy Yu, Julie Kientz, and Alexis Hiniker. 2018. What Makes Smartphone Use Meaningful or Meaningless?, In Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 22, 26 pages. https://doi.org/10.1145/3191754

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

[31] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–32. https://doi.org/10.1145/3359183

[32] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. *What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods*. Association for Computing Machinery, New York, NY, USA, pp. 1–18. https://doi.org/10.1145/3411764.3445610

[33] Philipp Mayring. 2020. Qualitative Inhaltsanalyse. In *Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren*, Günter Mey and Katja Mruck (Eds.). Springer Fachmedien Wiesbaden, Wiesbaden, 495–511. https://doi.org/10.1007/978-3-658-26887-9_52

[34] Thomas Mildner and Gian-Luca Savino. 2021. Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451659

[35] Alberto Monge Roffarello and Luigi De Russis. 2022. Towards Understanding the Dark Patterns That Steal Our Attention. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–7. https://doi.org/10.1145/3491101.3519829

[36] Christian Montag, Bernd Lachmann, Marc Herrlich, and Katharina Zweig. 2019. Addictive Features of Social Media/Messenger Platforms and Freemium Games against the Background of Psychological and Economic Theories. *International Journal of Environmental Research and Public Health* 16, 14 (Jan. 2019), 2612. https://doi.org/10.3390/ijerph16142612 Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[37] Jakob Nielsen. 1994. Usability inspection methods. In *Conference Companion on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 413–414. https://doi.org/10.1145/259963.260531

[38] OECD. 2022. Dark commercial patterns. , 96 pages.

[39] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142

[40] Sarita Yardi Schoenebeck. 2014. Giving up Twitter for Lent: how and why we take breaks from social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 773–782. https://doi.org/10.1145/2556288.2556983

[41] Holly B. Shakya and Nicholas A. Christakis. 2017. Association of Facebook Use With Compromised Well-Being: A Longitudinal Study. *American Journal of Epidemiology* 185, 3 (Feb. 2017), 203–211. https://doi.org/10.1093/aje/kww189

[42] Jorge Simões, Rebeca Díaz Redondo, and Ana Fernández Vilas. 2013. A social gamification framework for a K-6 learning platform. *Computers in Human Behavior* 29, 2 (March 2013), 345–353. https://doi.org/10.1016/j.chb.2012.06.007

[43] Jin-Liang Wang, Linda A. Jackson, James Gaskin, and Hai-Zhen Wang. 2014. The effects of Social Networking Site (SNS) use on college students' friendship and well-being. *Computers in Human Behavior* 37 (Aug. 2014), 229–236. https://doi.org/10.1016/j.chb.2014.04.051

[44] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy nudges for social media: An exploratory facebook study. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web* 01 (2013), 763–770.

[45] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania) *(SOUPS '11)*. Association for Computing Machinery, New York, NY, USA, Article 10, 16 pages. https://doi.org/10.1145/2078827.2078841

[46] L. Y.C. Wong and Jacquelyn Burkell. 2017. Motivations for Sharing News on Social Media. In *Proceedings of the 8th International Conference on Social Media & Society (#SMSociety17)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3097286.3097343

[47] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013)* (May 14-17). Society for the Advancement of the Science of Digital Games, Chania, Crete, Greece, 39–46. http://www.fdg2013.org/program/papers.html

A INDUCTIVE CODEBOOK INCLUDING DESCRIPTIONS

	Code	Description
1	Addictive Design	Features or elements that keep users hooked to content.
2	Auto Accept Third Party Terms	Unknowingly giving consent to share data with third parties per default settings.
3	Autoplay Content	Auto-playing content without further actions by the user.
4	Clinging To Accounts	Making the process of deleting an account unnecessarily difficult or reactivating accounts after deletion has already been initialised.
5	Decision Governing	Interface instances that navigate or steer users' decision-making.
6	Decision Uncertainty	Users do not know what they are left in confusion and what consequences their decision will have.
7	External Solution Search	Not able to find specific features or settings, users fall back to use search engines to find what they are looking for.
8	False Content Customisation	Shown content does not fit the users' preferences or followed accounts.
9	Fear Of Missing Out	Feeling pressured to (re)visit specific SNS features out of fear to miss something.
10	Forced Access Granting	Features requiring unnecessary access to special device hardware or local data.
11	Forced Dialogue Interaction	Prompting Text Boxes that require immediate attention with no option to dismiss them without interaction.
12	Forced Grace Period	When deleting content or accounts, users are forced to wait a minimum amount of days before changes become active (often 30 days).
13	Gamification	Playful elements that motivate users to do something by presenting artificial progress and/or rewards to get more data.
14	Hidden In Plain Sight	Crucial information is often obscured by attention-grabbing interface elements.
15	Infinite Scrolling	Users can infinitely scroll through content (often, more and more suggested content is shown the more they progress.
16	Labyrinth	Nested interface structures users get easily lost in.
17	Persuasive Language	Emotional pressuring language to push towards a specific direction that may not be in the users best interest.
18	Plain Evil	Interface elements suddenly change functionalities or are being exchanged for alternative ones.
19	Pull To Refresh	Pulling downward on a content-displaying interfaces will load new content. Sometimes, suggested content is shuffled into the feed to give users more to look at.
20	Reduced Friction	Purposefully making certain elements easier accessible and thus pushing alternatives into the back.
21	Regression Toward The Mean	Users are suggested topics which most people like and thus are likely to add to the already existing "popular main topics".
22	Social Connector	Requesting additional information to connect to friends/family and other social circles.

Table 3: Table containing all 22 codes from the inductive codebook from study 1, later used to create five themes deriving into the SNS-specific dark patterns.

PUBLICATION P4

Defending Against the Dark Arts: Recognising Dark Patterns in Social Media

Authors:
Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, & Rainer Malaka

The publication contributes to the following angles:

USER GUIDELINE

This publication builds on previous findings of P3 to study Social Networking Service (SNS) users' ability to recognise dark patterns across two studies. The first considers Human-Computer Interaction (HCI) experts to conduct cognitive walkthroughs across Facebook, Instagram, TokTok, and Twitter. The second online survey recruits 193 active users of SNS. Utilising the five dark pattern characteristics proposed by Mathur *et al.* (2019), this publication expands the provocation in P2 to formulise a procedure to assess dark patterns within interfaces.

Its contribution to the thesis is twofold. By considering SNS users' ability to recognise dark patterns, the publication contributes to the design angle, showing difficulties among users to effectively protect themselves. Building on these findings further, it proposes a process to assess dark patterns in interfaces, contributing to the guideline angle.

My contribution to this paper was the study design, data collection, and analysis of the results. Further, I interpreted the data to develop a process for assessing dark patterns in user interfaces. I drafted the manuscript and revised it before the final publication.

The contents of this chapter originally appeared in: Mildner, T., Freye, M., Savino, G.-L., Doyle, P. R., Cowan, B. R., and Malaka, R., "Defending Against the Dark Arts: Recognising Dark Patterns in Social Media," in *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA*, 2023. DOI: 1010.1145/3563657.3595964

Defending Against the Dark Arts: Recognising Dark Patterns in Social Media

Thomas Mildner
University of Bremen
Bremen, Germany
mildner@uni-bremen.de

Merle Freye
University of Bremen
Bremen, Germany
mfreye@uni-bremen.de

Gian-Luca Savino
University of St.Gallen
St.Gallen, Switzerland
gian-luca.savino@unisg.ch

Philip R. Doyle
University College Dublin
Dublin, Ireland
philip.doyle1@ucdconnect.ie

Benjamin R. Cowan
University College Dublin
Dublin, Ireland
benjamin.cowan@ucd.ie

Rainer Malaka
University of Bremen
Bremen, Germany
malaka@tzi.de

ABSTRACT

Interest in unethical user interfaces has grown in HCI over recent years, with researchers identifying malicious design strategies referred to as "dark patterns". While such strategies have been described in numerous domains, we lack a thorough understanding of how they operate in social networking services (SNSs). Pivoting towards regulations against such practices, we address this gap by offering novel insights into the types of dark patterns deployed in SNSs and people's ability to recognise them across four widely used mobile SNS applications. Following a cognitive walkthrough, experts ($N = 6$) could identify instances of dark patterns in all four SNSs, including co-occurrences. Based on the results, we designed a novel rating procedure for evaluating the malice of interfaces. Our evaluation shows that regular users ($N = 193$) could differentiate between interfaces featuring dark patterns and those without. Such rating procedures could support policymakers' current moves to regulate deceptive and manipulative designs in online interfaces.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; HCI theory, concepts and models; *Empirical studies in interaction design*; Interaction design theory, concepts and paradigms; • **Security and privacy** → *Usability in security and privacy*.

KEYWORDS

SNS, social media, social networking services, interface design, dark patterns, well-being, ethical interfaces

ACM Reference Format:
Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3563657.3595964

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9893-0/23/07.
https://doi.org/10.1145/3563657.3595964

1 INTRODUCTION

Among HCI researchers, interest in the ethical implications of how technology is designed has seen a noticeable increase over recent years. One of the more widely known topics within this work is research that focuses on unethical design strategies, referred to as "dark patterns". Cataloguing instances of dark patterns has led to a growing collection of interface artefacts that negatively affect users' ability to make informed decisions. A common example can be seen in cookie-consent banners that often visually elevate options allowing the tracking and storing of users' data over alternatives to denying such functionalities. Originating in e-commerce [7, 33], and other online websites [7, 19], dark patterns describe design strategies that coerce, steer, or obfuscate users into unfavourable actions that they may not have taken if they were fully informed [34]. Today, related work has identified a multitude of designs that fit this definition, including digital games [45], social networking sites (SNS) [23, 24, 35, 38], and mobile applications [4, 12, 19].

The adverse effects of dark patterns have drawn the attention of regulators worldwide. Examples aimed at better protecting users' privacy and autonomy can be seen in the California Consumer Privacy Act CCPA [29] or the planned Digital Service Act (DSA) of the European Union [9]. Regardless of the national background, regulating dark patterns faces common challenges, such as a missing taxonomy, the rapid development of new dark patterns, and difficulty identifying dark patterns that require legal interventions. We see that findings from human-computer interaction (HCI) can support the legal discussion and legislative efforts [20] in developing a taxonomy and providing the right tools to assess and regulate dark patterns. Therefore, it is crucial that research advances our understanding of the implications of dark patterns in as many domains as possible to enable regulators and legislators to create effective measures to protect users.

In this work, we take steps towards achieving this goal by (1) analysing the ability of experts and regular users of social media to identify dark patterns based on established definitions thereof and by (2) studying an alternative approach to classify interfaces based on high-level characteristics proposed by Mathur et al. [33, 34] to approach an easier evaluation. As this is a relatively new research area, knowledge about how people perceive dark patterns is still limited, with a handful of studies exploring this particular aspect of the topic [4, 12, 32]. In light of initial moves towards regulation and increased attention in the scientific literature, this work reflects

on the current state of the dark pattern research, investigates how applicable current taxonomies are in domains in which they were not first established, and whether current definitions can be utilised as evaluation tools. Before conducting this research, we collected 69 types of dark patterns from eight papers [5, 7, 11, 17, 18, 22, 33, 45], further included in Mathur et al.'s [34] literature review. While we are aware that recent work have updated the overall corpus of dark patterns [23, 36], which we could not include in our studies, the focus of this research is to aim for a simplified recognition tool to aid policy-makers' and regulators' efforts. For this endeavor, we turn towards SNSs as we still lack certain insights about how malicious interfaces manifest in this context. Additionally, the omnipresent nature of SNSs affords constant investigation as research repeatedly highlights negative effects posed on their users' well-being [2, 3]. Aiming to aid regulatory efforts, we address these research gaps based on two research questions:

RQ1 Can dark patterns taxonomies be used by experts to identify and recognise instances in SNSs?
RQ2 Are regular SNS users able to differentiate between interfaces with and without dark patterns?

We answer these questions through two studies. In the first, we conducted cognitive walkthroughs with six HCI researchers aimed at investigating whether current dark pattern taxonomies can be used to assess and identify dark patterns in novel interfaces. The four SNSs included in the study were Facebook, Instagram, TikTok, and Twitter. In a second study, we conducted an online survey to learn about the recognisability of dark patterns by regular SNS users. In contrast to the first study, we did not provide participants of the second study with the complete corpus of dark pattern research but instead relied on five questions adopting Mathur et al.'s [33, 34] high-level dark pattern characteristics with the aim of assessing the malice of a particular interface design. While this hinders an immediate comparison between both studies, our evaluation of this alternative process shows that regular users are able to generally recognise dark patterns. Conclusively, dark patterns were not rated to be very malicious (using Mathur et al.'s [33, 34] five high-level characteristics) but participants were able to successfully discern dark patterns from a selection of interface screenshots collected from Study 1, that either did or did not contain them. We also propose that a similar approach, one that is not fundamentally linked to specific examples of dark pattern design, could introduce more flexibility and practicality into current legislation processes and would better future-proof legislative efforts aiding the protection of users.

2 RELATED WORK

In this section, we will approach relevant research to identify, recognise, and regulate dark patterns from two directions. We will begin by establishing a taxonomy of dark pattern types resulting from the collaborative effort of prior research. This taxonomy is later used in our first study. Afterwards, we highlight work studying the perception and recognition of dark patterns, a necessary step towards successful regulation. We then outline the form of current approaches and strategies in the final paragraphs of this section.

2.1 Dark Pattern Taxonomy

Here, we attempt to provide a relatively comprehensive overview of the current dark patterns landscape. To provide a summary of the taxonomy used in our studies, Table 1 presents key contributions taken from Mathur et al.'s [34] earlier review on dark pattern literature. As we deem it important for our studies that the definitions for dark patterns should be the result of empirical research, we decided to limit the scope for the eight academic contributions part of Mathur et al.'s literature review [34]. Although more holistic guidelines exist, these are not included as they tend not to provide enough empirical evidence in their definitions. This left eight academic works that met our criteria, which collectively presented 69 different types of dark patterns that are outlined below in chronological order. Brignull [7], who first coined the term dark pattern, initialised the current body of work with twelve types that concern online design strategies. In a similar effort, Conti and Sobiesk [11] defined eleven types of malicious strategies based on a one-year data collection. Although their work was published before the term dark pattern gained the recognition it sees today, we refer to their results as dark patterns for the sake of conciseness. Offering seven game-specific dark patterns, Zagal et al. [45] studied tricks used in that industry to create, for example, competition or disparate treatment through unethical practices. In another work, Greenberg et al. [22] were interested in the possible exploitation of spatial factors when discussing dark patterns through the lens of proxemic theory. The result introduces eight types of proxemic dark patterns like speculative technologies targeting users with specific advertisements using public displays. Closely related to the Privacy by Design concept [25], and thus particularly interesting for our research, Bösch et al. [5] collected eight types of dark patterns enveloping schemes that target data collection and limitations of users' agency to customise their personal preferences.

Taking a different approach, Gray et al. [19] looked to investigate how dark patterns are created in the first place. Here, researchers analysed an image-based corpus of potential types of dark patterns using a qualitative approach while relying on Brignull's original taxonomy. They define five types of dark patterns that practitioners engage in when developing manipulative designs. Following this research, Gray et al. [17] applied content analysis on 4775 user-generated posts collected from the Reddit sub-forum *r/assholedesign*. Their result provides six properties "asshole designers" subscribe to. Interested in the number of web services embedding dark patterns, Mathur et al. [33] applied hierarchical clustering to identify that 11% of shopping websites employ text-based dark patterns based on a collection of more than 11k samples. Evaluation of their data generated twelve dark patterns embedded in shopping websites.

These works bring together 69 types of dark patterns. Noticeably, various domains have been investigated, widening our understanding of these strategies' origins. However, there is currently a potentially important gap regarding SNS-related platforms like Facebook, Instagram, TikTok, and Twitter – platforms that many people interact with frequently in their day-to-day lives. A growing body of research already illustrates problems with users accurately recollecting the amount of time they spend on SNSs and the frequency in which they use these services [13, 27, 39]. Concerns are also growing regarding alarming implications SNSs have on their

Brignull 2010 [7]	**Conti & Sobiesk** 2010 [11]	**Zagal et al.** 2013 [45]	**Greenberg et al.** 2014 [22]	**Bösch et al.** 2016 [5]	**Gray et al.** 2018 [19]	**Gray et al.** 2020 [18]	**Mathur et al.** 2019 [33]
· *Trick Questions*	· *Coercion*	· *Grinding*	· *Attention Grabber*	· *Privacy Zuckering*	· *Nagging*	· *Automating the User*	· *Countdown Timers*
· *Sneak Into Basket*	· *Distraction*	· *Impersonation*	· *Bait and Switch*	· *Hidden Legalese* Stipulations	· *Obstruction*	· *Two-Faced*	· *Limited-time* Messages
· *Roach Motel*	· *Forced Work*	· *Monetized Rivalries*	· *The Social Network* Of Proxemic Contracts Or Unintended Relationships	· *Shadow User Profiles*	· *Sneaking*	· *Controlling*	· *High-demand* Messages
· *Privacy Zuckering*	· *Manipulating Navigation*	· *Pay to Skip*	· *Captive Audience*	· *Bad Defaults*	· *Interface Interference*	· *Entrapping*	· *Activity Notifications*
· *Confirmshaming*	· *Restricting* Functionality	· *Playing by* Appointment	· *We Never Forget*	· *Immortal Accounts*	· *Forced Action*	· *Nickling-And-Diming*	· *Confirmshaming*
· *Disguised Ads*	· *Trick*	· *Pre-Delivered* Content	· *Disguised Data* Collection	· *Information Milking*		· *Misrepresenting*	· *Testimonials* of Uncertain Origins
· *Price Comparison Prevention*	· *Confusion*	· *Social Pyramid* Schemes	· *Making Personal* Information Public	· *Forced Registration*			· *Hard to Cancel*
· *Misdirection*	· *Exploiting Errors*		· *The Milk Factor*	· *Address Book* Leeching			· *Visual Interference*
· *Hidden Costs*	· *Interruption*						· *Low-stock Messages*
· *Bait and Switch*	· *Obfuscation*						· *Hidden Subscriptions*
· *Forced Continuity*	· *Shock*						· *Pressured Selling*
· *Friend Spam*							· *Forced Enrollment*

Table 1: This table shows 69 types of dark patterns described in eight related works. Columns are in chronological order in which these works were published.

users' well-being [3, 40, 43, 44]. Filling this gap, the research presented here considers the current discourse to review the presence of these described dark patterns in four major SNS platforms.

2.2 Perceiving Dark Patterns

Interested in the cognitive biases dark patterns exploit, Mathur et al. [33] analysed their dark patterns further and recognised five common characteristics in which these dark patterns operate: *asymmetric*; *restrictive*; *covert*; *deceptive*; and *information hiding*. In a follow-up effort, Mathur et al [34] applied these characteristics to prior dark pattern taxonomies while extending the framework to include a sixth characteristic named *disparate treatment*. Collectively, this framework promises an alternative and interesting tool to study dark patterns. To test its utility outside its original scope, our research applies this framework to recognise dark patterns in SNSs. Instead of focusing entirely on the identification of dark patterns, a multitude of works considers end-users' perspectives of dark patterns. In this sense, Di Geronimo et al. [12] sampled 240 popular applications from the Google Playstore and analysed each for contained dark patterns based on Gray et al.'s [19] taxonomy. Based on 10-minute cognitive walkthroughs, their results indicate that 95% of tested applications yield dark patterns. An ensuing online survey revealed that the majority of users fail to discern Dark Patterns in 30-second video recordings of mobile applications. However, their ability to identify harmful designs improves when educated on the subject. In line with prior research, including Maier and Harr's [32] confirmation of users' difficulty to recognise dark patterns [32], Bongard-Blanchy et al. [4] reinforce these implications through their online survey studying participants' ability to recognise dark patterns. Studying the effects browser modalities have on the number of dark patterns users are faced with, Gunawan et al [23] conducted a thematic analysis on recordings of various online services. Their work describes twelve previously not described dark patterns, including *extraneous badges* that describe nudging interface elements, like coloured circles, which provoke immediate interaction. Trying to understand Facebook users' control over ad-related settings, Habib et al. [24] demonstrate that the SNS does not meet users' preferred requirements. Considering dark patterns in their work, the authors discuss problematic interface structures limiting users' agency to choose settings efficiently and to their liking. This limitation is further discussed by Schaffner et al. [38], who demonstrate difficulties for users to successfully delete their accounts across 20 SNSs. Their success rate was additionally impacted by the modality in which a particular SNS is accessed.

Investigating persuasive designs, Utz et al. [42] demonstrate how nudging interfaces can shift users' decisions towards a preset goal. In a similar vein, Graßl et al. [21] showed evidence that nudges prevent informed decisions. In their experiments, users were either faced with banners visually promoting a privacy-diminishing option or a reverted interface where the option protecting users' privacy was promoted instead. Related efforts of this community highlight current shortcomings of the GDPR [8] to achieve its goals. Reviewing compliance of consent management platforms, Nouwens et al. [37] show that only 11.6% of websites from a corpus of 10k met the minimum requirements of European law. Reviewing the GDPR for its objectives to give users control over their data, Boyens et al. [6] find that users experience serious problems, leading to decreasing trust in institutions that should protect them.

These works collectively show that the responsibility to avoid dark patterns can and should not solely fall onto users. Additional protection needs to come from other sources, such as the better implementation of regulations, while research needs to foster our understanding of dark patterns' origins as well as exploited strategies. We contribute to the latter by turning towards SNSs. Unlike prior work, our study utilises Mathur et al.'s dark pattern characteristics as a framework to learn about users' ability to recognise dark patterns in this domain.

2.3 Regulating Dark Patterns

The advantages of interdisciplinary efforts between HCI and legal scholars have recently been shown in Gray et al.'s [20] work studying consent banners from multiple perspectives. The negative effects of dark patterns in online contexts are not a new phenomenon in law. Protecting users and consumers from manipulation, unfair practices, and imbalances has always been a subject of legislation.

Different laws can affect single design patterns, including data protection law, consumer law, and competition law, depending on their impact on consumers, traders, and personal data [28, 30]. Recently, attempts to regulate dark patterns as a whole have arisen. Especially the European Union started to draft legislation that specifically targets dark patterns. The Commission's proposal for a Digital Service Act [9] (DSA) and the Commission's proposal for the Data Act [10] explicitly provide a definition for dark patterns in their recitals.

A key challenge is to legislate patterns that are rapidly evolving while adopting new strategies to pass regulation, yet maintaining their malice. In the context of SNSs, our study draws attention to tools of HCI that could support legal decisions. Picking up on these works, legislators and regulators could utilise the existing knowledge about dark patterns to extend current approaches to protecting peoples' privacy on further problematic designs that potentially harm their well-being. In the presented work, we explore a novel approach to evaluate the malice of interfaces of four SNSs based on high-level characteristics proposed by Mathur et al. [34].

3 STUDY 1: COGNITIVE WALKTHROUGH

The purpose of this study is to see whether definitions of dark patterns can be used to recognise similar design strategies in domains other than the ones they were initially identified in. We, therefore, considered four SNSs (Facebook, Instagram, TikTok, and Twitter) where we had six HCI researchers review mobile applications in the form of cognitive walkthroughs [26]. Each researcher was asked to complete ten tasks designed for identifying and recording any instances of dark patterns on the SNSs' mobile applications. The decision to investigate exactly these four SNSs is based on their overall popularity [41], comparable features, and similar user bases. As the experiment was conducted during the COVID-19 pandemic, participants completed their walkthroughs without supervision. Study 1 aims to answer the following research question: Can dark patterns taxonomies be used by experts to identify and recognise instances in SNSs?

3.1 Reviewers

For this experiment, we recruited reviewers who have strong expertise in HCI and UX research and design. In a similar fashion to regulators who have to decide whether a problematic interface requires legal action or not, our participants needed to meet the necessary qualifications to identify dark patterns. Their knowledge of best practices in interface design and user experience makes them more susceptible to recognising potential issues compared to users without access to this particular expertise, as shown in prior research [4, 12]. Recruitment involved reaching out to researchers with backgrounds in cognitive science, computer science, and media science who also specialised in HCI research. Participation was on a voluntary basis. In total, we selected six participants (3 female, 3 male) from the authors' professional network. The average age of the panel was 28.33 years (SD = 1.63), with an average experience in HCI research of 3.83 years (SD = 1.47). All participants worked in academia in HCI-related research labs. Five are of German nationality, while one reviewer is Russian. While all participants had experience in interface design, except for one, none had prior knowledge of dark pattern academic research. Before conducting the study, each participant was provided with the necessary information on the topic before we obtained their consent. To protect them from the unethical consequences of dark patterns, we provided each participant with devices, new accounts for the SNSs, and data to be used during the study. This is further elaborated in subsection 3.2 Preparation.

3.2 Preparation

After receiving their consent for participating in this study, each reviewer received two smartphone devices, a factory reset iPhone X (iOS 14.5) and a Google Pixel 2 (Android 11), with the social media applications already installed to ensure the same version[1] was used by each participant. Both iOS and Android devices were used to distinguish between problematic interface designs caused by the applications and those linked to the operating systems. Also, each participant was provided with a new email account and phone number so they could create new user profiles for their assigned platforms. This was done to respect participants' privacy and to avoid customisation of accounts from previous usages that may impact participants' experience and, subsequently, their findings. Lastly, we stored some amount of media content on each device as part of the cognitive walkthrough, affording the participants to create and post content. Again, this ensured that participants did not have to share any personal information with the SNS.

3.3 Procedure

One key element of this study is an extracted dark pattern taxonomy based on Mathur et al.'s [34] work, including a review of the dark pattern landscape. The taxonomy, featuring 69 distinct types (see Table 1), was given to each reviewer after a one-hour-long introduction to the topic, followed by another hour to resolve unanswered questions mitigating inconsistencies in reviewers' expertise. Despite reviewers' backgrounds in HCI-related fields, this introductory session ensured a common understanding of current conceptualisations of dark patterns. After the introduction, each reviewer was handed informational material containing the presented information and the definitions of the 69 dark pattern types. This material is provided in the supplementary material of this paper. To maintain further consistency throughout the study, we created ten tasks reviewers were asked to complete during the cognitive walkthroughs [26]. Five of these tasks were adapted from research conducted by Di Geronimo et al. [12] that evaluated popular applications on the Google Play Store. Inspired by elements of their methodology, we increased the amount of time each SNS should be investigated to approximately 30 minutes based on a pre-study. This decision allows us to understand the interfaces of the four SNSs on a deeper level. Lastly, each reviewer was assigned two of the four SNSs ensuring that each application was reviewed three times by independent people on both iOS and Android operating systems. After a reviewer completed their walkthrough, we saved the stored recording data from the devices before setting them up

[1]Installed versions consistent throughout Study 1: Facebook (iOS: 321.0.0.53.119; Android: 321.0.0.37.119); Instagram: (iOS: 191.0.0.25.122; Android: 191.1.0.41.124); TikTok (iOS:19.3.0; Android: 19.3.4); Twitter (iOS: 8.69.2; Android: 8.95.0-release.00).

for the next session. Below are the ten tasks each reviewer performed. Tasks taken from or worded closely to Di Geronimo et al. [12] are highlighted by an asterisk. Items 1, 9, and 10 were added to improve the task flow, whilst items 4 and 5 were developed to address typical SNS activities such as creating and sharing personal content and networking.

1. *Turn on screen recording on each device.*
*2. *Open the app and create an account to log in and then out.*
*3. *Close and reopen the app.*
4. *Create any kind of content, post it, and delete it.*
5. *Follow and unfollow other accounts.*
*6. *Visit the personal settings.*
*7. *Visit the ad-related settings.*
*8. *Use the application for its intended use (minimum of five minutes):*
   I *Describe the natural flow of the app – what did you use it for?*
   II *Could you use the app as you wanted or did some features 'guide' your interactions?*
   III *how easy was it to get distracted and if so what distracted you?*
9. *Delete your account.*
10. *Turn off screen recording and save the recording.*

4 RESULTS OF STUDY 1

In this study, we considered a dark pattern taxonomy comprising 69 individual types of dark patterns (see Table 1) across mobile applications for the SNSs Facebook, Instagram, TikTok, and Twitter. Offering an answer to our first research question, the six participants identified a total of 548 dark pattern distinct instances from the considered 69 types that can be associated with descriptions contained within the taxonomy provided. Participants found $N_F = 232$ dark pattern instances in Facebook, $N_I = 96$ in Instagram, $N_{Ti} = 95$ in Twitter, and $N_{Tw} = 125$ in Twitter. Figure 1 presents four screenshots that demonstrate examples of dark patterns identified by participants across each of the four SNSs. Close inspection shows multiple types of dark patterns at play in each image. Although the four SNSs were selected based on similar functionalities and user bases, we do not compare results across platforms. Despite their similarities, each SNS contains unique features that distinguishes them from the others. Also, the number of functionalities between the SNSs varies considerably, with Facebook containing many more options for users to engage with than alternatives. Instead, we report descriptive statistics that will then be further elaborated on in the discussion section of this paper.

4.1 Recognised Types of Dark Patterns

Of the 69 types of dark patterns contained in the taxonomy participants were provided with at the beginning of this study, 31 distinct types were identified, leaving the remaining 55.07% unrecognised across any of the four SNSs. All recognised dark patterns can be seen in Figure 2. For brevity, only key illustrative instances are reported here, while the full analysis will be included in the supplementary material. Across the four SNSs, two dark pattern types stood out the most: With a total of 58 recognised instances, Gray et al.'s *Interface Interference* [19] (i.e. interfaces that privilege certain elements over others confusing users to make a particular choice) was most readily identified by participants, whilst Mathur et al.'s *Visual Interference* [33] (i.e. interfaces that deploy visual/graphical tricks to influence users' choices) was next most widely observed with 51 instances. The third most frequently identified dark pattern was Gray et al.'s *Obstruction* [19] dark pattern (interfaces that make certain actions unnecessarily difficult to demotivate users) recognised 47 times. Bösch et al.'s *Bad Defaults* [5] (privacy settings are pre-set to share users' personal information by default) came fourth with 44 instances, closely followed by 40 counts of Brignull's *Privacy Zuckering* [7] (tricks to deceive users into sharing more personal information than intended) dark pattern.

4.2 Types of Dark Patterns That Have Not Been Recognised

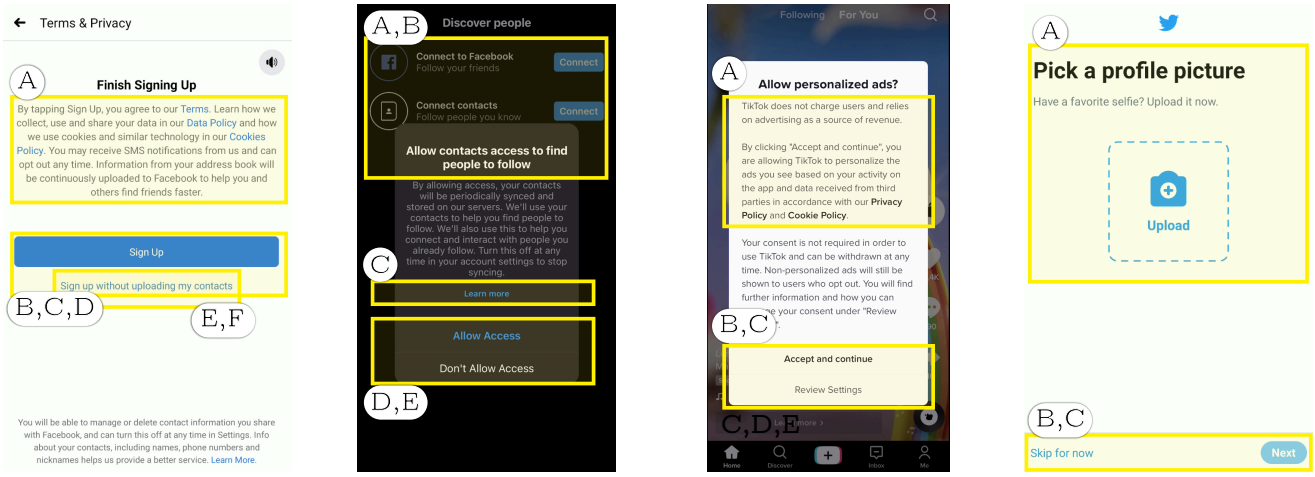
While 44.93% of dark pattern types were recognised during the cognitive walkthrough, the other 55.07% were not. Almost all dark pattern taxonomies contained some dark patterns that were recognised. However, the taxonomy by Zagal et al. [45], being video-game focused, did not contribute any specific dark patterns that were recognised. This result shows that not all dark pattern types are relevant for each domain. By adding new dark pattern types to the overall collection for each domain, regulators have increasingly more items to consider complicating their endeavour if they are to use them as guides.

4.3 Dark Patterns Co-Occurrences

To learn more about how dark patterns interact with each other, we also analysed them for co-occurrences. We used the software ATLAS.ti [16] to calculate the co-occurrence coefficient between any two dark patterns, which is based on the Jaccard similarity coefficient [15] returning a c-coefficient c. Interestingly, the data revealed that although two patterns are described differently, their working can be rather similar in the context of SNSs. Intersections between *Interface Interference* ∩ *Visual Interference* ($c = 0.85$, $N = 50$ co-occurrences), *Forced Action* ∩ *Forced Work* ($c = 0.89$, $N = 25$ co-occurrences), and *Roach Motel* ∩ *Hard to Cancel* ($c = 0.71$, $N = 17$ co-occurrences), for instance, follow this example. However, like the intersection between *Misrepresenting* ∩ *Immortal Accounts* ($c = 0.55$, $N = 12$ co-occurrences) or *Privacy Zuckering* ∩ *Bad Defaults* ($c = 0.35$, $N = 22$ co-occurrences), most co-occurrences are indications for interfaces yielding multiple distinct dark patterns simultaneously. Due to the overall co-occurrence data set is too large to be fully represented here, it has been included in the supplementary material.

5 STUDY 2: ONLINE SURVEY

Findings from Study 1 suggest existing taxonomies feature numerous types of dark patterns that are not applicable to SNSs and that some dark patterns employed by SNSs are not incorporated in earlier taxonomies. In this second study, we adopted a different approach to identifying dark patterns in interfaces. Instead of relying on fixed descriptions and definitions of existing dark patterns, we developed a questionnaire consisting of five questions based on dark pattern characteristics previously highlighted by Mathur et al. [34]. These higher-level characteristics go beyond dark pattern



(a) Screenshot from Facebook (b) Screenshot from Instagram (c) Screenshot from TikTok (d) Screenshot from Twitter

Figure 1: Example screenshots from Study 1. Figure 1a contains the dark patterns *Hidden-Legalese Stipulations (A), Misdirection (B), Interface Interference (C), Visual Interference (D), Privacy Zuckering (E),* and *Address Book Leeching (F).* Figure 1b contains the dark patterns *Privacy Zuckering (A), Address Book Leeching (B), Hidden-Legalese Stipulation (C), Interface Interference (D), and Visual Interference (E).* Figure 1c contains the dark patterns *Hidden-Legalese Stipulation (A), Interface Interference (B)* and *Visual Interference (C).* Figure 1d *Privacy Zuckering (A), Interface Interference (B), and Visual Interference (C).*

definitions by descriptively organising dark patterns from existing literature [34]. Following this approach, study 2 aims to address the following research question: Are regular SNS users able to differentiate between interfaces with and without dark patterns?

5.1 Screenshots

We used sixteen screenshots along with the aforementioned questionnaire to evaluate people's ability to recognise dark patterns within screenshots of the four SNSs. While eight of the sixteen screenshots contained dark patterns, the other eight did not and served as control. All screenshots were sampled from the previous study (see Figure 3 for four example images). Regarding those that contained dark patterns, two conditions had to be met: Screenshots had to (1) represent all five characteristics by Mathur et al. while (2) contained dark patterns had to be identified by at least two expert reviewers. Furthermore, we avoided using screenshots that contained dark patterns that only emerge through procedural interactions taken by users (e.g. *Roach Motel*). Consequently, two authors of this paper ensured to pick screenshots where the dark patterns were recognisable on a static image, for example by deploying visual/aesthetic (e.g.Visual Interference) or linguistic (e.g. Confirmshaming) manipulations.Screenshots that did not contain dark patterns were carefully selected by sampling situations where expert reviewers did not recognise any dark pattern. This was additionally validated by two authors of this paper to ensure no dark pattern had been accidentally overlooked. Using these screenshots, we test whether participants can generally recognise dark patterns and whether they can differentiate between screenshots with and without dark patterns.

5.2 Methodology

To investigate our research question, we conducted an online survey. The survey was divided into three parts: (1) screening for participants' SNS usage behaviour, (2) a dark pattern recognition task, and (3) a demographic questionnaire. In total, the survey featured 25 question items (included in supplementary material) and took on average 12:22 minutes (SD = 9:45) to complete. As we were interested if regular social media users could asses dark patterns in SNS, only participants who indicated previous and regular use of social media platforms were included in the sample. This was achieved using screening questions about previous social media usage. Before evaluating the sixteen screenshots, participants were provided with the following definition of dark patterns by Mathur et al.'s [34]: *"user interface design choices that benefit an online service by coercing, steering, or deceiving users into making decisions that, if fully informed and capable of selecting alternatives, they might not make"*. For each of the sixteen screenshots, participants had to first answer if they thought dark patterns were present in the screenshot based on the definition of dark patterns by Mathur et al.'s [34] with 'Yes', 'No' or 'Maybe'. In the next step, participants then had to answer if they saw dark patterns in the screenshot based on Mathur's dark pattern characteristics [34]. For this, we developed five questions adopting the characteristics [34], which participants rated based on a unipolar 5-point Likert-scale (see Table 2). Available responses ranged from "Not at all" to "Extremely". After assessing all five characteristics, they moved on to the next screenshot. Screenshots were delivered in a randomised order between participants. Once all screenshots were assessed, the survey concluded by collecting basic demographic data from each respondent, including age, gender, current country of residency, and an optional field to give feedback.

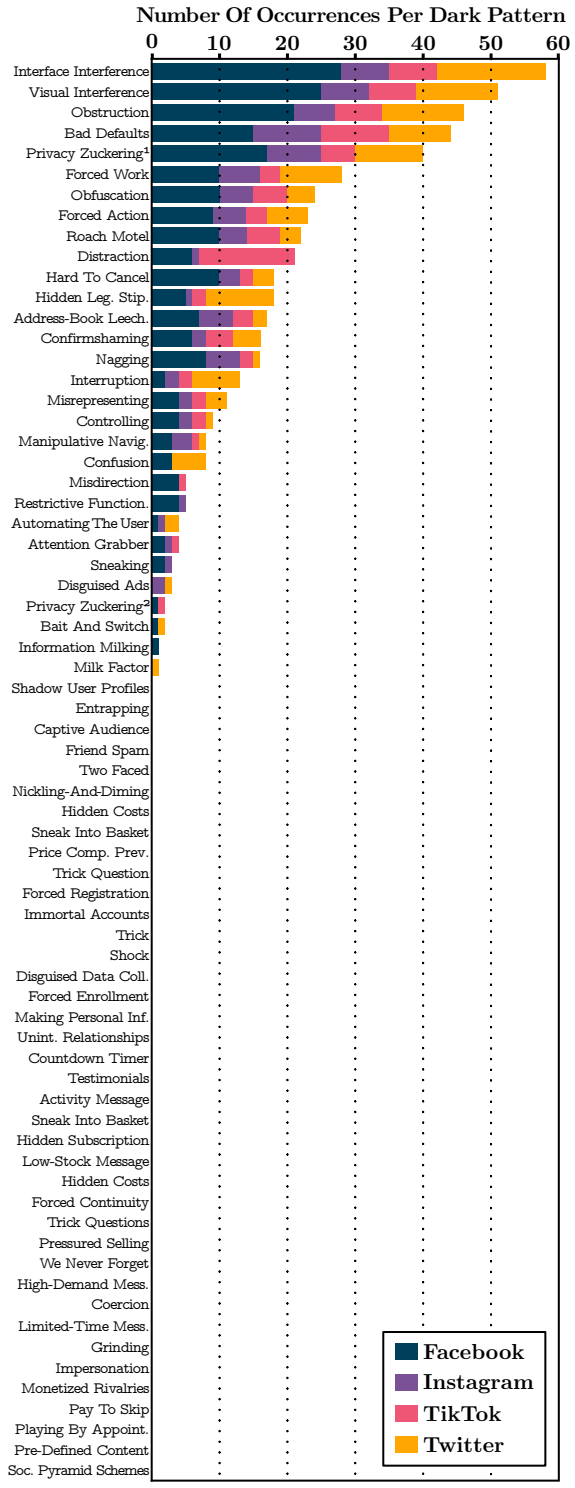


Figure 2: Summary of the occurrences of all 69 considered dark pattern types in four SNSs.Of the 69 types 31 were recognised. *Privacy Zuckering*[1] refers to Brignull's [7] description while *Privacy Zuckering*[2] refers to Bösch et al.'s defintion [5].

Mathur 2019 [33] Dark Pattern Characteristics	
Characteristic	Question
Asymmetric	Does the user interface design impose unequal weights or burdens on the available choices presented to the user in the interface?
Covert	Is the effect of the user interface design choice hidden from the user?
Deceptive	Does the user interface design induce false beliefs either through affirmative misstatements, misleading statements, or omissions?
Hides Information	Does the user interface obscure or delay the presentation of necessary information to the user?
Restrictive	Does the user interface restrict the set of choices available to users?

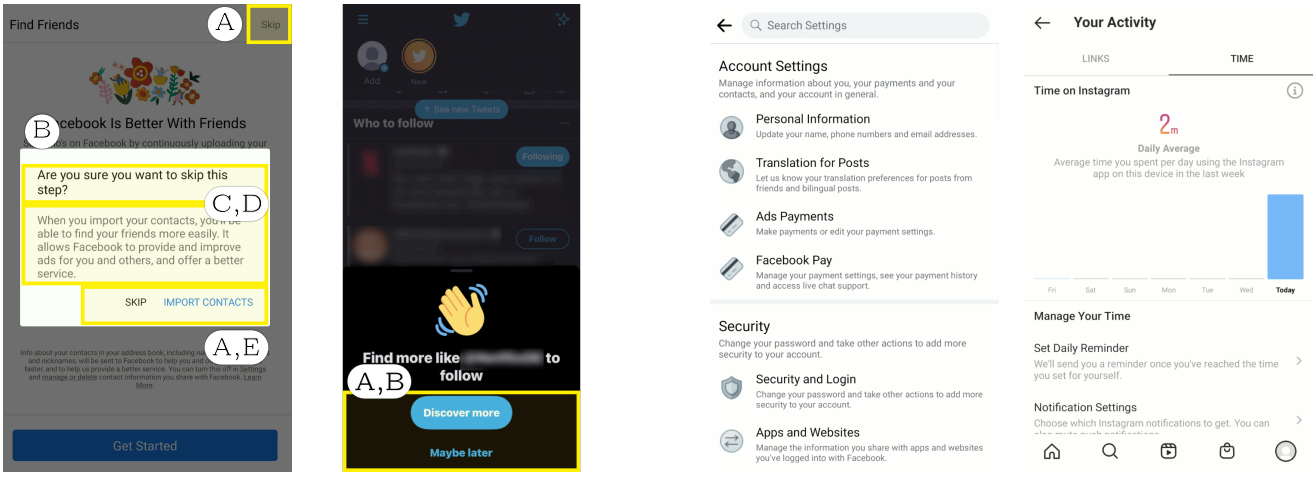
Table 2: This table lists the introductory questions Mathur et al. (2019) [33] gave for each dark pattern characteristic.

5.3 Participants

To calculate an appropriate sample size needed to answer our research questions, we conducted an *a priori* power analysis using the software G*Power [14]. Given our study design, to achieve a power of 0.8 and a medium effect size, the analysis suggested a total sample size of 166. Participants of this survey were recruited from two sources: (1) The Reddit forum *r/samplesize* [1] and (2) *Prolific* [31]. For redundancy, we invited 90 people, more than our power analysis suggested. After receiving their consent to participate in this study, 256 participants were recruited and completed the online survey. Of these 256 participants, 26 were recruited via Reddit [1] and 230 via Prolific [31]. Initially, we recruited participants from Reddit to assess the feasibility of our study design. After this was ensured and we successfully verified that the retrieved data was equal in quality to the data gained from Prolific, both sets were accumulated. Compensation for participating in this study was rewarded with £7.2 per hour, with individual compensation dependent on participants' time needed to complete the study (mean = 12.2 minutes, SD = 8.76 minutes). We excluded 63 data sets in total due to: failure to complete the questionnaire; failed attention checks (questions with a single true answer to measure participants' engagement); not meeting inclusion criteria; completing the questionnaire in unrealistic times based on *a priori* testing; and if they replied with the same option over 95% of instances. Eventually, data from a total of 193 participants were included in the analysis, thus satisfying the estimate of the power analysis.

6 RESULTS OF STUDY 2

In this section, we present the results of the online survey. The results are split into three parts: (1) demographic data on our participants, (2) results on whether participants can recognise dark



(a) Screenshot With Dark Patterns - Facebook

(b) Screenshot With Dark Patterns - Twitter

(c) Screenshot Without Dark Patterns - Facebook

(d) Screenshot Without Dark Patterns - Instagram

Figure 3: Four example screenshots used in study 2, sampled from study 1. Figure 3a contains the dark patterns *Interface Interference (A), Confirmshaming (B), Address-Book Leeching (C), Privacy Zuckering (D),* and *Visual Interference (E).* Figure 3b contains the dark patterns *Interface Interference (A),* and *Visual Interference (B).* Importantly, Figure 3a and Figure 3b were presented to participants without annotations. Neither Figure 3c nor Figure 3d contain any dark patterns. In total, sixteen screenshots were used in study 2 - eight containing dark patterns and eight that do not.

patterns based on the definition of dark patterns by Mathur et al. [34], and (3) whether they can differentiate between screenshots with and without dark patterns based on Mathur's dark pattern characteristics (see Table 2), as a recognition task including the 69 different individual dark pattern types would have exceeded the scope and purpose of this online survey. Instead, we relied on Mathur et al.'s high-level dark pattern characteristics. For each of the five dark pattern characteristics (*asymmetry*; *covert*; *deception*; *information hiding*; and *restriction*) participants rated on a 5-point Likert scale ("Not at all" - "Extremely"), how much the characteristic was present in the screenshot. For each screenshot, this resulted in an average rating. Figure 5 demonstrates how the screenshots were used to generate these ratings. This procedure allows us to compare participants' ratings between the different screenshots. Using this approach, the maximum rating for a screenshot featuring all dark pattern characteristics corresponds to $[4, 4, 4, 4, 4]$ and thus an average rating of 4, while a minimum rating for a screenshot without dark patterns corresponds to $[0, 0, 0, 0, 0]$ and thus an average rating of 0. In total, all 193 survey respondents rated $(193 * 16 = 3088)$ 3088 screenshots.

6.1 Demographic Information

The mean age across individuals was $\mu = 27.91$ years ($SD = 9.53$), with 155 identifying as female and 35 as male. The remainder (N=3) identified as either non-binary or with a third gender. When asked about their current country of residence, the participants replied as follows: Australia (4); Canada (35); France (1); Greece (1); Hong Kong - S.A.R. (1); Ireland (11); Japan (1); South Africa (2); Spain (1); United Kingdom of Great Britain and Northern Ireland (40); United States of America (96). In terms of how frequently participants used the internet, 189 self-reported using the internet on a daily basis, with the remainder (N=4) using it more than once per week. An inclusion criterion for participation was a previous experience with at least one of the four SNSs. Therefore, we asked participants about their usage of Facebook, Instagram, TikTok, and Twitter. Regarding Facebook, 138 participants reported actively using it, 20 do not use it, and 35 used to use it but not anymore. 167 participants currently use Instagram, while 15 do not use it, and 11 have used it but do not anymore. Looking at TikTok, 134 participants use it currently, 55 do not, and 4 have used it but do not anymore. Lastly, 112 participants actively use Twitter, 51 are not using it, whereas 30 used to but do not anymore.

6.2 Generally Recognising Dark Patterns

For the eight screenshots that did feature dark patterns, when asked if respondents notice any malicious interface elements in the screenshot, 426 screenshots received a "yes" rating, 408 a "maybe", and 710 a "no" rating. In contrast, for the eight screenshots that did not contain dark patterns, 143 received a "yes" rating, 269 a "maybe", and 1132 a "no" rating. A Wilcoxon signed rank test with continuity correction shows significant differences between the two groups of screenshot ratings ($V = 89253$, $p - value < 0.0001$, $R = 0.37$). Thus, we see that more people noticed malicious elements in screenshots that contained dark patterns.

6.3 Differentiating Between Screenshots With and Without Dark Patterns

Our previous results showed that people generally see differences between the two types of screenshots. We can thus test whether people rate screenshots differently when they show dark patterns

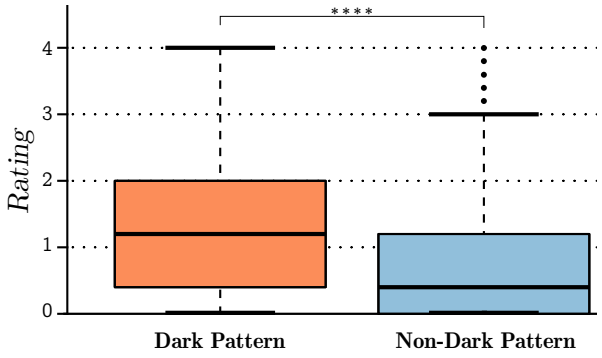


Figure 4: This box plot visualises the differences in which participants, who were provided with a definition for dark patterns, rated the screenshots after being asked if they noticed any malicious designs. The figure shows a significant difference between participants' ratings of screenshots containing dark patterns versus those that do not.

compared to screenshots with no dark patterns according to Mathur et al.'s [33] five characteristics. We thus calculated the median total rating for screenshots that featured dark patterns and the same for screenshots that did not feature dark patterns. Across all screenshots which featured dark patterns, we find a median rating of 1.2 ($mean$ = 1.26, SD = 1.02) compared to a median rating of 0.2 ($mean$ = 0.69, SD = 0.81) for screenshots without dark patterns (see Figure 4). A Wilcoxon signed-rank test results in a significant difference between the two ratings (V = 669900, p-value < 0.0001, R = 0.3). Given that non-dark pattern screenshots received a significantly lower median average rating than dark pattern screenshots, we conclude that people recognised a difference between screenshots containing dark patterns and those that did not base on questions adopting the five characteristics. We further observe a difference in participants' perceptions of the two types of screenshots. While the median rating of screenshots without dark patterns is 0.2, very close to 0 ("Not at all"), the median rating of screenshots with dark patterns is 1.2 ("A little bit"), relatively low considering a maximum rating of 4 ("Extremely"). This implies that while participants distinguish screenshots with and without dark patterns with a significant difference, based on the five characteristics, their rating is overall rather low.

6.3.1 Per Characteristic Rating. Based on participants' different ratings for dark pattern versus non-dark pattern screenshots, we gain a more detailed view of the applicability of the individual characteristics. We consider the median scores here because the data is not normally distributed. Overall, the median data indicates that across screenshots of the same kind, each characteristic contributed to the assessment, with a rating of 1 for screenshots that contain dark patterns and 0 for those not featuring dark patterns.

To further validate the five characteristics, we investigated their relationship to the malice rating from section 6.2. We performed a multiple linear regression to see how well the individual characteristics predict the malice rating. The result shows a F-statistic p-value of < 0.0001, suggesting that at least one of the five characteristics is significantly related to the malice score. Considering each t-statics,

Comparison of Five Characteristics					
Dark Pattern Screenshots					
	Asymmetry	Covert	Restrictive	Deceptive	Hides Info.
mean	1.42	1.21	1.40	1.02	1.27
median	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
SD	1.26	1.20	1.18	1.18	1.26
Non-Dark Pattern Screenshots					
mean	0.71	0.80	0.84	0.60	0.80
median	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
SD	1.03	1.08	1.12	0.99	1.11

Table 3: Overview of the mean, median, and standard deviation of participants' ratings of dark pattern and non-dark pattern screenshots according to Mathur et al.'s [33] five characteristics: *assymetric*, *covert*, *restrictive*, *deceptive*, and *information hiding*.

Comparison Of Screenshots								
Dark Pattern Screenshots								
	F1	F2	I1	I2	Ti1	Ti2	Tw1	Tw2
mean	1.40	1.42	1.45	1.21	1.76	1.14	**0.60**	1.12
median	1.40	1.40	1.40	1.20	1.80	1.00	**0.40**	1.00
SD	1.08	0.94	1.08	0.99	1.06	0.99	**0.73**	0.89
Non-Dark Pattern Screenshots								
	FA	FB	IA	IB	TiA	TiB	TwA	TwB
mean	**1.06**	0.66	0.45	0.54	0.69	**1.10**	0.39	0.56
median	**1.00**	0.20	0.00	0.20	0.40	**1.00**	0.00	0.20
SD	**0.99**	0.92	0.71	0.73	0.81	**0.99**	0.65	0.75

Table 4: Overview of the mean, median, and standard deviation of participants' ratings per dark pattern and non-dark pattern screenshot. Each of the four SNSs was represented with two screenshots containing dark patterns and two that did not. The letters in the screenshots' labels refer to a particular SNS: F = Facebook; I = Instagram; Ti = TikTok; Tw = Twitter.

further analysis revealed that the characteristics *asymmetric* (t < 0.001) and *restrictive* (t = 0.004) show a significant association with the malice score. The remaining characteristics *covert* (t = 0.053), *deceptive* (t = 0.081), and *hides information* (t = 0.074) do not yield such association, however. Thus, changes in those three characteristics do not significantly affect the malice score in our model.

6.3.2 Per Screenshot Rating. Considering the screenshots independently, we gain further insights into the differences between average scores. This allows us to notice the effectiveness and sensitivity with which this approach measures the malice in a single screenshot. Across the eight screenshots containing dark patterns, seven screenshots have median ratings >1, while the median rating for one screenshot is 0.4 (see Table 4, Tw1). Looking at the non-dark pattern screenshots, six were rated with a median <1, while two screenshots have a median rating of 1 (see Table 4, F1 and Ti2).

7 DISCUSSION

This work presents insights from two studies, widening our understanding of how dark patterns manifest in SNSs and exploring

a novel approach to evaluate the malice of interfaces. As online regulations have been shown to lack protection of users [6], we were interested in the effectiveness of current regulations that aim to shield users from dark patterns. Based on a comprehensive taxonomy, we let experienced HCI researchers apply dark patterns, by means of their descriptions, to four popular SNSs (Facebook, Instagram, TikTok, and Twitter). Although a range of dark patterns has been recognised, the results of the first study bear certain difficulties that hindered the process and thus highlight a necessity for more efficient approaches to recognising dark patterns. Exploring an alternative approach to evaluate the malice of interfaces, we defined five questions based on Mathur et al.'s [33] dark pattern characteristics. Letting regular users rate screenshots sampled from recordings of the first study, we found a potential measure in this approach that can be of aid for regulatory strategies. In this section, we discuss the applicability of dark pattern research as a tool to evaluate interfaces in relation to regulation.

7.1 A Taxonomy As Evaluation Tool

We acknowledge that the applied taxonomy, including entailed dark patterns from eight works, was not designed as a tool for the assessment of dark patterns and covers different scopes regarding their level of abstraction. While research on dark patterns moves forward, expanding our knowledge of the types of dark patterns that exist, we believe that it is important to reflect on the current status quo and consider the multitude of findings in new contexts. Study 1, therefore, tests the utility of dark patterns to identify their instances in SNSs. With the successful recognition of a range of these dark patterns in SNSs, the results of our first study imply that the chosen approach is suitable for identifying dark patterns in domains that may lie outside their original scope, offering an answer to our first research question. Tainting these results, however, we noticed certain issues that posed difficulties to the reviewers when executing their tasks.

Overall, 31 out of 69 considered dark patterns were recognised, leaving another 31 not applicable in the context of SNSs. Especially game-related dark patterns [45] and those inspired by proxemic theory [22] were not all or rarely noticed. In contrast, dark patterns by Gray et al. [19] were identified more frequently. This implies that expert reviewers found it easier to recognise dark patterns that were described more abstractly compared to domain-specific ones suggesting similar effectiveness in identifying dark patterns in regulatory contexts. A particular difficulty in this study emerged from dark patterns that shared the same names. Brignull's [7] *Confirmshaming* dark pattern, for instance, was carried over by Mathur et al. [33] who remained with its original definition, making it confusing as to which version should be applied when a related dark pattern is recognised. Other candidates - *Privacy Zuckering* by Brignull [7] and Bösch et al. [5] and *Bait and Switch* by Brignull [7] and Greenberg et al. [22] - were given distinct descriptions resulting in different applicability in SNSs. Contrary to this difficulty, the results of our co-occurrence tests show that dark patterns with different names apply in same interfaces. We see two possible explanations for this: (1) Provided descriptions of two dark patterns are too close, clouding distinct applications, at least in the context of SNSs. A high co-occurrence between *Interface Interference* [19] and *Visual Interference* [33] can be explained this way. Alternatively, (2) two different dark patterns complement each other creating particularly problematic situations. Here, *Privacy Zuckering* and *Bad Default* do not describe the same interface problems but *Privacy Zuckering* profits from the *Bad Default* dark pattern as the latter will often result in users sharing more data unknowingly.

7.2 Assessing the Malice of Interfaces

The results of study 1 indicate that abstract and distinct criteria are most efficient for evaluating the presence of dark patterns in interfaces. Study 2, therefore, explores an alternative approach by relying on Mathur et al.'s [33] five high-level characteristics to assess the malice of interfaces. Based on their framework, we developed five questions that we used to study regular users' ability to recognise dark patterns based on screenshots of the four SNSs. Answering our second research question, the results of this second study show that users were generally able to distinguish between screenshots featuring dark patterns and those that did not. However, ratings for the dark pattern screenshots indicate some difficulties as scores were considerably low (average median = 1.2), given that the maximum score a screenshot could receive is 4. Yet, participants' ability to differentiate screenshots based on these five characteristics suggests the promising effectiveness of this approach. Past work has found difficulties among participants in avoiding dark patterns [12, 32]. While our data suggest similar difficulties, our second study's results further support suggestions by Bongard-Blanchy et al. [4], who have shown that informing users about dark patterns helps to identify them.

This is further supported by the median ratings of each evaluated characteristic of the sixteen screenshots. We notice that across the eight dark pattern screenshots, each rating is 1 ("A little bit"), whereas the median rating for non-dark pattern screenshots is 0 ("Not at all"), as shown in Table 3. This consistency across participants implies that all characteristics contribute to the assessment of dark patterns in screenshots. Considering individual median ratings per screenshot (see Table 4), we see this consistency almost entirely confirmed. With regards to the dark pattern screenshots, participants were able to correctly identify malicious interfaces in seven out of eight instances (87.5%). In non-dark pattern screenshots, participants accurately determined no presence of dark patterns six out of eight times (75%). As neither the taxonomy nor Mathur et al.'s [33] characteristics were designed to identify or recognise dark patterns in SNSs, this attempt opens a possible pathway for future directions of dark pattern research. Relying on more abstract characteristics offers a promising approach to evaluating new interfaces. Figure 5 visually demonstrates this approach. If an interface is suspected of containing any number of dark patterns, it is evaluated using a 5-point Likert-scale ("Not at all" - "Extremely") according to the five questions adopting Mathur et al.'s [33] characteristics. The maliciousness of the interface can then be determined by considering each characteristic's rating based on their individual values or as an average calculated from all five. We gain further support for this model through the multiple linear regression showing a highly significant relationship between the questions and the malice score. Individually, two characteristics – *asymmetry* and *restrictive* – maintain this highly significant association while three do not, leaving

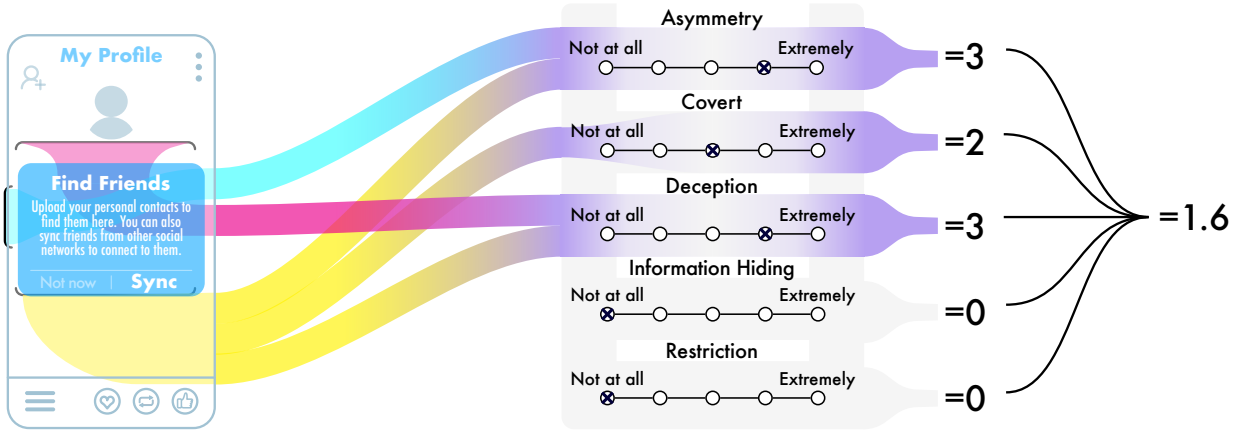


Figure 5: This figure demonstrates the approach to assess malice in interfaces by applying questions based on Mathur et al.'s [33] dark pattern characteristics. First, an interface is selected which is suspected of containing any amount of dark patterns. Using the five questions described in Table 2, the interface can then be evaluated using a Likert-scale from "Not at all" to "Extremely". In this example, we demonstrate this based on a five-item scale. The result are independent ratings for each characteristic, which can be averaged into a single digit.

room for future improvement. The nature of this study describes an experimental setup aiming to assess the malice in interfaces better. The general statistical significance of both users' ability to differentiate between malicious and harmless design as well as in our multiple linear regression affirms the utility of such characteristics and our model. This approach allows further insights into the types of dark patterns present in the interface by considering which characteristics they subscribe to. As participants of the second study only had to meet the criteria of being regular users of SNSs, we believe that more experienced evaluators could be able to evaluate interfaces more sensitively. Although this work utilises a total of 69 types of dark patterns, we acknowledge that our work has left new gaps for future work to consider SNS-specific types of dark patterns. Meanwhile, recent efforts have extended our knowledge of dark patterns in SNSs [23, 24, 36, 38], which leaves room for future updates of our research.However, while these prior efforts describe dark patterns that occur in SNSs based on qualitative approaches, to our knowledge, this research is among the first to quantitatively assess dark patterns in SNSs while considering both experts' and users' ability to recognise them in this environment. Moreover, we extend the current discourse with a possible measure to access the malice of interfaces, regardless of their origin, by not requiring a complete corpus after all. Instead, relying on wider characteristics enables users to assess this malice based on five simple yet extendable, high-level dimensions.

7.3 Paving The Way For Regulations

The variety of dark pattern types shows how far-stretched mischievous strategies in online domains can be. Still, they all have one thing in common: They harm users. Regulators and legislation already have powerful tools to ensure the protection of end-users. However, not all regulations are equally effective. To support this, findings from HCI research on dark patterns can aid existing approaches to protect peoples' privacy on problematic designs. The presented work has mainly two implications for legislative efforts regarding dark patterns. The first one addresses the problem that the law is prone to lag behind dark patterns evolution, suggesting alternative approaches are needed to protect users successfully. The regulation of dark patterns must, on the one hand, be concrete enough to address manipulative mechanisms and, on the other hand, abstract enough to capture future developments. Our findings show that research in HCI constantly explores new dark patterns resulting in diverse taxonomies, as depicted in Figure 1. Nevertheless, we see that recognising dark pattern characteristics on a meta-level is convincing and, referring to Mathur et al.'s high-level characteristics [33], might be a promising approach to achieving a shared conceptualisation. This suggests that generalisable definitions and characterisations are better suited and more future-proof to assess dark patterns in various domains. We argue that findings from HCI can support legislative efforts by providing dark pattern characteristics based on empirical research and offering a sustainable vocabulary helping lawmakers to get ahead of developments of unethical designs. Such characteristics could be a basis for a legal definition and a general ban on dark patterns. The second implication deals with recognising dark patterns in practice. Tools from HCI have the compelling potential for supporting courts and authorities since they could objectively measure the manipulation effect of a design (e.g. Figure 5). Offering authorities a tool to evaluate the malice of interfaces easily, the proposed score determines the degree to which a specific design is either harmless or contains malicious features based on empirical evidence. Here, the goal lies in the identification of a certain score within the sweet spot, or threshold, that most accurately distinguishes between interfaces with dark patterns from those without. Our results show that even regular users are able to correctly differentiate between malicious and harmless interfaces. Professionals and trained people would likely perform similar tasks with even better accuracy. Consequently, the findings and tools from HCI research can become a considerable and valuable instrument in the decision-making processes of authorities. Ultimately, HCI research can pave the way

for regulators to act on observed exploitation in interfaces that can, but are not limited to, target users' personal data or manipulate their decision space, provoking potentially harmful actions.

8 LIMITATIONS & FUTURE WORK

Both studies of this work yield certain limitations. Firstly, study 1 was conducted during the COVID-19 pandemic, which meant that the experiment was conducted without supervision. Although recordings do not suggest misunderstandings across reviewers, a present study supervisor can offer additional assistance. While we aimed to consider a range of SNSs, the number of platforms available today limited us to four applications with similar functionalities. Although the chosen SNSs present popular platforms, we neglected important services like YouTube or Twitch, featuring video-streaming platforms, but also messenger services like WhatsApp or Telegram, which each entail large user bases. Future work could consider alternative SNSs that were not in the scope of this work. As Mathur et al.'s [34] sixth *Disparate Treatment* characteristic was not applied at all during the reviews, meaning that none of Zagal et al.'s [45] dark patterns were recognised in SNSs, it would further be interesting to consider SNSs that offer paying users different experiences (e.g. LinkedIn, Twitch, or YouTube). Also, future work could include recording instances of users sharing their data in- and outside of SNSs, as we did not include such a task in our cognitive walkthroughs. Study 1 was further limited by the selection of dark patterns included in our taxonomy. Because we decided only to include dark patterns that resulted from empirical research, we excluded those part of guidelines and regulations. Furthermore, Gunawan et al. [23] propose twelve additional dark patterns that we did not include as our experiment was conducted at the time of their publication. Future work could include further types of dark patterns for gaining an even deeper understanding of dark patterns in SNSs. Moreover, our methodology proved fruitful gaining us important insights into dark patterns in SNSs. Future work could adopt this approach to utilising the existing corpus of dark pattern knowledge when investigating dark patterns in other domains.

In study 2, we tested our evaluation approach based on screenshots to assess the malice of interfaces. While results indicate certain accuracy in participants differentiating between screenshots containing dark patterns and those that do not, our results do not allow us to make any statements about how well participants identified specific dark patterns. Furthermore, the screenshots are limited to showing dark patterns within a single stage on a static image. While we made sure to choose dark patterns, which are recognisable on screenshots, this limitation excludes possible dark patterns that rather work on a procedural level during an interaction. To reach participants, we used the online research platform *Prolific* [31] to generate a convenience sample, restricted only to users who have prior experience with SNSs and are fluent in the English language, as screenshots were in English. However, we did not aim for a representative sample. Surprisingly, we noticed that 80,3% of the participants identified as females skewing the demographic. Although we did not notice any differences between individual participants' ratings, we acknowledge that the data set is biased towards females. Moreover, we decided to rely on regular users as participants for this study. As our findings suggest a novel approach to aid the regulation of dark patterns, it would be interesting the see how related professionals such as regulators and legal scholars recognise dark patterns in a similar study. This could further be enhanced by additional characteristics that better incorporate malicious interfaces currently not covered. Also, Gunawan et al. [23] suggest that dark patterns may exist in SNSs to a different extent in their desktop modality. While we identified a host in SNSs for existing dark patterns, this work considers dark patterns that are not specific to this domain. As many described dark patterns have their origin in online shopping websites, future work could investigate social media platforms to describe unique dark patterns here. This further includes the characteristics from Mathur et al. [33], which we used in our survey. Although the results of the multiple linear regression indicate a highly significant relationship between the questions and the malice score, only two out of five characteristics also yielded significant associations. This invites future research to advance our model and develop a suitable questionnaire for improved assessment.

9 CONCLUSION

In this paper, we examined four popular SNS platforms (Facebook, Instagram, TikTok, and Twitter) for dark patterns, advancing research in this context. Based on a cognitive walkthrough with six HCI experts, we learned which dark patterns occur in SNSs by considering a taxonomy based on prior findings in this field. Results of this study show that while this approach offers detailed insights, it lacks certain efficiency while posing difficulties to reviewers. Considering these results, we designed a novel approach to assess the malice of interfaces based on high-level characteristics. In a second study, we tested this alternative demonstrating a tool to recognise dark patterns in screenshots. Taking a legal perspective on current regulations for dark patterns, we discuss the findings of our second study, shining a light on how HCI research can aid the protection of SNS users.

ACKNOWLEDGMENTS

The research of this work was partially supported by the Klaus Tschira Stiftung gGmbH.

REFERENCES

[1] Reddit Inc © 2021. 2021. r/SampleSize: | Where your opinions actually matter! https://www.reddit.com/r/SampleSize/ (visited on 2021-08-25).
[2] Dohyun Ahn and Dong-Hee Shin. 2013. Is the social use of media for seeking connectedness or for avoiding social isolation? Mechanisms underlying media use and subjective well-being. *Computers in Human Behavior* 29, 6 (2013), 2453–2462.
[3] Ine Beyens, J Loes Pouwels, Irene I van Driel, Loes Keijsers, and Patti M Valkenburg. 2020. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports* 10, 1 (2020), 1–11.
[4] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I Am Definitely Manipulated, Even When I Am Aware of It. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086
[5] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proc. Priv. Enhancing Technol.* 2016, 4 (2016), 237–254.
[6] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. 2022. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3501947

[7] Harry Brignull. 2010. Deceptive Design – formerly darkpatterns.org. https://www.deceptive.design/ Visited on 2022-03-29.

[8] European Commission. 2016. GDPR-16 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

[9] European Commission. 2022. Proposal for a regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.html

[10] European Commission. 2022. Proposal for a regulation of the European Parliament and of the Council on harmonized rules on fair access to and use of data (Data Act). https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.html

[11] Gregory Conti and Edward Sobiesk. 2010. Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, Raleigh, North Carolina, USA, 271. https://doi.org/10.1145/1772690.1772719

[12] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[13] Sindhu Kiranmai Ernala, Moira Burke, Alex Leavitt, and Nicole B. Ellison. 2020. How Well Do People Report Time Spent on Facebook? An Evaluation of Established Survey Questions with Recommendations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376435

[14] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[15] Susanne Friese. 2019. *Qualitative data analysis with ATLAS. ti* (3 ed.). SAGE Publications Ltd, California, United States. 344 pages.

[16] ATLAS.ti Scientific Software Development GmbH. 2021. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[17] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300408

[18] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. *What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit*. Association for Computing Machinery, New York, NY, USA, 61–73. https://doi.org/10.1145/3357236.3395486

[19] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. *The Dark (Patterns) Side of UX Design*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[20] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. *Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective*. Association for Computing Machinery, New York, NY, USA, pp. 1–18. https://doi.org/10.1145/3411764.3445779

[21] Paul Graßl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. 2021. Dark and Bright Patterns in Cookie Consent Requests. *Journal of Digital Social Research* 3, 1 (Feb. 2021), 1–38. https://doi.org/10.33621/jdsr.v3i1.54 Number: 1.

[22] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. *Dark Patterns in Proxemic Interactions: A Critical Perspective*. Association for Computing Machinery, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[23] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2022. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. 5 (2022), 1–29. Issue CSCW2. https://doi.org/10.1145/3479521

[24] Hana Habib, Sarah Pearman, Ellie Young, Ishika Saxena, Robert Zhang, and Lorrie FaIth Cranor. 2022. Identifying User Needs for Advertising Controls on Facebook. 6 (2022), 1–42. Issue CSCW1. https://doi.org/10.1145/3512906

[25] Peter Hustinx. 2010. Privacy by design: delivering the promises. *Identity in the Information Society* 3, 2 (Aug. 2010), 253–255. https://doi.org/10.1007/s12394-010-0061-z

[26] Monique W.M. Jaspers, Thiemo Steen, Cor van den Bos, and Maud Geenen. 2004. The think aloud method: a guide to user interface design. *International Journal of Medical Informatics* 73, 11 (2004), 781–795. https://doi.org/10.1016/j.ijmedinf.2004.08.003

[27] Reynol Junco. 2013. Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior* 29, 3 (May 2013), 626–631. https://doi.org/10.1016/j.chb.2012.11.007

[28] Europäische Kommission, Generaldirektion Justiz und Verbraucher, F Lupiáñez-Villanueva, A Boluda, F Bogliacino, G Liva, L Lechardoy, and T Rodríguez de las Heras Ballell. 2022. *Behavioural study on unfair commercial practices in the digital environment : dark patterns and manipulative personalisation : final report*. Amt für Veröffentlichungen der Europäischen Union. https://doi.org/10.2838/859030

[29] California State Legislature. 2018. CCPA-18 2018. California Consumer Privacy Act of 2018 [1798.100 - 1798.199] (CCPA). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[30] M Leiser and M Caruana. 2021. Dark Patterns: Light to be found in Europe's Consumer Protection Regime. *Journal Of European Consumer And Market Law* 10(6) (2021), 237–251. Retrieved from https://hdl.handle.net/1887/3278362.

[31] Prolific Academic Ltd. 2021. Prolific | Online participant recruitment for surveys and market research. https://prolific.co/ (visited on 2021-08-25).

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

[33] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–32. https://doi.org/10.1145/3359183

[34] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. *What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods*. Association for Computing Machinery, New York, NY, USA, pp. 1–18. https://doi.org/10.1145/3411764.3445610

[35] Thomas Mildner and Gian-Luca Savino. 2021. Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451659

[36] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 192, 15 pages. https://doi.org/10.1145/3544548.3580695

[37] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376321

[38] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142

[39] Sarita Yardi Schoenebeck. 2014. Giving up Twitter for Lent: how and why we take breaks from social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 773–782. https://doi.org/10.1145/2556288.2556983

[40] Holly B. Shakya and Nicholas A. Christakis. 2017. Association of Facebook Use With Compromised Well-Being: A Longitudinal Study. *American Journal of Epidemiology* 185, 3 (Feb. 2017), 203–211. https://doi.org/10.1093/aje/kww189

[41] Statista. 2021. We Are Social, Hootsuite, DataReportal. (July 21, 2021). Most popular social networks worldwide as of July 2021, ranked by number of active users (in millions). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[42] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, New York, NY, USA, 973–990. https://doi.org/10.1145/3319535.3354212

[43] Jin-Liang Wang, Linda A. Jackson, James Gaskin, and Hai-Zhen Wang. 2014. The effects of Social Networking Site (SNS) use on college students' friendship and well-being. *Computers in Human Behavior* 37 (Aug. 2014), 229–236. https://doi.org/10.1016/j.chb.2014.04.051

[44] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania) *(SOUPS '11)*. Association for Computing Machinery, New York, NY, USA, Article 10, 16 pages. https://doi.org/10.1145/2078827.2078841

[45] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013)* (May 14-17). Society for the Advancement of the Science of Digital Games, Chania, Crete, Greece, 39–46. http://www.fdg2013.org/program/papers.html

PUBLICATION P5

Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users

Authors:
Thomas Mildner, Orla Cooney, Anna-Maria Meck, Marion Bartl, Gian-Luca Savino, Philip R. Doyle, Diego Garaialde, Leigh Clark, John Sloan, Nina Wenig, Rainer Malaka, & Jasmin Niess

The publication contributes to the following angles:
DESIGN USER GUIDELINE

This publication draws on previous provocations from P2 to explore ethical caveats of Conversational User Interface (CUI) design. Interviewing researchers, practitioners, and users of CUI systems, this work studies each cohort's perspective. Based on thematic analysis, the publication proposes five ethical caveats and guiding questions to support ethical CUI design. Conceptualising tensions between users and the design of their systems, the publication further contributes the CUI Expectation Cycle (CEC), a model to maintain realistic expectations.

Its contribution to the thesis is to the design angle. It landscapes the CUI domain for possible dark patterns and highlights risks of unique design features, contributing to the design angle. From a user-centred perspective, it explores roots for unrealistic and broken expectations resulting in the CEC, contributing to the user angle. Based on these findings, it warns about ethical caveats and provides guiding questions as contributions to the guideline angle.

My contribution to this paper was the design of the study and interviewing of participants, qualitative coding, execution of thematic analysis, and the development of the CEC. Best practice for qualitative research states that thematic analysis should be done between researchers. I administered the analysis together with co-authors of this paper. I drafted the manuscript and revised it before the final publication.

The contents of this chapter originally appeared in: Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R., and Niess, J., "Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642542

Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users

Thomas Mildner
mildner@uni-bremen.de
University of Bremen
Bremen, Germany

Orla Cooney
orla.cooney@ucdconnect.ie
University College Dublin
Dublin, Ireland

Anna-Maria Meck
anna.meck@googlemail.com
BMW Group
Germany

Marion Bartl
marion.bartl@ucdconnect.ie
University College Dublin
Dublin, Ireland

Gian-Luca Savino
gian-luca.savino@unisg.ch
University of St.Gallen
St.Gallen, Switzerland

Philip R. Doyle
phil.hmd.research@gmail.com
HMD Research
Ireland

Diego Garaialde
diego.garaialde@ucd.ie
University College Dublin
Dublin, Ireland

Leigh Clark
leighmhclark@gmail.com
Bold Insight UK
London, Great Britain

John Sloan
john.sloan.1@ucdconnect.ie
University College Dublin
Dublin, Ireland

Nina Wenig
nwenig@tzi.de
University of Bremen
Bremen, Germany

Rainer Malaka
malaka@tzi.de
University of Bremen
Bremen, Germany

Jasmin Niess
jasminni@uio.no
University of Oslo
Oslo, Norway

ABSTRACT

Advances in natural language processing and understanding have led to a rapid growth in the popularity of conversational user interfaces (CUIs). While CUIs introduce novel benefits, they also yield risks that may exploit people's trust. Although research looking at unethical design deployed through graphical user interfaces (GUIs) established a thorough understanding of so-called dark patterns, there is a need to continue this discourse within the CUI community to understand potentially problematic interactions. Addressing this gap, we interviewed 27 participants from three cohorts: researchers, practitioners, and frequent users of CUIs. Applying thematic analysis, we construct five themes reflecting each cohort's insights about ethical design challenges and introduce the CUI Expectation Cycle, bridging system capabilities and user expectations while considering each theme's ethical caveats. This research aims to inform future development of CUIs to consider ethical constraints while adopting a human-centred approach.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; **Interaction design theory, concepts and paradigms**; *Ubiquitous computing*; • **Social and professional topics** → *User characteristics.*

KEYWORDS

CUI, conversational user interfaces, conversational agents, voice agents, chatbots, thematic analysis, ethical design, deceptive design patterns, dark patterns

ACM Reference Format:
Thomas Mildner, Orla Cooney, Anna-Maria Meck, Marion Bartl, Gian-Luca Savino, Philip R. Doyle, Diego Garaialde, Leigh Clark, John Sloan, Nina Wenig, Rainer Malaka, and Jasmin Niess. 2024. Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642542

1 INTRODUCTION

Conversational systems have been around since the mid-20th century [17, 99]. However, recent technological advances in natural language processing and understanding have led to conversational user interface (CUI) interactions becoming common in many people's daily lives [50]. Today, CUIs come in two key forms: text-based chatbots, which are often used in commercial settings to fulfil customer service roles (e.g., IBM Watson [1], Amazon Lex [2]), and voice assistants, which are commonly integrated into smartphones and smart home devices (e.g., Apple's Siri [3], Amazon Alexa [4], Google

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642542

[1] https://www.ibm.com/watson
[2] https://aws.amazon.com/lex/
[3] https://www.apple.com/siri/
[4] https://alexa.amazon.com/

Assistant [5]). Following the claims of CUI developers, the technology promises novel ways to interact with service providers efficiently, intuitively, and seamlessly [87, 92]. The rapid growth of these technologies is mirrored by increased interest among researchers within the human-computer interaction (HCI) community [27, 83, 91].

Yet, many users experience frustration when engaging with CUIs [23] that negatively impacts their overall experience. According to the literature, frustrations stem from exaggerated expectations among users regarding communicative competence that current CUIs cannot live up to. For instance, anthropomorphic CUI design can lead to unrealistic expectations of near-human-level performance when interacting with such devices [26]. However, what users generally experience is a call-and-response type interaction rather than a free-flowing conversation. While recent efforts in generative artificial intelligence (AI) mitigate some of these shortcomings of traditional CUI interactions, recent work shows similar limitations remaining in large language models (LLM) [49]. In a similar vein, privacy and trust-related issues are said to limit the scope of tasks that users are willing to perform with a system [26, 62]. The lack of transparency around how personal data is recorded and used can also become an obstacle for users in taking full advantage of CUI capabilities, again limiting the contexts in which people are willing to use them [23, 84]. To mitigate these issues, recent work offers domain-specific design heuristics [56] or frameworks to guide user engagement [100]. Still, further guidelines are needed as the technology advances from its infancy. Suggesting a lack of sufficient guidelines, practitioners adopt graphical user interfaces (GUIs) best practices to account for the unique affordances that CUIs present [16, 46, 56]. However, we ought to be wary of the ethical implications as recent work in HCI voices serious concerns regarding unethical practices in contemporary GUIs – providing a taxonomy of deceptive design patterns often referred to as "dark patterns", which inhibit users' decision-making to benefit service providers [12]. Although malicious incentives prevail in CUIs [19, 79], we currently have a limited understanding of how unethical design manifests in the context of CUIs and which ethical caveats require consideration. We, therefore, see an opportunity to proactively address concerns akin to recently addressed deceptive design issues that required legislative and regulatory actions [18, 60, 81] to protect users in GUI contexts.

To that end, we aim to gain insights regarding ethical caveats for CUI design. Here, the dark pattern discourse supports this endeavour by providing a novel angle to understand unethical design in problematic CUI interactions. In the context of this work, we refer to *ethical caveats* as considerations practitioners need to make to avoid adopting design strategies or features that undermine users' autonomy or deceive them into making choices that are not necessarily in their best interests. To that end, this work explores ethical concerns around the design of current CUI systems, as well as potential issues that may arise in the future, among three specific cohorts: (1) researchers who focus their work on CUIs, (2) practitioners who develop CUIs, and (3) frequent users who engage with CUIs at least once a week. To our knowledge, this work presents the first to consider multiple perspectives to assess CUIs based on ethical caveats. In total, we interviewed 27 people to address the following research question:

RQ: Which ethical caveats should be considered when designing CUI interactions, and how should they be addressed?

In answering this research question, this work has two main contributions. Firstly, five themes outline ethical caveats in CUI design: *Building Trust and Guarding Privacy*, *Guiding Through Interactions*, *Human-like Harmony*, *Inclusivity and Diversity*, and *Setting Expectations*. Secondly, we introduce the CUI Expectation Cycle, a framework promoting ethical design considerations by incorporating our five themes. This framework responds to repeated demands for design guidelines to mitigate problematic interactions and negative user experiences expressed by practitioners while addressing users' concerns and expectations.

2 RELATED WORK

The related work of this paper encompasses the current development of ethical awareness in HCI and CUI research. The section begins with recent work from the CUI community, addressing design and technology limitations and showcasing a need for explicit guidelines for practitioners. We will then briefly outline work addressing unethical design practices (a.k.a. dark patterns), currently most widely discussed in GUI research. The lack of such work in CUI research emphasises a need to consider similar issues when developing chatbots and voice assistants and motivates our research. Lastly, we will review recent advances in developing CUI-related guidelines and best practices.

2.1 Design and Technology-based Limitations

Due to the distinct affordances they present, CUIs have to overcome design challenges that cannot simply be borrowed from other fields [70]. In particular, the use of language, uttered or typed, as a primary input for CUIs poses problems regarding the ease with which users can explore their functionalities [9, 15, 96] or restricting the assessment of a system's capabilities and boundaries [30, 47, 62]. This is affirmed by users' inability to recall the exact command needed for a response from their device [21, 79], limiting interactions to memorised prompts.

Further difficulties are linked to expectations users have toward CUIs. These expectations are steep as it is proposed that users can talk to their devices intuitively via a principal way of communication: language. However, in daily use, CUIs often fall short of their users' expectations [23, 26, 62, 95]. The degree of dissatisfaction this encourages among users has even led some experts in the field to suggest practitioners should rethink the "Conversational" part of "Conversational User Interfaces" altogether [86]. High user expectations also stem from anthropomorphic design features that encourage users to see CUIs as "social actors", a term coined by Nass and Brave in their Computers are Social Actors (CASA) paradigm [71]. Anthropomorphic design has measurable effects through users aligning their speech patterns when talking to CUIs, as explored by Cowan et al. [22]. Although anthropomorphic design can serve beneficial purposes such as increased technology acceptance [22, 54, 90, 96] and can influence the user experience positively [96], there are several usability [54] and ethical problems [4, 32] associated with overusing anthropomorphic design

[5] https://assistant.google.com/

features. Lacey and Cauwell [55], for instance, studied dark patterns in connection to social robots, which could exploit "cuteness" attributes to coerce users' decisions and are only possible through anthropomorphic design.

While Seymour and Van Kleek [90] found a link between relationship development and anthropomorphic CUI design, Kontogiorgos et al. [54] voice a note of caution: although an anthropomorphic social agent led to a higher degree of engagement and sociability compared to a smart speaker without human-like traits, they also found that task-completion time with an embodied anthropomorphic agent was 10% higher compared to a disembodied smart speaker. They link this to the anthropomorphic agent being perceived as a "socially present partner in conversation" [54, p 138]. In a similar vein, Lee et al. [59] note a preference in users towards physically present CUIs over invisible ones based on a user study describing their mental model when interacting with CUIs. In recent work, Dubiel et al. [28] spotlight a connection between trust and anthropomorphic traits of CUIs. Realising ethical caveats, the authors propose four design strategies to calibrate trust while avoiding anthropomorphic features to enhance user engagement. Hence, whilst anthropomorphic agents often seem to be preferred, they can lead to increased expectations and inadequate attributions of capabilities. This makes the design and implementation of human-like cues a delicate balancing act. Including human traits in CUI design can foster engagement and acceptance. However, concurrently, potential negative consequences for usability if users overestimate their CUI's abilities need to be considered.

Problems also arise from more general technical difficulties associated with datasets used to train LLMs that underpin CUI capabilities. Specifically, LLMs are said to contain and reproduce highly biased worldviews, leading them to "overrepresent hegemonic viewpoints" [4, p 610]. A lack of data representing marginalised social groups increases their marginalisation even further. This is the case, for instance, with older adults, users with speech impairments or pronounced idiosyncrasies, or speakers with a strong colloquial dialect, vernacular, and/or accent for whom speech recognition performs poorer than for standard language speakers [3, 44, 53, 65, 85]. As a result, diverse cultural experiences and identities are undercatered, leading to "othering" referring to the alienation of certain (social) groups resulting in marginalisation or exclusion in social contexts [65]. Where contemporary and commercially available CUIs initially promise natural language understanding and natural interactions, this promise only holds for a certain group of users. For example, African-American Vernacular English (AAVE) speakers need to adapt their speaking style and/or apply less natural speech patterns, like code-switching, to cater for a CUI's limited capabilities [44]. These technological and design limitations hint at a lack of user-centred approaches at the end-user's expense. This is highlighted in work done by Blair and Abdullah [5], who identify the challenges of deaf and hard-of-hearing individuals using smart assistants in daily contexts. If marginalised groups are neglected in development processes, many will be unable to use the resulting systems [5, 44], posing unethical consequences [45]. This work considers different perspectives and, in part, investigates how marginalised groups can be better addressed in the design phases of CUIs.

2.2 Summary of Dark Patterns

Research on dark patterns has illustrated a wide range of applications where problematic design occurs [37, 64, 69, 101] while demonstrating a limited focus in the scope of CUIs [19, 79]. Adjacent literature guides our work to understand design areas that require further attention from practitioners to avoid unethical design in CUIs. After Brignull coined the term "dark patterns" in 2010 [12] and initialised a first set of twelve dark patterns, numerous researchers have set out to describe unethical practices in GUI interfaces. In 2022, however, Brignull promoted the term "deceptive designs" instead of dark patterns to address the risk of the term being racially misappropriated [12]. Recently, the ACM Diversity, Equity, and Inclusion Council [1] added the term to their list of controversial terminologies. However, critique against the term "deceptive design" has been voiced as it is deemed too vague to describe the scope and precision of its predecessor, lacking reference to pattern language [2] while linking "dark" not to malicious intent but something that is hidden [76]. Another argument is that such interfaces do not only deceive but also obstruct, coerce or manipulate users [64, 76]. We acknowledge that, at the time of writing this paper, no perfect term exists to convey all unethical and problematic issues. As the community seeks a better term [39], we opted to use the term "dark patterns" in this work to maintain consistency and continuity with previous research.

Today, the dark patterns terrain has widened throughout numerous domains, including, but not limited to, mobile applications [24], e-commerce [37, 63], and social media [43, 64, 68, 69, 88]. Collectively, dark pattern research has produced a taxonomy of individual design strategies (for an overview, we refer to Gray et al.'s ontology [38]). In an attempt to capture the diverse nature of dark patterns, Mathur et al. [63] describe five characteristics: *asymmetry*, *covert*, *deceptive*, *hides information*, and *restrictive* (see Appendix A, Table 4 for a more detailed overview of the characteristics). The authors ground their work on prior dark pattern research and their study based on a large-scale investigation of over 11,000 shopping websites. Each characteristic is introduced as a dimension where dark patterns manifest, while a single dark pattern can contain aspects of multiple characteristics. Moreover, characteristics are described through distinct mechanisms that restrict informed decision-making (e.g. user interfaces promote specific choices over others or hide relevant information from the user). Our work utilises these characteristics during the interviews to learn about interviewees' views on potential unethical designs in CUIs.

Despite this growing field of research, most work focuses on GUI artefacts, and only limited work considers dark patterns in the context of CUIs. Surveying CUI users on twelve scenarios involving voice-based systems, Owen et al. [79] establish a ground for future work by highlighting the importance of considering dark patterns in this environment, which finds further support by the provocation by Mildner et al. [66] from the same year. The need for further research is also mirrored in a recent review by De Conca [19], who showcases the presence of known dark patterns in speech-based CUI systems. Thereby, the work not only highlights potential differences of dark patterns in CUIs compared to their GUI counterparts. The work further concerns necessary regulatory actions, emphasising the

need for more tailored considerations, as contemporary regulation mainly concerns GUI dark patterns.

Importantly, studies have repeatedly shown difficulty among users to recognise and identify dark patterns sufficiently [24], or with low accuracy [7], even if provided information about dark patterns [7, 67]. Users' inability to safeguard themselves places them in a vulnerable and exploitable situation that requires particular design considerations. In this context, it would be beneficial for practitioners to be guided by ethically aligned guidelines, which could aid in the development of user-centred systems that potentially empower users to make informed decisions.

Setting out to protect end-users of technologies, we aim to increase our understanding in this field by consulting practitioners, researchers, and frequent users. Led by the momentum of this discourse, we aim to draw attention to ethical caveats in CUI design as the technology becomes more ubiquitous. As practitioners often work under diverse ethical constraints, leading to the unconscious or unwilling deployment of unethical design [35], we aim to provide concepts to avoid implementing dark patterns to begin with.

2.3 Guidelines & Best-Practices

Spanning work including Nielsen's ten usability heuristics [73] and Friedman et al.'s Value Sensitive Design [31], HCI has produced a range of important frameworks, guidelines, and best practices to aid practitioner' efforts. Yet, often recorded frustration and negative experiences of CUI users indicate a lack of application or applicability of these aids in this domain. This issue has been addressed by various researchers [46, 56, 70]. Ghosh et al. [33] highlights a lack of accuracy for the system usability score [13, 14] — a prominent measure of subjective perceptions of usability of systems – when used to evaluate CUIs. To address this problem, recent work investigates the possibility of transferring concepts known to work for GUIs and other interfaces toward the context of CUIs. Langevin et al. [56], for instance, adapt Nielsen's heuristics within the context of CUI interaction. Similarly, Klein et al. [51] revise the widely used UEQ questionnaire [58] and extend it with scales for response behaviour, response quality, and comprehensibility to mirror aspects important for the user experience of CUIs. Indeed, various reviews of CUI research note the lack of validated measures of user perception as an ongoing problem that calls into question the reliability of this body of work due to a lack of continuity in how concepts are defined [16, 52, 89]. To date, there has been only one validated scale of this nature available to CUI researchers, known as the partner modelling questionnaire (PMQ) [25].

Still, guidelines tailored explicitly to CUIs are scarce. Additionally, they often exclude unique requirements of vulnerable groups who could especially benefit from interactions with hands-free and speech-based devices, as is the case for users with, for example, visual impairments [9]. Through our established themes and our framework, our work aims to provide some mindful guidance for practitioners and researchers to utilise a user-centred approach and reach a variety of different user groups.

3 METHOD

This study aimed to gain insights regarding CUI-specific ethical caveats and identify ways unethical practices manifest within CUIs. We, therefore, conducted a total of 27 semi-structured interviews split between three groups. The first group includes researchers focusing on CUI-related topics (N=9). The second cohort comprises practitioners who work in industries developing CUI technologies (N=8), whilst the third encompasses frequent users of CUI systems (N=10). The interviews were conducted online via video conference tools, which allowed for the recruitment of participants with an international scope.

3.1 Interview Protocol

The interview consisted of two parts. The first part focused on participants' general experience with CUIs. The second part is based on and inspired by Mathur et al.'s dark pattern characteristics, promising interesting insights into unethical design strategies by describing design choices and mechanisms that prohibit informed decision-making. As Mathur et al.'s definitions were constructed to convey similarities between certain groups of dark patterns, the original definitions were neither created for an interview context nor designed to account for CUI-based interaction. To address these limitations, we adapted the original questions to foster more relevant answers from our participants. While this enabled us to learn about participants' views on dark patterns in this context, we also added three questions targeting specific situations and demographics to gain a deeper understanding of circumstantial issues. The full interview protocol and the five dark pattern characteristics according to Mathur et al. [63] are included in the Appendix A. During the interview, participants were conditionally prompted to think about text-based and voice-based systems, both concerning their actual experiences and hypothetical future scenarios they might envisage.

3.2 Participants

In total, 27 participants were recruited for interviews. Researcher and practitioner cohorts were recruited from the authors' collective professional network and word of mouth, enlisting academic and commercial researchers, designers, and developers whose work focuses on CUIs. The third group, frequent users, were recruited via the online platform *Prolific* [61]. Recruitment criteria were used to ensure participants in this cohort were at least 18 years of age and used CUIs at least on a weekly basis, though no particular CUI system was stipulated. All participants were informed about the nature of the study, what participation involved, and their data rights before being asked to provide informed consent. Participants recruited via Prolific were rewarded with a £10 honorarium. This is in keeping with suggested hourly rates for ethical payments for research involving crowdworkers [57]. Researchers and practitioners participated voluntarily and were not furnished with an honorarium. Table 1 presents a full overview of all recruited participants. The interviews lasted an average of 42:30 minutes (SD = 11:29). Interviews with the researcher cohort took the longest (mean=48:00, SD = 08:48), while interviews with those held practitioners were slightly shorter on average (mean=44:47, SD = 13:50). The interviews conducted with frequent users were the shortest (mean=36:03, SD = 08:28).

Researchers. Eight academic researchers were recruited for our study (three female, four male, one preferred not to disclose their

Participant Table

ID	Age	Gender	Country of Residence	Occupation	Years of Experience
			Researcher		
R1	38	male	USA	Professor	16
R2	27	female	USA	Professor	11
R3	35	male	USA	PhD Candidate	10
R4	33	female	South Korea	Assistant Professor	5
R5	69	male	United Kingdom	Professor	48
R6	32	male	Germany	Postdoctoral Researcher	12
R7	30	female	Switzerland	PhD Candidate	4
R8	30	not disclosed	Ireland	PhD Candidate	5
Mean	= 36.75			Mean	= 13.88
SD	= 13.46			SD	= 14.4
			Practitioner		
P1	56	male	United Kingdom	Chief Science Officer	15
P2	49	female	USA	Manager	21
P3	40	male	USA	Educational Research	2
P4	32	male	UK	Tech. Consulting Manager	6
P5	40	female	Brazil	Research	9
P6	38	female	USA	Conversation Designer	10
P8	64	female	USA	Communication Consultant	35
P9	42	female	Ireland	Designer	15
P10	58	male	USA	Digital Business Executive	25
Mean	= 46.56			Mean	= 15.33
SD	= 10.74			SD	= 10.28
			Frequent User		
ID	Age	Gender	Country of Residence	Current Occupation	Highest Level of Education
F1	26	female	Mexico	ESL Teacher	Undergraduate
F2	26	female	South Africa	Customer Service Rep	Postgraduate (or higher)
F3	30	male	South Africa	Media Analyst	Secondary/Vocational
F4	33	male	South Africa	Construction	Secondary/Vocational
F5	24	male	South Africa	Student	Undergraduate
F6	25	female	South Africa	Student	Postgraduate (or higher)
F7	57	male	United Kingdom	Retired	Postgraduate (or higher)
F8	45	female	Italy	Remote Freelancer	Secondary/Vocational
F9	22	male	Poland	Frontend Developer	Undergraduate
F10	29	female	Mexico	Healthcare Entrepreneur	Undergraduate
Mean	= 31.7				
SD	= 11.02				

Table 1: This table presents our interview participants in three cohorts: researcher, practitioner, and frequent users. Notably, we did not have a participant P7. The associated participant cancelled the interview after IDs were already given to all participants. To avoid confusion, we retained our initial structure.

gender), four of whom were professors, one a postdoctoral researcher, and three who were PhD candidates. When conducting this study, the researchers' average age was 37.0 (SD = 13.0), and their average years of experience researching CUIs was 13.9 years (SD = 14.4).

Practitioners. Nine practitioners participated in our interviews (five female, four male) with roles in conversation design (2), executive management (4), research (2), and consultation (2). Practitioners' average age was 47.0 years (SD = 11.0), and they had an average of 15.3 years (SD = 10.3) of experience working in this space.

Frequent Users. Finally, we recruited ten individuals for the frequent users' cohort (five female, five male; mean age = 31.7 years,

$SD = 11$). To qualify as frequent users in this study, participants of this cohort had to use CUIs at least once per week basis. However, we did not require participants of this group to have long-term experience as the technology is still relatively young. We also asked them about the kinds of devices they used most often. Seven stated that they mostly accessed CUIs through their smartphones, whilst the other three mostly used smart speakers. CUIs used included Amazon's Alexa, Apple's Siri, Google Assistant, Microsoft's Cortana, Samsung Bixby, and text-based chatbots from online services. Six said they used CUIs daily, while four stated using CUIs more than once a week. The highest level of education among this cohort was: undergraduate (4); secondary/vocational (3); postgraduate or higher (3). Participants' occupations span construction (1), customer service (1), front-end developer (1), media analyst (1), medical laboratory scientist (1), self-employed healthcare entrepreneur (1), students (2), and teacher (1).

4 THEMATIC ANALYSIS

After completing all 27 interviews, we transcribed and prepared the recorded material for analysis, de-identifying any traceable or personal information. Data transcription was carried out by a professional UK-based service provider. Based on these data, we conducted a reflexive thematic analysis [11], which was divided into four phases: familiarisation; code generation; construction of themes; and revising and refining themes.

4.1 Positionality

Authors of this work lived, gained education, and worked in Central Europe most of their lives, with WEIRD (Western, Educated, Industrialised, Rich, and Democratic) [45] backgrounds. Their research backgrounds contain expertise in design, linguistics, computer science, and psychology, with scholarly work oriented toward social justice and well-being topics. The interview transcripts were coded by four authors who have previously engaged in scholarly work in human-computer interaction, computer science, and psychology, each contributing more than three years of experience in their respective domains. At the time of conducting the study, no personal or professional conflicts with CUI systems or their developers occurred among the authors. Participants of the researcher and practitioner cohorts were recruited through professional networks with some personal relations. Where an author had a rather close relationship with a participant, we ensured that another author would conduct the interview instead. Frequent users were recruited through Prolific [61], where no previous relationships existed. Focused on ethical considerations in CUI design, this work is partly inspired by Mathur et al.'s [63] dark pattern characteristics to analyse problematic CUI interactions and propose counter-measures. To conclude, we acknowledge potential bias based on our cultural, academic, and personal backgrounds.

4.2 Coding of the Transcripts and Identifying Themes

For the first round of data analysis, we selected two interviews per participant group to generate an initial set of inductive codes. Following advice from Braun et al. [11], two researchers coded these six interviews, discussed their strategies and results, and drafted an initial codebook encompassing 65 codes. Through axial coding, the codebook was then reduced to 46 codes. All interview transcripts were then split among four authors. Each interview was thus coded by a single author following Braun et al.'s proposed methodology [10]. Although coding was carried out independently, authors did meet to discuss the procedure once before the coding, once after half of the interviews had been coded, and one last time after coding had been completed. These discussions ensured coding strategies were aligned between authors, including resolving any issues authors may have had applying specific codes [11] and discussing any potential new discoveries in the data. Due to the collaborative nature of this process, indices of inter-rater reliability [11] are not provided.

Four authors then went on to identify themes for each cohort independently. Based on the codes, quotes, and annotations, each researcher applied affinity diagramming [6] to develop early iterations of these. During this process, these first themes were evaluated and refined by iteratively checking each for applicability to interview material. They were then discussed among researchers to establish agreement. Here, we followed Braun and Clarke's theme mapping [10] to collapse similar themes. If a theme was adapted, it was again revised against the whole dataset and other themes before being accepted. This last phase resulted in the construction of five themes and a corresponding ethical caveat for each.

5 FINDINGS

In this section, we describe the constructed themes based on reflexive thematic analysis. We highlight these themes across each of the interview cohorts individually. Notably, a degree of overlap is to be expected as themes tend to be intertwined, often in a supportive fashion. In total, we synthesised five high-level themes summarised in Table 2, where each theme is listed next to a statement for a particular ethical caveat. Additionally, we formulated guiding questions that refer to the ethical caveat to support the design of CUI interactions. Before each theme is addressed in detail, we present a high-level overview of our findings. The outline for each theme follows the same structure: a description followed by the perspectives of each of the three cohorts interviewed (researchers, practitioners, and users).

5.1 Building Trust and Guarding Privacy: Operating Extrinsic and Intrinsic Factors

This theme spotlights extrinsic and intrinsic challenges that result in trust and privacy deficits in CUI interactions. Across cohorts, a reoccurring theme echoed fears of untrustworthy handling of personal data, kindled by prior negative experiences and the reputation of companies and practitioners. A call from researchers to increase transparency to bridge users' concerns was underlined by users' mention of their own safeguarding strategies, which limit interactions to basic functionalities and do not require them to disclose private information. Practitioners acknowledged these problems but noted counterintuitive regulations and design limitations obscuring access to settings rooted in speech-based interactions.

Researcher. Researchers interviewed were aware of trust-related problems of CUIs, suggesting a need for greater transparency, "*it's really difficult for users to know how data travels across different*

Theme	Ethical Caveat	Guiding Question
Building Trust and Guarding Privacy	Users feel vulnerable to use CUIs, posing a need for CUI developers to prioritise transparency and control over data handling.	Does the system/interaction provide accessible and transparent information about personal data with easy control thereof?
Guiding Through Interactions	Guidelines and frameworks need to educate developers to develop accessible CUIs that empower users with diverse technological literacy to confidently interact with available features.	Does the system/interaction adequately inform users about its technical capabilities to enable full utilisation of its features?
Human-like Harmony	Anthropomorphic features should be implemented with care and in line with a CUI's capabilities to support intuitive and authentic interactions, preventing unrealistic expectations.	Does the system/interaction clarify the presence of anthropomorphic features to avoid misconceptions and unrealistic expectations?
Inclusivity and Diversity	The development and design of CUI interactions need to consider individual needs and characteristics of users, especially marginalised groups, ensuring equitable CUI interactions.	Does the system/interaction cater towards users with diverse needs, potentially through alternative interactions where otherwise inaccessible?
Setting Expectations	CUI capabilities should avoid deceptive interactions and, instead, be transparent to users to prevent frustration and mistrust.	Does the system/interaction handle user prompts truthfully, clarifying the scope of its capabilities to provide realistic expectations?

Table 2: This table summarises the five identified themes and design questions per ethical caveat.

platforms" (R4). Connecting this problem to the "*reputation of the companies behind each device*" one participant said that "*there's very little transparency [...] about how that data is being used and processed*" (R3). Another link is drawn to deceptions based on human-like features with which "*we are undermining the trust when we are projecting something which isn't real*" (R5). Meanwhile, there are "*ethical issues with data sharing and things like that, but [users] care more about: 'Can I get to what I need to get to fast enough?'*" (R4). Looking at users' uncertainty from a technological lens, another researcher discussed how "*people didn't know [the CUI] was listening to them all the time [...]. How does it know when I say 'Hey Alexa' or 'Hey Google' [...]? Is it already listening to me?*" (R2).

Practitioner. Practitioners also reflected on privacy and trust issues. The current situation users are in when engaging with CUIs was described as "*completely wild west, talking to a black hole*", because "*sometimes there is a lack of transparency, you just don't know what they're doing with your information and that's a problem for a lot of people*" (P8). CUI's present some inherent difficulties when users try to access their privacy settings: "*That info does seem to be kind of buried [...] you have to click in three or four different menus to get to the privacy stuff and I don't think it's explained*" (P6). It is later reflected that users "*don't get any kind of introduction to the app itself or the onboarding is very light*" (P6). Regarding a lack of technological literacy among users to understand and trust their CUIs, one practitioner asked the question: "*How do you prevent people from overhearing my voice and trying to imitate me, how does that thing know that it's not really me?* ". They went on to provide an answer: "*There's a simple explanation. It's too smart, it knows the difference. The technical explanation is, it gets into how it analyses your voice*" (P3).

Frequent User. Being the affected cohort, users mentioned various concerns regarding their privacy and why they hesitate to trust CUIs. One participant was "*not fully comfortable disclosing [their] personal details even if it's a virtual assistant*" (F5). Reflecting on potential trust connected to anthropomorphising CUIs, they later stated that "*it feels like you're speaking to a human and you're giving them your personal data*" (F5). Talking about giving consent to things they never fully read, one participant said: "*I don't think I've read them all, I just think that I know what I'm accepting. So I don't know any of the consequences that I might have with it*" (F3). Users are "*expected to know the risks to [their] privacy, to data, to enter into the real world. [Users] are expected to know it in advance.*" (F7). Some worry about exposing critical information publicly, for instance regarding their financial accounts: "*It puts me in danger of whoever that's around that can hear how much money I have. Maybe I'm at risk of being robbed*" (F2). Despite their concerns, the interviews outlined aspects of a privacy paradox. "*Today, there is no alternative [to commercial CUIs]. [...] It is bad, but I don't see [an] alternative*" (F8).

5.2 Guiding Through Interactions: Overcoming Knowledge Gaps

This theme illuminates the importance of providing users with informed guidance about possible CUI interactions, as a lack of technological literacy and limited experience results in difficulties in accessing available features. While researchers noted that

research findings need to be better aligned with industry development, practitioners echoed a desire for further guidelines.

Researcher. Researchers reflected on a lack of communication between stakeholders, as well as peculiarities of CUI interaction that lead to unique restrictions. A level of trust and limited collaboration was emphasised by researchers, who suggested work in academia seems to "*make very little impact on what big companies do. [...] If it is profitable, they're going to do it*" (R4). Another researcher speculated that there was a disconnect between the communities that design the components required to build CUIs, and the community of HCI researchers who explore and develop theories that explain how users interact with CUIs: "*What I see is a bunch of different technology CUIs just being demonstrated, and then in our particular space [design research] we've got toolkits, and the toolkits define what's possible. [...] I'm struggling really to understand how come there's this disconnect between these communities?*" (R5). It was noted that the communication between practitioners and users is also problematic "*if users don't care and only developers do*" (R4).

Practitioner. Practitioners again referred to a lack of best practices guidelines that would help them create better products. "*[...] There's very little that's actually focused on conversational user interfaces... even to the point that we don't have heuristics really. [...] To create these heuristics for CUIs, that hasn't been done, so there's this massive gap out there.*"(P6).When contrasted against comments by researchers, this further highlights the disconnect between these communities. Practitioners also highlighted the lack of standardised guidelines for users when it comes to how best to interact with these kinds of systems. They suggest users are essentially left to rely on what they know from graphic counterparts: "*For voice stuff, what's missing, there is no standard, like you know when you go to a website, and there's that little hamburger menu, those three lines.*" (P6). Instead, users have to compare their experiences with other, similar systems in the hopes that they operate the same way; "*[...] it's still fairly new and [...] design best practice is still kind of evolving. [...] Maybe if there was some more standardisation in how the experience works between different companies or between different interfaces, people will get more used to it*" (P4).

Frequent User. Frequent users echo the idea that the burden of figuring out the limitations of a speech agent falls largely on themselves. "*Most of the time you have to understand Google Assistant and what its limitations are when you ask it to do specific tasks* " (F5). Similarly, users "*have to use the correct terms*" and "*learn to adapt to [their] language set*" (F7). In reference to using a CUI on their smartphone, one participant reflects on the limited information available to them: "*You'd have to actually sit on your phone and actually test out everything to see what works and what doesn't*" (F3), whilst another suggests they draw expectations from smartphone experiences because "*they are able [...] and that's what we want from our home devices*" (F8).

5.3 Human-like Harmony: Providing Authentic Anthropomorphism

This theme describes the importance of implementing an appropriate amount of human-like features so that a system can support intuitive interaction without eliciting false beliefs that may leave users feeling deceived. Finding the right balance between humanness and technological limitations is challenged by the phenomena of anthropomorphism. That is the tendency to attribute human characteristics to non-human objects that most people engage into some degree [98] from early childhood [82]. Anthropomorphous behaviour appears significantly heightened in dialogue with technological devices endowed with gendered voices and names [34]. Other influences that might encourage anthropomorphous behaviour in this context include: expectations for social affordances implied by representations of speech interfaces in media and advertising [70]; the fact that these systems conduct tasks typically carried out by humans using human language [72]; and that language use itself might be inherently social and agentic [29, 48]. When developing CUIs, practitioners should conscientiously navigate the amount of human-like features to avoid inadvertently manipulating our bias for anthropomorphic characteristics.

Researcher. Researcher interviewees provided thought-provoking ideas to enhance transparency around the ontological nature of speech agents and demonstrated awareness of potential ethical concerns associated with endowing CUIs with human traits. Interviewees suggested CUIs should be designed in such a way "*that people are not fooled into thinking it's a social entity when it's just a machine*" (R7), and that artificial voices should be designed to enable users to differentiate between humans and machines easily (R1, R5, and R8). Some even went as far as to state that "*From an ethical standpoint [a CUI] should announce it's a machine*" (R5). This would undoubtedly provide clarity for some users who experience confusion due to "*not knowing what social situation [they are] in*" (R1). Overall, researchers agreed there should be increased transparency for users when engaging with CUI technologies.

Practitioner. The need to increase transparency was also acknowledged by individuals from the practitioner cohort, who suggested CUIs "*should flag the fact that it is a machine talking*" (P9), arguing for the introduction of a "*watermark [...] that indicate[s] that [the CUI] is fake*" (P8). Although "*some level of humanisation*" was thought to cause no harm, this could be context-dependent (P9). Practitioners felt the responsibility lies in "*how the bot identifies itself [...,] how human-like the conversation is*" and "*the visual representation"* (P9). However, there is a risk for users who "*still got [...] attached to [CUIs]*" (P2), even though the interaction was not particularly "*emotional or personal*". In this regard, a counterargument mentioned that "*people are lonely and [...] there is a benefit to having a virtual companion in some ways*" (P2). Another practitioner did not see any issues endowing CUIs with human-like characteristics, believing people can still easily distinguish between humans and a device that "*just speaks like a human*" (P10).

Frequent User. Many of the frequent users' observations echoed facets of the other cohorts' observations. Demonstrating the kind of bond some users are said to have with their devices, one participant admits that "*it really feels like you have your friend walking with you in your pocket*" (F5) when referring to the virtual assistant built into their mobile phone. Similarly, another participant finds that "*especially voice assistants [...] let users believe that they are human or, to an extent, friendly*" (F6). It was felt that a lack of awareness may lead users to "*interact [with CUIs] as [they] would with a human*

without knowing that it is not" (F1), with one participant recalling how elderly family members "treat the machine like a human being" and use it "*with courtesy*" (F10). However, some also highlighted limitations when interacting with certain CUIs, particularly those implemented in customer service. Expecting a human to help with a problem, the limited responses and understanding of queries exhibited by some systems were said to create frustration and exacerbate a desire to "talk to a person, a real person" (F10).

5.4 Inclusivity and Diversity: CUIs in the Wild

This theme delves into the unexpected design and interaction challenges emerging when CUIs are introduced to diverse user groups with distinct characteristics and needs. Participants spotlighted several groups susceptible to poor design choices or technical limitations within CUI interactions. They emphasised a regression towards the mean, acknowledging how CUIs are designed with an "average" user in mind, leading to the "othering" [65] – the marginalisation or exclusion of certain people – of individuals and groups that do not fit this profile. This encompasses users with stronger accents, colloquial dialects, second language users, people with deficits in speech or cognition, or users with lower technical literacy.

Researcher. Researchers note that in designing CUIs, one has to be aware of varying user abilities and behaviours; "*So certainly you have users with diminished capacity that are going to be particularly vulnerable.*" (R1). This was seen as particularly problematic around consent "*if you know that certain groups of users can't give their informed consent, yet they're still using certain technologies, then you'd have to be more careful.* " (R4)

Practitioner. Practitioners also suggested, "*[T]he use of these systems for vulnerable groups is problematic.*" (P1), whilst also identifying issues faced by minority groups: "*in terms of demographic, in terms of like accent, people who have a strong accent because the language that they're trying to interact with isn't their native language. Basically, anyone who doesn't fall into the kind of the centre of the bell curve, I expect, is more susceptible [to problematic CUI design]*" (P4).

Frequent User. Frequent users identified a number of groups they envision encountering issues with CUIs; "*people who are not technology literate or digitally literate, they might be vulnerable.*" (F5), "*there are certain people who are very gullible to [...] information in general, anything that is said online they absolutely believe.*" (F2). Frequent users also offered anecdotal examples of individuals that have been [other-ed] by CUIs, explaining that " *a user would need to have a fairly good memory to be able to use [CUIs] effectively. I have a friend who is severely disabled, and she just cannot use it because she can't remember the commands.*" (F2). Another participant shared that " *a friend of mine, her son is autistic [...] he has trouble with her Alexa, [...] sometimes he knows the answer but he'll ask the question for information. He'll expect it to conform exactly to what he knows, even using the correct words and he gets so frustrated when it doesn't*" (F7).

5.5 Setting Expectations: Transparency to Mitigate Frustration

This theme emphasises the pivotal role of transparency when designing CUI interactions to set realistic expectations. Unintended device reactions can result in disappointment and frustration in the user when a device appears more capable than it actually is. To deliver users an optimal experience, commercial incentives should be aligned with human-centred practices of HCI to avoid ethical concerns tied to deceptive and manipulative interactions. Instead, transparent and evident CUI capabilities should empower users to use the system autonomously and easily.

Researcher. The importance of setting the right expectations for what systems are capable of was noted among researchers for offering users a better experience that is more aligned with reality. Here it was stated that "*functionality expectations are not set right*" (R2) for contemporary CUIs, with introductions and manuals either not being readily available or ignored by users. Additionally, "*people bring their previous experiences to an interaction when they start*" (R7), suggesting users with prior knowledge have an advantage over those who are new to CUIs. If true, this highlights a lack of onboarding for novice users, with prior knowledge biasing experiences for frequent users. In this regard, "*lack of transparency [...] really makes it easy to misunderstand what the system does, or is capable of doing*" (R2).

Practitioner. Practitioners suggest a lack of best practices or guidelines for how and when to communicate a systems' features to users; "*getting across to the user what the system can and cannot do is one of the most difficult problems*" (P2). Current commercial CUIs are limited to presenting information synchronously on request. As the same metaphors working in GUI environments do not translate readily to CUIs, "*discovery [of functionalities] is very difficult right now because there's no directory to help us find it*" (P8). While multi-modal chatbots can mitigate certain problems, voice assistants without integrated displays cannot rely on the same design philosophies. "*specifically for voice stuff, what's missing, there is no standard*" (P6). It was further argued that "*when [CUIs] can only answer a fraction of questions it is inherently putting the burden on the user to rephrase their question over and over*" (P6). This difficulty is increased "*if there are a lot of options, it's probably more of a challenge to make people aware of the choices that they have*" (P4). In relation to human-like features, how CUIs "*are designed can make them seem more intelligent than they are*" (P9), creating unrealistic expectations that will eventually cause frustration in users.

Frequent User. Multiple users expressed frustrations they experience when using their CUIs; if a CUI "*fails to give you what you want it to do, you end up having to do it yourself*" (F5). An initial problem stems from raised expectations through advertisement of commercial products, where "*everything [...] runs smoothly [...] creat[ing] that expectation in your head as that's how it works*" (F1). Another participant explains that if "*you don't know what [the CUI] has heard [...], you have to go to the app*" (F10) as the device is incapable of providing insights into why errors occur. Limiting their expectations, one user adapted their behaviour to "*only [use voice-based systems] for easy communication because [...] they are not set yet for more complicated things*" (F8). Another user discussed the

importance of good memory, which "*users need to have [...] to be able to use [CUIs] effectively* and that they "*need to have read the literature for the [CUI] to know the limits*" (F7).

6 ADDRESSING EXPECTATIONS: THE CUI EXPECTATION CYCLE

Based on our findings, we constructed the CUI Expectation Cycle (CEC), a framework that practitioners and researchers might use to understand and serve users' needs better. To this end, the CEC aligns user expectations with system capabilities. It aims to enable successful CUI interactions that meet the goals of its users, mitigating unethical design strategies and voiced frustrations of our participants in line with related work [44, 100]. The CEC incorporates the themes identified in our qualitative analysis, including garnered ethical caveats. Further, it partly builds on existing frameworks – Norman's *Action Cycle* [75] and Oliver's Expectation Confirmation Theory [77, 78] – to inform about the variables that shape user expectations.

6.1 Constructing the CEC

Drawing from dark pattern literature, our interviews confirm the findings of existing CUI research while providing a detailed overview of the interplay of design challenges. While identifying our themes, we assessed their interrelationships and noticed a strong gap between users' expectations and CUI capabilities, as described by our participants. Focusing on this gap, we distilled the themes into a framework while consulting related work. To this end, the CEC (see Figure 1) links core issues of CUI design with a particular focus on anthropomorphism [25, 26], deception [19, 79], inclusivity [44], and trust [62, 93]. Moreover, the CEC builds on two established frameworks. On the one hand, it utilises Norman's Action Cycle [75] by adopting users' execution and evaluation gulfs; in the case of our framework, we refer to them as *bridges* as they help to cross the Expectation Gap. We were also inspired by Oliver's Expectation Confirmation Theory [77, 78], a cognitive framework that describes customer expectations and subsequent (dis)satisfaction with purchased goods. Supported by these widely accepted frameworks and related literature, the CEC focuses on creating realistic expectations extended by carefully considering ethical caveats to avoid problematic design in the area of CUIs. Ideally, this framework enables the development of transparent and trustworthy CUI interactions, allowing satisfying and successful user engagement.

6.2 The Elements of the CEC

Figure 1 illustrates the CEC and how the five themes relate to individual elements of the framework. Dividing user (at the top of the CEC) and CUI (at the bottom), the critical element of the CEC is the *Expectation Gap* (at the centre of Figure 1). However, the *Expectation Gap* can be overcome through two bridges that connect the CUI with the user (the *Evaluation Bridge*, left of the centre) and vice versa (the *Execution Bridge*, right of the centre). In jeopardising these bridges, two delimiters result from problematic or unethical design decisions (the *Deception Delimiter* on the outer left side and the *Distrust Delimiter* on the outer right, each pointing toward the bridges). Above the diagram are the five themes: *Building Trust and Guarding Privacy*, *Guiding Through Interactions*, *Human-like Harmony*, *Inclusivity and Diversity*, and *Setting Expectations*. Each theme is connected to the relevant elements of the diagram through arrows with differing dotted lines and colours. Informed by our themes, the elements of the CEC encompass the ethical caveats identified as considerations for the design of CUIs. Here, we describe each element more closely, beginning from the centre while highlighting elements and connected themes in *italic* font.

The *Expectation Gap* encompasses our *Setting Expectations* theme, which reflects the different perspectives across cohorts regarding technical capabilities and design decisions. As these decisions anchor user goals, it is pertinent to design accessible interactions that address users' needs depending on a CUI's purpose and its situational context. For instance, frequent users reported frustration with the complexity in which CUIs had to be prompted, while practitioners noted how many features were left unused. Consequently, reasonable bridges need to connect both sides. The CEC includes two such bridges (*Execution Bridge* and *Evaluation Bridge*) inspired by Norman's [75] gulfs of expectation and evaluation.

To the left of the CEC's centre, the *Evaluation Bridge* starts from the CUI and reaches toward the user. It thereby includes understandable and truthful communication of a CUI's capabilities to inform the user about possible interactions and set realistic expectations. This bridge is strongly connected to the *Guiding Through Interactions* theme, expressing a need to empower users to fully understand possible interactions and their consequences.

The *Execution Bridge*, right of the CEC's centre, starts at the user's side and links it to the CUI. The bridge describes how expectations are shaped into goals a user can realise. However, a variety of obstacles can hinder users from successfully executing interactions. In this regard, the *Inclusivity and Diversity* theme addresses the importance of accommodating users' individual characteristics in the design of CUIs.

In a well-designed system, both bridges support users in making realistic assumptions about a system and using it without frustration. In a sub-optimal system, users may encounter negative experiences and designs lacking ethical considerations, which could adversely affect their well-being, as dark pattern literature indicates [64, 69]. The CEC includes two delimiters fostering the negative influences of the two bridges. The *Deception Delimiter* (on the diagram's outer left side and pointing to the *Evaluation Bridge*) has a detrimental effect on the *Evaluation Bridge*, resulting in unrealistic expectations. Similarly, the *Distrust Delimiter* (on the diagram's outer right side and pointing toward the *Execution Bridge*) adversely affects the *Execution Bridge* by decreasing users' faith in the system and its responses. The two delimiters are informed by our *Building Trust and Guarding Privacy* and *Human-like Harmony* themes as they advocate transparent and authentic interactions. The different perspectives cast across our participants highlighted how each delimiter can be intrinsic and extrinsic in nature. To illustrate, previous individual experiences impact how a user engages with a CUI (i.e. intrinsic in nature), while design decisions influence how a CUI is perceived (i.e. extrinsic in nature). Importantly, either delimiter can be the source of unethical design practices commonly found in dark patterns.

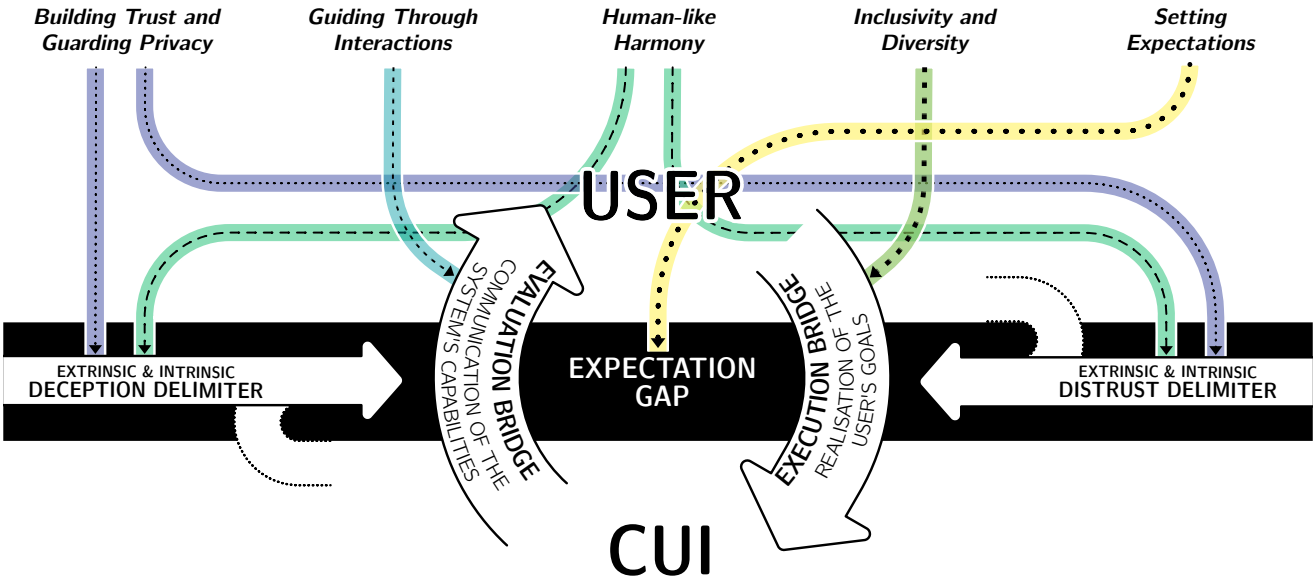


Figure 1: The CUI Expectation Cycle demonstrates how an expectation gap between user and system can be bridged by an evaluation and an execution bridge, inspired by Normans' two evaluation and execution gulfs [75]. However, deceptions can impact how users assess the system's capabilities, resulting in unrealistic expectations. Similarly, distrust limits users' faith in the system and its responses, influencing their execution of actions. Both delimiting factors can be of extrinsic as well as intrinsic nature. The diagram further features the five themes and connects them to respective sections of the CUI Expectation Cycle.

6.3 Using the CUI Expectation Cycle

Although previous research has repeatedly addressed a gap between users' expectations and CUI devices [26, 56, 62], our participants, especially practitioners, requested design guidelines to overcome related issues. The CEC responds to these demands through a user-centred approach, embedding ethical caveats to avoid problematic designs. While the CEC could be used as a standalone framework, we believe it is best utilised next to contemporary efforts in CUI research to provide additional design considerations and constraints as alternatives to GUI-related best practices. To support practitioners' work, we created questions derived from the ethical caveats that allow tracing of related limitations (see Table 2). Following the CEC (Figure 1), practitioners can consult each theme in the corresponding design areas to ensure that ethical caveats are respected.

Generally, CUIs promise benefits for people with diverse abilities, like visual impairments [9]. However, studies demonstrating difficulties among marginalised demographics [44] suggest that these benefits are not exhausted, addressed in our *Inclusivity and Diversity* theme. CUIs, such as Amazon Alexa, often require additional GUI input to control device settings or personal data, restricting access. To cater to diverse user groups, practitioners could consider including alternative interactions that allow simple execution of planned goals. Asking our guiding question – "Does the system/interaction cater towards users with diverse needs, potentially through alternative interactions where otherwise inaccessible" – would remind practitioners about the importance of accessibility in their systems.

Taking the *Human-like Harmony* theme, for example, anthropomorphism can impact the deception and distrust delimiters. To mitigate deriving consequences, chatbot systems should clarify if no human is in the loop and state the scope of possible prompts. Design borrowing from other chat interfaces can cause additional confusion but could easily be avoided through a distinctive interface design. Similarly, anthropomorphic voice-based CUIs may confuse some users, particularly those with little technical literacy, leveraging their expectations. Appropriate voice design could clarify a CUI's artificial nature, clarifying related misconceptions. Asking our guiding question – "Does the system/interaction clarify the presence of anthropomorphic features to avoid misconceptions and unrealistic expectations?" – helps to unravel otherwise deceiving design choices.

In line with prior work [56], our findings demonstrate that simple adoption of GUI best practices do not readily translate to CUI interactions as heuristics differ. While GUIs utilise graphical icons, buttons, or links [79], with some exceptions, CUIs' general input lie in uttered or typed-in commands. Though CUIs, including a GUI, can use these interactions, voice-based devices often require prompts to be spoken only. In most current CUIs, however, interactions require users to remember the correct prompts for the expected response, as recalled by some participants. Consequently, practitioners should account for human errors, support them better during interactions, and offer suitable alternatives. Adequate onboarding of the user is a crucial step to lower barriers and set expectations. While the CEC cannot account for malintent, we envision following its recommendations facilitates a better user experience and helps mitigate risks for users in consideration of our proposed ethical caveats.

6.4 Retrospective Application of the CEC

To assess the guiding utility of the CEC, we apply our framework retrospectively on a selection of four high-quality, contemporary papers relevant to CUI research while sharing topical overlaps with our contribution. Papers were chosen based on their research quality and rigour from esteemed conferences. Consequently, we apply the CEC to these papers' findings to illustrate its relevance while demonstrating how it can be used.

In their work, Yeh et al. [100] discuss common pitfalls of guiding users effectively in task-oriented chatbot interactions. A key finding of their study describes how lack of transparency segues into frustration. In addressing the complexity of making possible interactions evident to the user, the authors describe guidance strategies that inform users about a chatbot's capabilities adequately. In line with the CEC, the authors unravel the need to align expectations (*Expectation Gap*) between user and system through guidance as featured in our *Guiding Through Interactions* theme. Incorporated in the *Evaluation Bridge*, our *Guiding Through Interactions* theme helps overcome these difficulties.

Following the potential advantages of LLM-supported CUIs, Jo et al. [49] studied users' experience with a chatbot as an empathetic conversation partner in care environments. However, users reported a lack of emotional support, which they linked to limited personalisation and missing health history. Moreover, particular worries accompanied users' experience, for example, a fear that personalised and empathetic CUIs may reduce peoples' desire to engage in social activities. The authors identified problems and foreshadowed solutions akin to our *Inclusivity and Diversity* theme and reflected by our Deceptive and Trust Delimiters. As the *Setting Expectations* theme addresses requirements to bridge an *Expectation Gap*, our *Building Trust and Guarding Privacy* and *Human-like Harmony* themes could answer user's worries by making the system more transparent and trustworthy, increasing their experience.

Voice-based systems introduce additional layers to the interaction that require consideration. Language and speech barriers become obstacles for many marginalised groups that require special attention. Based on a study run in the U.S.A., Harrington et al. [44] showed how Black adults from lower-income households struggle when using voice-based CUIs. Participants mentioned falling back to code-switching, mentioning the fear of being misunderstood. Additionally, the researchers captured distrust toward Google Home in the health-related context of their study. These trust issues were amplified by a lack of knowledge and technological literacy, hindering users from experiencing the CUI's full potential. As a source for these problems, the authors identify a host in the ignorance of developers who neglect the cultural identity of their users. Providing practitioners with further insights, our *Inclusivity and Diversity* theme spotlights the importance of catering to diverse users, as ignorance and negligence can swiftly lead to unethical implications. As Harrington et al. describe, CUIs should foster trust and transparency (i.e. avoid *Deception* and *Distrust Delimiters*). Their study highlights the importance of careful and inclusive research design, given the WEIRD (western, educated, industrialized, rich, and democratic) [45] context in which many contemporary studies in HCI are conducted. It further demonstrates that some issues cannot be mitigated through LLMs alone.

Generally, LLMs open promising avenues for CUIs whether deployed through text or voice-based systems. Exploring these advantages to make mobile GUIs more accessible through an assistive CUI, Wang et al. [97] describe prompting techniques that allow users to engage with interfaces through speech, particularly helping marginalised and vulnerable groups to access otherwise unattainable content. Their work is an excellent example of how the CEC's *Expectation Gap* can be bridged by showcasing how conversations could be customised around special needs, thereby increasing the inclusivity of this technology in the future in line with our *Inclusivity and Diversity* theme.

Collectively, these works illustrate relevant design challenges that the CEC emphasises. The addition of ethical caveats, elevating user experience, finds support as our *Inclusivity and Diversity* theme resonates in three of the four selected papers [44, 49, 97]. Moreover, the studies foreshadow benefits gained from aligning user and CUI expectations [97, 100] while informing about system capabilities [49] and making interactions transparent to increase trust [100]. We hope our framework can guide future research and design of CUIs in a human-centred manner.

7 DISCUSSION & FUTURE WORK

Our research aimed to gain insights into the ethical caveats and design considerations faced in current and future CUI research and development. Motivated by dark pattern scholarship, particularly Mathur et al. [63] dark pattern characteristics, we conducted 27 semi-structured interviews and identified five themes providing answers to our research question. Derived from these themes, we propose the CUI Expectation Cycle (CEC), a framework to guide future work in bridging CUI capabilities with users' expectations and goals. In this section, we return to our research question and discuss the different perspectives held by the interviewed cohorts. Lastly, we point toward some directions for future work.

7.1 Revisiting Our Research Question

With the main aim of this study to identify ethical caveats in CUI design, our findings answer our research question: Which ethical caveats should be considered when designing CUI interactions, and how should they be addressed? Our themes illustrate how the unique nature of CUIs cannot simply carry over expertise readily available for other domains. To fill this gap, the subordinate ethical caveats contain design considerations that, if disregarded, compromise user experience. These include the discoverability of features and possible interactions [9, 15, 30, 47, 96], often further limited to diverse and marginalised users [44], but also deceptive designs that lead to unrealistic expectation [62] and even unwanted interactions – as seen in dark pattern literature [37, 63]. Incorporating these themes, the CEC builds on established theory to provide practitioners and scholars with means to mitigate these effects.

7.2 Perceptions Across Researchers, Practitioners, and Users

Each developed theme considers the perspectives of each cohort, highlighting similarities and differences in the diverse code groups (the full codebook is included in the supplementary material of this

	Themes				
	Building Trust and Guarding Privacy	**Guiding Through Interactions**	**Human-like Harmony**	**Inclusivity and Diversity**	**Setting Expectations**
Dark Patterns	· Bad Defaults [8] · Captive Audience [41] · Disguised Ads [12] · Disguised Data Collection [41] · Hidden Legalese Stipulations [8] · Hides Information [63, 64] · Making Personal Information Public [41] · Privacy Zuckering [12]	· Confusion [20] · Exploiting Errors [20] · Manipulating Navigation [20] · Misrepresenting [36] · Obfuscation [20] · Obstruction [37] · Roach Motel [12] · Trick [20]	· Deceptive [63, 64] · Hides Information [63, 64] · Misrepresenting [36] · Sneaking [37] · Trick [20]	· Covert [63, 64] · Disparate Treatment [64] · Making Personal Information Public [41] · Social Network Of Proxemic Contacts Or Unintended Relationships [41]	· Asymmetry [64] · Automating The User [36] · Bait & Switch [8, 41] · Forced Actions [37] · Forced Work [20] · Hidden Legalese Stipulations [8] · Interface Interference [37] · Obfuscation [20] · Obstruction [37] · Restricting Functionalities [20]

Table 3: This table summarises previously captured dark patterns within each theme.

paper). Although researchers and practitioners often shared similar expertise about the involved technologies, we noticed worries among users, illustrating a lack of the same technical understanding. Interestingly, this was further mirrored by similar views held among researchers and practitioners who – in sum – recognised a need to address users' concerns and develop CUIs that are more accessible and inclusive.

While all cohorts agree that CUIs should be more transparent about their capabilities and how data is handled, suggestions as to how this could be addressed differ. Researchers and frequent users pointed out that CUI design should reflect capabilities as anthropomorphic features overshadow limitations. Practitioners, noticing some relevance to announcing a CUI's artificial nature, argued for the benefits of CUIs offering companionship to lonely users. This is in line with previous work (e.g. [74, 94], and mirrored by users admitting emotional bonds with their devices. However, both frequent users and practitioners expressed frustration regarding user experience. Interestingly, though, they are divided as to which group should take responsibility. While frequent users are frustrated about the lack of discoverability in hard-to-navigate interfaces, practitioners argue that users are not exhausting their CUI's technical potential. This disparity is one example of the gap between users and practitioners. A potential solution was presented across cohorts when discussing the need for improved onboarding for setting realistic expectations. Here, researcher and practitioner cohorts were aware of a need to address diverse and marginalised users better, while practitioners, in particular, described current technical limitations. A general need for guidelines is voiced by both researchers and practitioners. While the latter echoed this sentiment but missed standardised guidelines for CUIs, the former also noticed a need for better communication as contemporary research aims to fill this gap.

7.3 Paving the Way for Future Work

Our decision to listen to different voices allowed us to garner insights from three perspectives with partially contrasting incentives. Although related work has outlined similar concerns raised by our participants [9, 19, 26, 80, 100], our approach to connecting the discourse to unethical design and dark pattern literature has led to novel findings, resulting in further design considerations and bridging expectations between cohorts. Future work could aim to lift the tension between user and CUI and develop interfaces that meet their expectations. Thus, increasing user experience would be a success for the practitioners' work.

Our work can serve as a foundation for ethical design in CUI contexts. As the interview questions adopted Mathur et al.'s [63] dark pattern characteristics, we reviewed each theme as a potential host for dark patterns. For this, we follow Mildner et al. [69] and draw from contemporary typologies that collectively describe over 80 types of dark patterns [8, 12, 20, 36, 37, 41, 43, 63, 64, 69, 101]. Table 3 allocates identified dark patterns, suggesting that each theme features nefarious opportunities for manipulative interactions. Notably, most dark pattern types have been described in GUI contexts. Introducing this body of work to CUI research offers valuable insights. However, our themes foreshadow additional patterns unique to CUI interactions, extending previously described instances in CUI contexts [19, 79]. As the technology is still in its relative infancy, understanding unethical practices early can present an important head start in protecting users. Future work could build on our findings to gain a better understanding of the underlying technologies of CUIs that could be exploited to deploy novel types of dark patterns.

8 LIMITATIONS

Although we were careful when designing and conducting this study, our work has limitations. The interviews were conducted using an online format, which, in individual cases, led to connectivity problems, prolonging some interviews. Although we ensured

that all interviewees were asked the same questions, technological obstacles may have influenced participants' responses.

Sample size limitations should also be acknowledged, with a relatively small sample representing each cohort. That said, steps were taken to ensure cohorts were relatively balanced in size and gender and to ensure practitioners and researchers represented a relatively broad range of backgrounds and experiences. The size of each cohort sample is also common to qualitative research in the field, as similar qualitative studies estimate high saturation of codes within 7-12 interviews [42]. Nonetheless, this may pose limitations concerning the breadth of opinions expressed, which may not be representative across professions. Additionally, many practitioners have experience in research areas with multiple participants holding a Ph.D. For frequent users, we were mainly interested in people who use CUI devices weekly. However, we acknowledge that we did not further inquire about the duration of participants' interactions with CUIs. Moreover, the fact that participants could afford devices may suggest a selection bias in this sample. While we aimed to sample frequent users from across the world, the size of this study hinders a representative sample. We used Prolific [61] to recruit matching participants, where we had to trust the platform and participants' self-evaluation to meet our criteria.

With regard to our themes, we recognise two limitations. First, participants drew heavily from their experience with voice-based interfaces. While we prompted each interviewee to consider both voice-based and text-based CUIs, replies from frequent users indicate insufficient experience with the latter. We were careful to frame our themes around all kinds of CUIs, but the lack of chatbot experience among the frequent user group may have introduced a bias toward voice-based systems. Second, not every design issue in our themes necessarily implies malicious intent. The current discourse accompanying dark patterns and malicious designs discusses whether the malevolent intents of practitioners are relevant or if any potential harm suffices to fulfil its criteria [40, 43, 76]. Future work could address this gap to study potential links between intents and dark patterns to aid regulators and create guidelines for ethically aligned user interfaces.

9 CONCLUSION

Recent interest in HCI has raised awareness of unethical design in technologies. In this paper, we identified ethical caveats related to CUI technologies by conducting 27 interviews between researchers, practitioners, and frequent users of CUI systems. Based on our analysis, we identified five themes covering each group's perspectives and informing about exploitative designs' unethical consequences. In line with prior research, we noticed broken expectations between users and their devices' capabilities and learned about the underlying problems. We hope that the five ethical caveats, derived from our themes, can be used to support user-centred designs and create systems that are better aligned with service providers' intentions and users' expectations. To mitigate users' frustration and ensure more transparency and ethically aligned interactions, this work contributes a framework, the CUI Expectation Cycle, to connect users' expectations with CUIs' capabilities.

ACKNOWLEDGMENTS

Special thanks go to Justin Edwards, whose great support has aided this work significantly throughout the project's course. The research of this work was partially supported by the Klaus Tschira Stiftung gGmbH. This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289_P2. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] ACM. 2023. Words matter: Alternatives for charged terminology in the computing profession. https://www.acm.org/diversity-inclusion/words-matter

[2] C. Alexander, S. Ishikawa, and M. Silverstein. 1977. *A Pattern Language: Towns, Buildings, Construction.* OUP USA. https://books.google.ch/books?id=hwAHmktpk5IC

[3] Saminda S. Balasuriya, Laurianne Sitbon, Andrew A. Bayor, Maria Hoogstrate, and Margot Brereton. 2018. Use of voice activated interfaces by people with intellectual disability. In *OzCHI '18: Proceedings of the 30th Australian Conference on Computer-Human Interaction.* ACM, Melbourne Australia, 102–112. https://doi.org/10.1145/3292147.3292161

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmitchell Shmargaret. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[5] Johnna Blair and Saeed Abdullah. 2020. It Didn't Sound Good with My Cochlear Implants: Understanding the Challenges of Using Smart Assistants for Deaf and Hard of Hearing Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 118 (dec 2020), 27 pages. https://doi.org/10.1145/3432194

[6] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. *Synthesis Lectures on Human-Centered Informatics* 9, 1 (April 2016), 1–115. https://doi.org/10.2200/S00706ED1V01Y201602HCI034 Publisher: Morgan & Claypool Publishers.

[7] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "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* (Virtual Event, USA) *(DIS '21).* Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[8] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proc. Priv. Enhancing Technol.* 2016, 4 (2016), 237–254.

[9] Stacy M. Branham and Antony R. Mukkath Roy. 2019. Reading Between the Guidelines: How Commercial Voice Assistant Guidelines Hinder Accessibility for Blind Users. In *ASSETS '19: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility.* ACM, Pittsburgh USA, 446—-458. https://doi.org/10.1145/3308561.3353797

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[11] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, Pranee Liamputtong (Ed.). Springer Singapore, Singapore. https://doi.org/10.1007/978-981-10-5251-4_103

[12] Harry Brignull. 2010. Deceptive Design – formerly darkpatterns.org. https://www.deceptive.design/ Visited on 2023-01-24.

[13] John Brooke. 2013. SUS : A Retrospective. *Journal of Usability Studies* 8, 2 (2013), 29–40. http://delivery.acm.org/10.1145/2820000/2817913/p29-brooke.pdf?ip=193.1.166.20&id=2817913&acc=ACTIVESERVICE&key=846C3111CE4A4710.FFC467975145E027.4D4702B0C3E38B35.4D4702B0C3E38B35&__acm__=1553781942_85763446e3705c872a65d0c0431bbc79

[14] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[15] Mei-Ling Chen and Hao-Chuan Wang. 2018. How Personal Experience and Technical Knowledge Affect Using Conversational Agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (Tokyo, Japan) *(IUI '18 Companion).* Association for Computing Machinery, New York,

NY, USA, Article 53, 2 pages. https://doi.org/10.1145/3180308.3180362

[16] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016 arXiv:https://academic.oup.com/iwc/article-pdf/31/4/349/33525046/iwz016.pdf

[17] Kenneth Mark Colby. 1981. Modeling a paranoid mind. *Behavioral and Brain Sciences* 4, 4 (1981), 515–534. https://doi.org/10.1017/S0140525X00000030

[18] European Commission. 2016. GDPR-16 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

[19] Silvia De Conca. 2023. The present looks nothing like the Jetsons: Deceptive design in virtual assistants and the protection of the rights of users. *Computer Law & Security Review* 51 (Nov. 2023), 105866. https://doi.org/10.1016/j.clsr.2023.105866

[20] Gregory Conti and Edward Sobiesk. 2010. Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, Raleigh, North Carolina, USA, 271. https://doi.org/10.1145/1772690.1772719

[21] Eric Corbett and Astrid Weber. 2016. What Can I Say? Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 72––82. https://doi.org/10.1145/2935334.2935386

[22] Benjamin R. Cowan, Holly P. Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human-Computer Studies* 83 (2015), 27–42 pages. https://doi.org/10.1016/j.ijhcs.2015.05.008

[23] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, Vienna Austria, 1–12. https://doi.org/10.1145/3098279.3098539

[24] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[25] Philip R Doyle, Leigh Clark, and Benjamin R. Cowan. 2021. What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 244, 14 pages. https://doi.org/10.1145/3411764.3445206

[26] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, Taipei Taiwan, 1–12. https://doi.org/10.1145/3338286.3340116

[27] Philip R Doyle, Daniel John Rough, Justin Edwards, Benjamin R. Cowan, Leigh Clark, Martin Porcheron, Stephan Schlögl, María Inés Torres, Cosmin Munteanu, Christine Murad, Jaisie Sin, Minha Lee, Matthew Peter Aylett, and Heloisa Candello. 2021. CUI@IUI: Theoretical and Methodological Challenges in Intelligent Conversational User Interface Interactions. In *26th International Conference on Intelligent User Interfaces - Companion* (College Station, TX, USA) *(IUI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 12–14. https://doi.org/10.1145/3397482.3450706

[28] Mateusz Dubiel, Sylvain Daronnat, and Luis A. Leiva. 2022. Conversational Agents Trust Calibration: A User-Centred Perspective to Design. In *Proceedings of the 4th Conference on Conversational User Interfaces*. ACM, Glasgow United Kingdom, 1–6. https://doi.org/10.1145/3543829.3544518

[29] Caitlin Fausey, Bria Long, Aya Inamori, and Lera Boroditsky. 2010. Constructing Agency: The Role of Language. *Frontiers in Psychology* 1 (2010). https://doi.org/10.3389/fpsyg.2010.00162

[30] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

[31] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman (Eds.). Vol. 16. Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4 Series Title: Philosophy of Engineering and Technology.

[32] Radhika Garg and Subhasree Sengupta. 2020. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 11 (mar 2020), 24 pages. https://doi.org/10.1145/3381002

[33] Debjyoti Ghosh, Pin Sym Foong, Shan Zhang, and Shengdong Zhao. 2018. Assessing the Utility of the System Usability Scale for Evaluating Voice-Based User Interfaces. In *Proceedings of the Sixth International Symposium of Chinese CHI* (Montreal, QC, Canada) *(ChineseCHI '18)*. Association for Computing Machinery, New York, NY, USA, 11–15. https://doi.org/10.1145/3202667.3204844

[34] Li Gong and Clifford Nass. 2007. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human communication research* 33, 2 (2007), 163–193.

[35] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. https://doi.org/10.1145/3290605.3300408

[36] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. *What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit*. Association for Computing Machinery, New York, NY, USA, 61–73. https://doi.org/10.1145/3357236.3395486

[37] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. *The Dark (Patterns) Side of UX Design*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[38] Colin M. Gray, Cristiana Santos, and Nataliia Bielova. 2023. Towards a Preliminary Ontology of Dark Patterns Knowledge. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 286, 9 pages. https://doi.org/10.1145/3544549.3585676

[39] Colin M. Gray, Cristiana Teixeira Santos, Nicole Tong, Thomas Mildner, Arianna Rossi, Johanna T. Gunawan, and Caroline Sinders. 2023. Dark Patterns and the Emerging Threats of Deceptive Design Practices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 510, 4 pages. https://doi.org/10.1145/3544549.3583173

[40] Paul Graßl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. 2021. Dark and Bright Patterns in Cookie Consent Requests. *Journal of Digital Social Research* 3, 1 (Feb. 2021), 1–38. https://doi.org/10.33621/jdsr.v3i1.54 Number: 1.

[41] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. *Dark Patterns in Proxemic Interactions: A Critical Perspective*. Association for Computing Machinery, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[42] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (Feb. 2006), 59–82. https://doi.org/10.1177/1525822X05279903

[43] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2021. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–29. https://doi.org/10.1145/3479521

[44] Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3501995

[45] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (2010), 29–29.

[46] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*. ACM, BELFAST United Kingdom, 207–214. https://doi.org/10.1145/3335082.3335094

[47] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3174042

[48] Haiyan Jia, Mu Wu, Eunhwa Jung, Alice Shapiro, and S. Shyam Sundar. 2013. When the Tissue Box Says "Bless You": Using Speech to Build Socially Interactive Objects. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 1635–1640. https://doi.org/10.1145/2468356.2468649

[49] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI

Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581503

[50] Hyunhoon Jung, Hyeji Kim, and Jung-Woo Ha. 2020. Understanding Differences between Heavy Users and Light Users in Difficulties with Voice User Interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 51, 4 pages. https://doi.org/10.1145/3405755.3406170

[51] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2020. Construction of UEQ+ scales for voice quality: measuring user experience quality of voice interaction. In *Mensch und Computer 2020 - Tagungsband*, Florian Alt, Stefan Schneegass, and Eva Hornecker (Eds.). ACM, New York, 1—-5. https://doi.org/10.1145/3404983.3410003

[52] A Baki Kocabalil, Liliana Laranjo, and Enrico Coiera. 2018. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*. 1–12.

[53] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, D Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *PNAS* 117, 14 (2020), 7684–7689.

[54] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. 2019. The Effects of Anthropomorphism and Non-Verbal Social Behaviour in Virtual Assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 133—-140. https://doi.org/10.1145/3308532.3329466

[55] Cherie Lacey and Catherine Caudwell. 2019. Cuteness as a 'Dark Pattern' in Home Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, Korea (South), 374–381. https://doi.org/10.1109/HRI.2019.8673274

[56] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R. Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic Evaluation of Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445312

[57] Laura Lascau, Sandy J. J. Gould, Anna L. Cox, Elizaveta Karmannaya, and Duncan P. Brumby. 2019. Monotasking or Multitasking: Designing for Crowdworkers' Preferences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300649

[58] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society* (Graz, Austria) *(USAB 2008)*. 63–76. https://doi.org/10.1007/978-3-540-89350-9_6

[59] Sunok Lee, Minji Cho, and Sangsu Lee. 2020. What If Conversational Agents Became Invisible? Comparing Users' Mental Models According to Physical Entity of AI Speaker. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 88 (sep 2020), 24 pages. https://doi.org/10.1145/3411840

[60] California State Legislature. 2018. CCPA-18 2018. California Consumer Privacy Act of 2018 [1798.100 - 1798.199] (CCPA). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[61] Prolific Academic Ltd. 2023. Prolific | Online participant recruitment for surveys and market research. https://prolific.co/ (visited on 2023-01-24).

[62] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[63] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–32. https://doi.org/10.1145/3359183

[64] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[65] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence* 4 (2021), 169.

[66] Thomas Mildner, Philip Doyle, Gian-Luca Savino, and Rainer Malaka. 2022. Rules Of Engagement: Levelling Up To Combat Unethical CUI Design. In *4th Conference on Conversational User Interfaces*. ACM, Glasgow United Kingdom, 1–5. https://doi.org/10.1145/3543829.3544528

[67] Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 2362–2374. https://doi.org/10.1145/3563657.3595964

[68] Thomas Mildner and Gian-Luca Savino. 2021. Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451659

[69] Thomas Mildner, Gian-Luca Savino, Philip Doyle, Benjamin Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (April 23-28, 2023) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3580695

[70] Christine Murad and Cosmin Munteanu. 2020. Designing Voice Interfaces: Back to the (Curriculum) Basics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376522

[71] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81-103 pages. https://doi.org/10.1111/0022-4537.00153

[72] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[73] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation Of User Interfaces. In *Proceedings of the 1990 CHI Conference on Human Factors in Computing Systems* (April). ACM, Seattle, Washington, USA, 249–256.

[74] Jasmin Niess and Paweł W. Woźniak. 2020. Embracing Companion Technologies. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 31, 11 pages. https://doi.org/10.1145/3419249.3420134

[75] Donald A. Norman. 2013. *The design of everyday things* (revised and expanded edition ed.). Basic Books, New York, New York.

[76] Ikechukwu Obi, Colin M. Gray, Shruthi Sai Chivukula, Ja-Nae Duane, Janna Johns, Matthew Will, Ziqing Li, and Thomas Carlock. 2022. Let's Talk About Socio-Technical Angst: Tracing the History and Evolution of Dark Patterns on Twitter from 2010-2021. http://arxiv.org/abs/2207.10563 arXiv:2207.10563 [cs].

[77] Richard L. Oliver. 1977. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology* 62, 4 (1977), 480–486. https://doi.org/doi.org/10.1037/0021-9010.62.4.480

[78] Richard L. Oliver. 1980. A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research* 17, 4 (1980), 460–469. https://doi.org/10.1177/002224378001700405

[79] Kentrell Owens, Johanna Gunawan, David Choffnes, Pardis Emami-Naeini, Tadayoshi Kohno, and Franziska Roesner. 2022. Exploring Deceptive Design Patterns in Voice Interfaces. In *Proceedings of the 2022 European Symposium on Usable Security* (Karlsruhe, Germany) *(EuroUSEC '22)*. Association for Computing Machinery, New York, NY, USA, 64—-78. https://doi.org/10.1145/3549015.3554213

[80] Kentrell Owens, Johanna Gunawan, David Choffnes, Pardis Emami-Naeini, Tadayoshi Kohno, and Franziska Roesner. 2022. Exploring Deceptive Design Patterns in Voice Interfaces. In *Proceedings of the 2022 European Symposium on Usable Security* (Karlsruhe, Germany) *(EuroUSEC '22)*. Association for Computing Machinery, New York, NY, USA, 64–78. https://doi.org/10.1145/3549015.3554213

[81] European Parliament. 2022. Digital Servicess Act***I. European Parliament [A9-0356/2021]. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.html

[82] J. Piaget. 1997. *The Child's Conception of the World*. Routledge, London, UK. https://books.google.de/books?id=wxWd6bY2FAkC

[83] Martin Porcheron, Leigh Clark, Matt Jones, Heloisa Candello, Benjamin R. Cowan, Christine Murad, Jaisie Sin, Matthew P. Aylett, Minha Lee, Cosmin Munteanu, Joel E. Fischer, Philip R. Doyle, and Jofish Kaye. 2020. CUI@CSCW: Collaborating through Conversational User Interfaces. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 483–492. https://doi.org/10.1145/3406865.3418587

[84] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174214

[85] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *CHI '18: Proceedings of the 2018 CHI Conference on Human*

Factors in Computing Systems. ACM, Montréal Canada, 102–112. https://doi.org/10.1145/3173574.3174033

[86] Stuart Reeves. 2019. Conversation Considered Harmful?. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 10, 3 pages. https://doi.org/10.1145/3342775.3342796

[87] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Transactions on Computer-Human Interaction* 29, 3 (June 2022), 1–41. https://doi.org/10.1145/3484221

[88] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142

[89] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4, Article 81 (may 2021), 43 pages. https://doi.org/10.1145/3386867

[90] William Seymour and Max Van Kleek. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 371 (oct 2021), 16 pages. https://doi.org/10.1145/3479515

[91] Jaisie Sin, Heloisa Candello, Leigh Clark, Benjamin R. Cowan, Minha Lee, Cosmin Munteanu, Martin Porcheron, Sarah Theres Völkel, Stacy Branham, Robin N. Brewer, Ana Paula Chaves, Razan Jaber, and Amanda Lazar. 2023. CUI@CHI: Inclusive Design of CUIs Across Modalities and Mobilities. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 341, 5 pages. https://doi.org/10.1145/3544549.3573820

[92] Alain D. Starke and Minha Lee. 2022. Unifying Recommender Systems and Conversational User Interfaces. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 7, 7 pages. https://doi.org/10.1145/3543829.3544524

[93] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society*. ACM, Washington DC USA, 1–6. https://doi.org/10.1145/3183654.3183691

[94] Alexandra Voit, Jasmin Niess, Caroline Eckerth, Maike Ernst, Henrike Weingärtner, and Paweł W. Woźniak. 2020. 'It's Not a Romantic Relationship': Stories of Adoption and Abandonment of Smart Speakers at Home. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia* (Essen, Germany) *(MUM '20)*. Association for Computing Machinery, New York, NY, USA, 71–82. https://doi.org/10.1145/3428361.3428469

[95] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 254, 15 pages. https://doi.org/10.1145/3411764.3445536

[96] Katja Wagner and Hanna Schramm-Klein. 2019. Alexa, Are You Human? Investigating Anthropomorphism of Digital Voice Assistants - A Qualitative Approach. In *Proceedings of the 40th International Conference on Information Systems, ICIS 2019, Munich, Germany, December 15-18, 2019*, Helmut Krcmar, Jane Fedorowicz, Wai Fong Boh, Jan Marco Leimeister, and Sunil Wattal (Eds.). Association for Information Systems, Munich, Germany, 1–17. https://aisel.aisnet.org/icis2019/human_computer_interact/human_computer_interact/7

[97] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3580895

[98] Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science* 5, 3 (2010), 219–232. https://doi.org/10.1177/1745691610369336 PMID: 24839457.

[99] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[100] Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to Guide Task-oriented Chatbot Users, and When: A Mixed-methods Study of Combinations of Chatbot Guidance Types and Timings. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3501941

[101] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013)* (May 14-17). Society for the Advancement of the Science of Digital Games, Chania, Crete, Greece, 39–46. http://www.fdg2013.org/program/papers.html

A INTERVIEW QUESTIONNAIRE

Here, we provide a Table 4 including Mathur et al.'s [63] original characteristics as well as adapted questions. Below the table is the full script of all nine interview questions. We conditionally prompted interviewees to talk about graphical-based systems and voice-based systems as well as contemporary and future designs to get multi-faceted answers.

Original and adapted questions from the dark pattern characteristics by Mathur et al. [63]	
Characteristic	Question (Originals are *italic* whereas interview questions are not)
Asymmetry	*Does the user interface design impose unequal weights or burdens on the available choices presented to the user in the interface?* Can you tell me about how the design of conversational user interfaces may place specific burdens on people in terms of understanding what choices are available to them during interactions?
Covert	*Is the effect of the user interface design choice hidden from the user?* How clear are the consequences of making particular choices to people when using conversational user interfaces?
Deceptive	*Does the user interface design induce false beliefs either through affirmative misstatements, misleading statements, or omissions?* Can you tell me about how the design of conversational user interface may induce false beliefs?
Hides Info.	*Does the user interface obscure or delay the presentation of necessary information to the user?* How could aspects of the design of conversational user interface obscure or delay important information from the user?
Restrictive	*Does the user interface restrict the set of choices available to users?* Do you feel that all available choices are clear to people when using conversational user interfaces?

Table 4: This table lists the descriptive questions by Mathur et al. (2019) [63] for their dark pattern characteristics. It further contains the adapted interview questions.

(1) Could you please describe your own experience with CUIs?
(2) Can you tell me about how the design of conversational user interfaces may place specific burdens on people in terms of understanding what choices are available to them during interactions?
(3) How clear are the consequences of making particular choices to people when using conversational user interfaces?
(4) Do you feel that all available choices are clear to people when using conversational user interfaces?
(5) Can you tell me about how the design of conversational user interface may induce false beliefs? (these might include false beliefs about the system or in terms of the information it provides.)
(6) How could aspects of the design of conversational user interface obscure or delay important information from the user?
(7) Can you tell me about some inherent limitations of conversational user interfaces that are not always apparent to people who use them?
(8) Can you tell me about situational contexts that might make people more vulnerable to problematic design in CUI interactions?
(9) Can you tell me about specific groups of people who might be particularly vulnerable to problematic design in CUI interactions?

PUBLICATION P6

An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action

Authors:
Colin M. Gray, Nataliia Bielova, Cristiana Santos, & Thomas Mildner

The publication contributes to the following angles:
DESIGN

This publication unites contemporary streams of dark pattern research within a shared ontology. Aiming to support future dark pattern scholarship, the work situates diverse dark pattern types and strategies across a three-layered, granulated structure, including high-level, meso-level, and low-level types. Based on a case study, the publication includes a demonstration for future work to extend the ontology and illustrates how this work can support transdisciplinary efforts.

Its contribution to the thesis is to the design angle. Through landscaping various contributions to the dark pattern scholarship, it offers a joined framework for this future work and points to new pathways for transdisciplinary collaboration.

My contribution to this paper in defining high-, meso-, and low-level types for the presented dark pattern ontology. Moreover, I contributed three case studies demonstrating the extension of ontology with new high-, meso-, and low-level types. I contributed to the findings' interpretation and the writing of the manuscript with an emphasis on the extension of the ontology. I approved the final draft before it was submitted by the first author.

The contents of this chapter originally appeared in: Gray, C. M., Santos, C. T., Bielova, N., and Mildner, T., "An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642436

An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building

Colin M. Gray
comgray@iu.edu
Indiana University
Bloomington, Indiana, USA

Cristiana Teixeira Santos
c.teixeirasantos@uu.nl
Utrecht University
The Netherlands

Nataliia Bielova
nataliia.bielova@inria.fr
Inria Centre at Université Côte d'Azur
France

Thomas Mildner
mildner@uni-bremen.de
University of Bremen
Germany

ABSTRACT

Deceptive and coercive design practices are increasingly used by companies to extract profit, harvest data, and limit consumer choice. Dark patterns represent the most common contemporary amalgamation of these problematic practices, connecting designers, technologists, scholars, regulators, and legal professionals in transdisciplinary dialogue. However, a lack of universally accepted definitions across the academic, legislative, practitioner, and regulatory space has likely limited the impact that scholarship on dark patterns might have in supporting sanctions and evolved design practices. In this paper, we seek to support the development of a shared language of dark patterns, harmonizing ten existing regulatory and academic taxonomies of dark patterns and proposing a three-level ontology with standardized definitions for 64 synthesized dark pattern types across low-, meso-, and high-level patterns. We illustrate how this ontology can support translational research and regulatory action, including transdisciplinary pathways to extend our initial types through new empirical work across application and technology domains.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**.

KEYWORDS

dark patterns, deceptive design, regulation, ontology

ACM Reference Format:
Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3613904.3642436

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642436

1 INTRODUCTION

Deceptive design practices are increasingly common in digital environments, impacting digital experiences on social media [42, 51], e-commerce [40], mobile devices [29], cookie consent banners [26], and gaming [58], among others. An increasingly dominant framing of these deceptive practices is known as "dark patterns"[1]—describing instances where design choices subvert, impair, or distort the ability of a user to make autonomous and informed choices in relation to digital systems regardless of the designer's intent [11, 14, 21].

While the origins of dark patterns as a concept to describe manipulative design practices goes back over a decade to when the term was coined by practitioner and scholar Harry Brignull [6], in the past five years there has been growing momentum in the use of the term to unite scholars, regulators, and designers in transdisciplinary dialogue to identify problematic practices and find ways to prevent or discourage the use of these patterns. In 2021, Mathur and colleagues [41] published a paper at CHI beginning this work in uniting the community by outlining the general scope of the term "dark patterns," proposing common attributes, and identifying methods for identifying and characterizing dark patterns—a paper which has since supported regulatory and legal action relating to dark patterns. This momentum is also borne out on social media; according to a recent study of the historical evolution of #darkpatterns on Twitter (since renamed to "X") by Obi and colleagues [47], since 2019, conversations have included stakeholders not only from design and technology but also social scientists, lawyers, journalists, lawmakers, and members of regulatory bodies and consumer protection organizations.

Within the regulatory space, in 2022 alone, the term "dark patterns" was codified into EU law in the Digital Services Act [14], the Digital Markets Act [13], and the Data Act proposal [12], and into US law in the California CPRA [11]. Regulatory bodies such as the US Federal Trade Commission (FTC), the UK Competition and Market Authority (CMA), the EU Commission, the European Data Protection Board (EDPB) and the the Organisation for Economic

[1]We use this term to connect our efforts to prior scholarship and legal statute, recognizing that other terms such as "deceptive design" or "manipulative design" are sometimes used to describe similar tactics. While the ACM Diversity and Inclusion Council has included dark patterns on a list of potentially problematic terms, there is no other term currently in use that describes the broad remit of dark patterns practices that include deceptive, manipulative, and coercive patterns that limit user agency and are often hidden to the user.

Co-operation and Development (OECD) have released guidance on specific types of dark patterns with various levels of overlap with definitions from academic scholarship [10, 15, 16, 21, 48]. In late Summer 2023, the Department of Consumer Affairs in India also released draft guidelines regarding dark patterns [33] which were finalized in November 2023 [1]. In addition, the concept of dark patterns has been leveraged in sanctions against companies that have relied upon manipulative practices. Recent actions include a $245 million USD judgment against Fortnite, a product from Epic Games, for their use of manipulative practices to encourage the purchase of content [56] and multiple settlements by various US states against Google for their use of dark patterns to obtain location data [49, 55]. In the EU, both Data Protection Authorities (DPAs) and court decisions have forbidden certain practices related to dark patterns, including: pre-selection of choices [9]; refusing consent if it is more difficult than giving it [19, 20]; and misinforming users on the purposes of processing data and how to reject them [20, 39].

As part of this convergent discourse, HCI scholars have addressed the threat of dark patterns in a wide range of publications, proposing definitions and types of dark patterns [4, 24, 37, 40, 41]. However, the specific forms that dark patterns can take, the role of context, the ubiquity of the practices, the technologies used or application area, the comparative harms of different patterns, remedies, and the role of user education and countermeasures are still a topic of ongoing research. The consequence of this dynamic topic is of an ever-expanding list of categories and variants whose scale continues to grow.

Two large challenges face an ongoing transdisciplinary engagement with the concept of dark patterns. First, the literature has grown quickly and tends to be siloed, often lacking accurate citation provenance trails of given typologies and definitions, making it difficult to trace where new or more detailed types of patterns emerged and under which conditions. For instance, some patterns were originally coined in particular framings of user interaction such as privacy (e.g., Bösch's "immortal accounts") or rely on domain-specific characteristics (e.g., Brignull's "sneak into basket" which is strongly associated with e-commerce). Without these original contexts in mind, it can be difficult to understand how the use of a pattern and its associated definition can be productively (or problematically) applied to a new domain. In parallel, the space that dark patterns scholars have sought to cover is also vast, with important research occurring in specific domains (e.g., games, e-commerce, privacy and data protection) and across different technologies and modalities (e.g., mobile, desktop, conversational user interfaces (CUIs), AR/VR), as shown in a recent systematic review of dark patterns literature [22]. This diversity of research has led some scholars to propose fragmentary, domain-specific typologies without necessarily finding commonalities across domains—resulting in extra work and often needlessly strengthening scholarly siloes. Second, regulators and policy makers have been interested in the scholarly conversation regarding dark patterns, but have in some cases created wholly new domain-related terminology to describe types already known in the academic literature (e.g., the EDPB social media guidelines [16, 17], which included many previously known dark patterns that were described by wholly new names, severing connections to other relevant literature) in their pursuit of providing legal guidance on emergent issues relating to dark patterns (e.g., [17]. In other cases, regulators and policymakers have inconsistently cited academic sources (e.g., [17, 21]) making connections across the regulatory, legal, and academic spaces fraught—and making academic and practitioner work that connects these domains difficult to broker.

We seek to support these challenges and ongoing conversations by building the foundation for a common ontology of dark patterns. This effort is directly motivated by multiple years of engagement by the research team—including discussions with participants at numerous international conference presentations, workshops, and symposia, alongside interactions with regulators, legal scholars, and engagement as expert witnesses on legal cases relating to dark patterns. Through these encounters, we have confronted both the challenges of conducting work in an emergent space where there is broad consensus on the key components of dark patterns but not necessarily a shared language (as extensively described by Mathur and colleagues [41]) and the promise of synergies with other interdisciplinary partners when this shared language is realized. In particular, there has been broad interest in using a consolidated set of terms to describe the types of dark patterns, their presence, and their impacts—connecting the design of digital systems, social scientists that study the implications of these systems on users, and regulators that seek to rein in unfair, deceptive, or coercive business practices—regardless of the domain or context in which these practices occur.

By taking the first steps towards building an ontology, we seek to create a shareable, extendable, and reusable knowledge representation of dark patterns which is hosted at **https://ontology.darkpatternsresearchandimpact.com**. This groundwork for an ontology is both domain and application agnostic though it has potential utility in domain or context-specific instances as well. For instance, the *Bad Defaults* dark pattern is often embedded in settings menus, pre-set so that users share personal information on social media platforms or accepting to receive advertising content on online shopping sites unknowingly. Such context-specific instances are enabled through *Interface Interference*—a domain-agnostic strategy used to manipulate interfaces, privileging certain actions and, thus, limiting discoverability of alternatives. As noted by Fonseca [18], ontologies can be useful in supporting social science research by "creating better conceptual schemas and applications." We build upon this argument from Fonseca, arguing that our ontology also supports alignment across social science researchers, legal scholars, regulators, and designers—supporting these stakeholders with a shared vocabulary which they can use to discuss existing and emergent concerns relating to dark patterns across a variety of domains and contexts. To create this preliminary ontology, we build upon ten contemporary taxonomies of dark patterns from both the academic and regulatory literature, and thereafter we identify three levels of hierarchy for pattern types. Hence we harmonize concepts across these taxonomies to provide a consistent and consolidated, shared, and reusable dark patterns ontology for future research, regulatory action, and sanctions.

We make four contributions in this paper. First, we introduce the hierarchical concepts of low-level, meso-level, and high-level dark patterns to the literature, disambiguating UI-level patterns that may lead to opportunities for detection (low-level) and strategies

that may be targeted by policy and legislation (meso- and high-level). Second, by analysing the provenance of dark patterns from academic and regulatory sources, we identify when patterns first emerged and how naming has evolved over time and across sources. Third, we describe a common definition syntax, set of definitions, and hierarchy of dark patterns that aligns disparate terminology from scholars and regulators. Fourth, we demonstrate how the ontology can be strengthened and extended through additional empirical work, and how the ontology can effectively be utilized by practitioners, scholars, regulators, and legal professionals to support transdisciplinary action.

2 MOTIVATION & BACKGROUND

Since the initial set of a dozen types of dark patterns proposed by Brignull in the 2010s, research has focused on related issues from multiple angles including, but not limited to, e-commerce, games, social media, and IoT [22]. While this scholarship contributes significant insights to the discourse, we noticed varying approaches to adopt existing descriptions, defining novel scenarios in which users are harmed. Meanwhile, the specification of individual typologies creates a certain ambiguity within the overall discourse on the matter. In developing this ontology, we confront numerous timely issues relating to the description of dark patterns, the study of dark patterns and their harms through empirical work, and the leveraging of this scholarship to support legal and regulatory action.

Dark patterns are known to be ubiquitous; however, most pattern types have been explored in relatively narrow contexts, cultures, or domains with more scholarship needed to fully define causal links, harms, and impacted populations [22]. The HCI community has been engaged and interested in impacting society and the future of technology practices relating to dark patterns [23, 27, 38]—and indeed, HCI scholars have been central in the study of dark patterns, revealing insights relating to the harm and severity of dark patterns that then support enforcement action and regulation. However, we currently lack a shared landscape of definitions, types, and language to unify the study of dark patterns. Without this shared landscape, research has become (and will continuously be) fragmented by domain, context, and technology type—which if not addressed, may lead to duplicated effort by scholars working on similar issues in different domains, and additionally may hamper regulatory enforcement due to lack of precision and shared language regarding precisely what dark patterns are used and with what effect. Such lack of a shared ontological framework may also restrict traceability and searchability of dark patterns.

Our work unifies practitioner, scholarly, and regulatory efforts that describe the range of dark patterns, leading to a shared vocabulary and ontology that allows for coordination of efforts across diverse contexts (e.g., technologies, specific functionality, areas of technology use) and stakeholders (e.g., regulators, legal scholars, social scientists, practitioners). This ontology will support not only the advancement of scholarship, but also translational and transdisciplinary efforts that connect scholarship to legal sanctions and regulatory frameworks. For instance, there are now high-level prohibitions of dark patterns by regulatory authorities and legal statute; however, the specific low-level practices that should be deemed illegal under these prohibitions are not yet detailed in enforcement action or case law. This paper connects these different strands of work by harmonizing regulatory and academic work into a single ontology, enabling future scholars and practitioners from all disciplines to utilise our structures and definitions to support their work.

3 METHODOLOGY

We used a qualitative content analysis approach [31] to identify and characterize elements of existing dark patterns taxonomies using the method described in Figure 1[2].

As a research team, we leveraged our collective experiences in human-computer interaction, design, computer science, law, and regulation. Specifically, our team included established dark patterns scholars, including one with a focus on human-computer interaction and design (Gray and Mildner), one with a focus on computer science and web measurement and experience in regulation (Santos), and one with a background in computer science and data protection law (Bielova). Across these perspectives, in accordance with previous scholarship, we sought to characterize dark patterns in a transdisciplinary way, drawing on multiple disciplinary perspectives that provide differing views on the origins and types of dark patterns [26]. However, these backgrounds also introduce gaps, tensions, and opportunities that relate to the unique experience and academic training of each author. To account for this difference in perspective, each dark pattern type was initially reviewed by each author independently before engaging in conversation amongst the researchers that led to the final agreement on the harmonized type and definition.

3.1 Data Collection

We collected dark patterns taxonomies in Fall 2022 from a total of 10 sources, including:

(1) A set of patterns shared on https://darkpatterns.org since 2010 by Harry Brignull[3].
(2) Scholarly academic sources that were highly cited[4], present a distinct and comprehensive taxonomy, and have had one or more component dark patterns types included in a regulatory reports (with or without citation) [4, 24, 37, 40].
(3) Public reports from stakeholders and regulators in the EU, UK, and USA that include a dark patterns taxonomy [10, 15, 16, 21, 48].

The selection of these sources encompass, at the time of our data collection in Fall 2022: i) the "classic" set of patterns shared on darkpatterns.org for over a decade by Brignull; ii) the most commonly cited taxonomies in the research literature (which were also referenced in regulatory taxonomies in a direct or indirect way, likely due to their prominence), and iii) the most comprehensive set

[2]This work builds upon an extends a previous draft version of this ontology published at CHI 2024 as a late-breaking work. [25]
[3]This collection of dark patterns was moved to https://www.deceptive.design in 2022, but the 12 patterns we drew on have been stable since 2018 when the final pattern, "confirmshaming," was added. In 2023, this website was updated to include additional pattern types, resulting in a modified collection of 16 types.
[4]As of December 2023, Google Scholar reports citations of these sources as follows: Gray et al. [24] (657); Mathur et al. [40] (454); Luguri and Strahilevitz [37] (260); and Bösch et al. [4] (259).

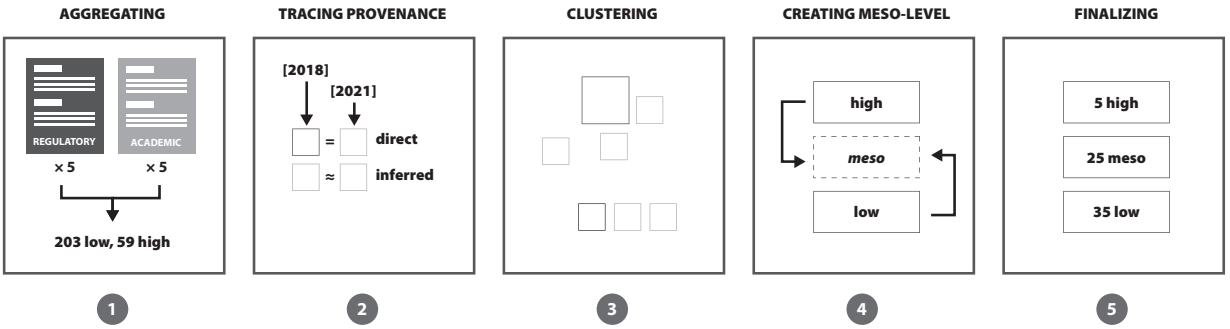


Figure 1: Our method for creating the ontology, mapping to the steps in Section 3.2.1.

of regulatory literature from countries that had produced reports specific to dark patterns at this time. Taken as a set, these academic, practitioner, and regulatory taxonomies provide a strong foundation for our ontology, setting the stage for inclusion of other domain- or context-specific taxonomies in the future (which we outline in Section 6.

3.2 Data Analysis

Once we gathered the set of taxonomies, we began our analysis by identifying the constitutive components of each taxonomy without considering overlaps across sources through a bottom-up approach.

Quantification of dark pattern types Across the ten taxonomies from academic and regulatory sources collected in Fall 2022, we identified 186 low-level and 59 high-level patterns (a total of 245 patterns).

After our initial analysis, the patterns used on Brignull's site (https://www.deceptive.design) were substantially updated in the Summer 2023, and we collected the additional set of patterns for that source—resulting in 11 total sources. Also, the EDPB regulatory report was made final in February 2023, and we used its final taxonomy in this paper after completing our initial mapping in the Fall 2022 based on the draft report taxonomy. Based on the updates to the EDPB guidelines and Brignull's site in the Spring and Summer 2023, the total number of patterns we analyzed included 203 low-level (adding 1 new pattern from the revised EDPB guidelines and 16 patterns from the updated Brignull site) and 59 high-level patterns—**a total of 262 patterns** (see Tables 2 and 3 in the supplemental material). All taxonomy elements are included in supplemental material for other scholars to build upon.

Rationale underlying the high number of dark pattern types This large number of discrete elements is perhaps unsurprising, since each typology author has used a different point of focus and categorization based on the sector they sought to describe or support. For instance, Mathur et al. [41] and the CMA [10] focus on e-commerce; the EDPB focuses on data protection practices within social media platform interfaces [16], and the FTC [21] and EU Commission [15] focus on guidance specific to their jurisdictions and underlying legal authority. The types themselves also evolved in one case due to input from the practitioner and regulatory community, which is the case of the EDPB naming of patterns changed slightly from the 2022 draft report to the final 2023 report, with one high-level strategy "hindering" changing to "obstructing" to bring it into better alignment with academic taxonomies.

3.2.1 Creating the Ontology Framework. We used the following procedure to carefully identify existing taxonomy components, their source, relationships and similarities between components across taxonomies, visualized in Figure 1:

(1) **Aggregating existing patterns.** We first listed all high- and low-level patterns verbatim in the structure originally indicated in the textual source. *High-level patterns* include any instances where the pattern is denoted as a category, strategy, goal, intention, or other parent in a parent-child relationship. *Low-level patterns* indicate specific patterns that are included as a child in a parent-child relationship, or are otherwise undifferentiated in hierarchy (e.g., Brignull's patterns).
(2) **Identifying provenance through direct citations and inference.** Based on citations provided in the source-document, we indicated any instances where patterns were *directly cited* or otherwise duplicated from previous sources. Because many patterns were uncited—particularly in regulatory reports—we also relied upon citations elsewhere in the document or explicit use of existing pattern vocabulary and definitions from previously published sources, which we indicate as *inferential*. We used these direct and inferential citation patterns to identify where patterns were first introduced, even if they appeared alongside other patterns that had been published previously. This allowed us to map the historical progression of high- and low-level types over time.
(3) **Clustering similar patterns.** We grouped patterns that appeared either to be identical or similar (in a *is-a* or *equivalent-to* relationship) on Miro (see Figure 2), using definitions to identify affinities among patterns that did not have identical names. This portion of the analysis was the most extensive, including in depth conversations between an HCI and legal scholars and a careful reading of the definitions as they might be understood by designers and lawyers. We tried out numerous different groupings based on what we understood

Figure 2: A screenshot of our Miro workspace where we organized and clustered elements of the ten source taxonomies. Columns indicate an entire structure of meso- and low-level patterns underneath a high-level pattern and yellow Post-It notes indicate draft meso-level patterns. The elements are color-coded based on which taxonomy they came from. A full version of this workspace is included as a supplemental material.

to be the main focus of each pattern and then sought to characterize what level of pattern each represented.

(4) **Creating meso-level patterns.** From the findings of this visually-organized analysis procedure, we recognized that there were not only low- and high-level patterns present, but also a *"meso"* level of pattern knowledge. By recognizing similarities among low-level patterns, we introduced *meso-level patterns* into our analysis, identifying these patterns by using the names or elements of existing taxonomies where possible, or coining new names to characterize the low-level patterns we grouped together. If the pattern cluster was specific to low-level UI concerns, we sought to identify a meso-level pattern name that was more abstract and could contain the low-level pattern. If the pattern represented a meso-level abstraction, we did not seek to identify specific low-level instantiations—instead leaving that task for future scholarship efforts in domain- and technology-specific areas.

(5) **Finalizing the ontology.** Across these three levels of hierarchy, we grouped 233 of the 245 taxonomy elements[5]. After evaluating the changes to the EDPB guideline taxonomy and updated Brignull taxonomy in Spring and Summer 2023, we updated our mapping of 262 patterns, which resulted in no additional novel pattern types. The final ontology includes 5 high-level patterns, 25 meso-level patterns, and 35 low-level patterns—**a total of 65 patterns**.

3.2.2 Harmonizing Definitions of Dark Patterns Types. Building on this ontology framework, we then proceeded to create a definitional syntax across the three levels of the ontology and then created definitions for each final pattern using the following approach:

(1) **Creating definition syntax.** We evaluated the range of approaches to definitions in the existing taxonomies.

[5]Four ungrouped elements were from the CMA report [10] in Fall 2022 and described generic elements of digital systems which were not explicitly framed as deceptive or manipulative: Choice Structure, Choice Information, Feedback, and Messengers. All eight high-level patterns from Bösch [4] were also excluded since they were not reiterated in any downstream literature.

- *Short vs long definitions.* Some definitions were very short (e.g., the EU Commission's definition for *forced registration*: "Consumer tricked into thinking registration is necessary") while other definitions were more elaborate (e.g., the FTC's definition for *baseless countdown timer*: "Creating pressure to buy immediately by showing a fake countdown clock that just goes away or resets when it times out. Example: 'Offer ends in 00:59:48"'; the EU Data Protection Board's definition for *longer than necessary*: "When users try to activate a control related to data protection, the user journey is made in a way that requires more steps from users, than the number of steps necessary for the activation of data invasive options. This is likely to discourage them from activating such control.").
- *Description of the definitions.* Most definitions were based in a description of user interaction with a system, like the examples above; however, Brignull's 2018 definitions were written in first-person language demonstrating how a user would experience a dark pattern (e.g., the definition for *roach motel*: "You get into a situation very easily, but then you find it is hard to get out of it (e.g. a premium subscription)."). Interestingly, Brignull's 2023 language appears to model other taxonomies with all definitions beginning with "The user struggles...," "The user expects..." or similar structures.
- *Definition structure and syntax.* We used an iterative process where two authors independently and collaboratively tested different definition structures. Based on these efforts and through discussion, we finalized sample definition structures and syntax that captured the relevant type of knowledge (e.g., strategy, angle of attack, means of execution). For instance, all high-level patterns included the interplay of an undesired action and a limitation of their decision-making or free choice. Meso-level patterns addressed a mismatch in users' expectations of a system and the relevant impact. Low-level patterns identified how they manifest their parent high- and meso-level pattern in relation to one or more elements of the UI and a mismatch of expectation and resulting effect on the user experience.

(2) **Creating and member-checking high- and meso-level pattern definitions.** We then drafted definitions for all high- and meso-level patterns, iterating on the structure until we found a syntax that appeared to address all critical elements of the existing definitions and allow us to clearly indicate how the pattern subverted user autonomy and manifest as deceptive or coercive. We began with definitions at these levels since low-level patterns were already grounded in specific UI examples, and thus more effort was needed to identify what components a definition at a higher level of abstraction should include. To support member-checking, our set of 30 definitions and the draft definition structures were then shared via a Google Doc with members of a large Slack community focused on research and enforcement action relating to dark patterns. This Slack channel was initiated in 2021 to grow and foster a community of dark pattern researchers after a successful CHI workshop on the topic [38]. The community has since grown into a transdisciplinary network with over 100 participants around the world, including dark patterns scholars, practitioners, legal experts, regulators, and representatives of non-profits seeking to combat deceptive design practices. We asked this community for feedback on the utility of the definitions, the completeness of the definition structures, and the ability of these definitions to leave as open-ended the many different low-level manifestations of dark patterns. More than two dozen community members viewed the draft materials (evidenced through comments or reactions), and over ten gave us feedback. Most interactions were quite short (particularly on Slack), while others involved threaded messages with replies to clarify meaning. For instance, there were discussions of how central "deception" should be in the definitions, requests for more information on how the different levels of definitions functioned, and specific words that can or should be used to indicate curtailing of user autonomy. Regardless of the form of interaction, the feedback was overwhelmingly positive, with respondents mentioning the utility of the patterns and definitions in supporting future work and validating our approach to focus on mechanisms that support the power of dark patterns rather than overly focusing on intent. This positive feedback mirrored what we have heard from scholars and regulators when presenting draft versions of the ontology in various international symposia, workshops, and conferences for almost a year.
While more formal evaluation of specific definitions in particular contexts (e.g., design, regulation, research) will be useful, the community members who stand to benefit the most from the ontology have ensured that the definitions have face validity. Since this initial evaluation in Summer 2023, the ontology has also been used to support an emerging collection of "fair patterns" as well, wherein meso- and high-level dark patterns types from our ontology are linked to specific countermeasures that could be considered by designers and regulators.[6]

(3) **Finalizing low-level pattern definitions.** After mapping out the initial 30 definitions, we created definitions for the 34 low-level patterns that were grounded in the specifics of the UI execution. These patterns were easier to write since many taxonomy definitions (in particular those from Brignull [7], Gray [24], and the FTC [21]) included richer detail for patterns that pointed towards a real-world implementation. As a research team, we read and edited the definitions until we were satisfied with their level of consistency and relationships to the higher-level categories in which they belonged. All definitions are included in the appendix of this paper and supplemental materials to support future work.

4 MAPPING THE EVOLUTION OF DARK PATTERNS

Pattern names have largely stabilized in the past five years, including high-level pattern types (e.g., nagging, obstruction, sneaking, interface interference, forced action) and low-level patterns (including Brignull's [6, 7] and those introduced by Gray et al. [24] and

[6] https://fairpatterns.com/what-are-dark-patterns/

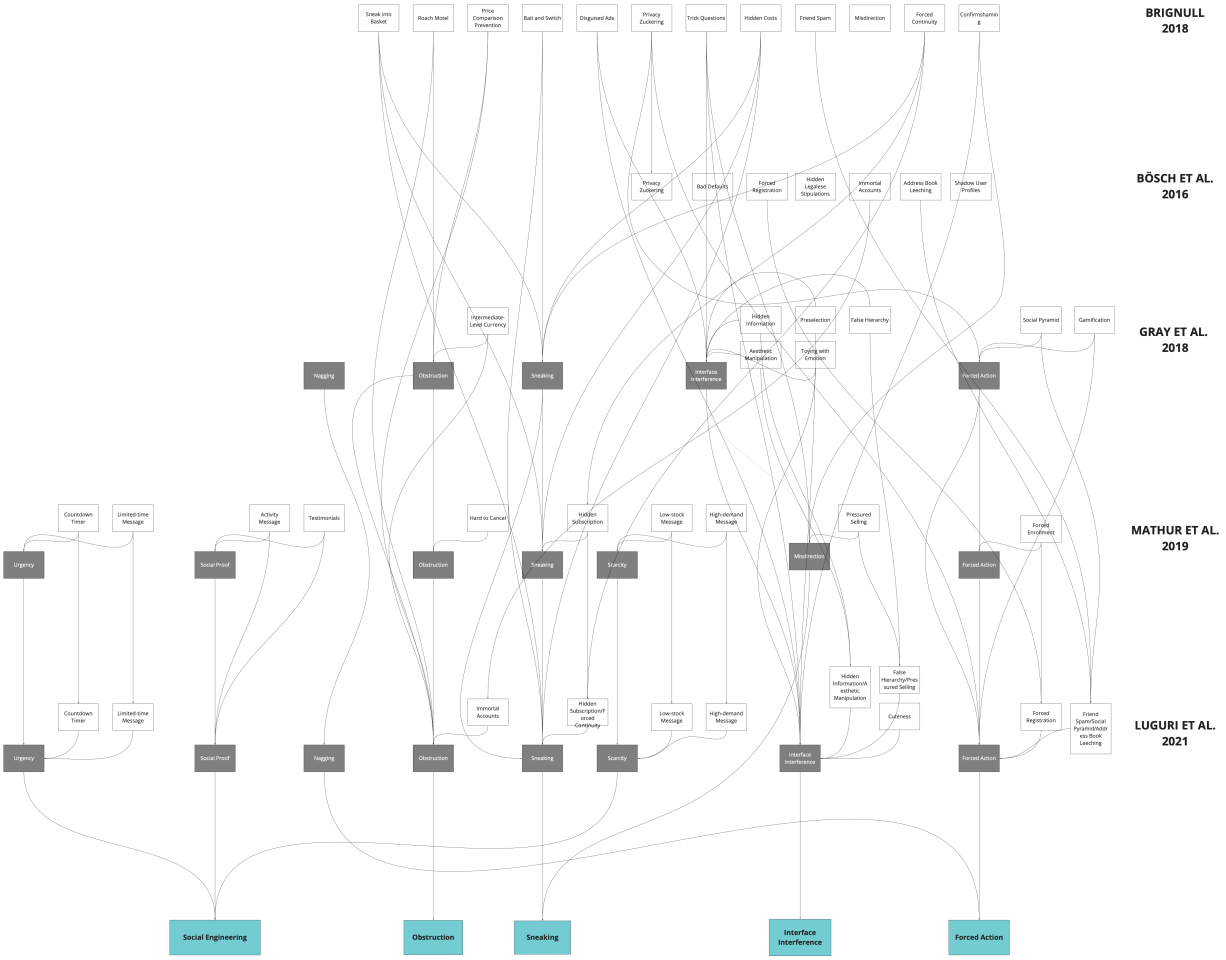


Figure 3: A visual mapping of the evolution of dark patterns in the academic taxonomies we analyzed from 2018-2021. Each row includes elements of the related taxonomy by year and source, and connecting lines indicate relationships between or reiterations of different patterns over time. Pattern names in gray boxes are high-level patterns, pattern names in white boxes are low-level patterns or otherwise lack hierarchy, and pattern names at the bottom are the final high-level patterns we adopt in our ontology. A full version of this mapping is included as a supplemental material.

Mathur et al. [40]). A mapping of these patterns over time across the academic and practitioner sources we considered is included in Figure 3.

High-level patterns were most likely to co-occur across multiple sources. For instance, Gray et al.'s [24] original five high-level "dark pattern strategies" were found across multiple other sources, even if they were not consistently cited: nagging [15, 37], obstruction [15, 21, 37, 40], sneaking [15, 21, 37, 40], interface interference [15, 21, 37], and forced action [15, 21, 37, 40] (FTC uses "coerced action" instead). As shown in Figure 3, virtually all of the high level patterns proposed by Gray et al. in 2018 were carried forward in other academic taxonomies. In Brignull's 2023 changes to https://www.deceptive.design, multiple high-level strategies from Gray et al.'s [24] taxonomy were added to the website (nagging, obstruction, sneaking, forced action, visual interference)—however, these changes were not cited and Brignull continued his practice of not providing direct citations or hierarchical structure to his patterns. After their introduction in Mathur et al. [40], newly introduced categories relating to social psychology or behavioral economics also became common: urgency [15, 21, 37], scarcity [21, 37], and social proof [15, 21, 37] (the FTC bundles "Endorsements" with "social proof"). We have grouped these types together as part of a sixth high-level pattern of "social engineering."

Domain or context-specific patterns. The most volatility has occurred in relation to *domain-* or *context-specific patterns*. These include expansions of Mathur et al.'s [40] high-level patterns of "social proof" and "scarcity," which have since been reiterated by the EU Commission [15] and OECD [48] and extended by the CMA [10] and FTC [21] taxonomies. In addition, the EDPB guidance on dark patterns in social media [16] included a wholly new set of 6 high-level and 15 low-level patterns, although the majority of these could be inferred as similar to already existing patterns proposed in the academic literature. Importantly, though, the EDPB taxonomy included multiple patterns which we found to be new low-level or meso-level additions, including "privacy maze," "dead end," "conflicting information," "information without context" (which we renamed from the EDPB pattern "decontextualizing"), and "visual prominence" (which we renamed from the EDPB pattern "look over there"). Similarly, the CMA taxonomy focused on choice architecture as a guiding structure with three categories focused on choice "structure," "information," and "pressure." This taxonomy structure also yielded new patterns, including "bundling," "complex language," and "personalization."

Our analysis demonstrates the value in classifying or generating context-specific patterns that illuminate gaps in current taxonomies, and also the benefit of mapping these patterns within larger ontologies to identify abstractions of patterns that may apply across many domains, contexts, and legal fields. Our final ontology mapping is included in Figures 4 and 5 and can also be found in the supplementary materials.

5 CREATING A DEFINITIONAL STRUCTURE BY ONTOLOGY LEVEL

As described in Section 3.2.1, our ontology includes three different levels of hierarchy:

- **High-level patterns** are the most abstracted form of knowledge, including general *strategies* that characterize the inclusion of manipulative, coercive, or deceptive elements that might limit user autonomy and decision making. These patterns are context-agnostic and can be employed through a range of modalities and technologies (e.g., desktop, mobile, VUIs, VR/AR) and application types (e.g., e-commerce, gaming, social media).
- **Meso-level patterns** bridge high- and low-level forms of knowledge and describe an *angle of attack* or specific approach to limiting, impairing, or undermining the ability of the user to make autonomous and informed decisions or choices. These patterns are content-agnostic and may be interpreted in a contextually-appropriate way based on the specific context of use or application type.
- **Low-level patterns** are the most situated and contextually dependent form of knowledge, including specific *means of execution* that limits or undermines user autonomy and decision making, is described in visual and/or temporal form(s), and is likely to be detectable through algorithmic, manual, or other technical means.

To create a definitional structure for each level, we first used a subset of approximately ten dark patterns types and definitions in order to "play-test" a combined and unified definition for dark patterns types at multiple levels of granularity (i.e., high, meso, low). Through this process, we considered not only the level of abstraction inherent in dark patterns at differing levels, but also the interaction between: the user's expectations of what should or would be likely to occur (i.e., manipulation of the gulf of execution); the user's identification that something had occurred that they did not wish to happen (i.e., manipulation of the gulf of evaluation); and the mechanisms used to inform or execute manipulation in either of these prior elements. We also considered cases where the deception or manipulation was likely to be hidden to the user (e.g., cases of sneaking, obstruction, or interface interference) as well as cases where deception or coercion was overt and known to the user (e.g., forced action). Based on this iterative generation of a definitional structure, we created a standardized syntax for each dark pattern level, described below. All 65 final definitions are included as a supplemental material.

5.1 High-Level Patterns

> **{HIGH-LEVEL DARK PATTERN}** is a strategy which **{UNDESIRED ACTION}** that [optionally, if known to users, would] **{DISTORT/SUBVERT/IMPEDE/OTHERWISE LIMIT USERS' AUTONOMY, DECISION-MAKING, OR FREE CHOICE}**.

Across our 5 high-level pattern definitions, we considered *undesired actions* such as: hiding, disguising, delaying, redirecting, repeating, impeding, privileging, or requiring actions. We also considered a range of *mechanisms* that could be used to limit users' autonomy, decision-making, or free choice such as: foregrounding unrelated tasks, dissuading a user from taking an action, confusing the user, limiting discoverability of action possibilities, causing a user to unintentionally take an action they would likely object to, or forcing a user to take an action they would not otherwise take. Most of these definitions placed a focus on mechanisms which were primarily hidden, resulting in the user being deceived, such as: "*Interface Interference* is a strategy which privileges specific actions over others through manipulation of the user interface, thereby confusing the user or limiting discoverability of relevant action possibilities." However, the definition for *Forced Action* was focused more on the coercive nature of the interaction which may involve users' awareness they are being manipulated: "*Forced Action* is a strategy which requires users to perform an additional and/or tangential action or information to access (or continue to access) specific functionality, preventing them from continuing their interaction with a system without performing that action."

5.2 Meso-Level Patterns

> **{MESO-LEVEL DARK PATTERN}** subverts the user's expectation that **{EXPECTATION}**, instead producing or informing **{DIFFERENT EFFECT ON USER}**.

Across our 24 meso-level pattern definitions, we considered a range of *user expectations* such as: presence of relevant and timely information, match between user goal and action, completeness and truthfulness of information provided, and the ability to change one's mind and reverse a decision. We also considered a range of *potential*

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)	–

Figure 4: Our ontology of dark patterns organized by level of pattern. "D" indicates a *direct* use of the pattern language in the original source(s) and "I" indicates an *inferred* similarity between different terminology used across two or more pattern types. Sources are indicated by abbreviation and are colored cyan if they are regulatory reports or magenta if they are academic or practitioner sources. "Br" indicates his 2018 patterns and "Br23" indicates his 2023 patterns. *Italized pattern names* indicate new pattern types introduced in this paper while all other text relies upon the sources indicated. Underlined sources indicate the earliest mention of that pattern or patterns in the sources we analyzed. A full description of the inferred pattern names is included in supplemental material to support future work.

High-Level Pattern	Meso-Level Pattern	Low-Level Pattern
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)
		Low Stock (D: Ma Lu FTC EUCOM OECD)
	Social Proof (D: Ma Lu EUCOM OECD; I: Br23)	Endorsements and Testimonials (D: Ma Lu FTC EUCOM OECD)
		Parasocial Pressure (I: FTC)
	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 5: Ontology of dark patterns organized by level of pattern, continued.

negative effects on the user, such as: unexpected or unanticipated outcomes, confusion or pressure, being prevented from locating relevant information, or making a different choice than they would otherwise make. Meso-level definitions as a set touched on many different aspects of the user experience, with some pointing more towards static moments in the user journey and others describing temporal effects that might be realized over a longer portion of the user journey. For instance, these two patterns represent instances where the focus was primarily on static UI elements or a particular moment of interaction:

- "*Manipulating Choice Architecture* subverts the user's expectation that the options presented will support their desired goal, instead including an order or structure of options that makes another outcome more likely."
- "*Scarcity or Popularity Claims* subverts the user's expectation that information provided about a product's availability or desirability is accurate , instead pressuring the user to purchase a product without additional reflection or verification."

In contrast, other patterns represented instances where the full effect of the pattern was felt over time and might involve multiple interactions with a system that accumulate to achieve the overall effect:

- "*Roach Motel* subverts the user's expectation that an action will be as easy to reverse as it is to make, instead creating a situation that is easy to get into, but difficult to get out of."
- "*Hiding Information* subverts the user's expectation that all relevant information to make an informed choice will be available to them, instead hiding information or delaying

the disclosure of information until later in the user journey that may have led to them making another choice."

5.3 Low-Level Patterns

{LOW-LEVEL DARK PATTERN} uses **{RELATED HIGH- AND MESO-LEVEL DARK PATTERN}** to **{ELEMENT OF UI ALTERED}**. As a result, **{INCORRECT USER EXPECTATION}** leads to **{UNDESIRED EFFECT ON USER}**.

Across our 35 low-level definitions, we considered a range of *means of execution* in the UI or user experience, such as: provision of information that is conflicting, prohibiting certain kinds of interactions, adding items without a user's knowledge, providing incomplete or misleading information, distracting a user through extraneous cues, or using social or other extrinsic pressure to steer user's decisions. These means of execution were supported by a wide range of *incorrect user expectations and related undesired effects*, including: preventing a user from making an informed choice about their privacy or purchase of a product, disclosing incomplete or misleading information that leads to choices the user would not otherwise make, or distracting a user and thus preventing them from discovering information that would be relevant to their decision. Low-level patterns all exploit the user experience in direct ways, but address different aspects of the experience:

- Focus on specific user interactions that are limited (e.g., "*Price Comparison Prevention* Creates Barriers and uses Obstruction by excluding relevant information, limiting the ability of a user to copy/paste, or otherwise inhibiting a user from comparing prices across two or more vendors. As a result, the user cannot make an informed decision about where to buy a product or service.")
- Focus on a coordinated set of user interactions that produce the desired effect (e.g., "*Privacy Mazes* Add Steps and use Obstruction to require a user to navigate through many pages a result, the user is prevented from easily discovering relevant information or action possibilities, leaving them unable to make informed decisions regarding their privacy.")
- Focus on discrete UI elements (e.g., "*False Hierarchy* Manipulates the Choice Architecture, using Interface Interference to give one or more options visual or interactive prominence over others, particularly where items should be in parallel rather than hierarchical. As a result, the user may misunderstand or be unable to accurately compare their options, making a selection based on a false or incomplete choice architecture.")
- Focus on user comprehension of the interface (e.g., "*Wrong Language* leverages Language Accessibility, using Interface Interference to provide important information in a different language than the official language of the country where users live. As a result, the user will not have access to relevant information about their interaction with the system and their ability to choose, leading to uninformed decisions.")

6 EXTENDING THE ONTOLOGY BASED ON CURRENT AND FUTURE SCHOLARSHIP

Dark patterns researchers have addressed the impact of manipulative, deceptive, and coercive design in a range of technological domains. While these efforts are important in protecting online users and identifying areas for regulatory or legal impact, the novelty and breadth of this work potentially hinders an exhaustive mapping of dark patterns onto our ontology. Building on our proposed ontology, we identify pathways for many stakeholders to contribute to the growth of ontology elements—both through the addition of new patterns and strengthening contextual or domain-specific examples of existing patterns. This extension can help not only to anchor instances of patterns from future studies in existing literature, but also to enable the scholarly community to extend or further characterize these pattern types. The ontology's stratification allows anyone to extend the current framework by following the structure and syntax given for each high, meso, and low level dark pattern type.

To perform this mapping and extension exercise, we sought to identify existing alignment between proposed dark patterns and the ontology. To this end, we consider how a source might offer new perspectives for existing or examples of novel dark patterns. The method we used to extend the ontology involves three steps:

(1) We analyzed the dark pattern definition included by the author and, if provided, considered any cited relationships to other dark patterns and related terminologies.
(2) We then aligned the author's definition with the syntax of the high, meso, and low levels, placing the dark pattern at the most logical level of abstraction.
(3) Finally, we considered how the addition of the type informs a revision of the ontology. A type could reiterate an existing type in the ontology (leaving the core ontology unchanged), extend an existing type in the ontology (providing rationale for a more expansive definition of an existing type), or identify the presence of a wholly new type (adding a type to the core ontology).

This section demonstrates how we envision for the community to extend the ontology by drawing examples from three contemporary studies defining dark patterns from domain and context-specific areas, underlining the decision behind selecting these relevant works. These examples extend across multiple emergent areas of the state-of-the-art in dark patterns literature, encompassing some of the first examples of studies addressing: dark patterns in Japanese apps (Section 6.1), dark patterns experienced across multiple modalities (Section 6.2), and dark patterns experienced on prominent social media apps (Section 6.3). We also show how the ontology can be extended to map legislation and case law relating to dark patterns. Table 1 summarizes how three different sources were compared to our ontology through this method, demonstrating how the community could extend the ontology over time. The ontology can be accessed at **https://ontology.darkpatternsresearchandimpact.com**, which includes a current state of the ontology and community-vetted changes over time that follow this process in a public, deliberative manner. Initial and future iterations of the ontology will be versioned and include a version history for citation accuracy.

Extending the Ontology			
Name	**Definition from the Sources**	**Mapping to Ontology**	**Level**
Linguistic Dead-End [30]	"[D]esign patterns wherein language use prevents or makes it very difficult for the user to understand crucial functionality [...]".	*Language Inaccessibility*	extends meso-level
Untranslation [30]	"[D]esign patterns in which part or all of the app is in a language unfamiliar to the people using it, even if the app is stated as available in the local language in the store".	*Wrong Language*	extends low-level
Alphabet Soup [30]	"[D]esign pattern language use prevents or makes it very difficult for the user to understand crucial functionality [...]".	*Language Inaccessibility*	new low-level
Extraneous Badges [29]	"[D]esign elements — often tiny, brightly colored circles—that visually highlight UI elements that require immediate user attention".	*Aesthetic Manipulation*	new low-level
Account Deletion Roadblocks [29]	"*Unclear deactivation/deletion options* covers cases where a service insufficiently communicates what will happen if a person deactivates or deletes their account."	*Roach Motel*	new low-level
	"*Time-Delayed Account Deletion* covers cases where a service will only initiate the account deletion process after a cool-off period, rather than instantaneously."	*Roach Motel*	new low-level
Engaging Strategies [43]	"[D]ark patterns where the goal is to keep users occupied and entertained for as long as possible".	*Social Engineering*	extends high-level
Governing Strategies [43]	Dark patterns "that navigate users' decision-making towards the designers' and/or platform providers' goals".	*Obstruction*	extends high-level
Labyrinthine Navigation [43]	"[N]ested interfaces that are easy to get lost in, disabling users from choosing preferred settings".	*Privacy Maze*	extends low-level

Table 1: This table presents an overview of selected dark patterns from Hidaka et al. [30], Gunawan et al. [29], and Mildner et al. [43] to demonstrate extending the dark pattern ontology.

6.1 Dark Patterns In Japanese Apps

Hidaka et al. [30] studied dark patterns in Japanese apps and identified two dark pattern types—*Unstranslation* and *Alphabet Soup*—which are sub-types of a novel *Linguistic Dead-End* dark pattern. They specifically motivated their work as one of the first studies of dark patterns in a non-Western context. We closely evaluated the authors' definition of *Linguistic Dead-End*, where the use of a foreign language hinders users from understanding the consequences of their interactions. When comparing these three patterns to our ontology, the high-level pattern *Linguistic Dead-End* appears to fit within the existing meso-level dark pattern *Language Inaccessibility* while extending its coverage. The remaining two low-level patterns, *Untranslation* and *Alphabet Soup*, can then be nested as two low-level types underneath the same meso-level dark pattern, with *Untranslation* mapping to and extending the existing *Wrong Language* dark pattern and *Alphabet Soup* forming a new low-level pattern. In this case, the three dark patterns extend and further support a distinct area of the ontology, demonstrating how novel contexts help to usefully supplement existing dark patterns and identify new low-level means of execution. Additionally, this study demonstrates that dark patterns exist across multiple cultures and areas of the world, but may take different forms depending on local design norms.

6.2 Contextual Dark Patterns in Different Screen Modalities

Gunawan et al. [29] investigated the presence of dark patterns across different screen modalities, describing eight novel dark pattern types which limit the choices of users depending on the device used. In the provided definitions for each proposed dark pattern, the authors included links to previously defined dark patterns—linking these patterns to elements of the ontology, thus providing an easy

mapping path. The *Extraneous Badges* dark pattern, for example, is indicated as related to *Aesthetic Manipulation* [24] as a form of *Interface Interference*, and would result in this dark pattern being included as a new low-level type in the ontology. Similarly, using the authors' definitions and identification of mapping in the paper text, *Account Deletion Roadblocks* could extend *Roach Motel* through two specific new low-level types focusing variously on insufficient communication and time delay: *Unclear Deactivation/Deletion Options* and *Time-Delayed Account Deletion*. These examples illustrate how contextual and situational links to previously defined dark patterns support the ontology, describing specific situations that strengthen established dark patterns and identify new low-level means of execution.

6.3 Domain-Specific Dark Patterns in Social Media Applications

Mildner et al. [43] investigated dark patterns on social media platforms, proposing five dark patterns across two strategies. As with Hidaka et al., the granularity of their definitions implies a mapping on multiple levels of the ontology. We began by drawing from the authors' definitions of *Engaging Strategies* and *Governing Strategies*. The authors describe the aim of *Engaging Strategies* as entertaining users for as long as possible, related to *Attention Capture* [44], which is already included in the ontology as a meso-level pattern under *Forced Action*. However, some elements of the original definition (e.g., occupying and entertaining) fit more closely within concepts of *Social Engineering*. Similarly, *Governing Strategies* can be partially linked to multiple patterns in the ontology. For example, as the authors originally suggest, the strategy can be enabled through *Interface Interference*. However, *Governing Strategies* also offers a high-level focus to inspect *Obstruction* with *Labyrinthine Navigation*, presenting an interesting adaption of *Privacy Maze* already present in the ontology. These examples indicate how the authors make their dark pattern types distinct from prior ones, functioning as a lens that might invite reinspection of dark patterns in the ontology and perhaps indicate opportunities for further development of low-level patterns.

6.4 Dark Patterns in Legislation and Case Law

An alignment between legislation, the ontology, and case law shows that it could also be a robust and reliable artifact for regulators and policy makers to use in their compliance monitoring and enforcement actions.

Mapping the ontology to case law Dark patterns have been detected in regulatory cases by enforcers, such as Data Protection Authorities (DPAs) and Consumer Protection Authorities, for more than a decade [15, 21]. However few cases explicitly designate dark patterns as such.[7] Decisions analyse several practices that are related to dark patterns, but without qualifying each practice into a concrete granular type of dark pattern. Current case law descriptions of the use of dark patterns often report infringements only at a general level, but without qualifying each practice as a concrete type of dark pattern [50]. In doing so, case law could miss lower-level granularity that may translate across domains. A recent example shows that a EU regulator, the Italian DPA, used the concept of dark patterns related to certain consent practices for the first time in an official EU legal decision [34]. By mapping case law to the ontology, regulators can gain additional knowledge identifying where dark patterns practices at multiple levels and in multiple combinations are at play, and were deemed to be illegal per jurisdiction [36], enhancing legal certainty about dark patterns practices. For example, the EU Court of Justice has ruled that the practice called "pre-selection" violates the GDPR [9], which maps to the meso-level dark pattern *"Bad Defaults"* in our ontology.

Further, the ontology has the potential to support enforcement decisions since it can test and confirm the traceability of concrete dark patterns-related practices. For instance, the Italian Data Protection Authority has already added the keyword "dark pattern" to the available tags of their online database[8]—a useful effort that should be extended to official and unofficial searchable databases of enforcers' decisions. Connecting case law to multiple levels of dark patterns in our proposed ontology has the potential to inform enforcers of different jurisdictions in the EU/US and reduce the risks of gaps or overlaps.

Mapping the ontology to legislation The proposed ontology can also help regulators across different jurisdictions to understand relationships between different definitions of dark patterns, including high-, meso- and low level dark patterns, including when such definitions map to existing and upcoming legislation. The recent EU Digital Service Act (DSA)[14, Art.25(3)(b), recital 67] explicitly prohibits user manipulation and specifies that further guidelines will be given on a specific practice, where "repeatedly requesting that the recipient of the service make a choice where that choice has already been made, especially by presenting pop-ups that interfere with the user experience"; this example maps well to the proposed *Nagging* dark pattern in our ontology. Because new legislation, such as the DSA[14], Data Market Act (DMA)[13], Data Act [12], and California CPRA [11] contain dark patterns specific prohibitions, we believe the proposed ontology has the capability to ensure a precise mapping between the concepts of dark patterns in research literature and the legally-binding provisions. When the concepts of the ontology are mapped to a legal concept, then it is easier for regulators to link a specific dark pattern to a concrete binding legislative provision. Consequently, the ontology will help to conclude the normative value of such practice—whether a specific dark pattern is illegal or legal—and what relevant obligations and rights are derived from the law and must be enforced. If regulators and policy-makers across jurisdictions rely on the same definitions of dark patterns, this can assure an easier re-use of case law for future legal cases.

7 USING THE ONTOLOGY TO SUPPORT TRANSDISCIPLINARY ENGAGEMENT

In this ontology, we seek to synthesize and harmonize existing academic and regulatory taxonomies while adding useful and consistent structure to allow for other stakeholders to build upon and derive benefit from a shared description of dark patterns knowledge. This paper lays the foundation for shared action, which includes many different stakeholders with differing aims. In this section,

[7] Case law and legal frameworks have recently been added to the https://deceptive.design site, which includes mappings to specific dark patterns [2].

[8] https://www.garanteprivacy.it/temi/internet-e-nuove-tecnologie/dark-pattern

we outline key opportunities for future transdisciplinary engagement, identifying opportunities for scholars to continue building knowledge about dark patterns and their harms, for regulators and other enforcement agencies to better detect and thus sanction dark patterns, and for legal scholars and legislators to address current and future consequences of dark patterns that can inform further action.

7.1 Challenges in Evolving the Ontology

Not all of our mappings were clear-cut and some may be productively extended or disputed in future versions of this ontology. Through dialogue, we sought to locate existing patterns within our ontology based on our best understanding of the pattern as described by its name and definition in the source taxonomy. One challenge we faced was that some combinations of patterns have evolved over time. For instance, Mathur et al.'s [40] high-level pattern "social proof" originated with two sub-patterns, "activity messages" and "testimonials." Later, the FTC created new low-level patterns, introducing "endorsements" (we bundled it with testimonials as one low-level pattern) and more specific types of endorsement or testimonials (e.g., "deceptive celebrity endorsements," "false activity messages"). Future work could identify the most useful level of abstraction for these patterns.

Additionally, the use of novel names for patterns (particularly by the EDPB and CMA) or the use of patterns in specific contexts (e.g., e-commerce, social media) caused us to consider both the presence of granular low-level patterns and the relation of these low-level patterns to inferred meso-level patterns. In particular, the use of novel names for patterns types and definitions was a challenge from an analytic perspective, resulting in: i) instances where a wholly new pattern was introduced (e.g., CMA's "information overload" which we leveraged to create a new meso-level pattern of "choice overload"); ii) instances where a new high-level strategy was highly similar to an existing high-level strategy (e.g., EDPB's "skipping" which we subsumed within "sneaking"); and iii) instances where existing patterns included both a generalizable pattern and domain-specific information which may need to be captured in specific low-level patterns in future work (e.g., EDPB's "left in the dark" is a form of "hidden information" but implies specific low-level patterns that are specific to data protection).

These observed challenges point towards the value of a shared ontology that includes a consistent vocabulary, but also points to opportunities to generate more specific knowledge that is linked to particular contexts and technologies. For instance, low-level patterns could be tagged based on how well they relate to specific contexts (e.g., e-commerce, social media), technologies (e.g., CUIs, VR/AR, robots), or application domains (e.g., health, travel) as indicated by a recent systematic review of dark patterns literature [22].

Finally, formal evaluation of the definitions and ontology structure we have proposed will strengthen our understanding of how various stakeholders consider, interpret, and use the ontology to support their work—within and across technology contexts. For instance, the language specificity demanded by a legal or regulatory professional from a given definition within the ontology may require different kinds of analytic precision as compared to the generative or evaluative use of the same definitions by a designer performing an audit of dark patterns on digital systems for their company. Future work should address both the utility and the rigor of various components of our ontology for differing purposes, including expert evaluation, gathering of evidence for legal and regulatory action, operationalization of dark patterns for social science research, and use by designers to avoid inscribing dark patterns into their design work.

7.2 Activating Transdisciplinary Pathways

As we have outlined, work relating to dark patterns has connected many different disciplinary communities toward shared goals, including social scientists studying the presence and harms of dark patterns, legal scholars linking instances of dark patterns to relevant consumer protection or data protection legal frameworks, legislators targeting specific legal provisions about dark patterns to support new obligations and/or future sanctions, and regulators detecting legal violations related to dark patterns to support enforcement sanctions. We consider multiple opportunities for collaboration within and across these stakeholder groups:

- **Social Scientists** Scientists studying dark patterns can use the ontology to better map the impact triggered by certain dark patterns in concrete contexts in ways that support shared knowledge building and reduce duplication. This approach has been applied for specific low-level patterns by various empirical studies that evaluated the impact of dark pattern design on the outcome of users' consent decisions [45], but could be scaled up substantially using the ontology as a means of producing and sharing these mappings.
- **Social Scientists + Computer Scientists** The detection of dark patterns could also be more robustly supported by our ontology, with our assertion that low-level patterns show the most promise in being detectable. Existing detection efforts (e.g., [5, 8, 35, 40, 46, 52–54, 57]) have shown that higher-level patterns are difficult or impossible to detect at scale due to their abstract nature that requires interpretation, while low-level UI elements with discrete and known qualities (e.g., cookie consent banners, elements of the checkout process) are more detectable using software tools for automated detection. Our ontology of low-level patterns and gaps creates a foundation for future detection efforts, allowing computer science scholars to focus on pattern types which are most likely to be detectable and measurable.
- **Social Scientists + Regulators** Bielova et al. [3] have recently compared the results of such empirical studies and designs recommended by EU regulators and found multiple gaps and contradictions relating to instances of dark patterns, showing that empirical studies bring important insights not only in the research community but also for the regulators and policy-makers. This effort demonstrates an opportunity for regulators and social scientists to work more closely—commissioning studies where user experience of dark patterns is unknown or unclear (particularly with relation to causal mechanisms) while deprioritizing studies

that address design choices that are already illegal under statute.

- **Social Scientists + Legal Scholars** The ontology can be extended to consider potential harms in relation to specific dark patterns types [28]. For example, the meso-level dark pattern *Nagging* can arguably trigger "attentional theft," thus harming consumer welfare, and can lead to indirect harms such as increased vulnerability to privacy violations, and finally, to anti-competitive harms [32]. A mapping of harms to specific types of dark patterns in the ontology may support connections to avenues for legal remedies, as well as aid in identifying areas where additional research is needed.
- **Legal Scholars + Regulators** The ontology may also be extended to refer to concrete enforcement cases already consolidated in a database of dark patterns case law, such as those on Brignull's updated site [2]. This will allow for case law to inform future legal sanctions, identify which elements of the ontology connect to existing legal frameworks, and lay the groundwork for future legislative action to allow for sanctioning of novel patterns that are not well addressed through existing laws.

8 CONCLUSION

To support the development of a shared language of dark patterns, in this paper we present our analysis of ten existing regulatory and academic taxonomies of dark patterns and propose a three-level ontology with standardized definitions for 65 synthesized dark pattern types across low-, meso-, and high-level patterns. Building on our analysis, future scholars, regulators, and legal professionals can benefit from our hierarchical organization of dark patterns types to indicate links to existing and similar concepts. This description encourages the establishment of provenance in future work, allowing scholars and regulators to identify pattern types and their origins and provide an audit trail to connect specific contextually-bound instances with broader categorizations. This ontology creates a foundation for a shared and reusable knowledge source, allowing many stakeholders to work together in building a shared, explicit and precise conceptualization of what is already known in the literature and which can be further refined and extended. Finally, we illustrate how this ontology can support translational research and regulatory action, by extending the ontology from three contemporary studies defining dark patterns from domain and context-specific areas, as well as ontology extension to map legislation and case law.

ACKNOWLEDGMENTS

This work is funded in part by the National Science Foundation under Grant No. 1909714 and the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR. The research of this work was partially supported by the Klaus Tschira Stiftung gGmbH.

REFERENCES

[1] 2023. Central Consumer Protection Authority issues 'Guidelines for Prevention and Regulation of Dark Patterns, 2023' for prevention and regulation of dark patterns listing 13 specified dark patterns. https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1983994

[2] 2023. Deceptive patterns - Legal Cases. https://www.deceptive.design/cases Accessed: 2023-9-14.

[3] Nataliia Bielova, Cristiana Santos, and Colin M Gray. 2024. Two worlds apart! Closing the gap between regulating EU consent and user studies. *Harvard Journal of Law & Technology* 37 (2024).

[4] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* 2016, 4 (2016). https://doi.org/10.1515/popets-2016-0038

[5] Ahmed Bouhoula, Karel Kubicek, Amit Zac, Carlos Cotrini, and David Basin. 2023. Automated, Large-Scale Analysis of Cookie Notice Compliance. In *USENIX Security Symposium*.

[6] Harry Brignull. 2018. Deceptive Patterns: User Interfaces Designed to Trick People. http://darkpatterns.org/

[7] Harry Brignull. 2023. Deceptive Patterns. https://www.deceptive.design

[8] Jieshan Chen, Jiamou Sun, Sidong Feng, Zhenchang Xing, Qinghua Lu, Xiwei Xu, and Chunyang Chen. 2023. Unveiling the Tricks: Automated Detection of Dark Patterns in Mobile Applications *(UIST '23 Adjunct)*. Association for Computing Machinery, San Francisco, CA, USA, 1–20. http://arxiv.org/abs/2308.05898 arXiv:2308.05898 [cs].

[9] CJEU-Planet49-19 2019. Judgment in Case C-673/17 Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband eV v Planet49 GmbH. http://curia.europa.eu/juris/documents.jsf?num=C-673/17.

[10] CMA2022 2022. *Evidence review of Online Choice Architecture and consumer and competition harm.* Technical Report. https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers/evidence-review-of-online-choice-architecture-and-consumer-and-competition-harm Accessed: 2022-4-13.

[11] CPRA 2022. California Privacy Rights Act. https://cppa.ca.gov/meetings/materials/20220608_item3.pdf

[12] Data-Act-proposal 2022. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

[13] DMA 2022. Digital Markets Act - Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance). http://data.europa.eu/eli/reg/2022/1925/oj

[14] DSA2022 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[15] 2022. *Behavioural study on unfair commercial practices in the digital environment : dark patterns and manipulative personalisation : final report.* Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/606365bc-d58b-11ec-a95f-01aa75ed71a1/language-en/format-PDF/source-257599418

[16] European Data Protection Board. 2022. *Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them.* Technical Report Version 1.0. https://edpb.europa.eu/system/files/2022-03/edpb_03-2022_guidelines_on_dark_patterns_in_social_media_platform_interfaces_en.pdf

[17] European Data Protection Board. 2023. *Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them.* Technical Report Version 2.0. https://edpb.europa.eu/system/files/2023-02/edpb_03-2022_guidelines_on_deceptive_design_patterns_in_social_media_platform_interfaces_v2_en_0.pdf

[18] Frederico Fonseca. 2007. The double role of ontologies in information science research. *Journal of the American Society for Information Science and Technology* 58, 6 (April 2007), 786–793. https://doi.org/10.1002/asi.20565

[19] French DPA (CNIL). 2021. Deliberation of the restricted committee No. SAN-2021-023 of 31 December 2021 concerning GOOGLE LLC and GOOGLE IRELAND LIMITED. https://www.cnil.fr/sites/default/files/atoms/files/deliberation_of_the_restricted_committee_no._san-2021-023_of_31_december_2021_concerning_google_llc_and_google_ireland_limited.pdf.

[20] French DPA (CNIL). 2022. Deliberation of the restricted committee No. SAN-2021-024 of 31 December 2021 concerning FACEBOOK IRELAND LIMITED. https://www.cnil.fr/sites/default/files/atoms/files/deliberation_of_the_restricted_committee_no._san-2021-024_of_31_december_2021_concerning_facebook_ireland_limited.pdf.

[21] FTC2022 2022. *Bringing Dark Patterns to Light Staff Report.* Technical Report. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report%209.14.2022%20-%20FINAL.pdf

[22] Colin M Gray, Lorena Sánchez Chamorro, Ike Obi, and Ja-Nae Duane. 2023. Mapping the Landscape of Dark Patterns Scholarship: A Systematic Literature Review. In *Designing Interactive Systems Conference (DIS Companion '23)* (Pittsburgh, PA, USA), Vol. 1. Association for Computing Machinery. https://doi.org/10.1145/3563703.3596635

[23] Colin M Gray, Johanna Gunawan, René Schäfer, Nataliia Bielova, Lorena Sánchez Chamorro, Katie Seaborn, Thomas Mildner, and Hauke Sandhaus. 2024. Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA'24)*. Association for Computing Machinery. https://doi.org/10.1145/3613905.3636310

[24] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. dl.acm.org, New York, NY, USA, 534:1–534:14. https://doi.org/10.1145/3173574.3174108

[25] Colin M Gray, Cristiana Santos, and Nataliia Bielova. 2023. Towards a Preliminary Ontology of Dark Patterns Knowledge. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. https://doi.org/10.1145/3544549.3585676

[26] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. 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)*. ACM Press. https://doi.org/10.1145/3411764.3445779

[27] Colin M Gray, Cristiana Santos, Nicole Tong, Thomas Mildner, Arianna Rossi, Johanna Gunawan, and Caroline Sinders. 2023. Dark Patterns and the Emerging Threats of Deceptive Design Practices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. https://doi.org/10.1145/3544549.3583173

[28] Gunawan, Santos, and Kamara. 2022. Redress for Dark Patterns Privacy Harms? A Case Study on Consent Interactions. *2022 ACM Symposium on* (2022). https://johannagunawan.com/assets/pdf/gunawan-22-cslaw.pdf

[29] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2021. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–29. https://doi.org/10.1145/3479521

[30] Shun Hidaka, Sota Kobuki, Mizuki Watanabe, and Katie Seaborn. 2023. Linguistic Dead-Ends and Alphabet Soup: Finding Dark Patterns in Japanese Apps. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–13. https://doi.org/10.1145/3544548.3580942

[31] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (Nov. 2005), 1277–1288. https://doi.org/10.1177/1049732305276687

[32] Alison Hung. 2021. KEEPING CONSUMERS IN THE DARK: ADDRESSING "NAGGING" CONCERNS AND INJURY. *Columbia Law Review* 121, 8 (2021), 2483–2520. https://www.jstor.org/stable/27093855

[33] india-2023-ek 2023. Draft Guidelines for Prevention and Regulation of Dark Patterns. https://consumeraffairs.nic.in/sites/default/files/file-uploads/latestnews/Draft%20Guidelines%20for%20Prevention%20and%20Regulation%20of%20Dark%20Patterns%202023.pdf

[34] ItalianDPAvsEdiscom2023 2023. Provvedimento prescrittivo e sanzionatorio nei confronti di Ediscom S.p.A. - 23 febbraio 2023 [9870014]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870014

[35] Simon Koch, Benjamin Altpeter, and Martin Johns. 2023. The OK is not enough: Large Scale Study of Consent Dialogs in Smartphone Applications. In *USENIX Security Symposium*.

[36] Mark Leiser. 2020. *'Dark Patterns': the case for regulatory pluralism*. LawArXiv ea5n2. Center for Open Science. https://doi.org/10.31219/osf.io/ea5n2

[37] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a Light on Dark Patterns. *Journal of Legal Analysis* 13, 1 (March 2021), 43–109. https://doi.org/10.1093/jla/laaa006

[38] Kai Lukoff, Alexis Hiniker, Colin M Gray, Arunesh Mathur, and Shruthi Sai Chivukula. 2021. What can CHI do about dark patterns?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan). ACM, New York, NY, USA. https://doi.org/10.1145/3411763.3441360

[39] Luxembourg DPA. 2021. Decision regarding Amazon Europe Core S.À RL. https://cnpd.public.lu/fr/actualites/international/2021/08/decision-amazon-2.html.

[40] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), Article No. 81. https://doi.org/10.1145/3359183

[41] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445610

[42] Thomas Mildner and Gian-Luca Savino. 2021. Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21, Article 464)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451659

[43] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580695

[44] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. 2023. Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3580729

[45] Nataliia Bielova. 2023. A survey of user studies as evidence for dark patterns in consent banners. https://backoffice.cnil.fr/sites/default/files/atoms/files/full_2022-12-02_v2.pdf, accessed on 7 September 2023..

[46] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376321

[47] Ikechukwu Obi, Colin M Gray, Shruthi Sai Chivukula, Ja-Nae Duane, Janna Johns, Matthew Will, Ziqing Li, and Thomas Carlock. 2022. Let's Talk About Socio-Technical Angst: Tracing the History and Evolution of Dark Patterns on Twitter from 2010-2021. (July 2022). arXiv:2207.10563 [cs.SI] http://arxiv.org/abs/2207.10563

[48] OECD. 2022. *Dark commercial patterns*. Technical Report. https://doi.org/10.1787/44f5e846-en

[49] Press-Release2022-fq 2022. Press Release: AG Racine announces Google must pay $9.5 million for using "dark patterns" and deceptive location tracking practices that invade users' privacy. https://thedcline.org/2022/12/30/press-release-ag-racine-announces-google-must-pay-9-5-million-for-using-dark-patterns-and-deceptive-location-tracking-practices-that-invade-users-privacy/. https://thedcline.org/2022/12/30/press-release-ag-racine-announces-google-must-pay-9-5-million-for-using-dark-patterns-and-deceptive-location-tracking-practices-that-invade-users-privacy/ Accessed: 2022-12-31.

[50] Cristiana Santos and Arianna Rossi. 2023. The emergence of dark patterns as a legal concept in case law. https://policyreview.info/articles/news/emergence-of-dark-patterns-as-a-legal-concept

[51] Brennan Schaffner, Neha A Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142

[52] Than Htut Soe, Cristiana Teixeira Santos, and Marija Slavkovik. 2022. Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way. (April 2022). arXiv:2204.11836 [cs.LG] http://arxiv.org/abs/2204.11836

[53] Ioannis Stavrakakis, Andrea Curley, Dympna O'Sullivan, Damian Gordon, and Brendan Tierney. 2021. A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically. (2021). https://doi.org/10.21427/20g8-d176

[54] Marieke Van Hofslot, Almila Akdag Salah, Albert Gatt, and Cristiana Santos. 2022. Automatic Classification of Legal Violations in Cookie Banner Texts. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 287–295. https://aclanthology.org/2022.nllp-1.27

[55] Jess Weatherbed. 2022. Google is paying a $85m settlement to Arizona to end user-tracking suit. https://www.theverge.com/2022/10/5/23389331/google-settlement-arizona-user-tracking-privacy-suit. https://www.theverge.com/2022/10/5/23389331/google-settlement-arizona-user-tracking-privacy-suit Accessed: 2023-1-4.

[56] Shoshana Wodinsky. 2022. The 'dark patterns' in Fortnite that led to the largest FTC penalties ever. https://www.marketwatch.com/story/the-dark-patterns-in-fortnite-that-led-to-the-largest-ftc-penalties-ever-11671488228. https://www.marketwatch.com/story/the-dark-patterns-in-fortnite-that-led-to-the-largest-ftc-penalties-ever-11671488228 Accessed: 2022-12-20.

[57] Yuki Yada, Jiaying Feng, Tsuneo Matsumoto, Nao Fukushima, Fuyuko Kido, and Hayato Yamana. 2022. Dark patterns in e-commerce: a dataset and its baseline evaluations. In *2022 IEEE International Conference on Big Data (Big Data)*. 3015–3022. https://doi.org/10.1109/BigData55660.2022.10020800

[58] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Foundations of Digital Games*. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1043332&dswid=1018

A FINAL ONTOLOGY DEFINITIONS

- **Sneaking** is a strategy which hides, disguises, or delays the disclosure of important information that, if made available to users, would cause a user to unintentionally take an action they would likely object to.
  - **Bait and Switch** subverts the user's expectation that their choice will result in a desired action, instead leading to an unexpected, undesirable outcome.
    - **Disguised Ads** *Bait and Switch* and use *Sneaking* to style interface elements so they are not clearly marked as an advertisement or other biased source. As a result, users are induced into clicking on the interface element because they assume that it is a relevant and salient interaction, leading to unwitting interaction with advertising content.
  - **Hiding Information** subverts the user's expectation that all relevant information to make an informed choice will be available to them, instead hiding information or delaying the disclosure of information until later in the user journey that may have led to them making another choice.
    - **Sneak into Basket** *Hides Information* and uses *Sneaking* to add unwanted items to a user's shopping cart without their consent. As a result, a user assumes that only the items they explicitly added to their cart will be purchased, leading to unintentional purchase of additional items.
    - **Drip Pricing, Hidden Costs, or Partitioned Pricing** *Hides Information* and uses *Sneaking* to reveal new charges or costs, present only partial price components, or otherwise delay revealing the full price of a product or service through late or incomplete disclosure. As a result, the user is misled about the total or complete price of the product or service, leading to them to make a purchase decision after they have expended effort on false pretenses.
    - **Reference Pricing** *Hides Information* and uses *Sneaking* to include a misleading or inaccurate price for a product or service that makes a discounted price appear more attractive. As a result, the user is misled into believing that the price they pay is discounted, leading them to make a decision to purchase a product or service on false pretenses.
  - **(De)contextualizing Cues** subverts the user's expectation that provided information will guide the user to making an informed choice, instead confusing the user and/or preventing them from locating relevant information due to the context where information is presented.
    - **Conflicting Information** uses *(De)contextualizing Cues* and *Sneaking* to include two or more sources of information that conflict with each other. As a result, the user is unsure what the consequences of their actions will be and will be more likely to accept default settings that may not be in their best interest.
    - **Information without context** uses *(De)contextualizing Cues* and *Sneaking* to alter the relevant information or user controls to limit discoverability. As a result, the user is unlikely to find the information or action possibility they are interested in.
- **Obstruction** is a strategy which impedes a user's task flow, making an interaction more difficult than it inherently needs to be, dissuading a user from taking an action.
  - **Roach Motel** subverts the user's expectation that an action will be as easy to reverse as it is to make, instead creating a situation that is easy to get into, but difficult to get out of.
    - **Immortal Accounts** create a *Roach Motel* and use *Obstruction* to make it difficult or impossible to delete a user account once it has been created. As a result, the user may create an account or share data with the false assumption that they can later delete this information, even though that account and/or data are then unable to be removed by the user.
    - **Dead Ends** create a *Roach Motel* and use *Obstruction* to prevent users from finding information through inactive links or redirections that limit or completely prevent the display of relevant information. As a result, the user may seek to find relevant information or action possibilities but instead be left unable to achieve their goal.
  - **Creating Barriers** subverts the user's expectation that relevant user tasks will be supported by the interface, instead preventing, abstracting, or otherwise complicating a user task to disincentive user action.
    - **Price Comparison Prevention** *Creates Barriers* and uses *Obstruction* by excluding relevant information, limiting the ability of a user to copy/paste, or otherwise inhibiting a user from comparing prices across two or more vendors. As a result, the user cannot make an informed decision about where to buy a product or service.
    - **Intermediate Currencies** *Create Barriers* and use *Obstruction* to hide the true cost of a product or service by requiring the user to spend real money to purchase a virtual currency that is then used to purchase a product or service. As a result, the user is unable to easily ascertain the true monetary cost of a product or service, leading them to make an uninformed purchase decision based on an obscured cost.
  - **Adding Steps** subverts the user's expectation that a task will take as few steps as technologically needed, instead creating additional points of unnecessary but required user interaction to perform a task.
    - **Privacy Mazes** *Add Steps* and use *Obstruction* to require a user to navigate through many pages to obtain relevant information or control without a comprehensive and exhaustive overview. As a result, the user is prevented from easily discovering relevant information or action possibilities, leaving them unable to make informed decisions regarding their privacy.
- **Interface Interference** is a strategy which privileges specific actions over others through manipulation of the user interface, thereby confusing the user or limiting discoverability of relevant action possibilities.

  - **Manipulating Choice Architecture** subverts the user's expectation that the options presented will support their desired goal, instead including an order or structure of options that makes another outcome more likely.
    * **False Hierarchy** *Manipulates the Choice Architecture,* using *Interface Interference* to give one or more options visual or interactive prominence over others, particularly where items should be in parallel rather than hierarchical. As a result, the user may misunderstand or be unable to accurately compare their options, making a selection based on a false or incomplete choice architecture.
    * **Visual Prominence** *Manipulates the Choice Architecture,* using *Interface Interference* to place an element relevant to user goals in visual competition with a more distracting and prominent element. As a result, the user may forget about or be distracted from their original goal, even if that goal was their primary intent.
    * **Bundling** *Manipulates the Choice Architecture,* using *Interface Interference* to group two or more products or services in a single package at a special price. As a result, the user may incorrectly assume that these items must be purchased as a bundle or be unaware of the unbundled price for the component elements, possibly leading to an uninformed purchasing decision.
    * **Pressured Selling** *Manipulates the Choice Architecture,* using *Interface Interference* to preselect or use visual prominence to focus user attention on more expensive product options . As a result, the user may be unaware that a lower price is available or even desirable for their needs , steering the user into making a more expensive product selection than they otherwise would have.
  - **Bad Defaults** subverts the user's expectation that default settings will be in their best interest, instead requiring users to take active steps to change settings that may cause harm or unintentional disclosure of information.
  - **Emotional or Sensory Manipulation** subverts the user's expectation that the design of the site will allow them to achieve their goal without manipulation, instead altering the language, style, color, or other design elements to evoke an emotion or manipulate the senses in order to persuade the user into a particular action.
    * **Cuteness** uses *Emotional or Sensory Manipulation* and *Interface Interference* to embed attractive cues in the design of a robot interface or form factor. As a result, a user may place undue trust in the robot, leading the user to inaccurately or incompletely assess the risks of interacting with the robot.
    * **Positive or Negative Framing** uses *Emotional or Sensory Manipulation* and *Interface Interference* to visually obscure, distract, or persuade a user from important information they need to achieve their goal. As a result, the user may assume that the system is providing equal access to relevant information, leading the user to be distracted by positive or negative aesthetic cues that distract them from important information or action possibilities or otherwise convince them to pursue a different goal.
  - **Trick Questions** subvert the user's expectation that prompts will be written in a straightforward and intelligible manner, instead using confusing wording, double negatives, or otherwise leading language or interface cues to manipulate a user's choice.
  - **Choice Overload** subverts the user's expectation that the choices they make should be understandable and comparable, instead providing too many options to compare or encouraging users to overlook relevant information due to the volume of choices provided.
  - **Hidden Information** subverts the user's expectation that relevant information will be made accessible and visible, instead disguising relevant information or framing it as irrelevant.
  - **Language Inaccessibility** subverts the user's expectation that guidance will be provided in a way that is understandable and intelligible, instead using unnecessarily complex language or a language not spoken by the user to decrease the likelihood the user will make an informed choice.
    * **Wrong Language** leverages *Language Accessibility*, using *Interface Interference* to provide important information in a different language than the official language of the country where users live. As a result, the user will not have access to relevant information about their interaction with the system and their ability to choose, leading to uninformed decisions.
    * **Complex Language** leverages *Language Accessibility*, using *Interface Interference* to make information difficult to understand by using obscure word choices and/or sentence structure. As a result, the user will not be able to comprehend relevant information about their interaction with the system and their ability to choose, leading to uninformed decisions.
  - **Feedforward Ambiguity** subverts the user's expectation that their choice will be likely to result in an action they can predict, instead providing a discrepancy between information and actions available to users that results in an outcome that is different from what the user expects.
- **Forced Action** is a strategy which requires users to knowingly or unknowingly perform an additional and/or tangential action or information to access (or continue to access) specific functionality, preventing them from continuing their interaction with a system without performing that action.
  - **Nagging** subverts the user's expectation that they have rational control over the interaction they make with a system, instead distracting the user from a desired task the user is focusing on to induce an action or make a decision the user does not want to make by repeatedly interrupting the user during normal interaction.
  - **Forced Continuity** subverts the user's expectation that a subscription created in the past will not auto-renew or otherwise continue in the future, instead causing undesired charges, difficulty to cancel, or lack of awareness that a subscription is still active.
  - **Forced Registration** subverts the user's expectation that they can complete an action without registering or creating an account, instead tricking them into thinking that registration is required, often resulting in the sharing of unneeded personal data.

  - **Forced Communication or Disclosure** subverts the user's expectation that a system will only request information needed to complete their desired goals, instead tricking them into sharing more information about themselves or using their information for purposes that they do not desire.
    * **Privacy Zuckering** uses *Forced Communication or Disclosure* as a type of *Forced Action* to trick users into sharing more information about themselves than they intend to or would agree to if fully informed. As a result, the user assumes that information they are requested to provide is vital for use of the service, even while this information is used or sold for other purposes.
    * **Friend Spam** uses *Forced Communication or Disclosure* as a type of *Forced Action* to collect information about other users through extractive means that results in unwanted contact from the service. As a result, the user assumes that information about their friends or social network is vital for use of the service, even while this information is used to spam other users.
    * **Address Book Leeching** uses *Forced Communication or Disclosure* as a type of *Forced Action* to collect information about other users through extractive means, which are often hidden to the user and/or conducted under false pretenses. As a result, the user assumes that only vital information will be collected when signing up for or using a service, even while this information is used to gain knowledge of other users or inform other purposes that have not been initially declared.
    * **Social Pyramid** uses *Forced Communication or Disclosure* as a type of *Forced Action* to manipulate existing users into recruiting new users to use a service, often by tying this recruitment to additional functionality or other benefits. As a result, the user assumes that social recruiting is necessary to continue to use aspects of the service, even while this information is primarily used to build the service's user base.
  - **Gamification** subverts the user's expectation that system functionality is based on alignment with user goals and needs, instead coercing them into gaining access to aspects of a service through repeated (and perhaps undesired) use of aspects of the service.
    * **Pay-to-Play** uses *Gamification* as a type of *Forced Action* to initially claim that aspects of a service or product are available via purchase or download, but then later charging users to actually obtain that functionality. As a result, the user incorrectly assumes that a service or product will allow them certain functionality, leading to them downloading or purchasing the product or service under false pretenses.
    * **Grinding** uses *Gamification* as a type of *Forced Action* to require repeated, often cumbersome and labor-intensive actions over time in order to obtain certain relevant functionality. As a result, the user may seek to avoid these repetitive actions, leading to them making unwanted additional in-app purchases to unlock the same functionality without "grinding" over an extended period of time.
  - **Attention Capture** subverts the user's expectation that they have rational control over the time they spend using a system, instead tricking them into spending more time or other resources to continue use for longer than they otherwise would.
    * **Auto-Play** uses *Attention Capture* as a type of *Forced Action* to automatically play new video after an existing video has completed. As a result, the user may lose control over their viewing experience, leading them to watch more content than they intended or result in them watching content that is unexpected or harmful.
- **Social Engineering** is a strategy which presents options or information that causes a user to be more likely to perform a specific action based on their individual and/or social cognitive biases, thereby leveraging a user's desire to follow expected or imposed social norms.
  - **Scarcity or Popularity Claims** subverts the user's expectation that information provided about a product's availability or desirability is accurate , instead pressuring the user to purchase a product without additional reflection or verification.
    * **High Demand** uses *Scarcity and Popularity Claim*s as a type of *Social Engineering* to indicate that a product is in high-demand or likely to sell out soon, even though that claim is misleading or false. As a result, the user may assume that demand is high when it is not, leading to their uninformed purchase of a product or service.
  - **Social Proof** subverts the user's expectation that the indicated behavior of others in a specific situation is correct or desirable, instead accelerating user decision-making and encouraging the user to trust flawed implications through provided information.
    * **Low Stock** uses *Social Proof* as a type of *Social Engineering* to indicate that a product is limited in quantity, even though that claim is misleading or false. As a result, the user may assume that a product is desirable due to demand, leading to undue or uninformed pressure to buy the product immediately.
    * **Endorsements and Testimonials** use *Social Proof* as a type of *Social Engineering* to indicate that a product or service has been endorsed by another consumer, even though the source of that endorsement or testimonial is biased, misleading, incomplete, or false. As a result, the user may assume that the endorsement or testimonial is accurate and unbiased, leading to their uninformed purchase of a product or service.
    * **Parasocial Pressure** uses *Social Proof* as a type of *Social Engineering* to indicate that a product or service has been endorsed by a celebrity, influencer, or other entity that the user trusts, even though the source of that endorsement is biased, misleading, incomplete, or false. As a result, the user may assume that the endorsement is accurate and unbiased, leading to their uninformed purchase of a product or service.
  - **Urgency** subverts the user's expectation that information provided about discounts or a limited-time deal for a product is accurate, instead accelerating the user's decision-making process by demanding immediate or timely action.
    * **Activity Messages** use *Urgency* as a type of *Social Engineering* to describe other user activity on the site or service, even though the data presented about other users' purchases, views, visits, or contributions are misleading or false. As a result, the user may

falsely feel a sense of urgency, assuming that others users are purchasing or otherwise interested product or service, leading to their uninformed purchase of a product or service.

  - **Countdown Timers** use *Urgency* as a type of *Social Engineering* to indicate that a deal or discount will expire by displaying a countdown clock or timer, even though the clock or timer is completely fake, disappears, or resets automatically. As a result, the user may feel undue urgency and purchasing pressure, leading to their uninformed purchase of a product or service.
  - **Limited Time Messages** use *Urgency* as a type of *Social Engineering* to indicate that a deal or discount will expire soon or be available only for a limited time, but without specifying a specific deadline. As a result, the user may feel undue urgency and purchasing pressure, leading to their uninformed purchase of a product or service.

- **Personalization** subverts the user's expectation that products or service features are offered to all users in similar ways, instead using personal data to shape elements of the user experience that manipulate the user's goals while hiding other alternatives.
  - **Confirmshaming** uses *Personalization* as a type of *Social Engineering* to frame a choice to opt-in or opt-out of a decision through emotional language or imagery that relies upon shame or guilt. As a result, the user may be convinced to change their goal due to the emotionally manipulative tactics, resulting in being steered away from making a choice that matched their initial goal.

B ANALYZED TAXONOMIES OF DARK PATTERNS

Table 2: Academic taxonomies of dark patterns.

	High-Level Pattern	Low-Level Pattern
Brignull 2018-2022 [6]	—	Sneak into Basket, Bait and Switch, Roach Motel, Price Comparison Prevention, Disguised Ads, Privacy Zuckering, Trick Questions, Hidden Costs, Confirmshaming, Friend Spam, Forced Continuity, Misdirection
Brignull 2023 [7]	—	Comparison Prevention, Confirmshaming, Disguised Ads, Fake Scarcity, Fake Social Proof, Fake Urgency, Forced Action, Hard to Cancel, Hidden Costs, Hidden Subscription, Nagging, Obstruction, Preselection, Sneaking, Trick Wording, Visual Interference
Bösch et al. [4]	Obscure	Privacy Zuckering, Immortal Accounts, Hidden Legalese Stipulations, Bad Defaults
	Maximize	Shadow User Profiles, Address Book Leeching, Forced Registration
	Deny	Immortal Accounts
	Preserve	Shadow User Profiles, Address Book Leeching
	Centralize	Shadow User Profiles
	Publish, Violate, Fake	—
Gray et al. [24]	Nagging	—
	Sneaking	Intermediate-Level Currency, Roach Motel, Price Comparison Prevention
	Obstruction	Bait and Switch, Sneak into Basket, Hidden Costs, Forced Continuity
	Interface Interference	Toying with Emotion, Aesthetic Manipulation, Trick Questions, Preselection, Disguised Ad, Hidden Information, False Hierarchy
	Forced Action	Gamification, Privacy Zuckering, Social Pyramid
Mathur et al. [40]	Sneaking	Sneak into Basket, Hidden Costs, Hidden Subscription
	Urgency	Limited-time Message, Countdown Timer
	Misdirection	Confirmshaming, Visual Interference, Trick Questions, Pressured Selling
	Social Proof	Activity Message, Testimonials
	Scarcity	Low-stock Message, High-demand Message
	Obstruction	Hard to Cancel
	Forced Action	Forced Enrollment
Luguri et al. [37]	Nagging	—
	Social Proof	Testimonials, Activity Messages
	Obstruction	Immortal Accounts, Intermediate-Level Currency, Roach Motel, Price Comparison Prevention
	Sneaking	Bait and Switch, Sneak into Basket, Hidden Costs, Hidden Subscription / Forced Continuity
	Interface Interference	Cuteness, False Hierarchy / Pressured Selling, Toying with Emotion, Trick Questions, Preselection, Disguised Ad, Hidden Information / Aesthetic Manipulation, Confirmshaming
	Forced Action	Friend spam/social pyramid/address book leeching, Privacy Zuckering, Gamification, Forced Registration
	Scarcity	High Demand Message, Low Stock Message
	Urgency	Countdown Timer, Limited Time Message

Table 3: Regulatory taxonomies of dark patterns.

	High-Level Pattern	Low-Level Pattern
EDPB [17]	Overloading	Continuous Prompting, Privacy Maze, Too Many Options
	Skipping	Deceptive Snugness, Look Over There
	Stirring	Emotional Steering, Hidden in Plain Sight
	Obstructing	Dead End, Longer than Necessary, Misleading Action
	Fickle	Lacking Hierarchy, Decontextualizing, Language Discontinuity, Inconsistent Interface
	Left in the Dark	Conflicting Information, Ambiguous Wording or Information
EU Com. (EC) [15]	Nagging	—
	Social Proof	Testimonials, Activity Messages
	Obstruction	Intermediate-Level Currency, Roach Motel / Difficult Cancellations, Price Comparison Prevention
	Sneaking	Bait and Switch, Sneak into Basket, Hidden Costs, Hidden Subscription / Forced Continuity
	Interface Interference	Toying with Emotion, Trick Questions, Preselection (default), Disguised Ad, Hidden Information / False Hierarchy, Confirmshaming
	Forced Action	Forced Registration
	Urgency	Countdown Timer / Limited TIme Message, Low Stock / High Demand Message
OECD [48]	Forced Action	Forced Registration, Forced Disclosure / Privacy Zuckering, Friend Spam / Social Pyramid / Address Book Leeching, Gamification
	Interface Interference	Hidden Information, False Hierarchy, Preselection, Misleading Reference Pricing, Trick Questions, Disguised Ads, Confirmshaming / Toying with Emotion
	Nagging	Nagging
	Obstruction	Hard to Cancel or Opt Out / Roach Motel / Click Fatigue / Ease, (Price) Comparison Prevention, Immortal Accounts, Intermediate Currency
	Sneaking	Sneak into Basket, Hidden Costs / Drip Pricing, Hidden Subscription / Forced Continuity, Bait and Switch (including Bait Pricing)
	Social Proof	Activity Messages, Testimonials
	Urgency	Low Stock / High Demand Message, Countdown Timer / Limited Time Message
UK CMA [10]	Choice Structure	Defaults, Ranking, Partitioned Pricing, Sludge, Bundling, Dark nudge, Choice overload and decoys, Virtual currencies in gaming, Sensory manipulation, Forced outcomes
	Choice Information	Drip pricing, Reference pricing, Framing, Complex language, Information overload
	Choice Pressure	Scarcity and popularity claims, Prompts and reminders, Messengers, Commitment, Feedback, Personalisation
US FTC [21]	Endorsements (Social Proof)	False Activity Messages, Deceptive Consumer Testimonials, Deceptive Celebrity Endorsements, Parasocial Relationship Pressure
	Scarcity	False Low Stock Message, False High Demand Message
	Urgency	False Discount Claims, False Limited Time Message, Baseless Countdown Timer
	Obstruction	Immortal Accounts Roadblocks to Cancellation, Price Comparison Prevention
	Sneaking or Information Hiding	Intermediate Currency, Hidden Subscription or Forced Continuity, Drip Pricing, Hidden Costs, Hidden Information, Sneak-into-Basket
	Interface Interference	Bait and Switch, Disguised Ads, False Hierarchy or Pressured Upselling, Misdirection
	Coerced Action	Friend Spam, Social Pyramid Schemes, and Address Book Leeching, Pay-to-Play or Grinding, Forced Registration or Enrollment, Nagging, Auto-Play, Unauthorized Transactions
	Asymmetric Choice	Subverting Privacy Preferences, Preselection, Confirm Shaming, Trick Questions

PUBLICATION P7

Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow"

Authors:
Colin M. Gray, Thomas Mildner, & Nataliia Bielova

The publication contributes to the following angles:

DESIGN

This publication describes the ubiquitous nature of dark patterns across various fractions of digital interfaces and derives the necessity to consider the sequential and co-occurring dimensions of dark patterns. To this end, the publication introduces Temporal Analysis of Dark Patterns (TADP), a methodology to analyse dark patterns in these veins. A case study based on Amazon's cancellation process illustrates the potential effectiveness of this method.

Its contribution to the thesis is to the design angle. The work spotlights limitations of contemporary dark pattern research when only considering isolated design artefacts. To this end, the work introduces an extending method (TADP) to analyse sequential and layered dimensions of interfaces.

My contribution to this paper ideating dark patterns as temporal processes as well as the coding, and analysis of the case study. I contributed to the findings' interpretation and the writing of the manuscript with an emphasis on the introduction, results, and discussion of this work. I approved the final draft before it was submitted by the first author.

The contents of this publication are currently under review but pre-published in: Gray, C. M., Mildner, T., and Bielova, N., "Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow"," arXiv:2309.09635, 2023

Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow"

COLIN M. GRAY, Indiana University, USA
THOMAS MILDNER, University of Bremen, Germany
NATALIIA BIELOVA, Inria Centre at Université Côte d'Azur, France

Dark patterns are ubiquitous in digital systems, impacting users throughout their journeys on many popular apps and websites. While substantial efforts from the research community in the last five years have led to consolidated taxonomies of dark patterns, including an emerging ontology, most applications of these descriptors have been focused on analysis of static images or as isolated pattern types. In this short paper, we present a case study of Amazon Prime's "Iliad Flow" to illustrate the interplay of dark patterns across a user journey, grounded in insights from a US Federal Trade Commission complaint against the company. We use this case study to lay the groundwork for a methodology of Temporal Analysis of Dark Patterns (TADP), including considerations for characterization of individual dark patterns across a user journey, multiplicative effects of multiple dark patterns types, and implications for expert detection and automated detection.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in HCI**.

Additional Key Words and Phrases: dark patterns, temporal analysis, detection, methodology

ACM Reference Format:
Colin M. Gray, Thomas Mildner, and Nataliia Bielova. 2023. Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the "Iliad Flow". 1, 1 (September 2023), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Draft: September 15, 2023

1 INTRODUCTION

After over a decade on research, interest in dark patterns[1] research is still growing—impacting not only HCI scholarship, but also connections to policy, law, and design [17]. The study of dark patterns was originally focused on design practitioners with the goal of "naming and shaming" companies into providing better user experiences [10], but has since captured a broad range of patterns that deceive, coerce, or manipulate users. Researchers have developed a growing knowledge of dark pattern instances in specific domains, such as e-commerce [26], games [32], and social media [29]. These efforts have supported the development of a domain-agnostic ontology [19], which has categorized

[1]We use this term to connect our efforts to prior scholarship and legal statute, while recognizing that other terms such as "deceptive design" or "manipulative design" are sometimes used to describe similar tactics. While the ACM Diversity and Inclusion Council has included dark patterns on a list of potentially problematic terms, there is no other term currently in use that describes the broad remit of dark patterns practices that include deceptive, manipulative, and coercive patterns that limit user agency and are often hidden to the user.

Authors' addresses: Colin M. Gray, comgray@iu.edu, Indiana University, Bloomington, Indiana, USA; Thomas Mildner, mildner@uni-bremen.de, University of Bremen, Germany; Nataliia Bielova, nataliia.bielova@inria.fr, Inria Centre at Université Côte d'Azur, France.

Manuscript submitted to ACM

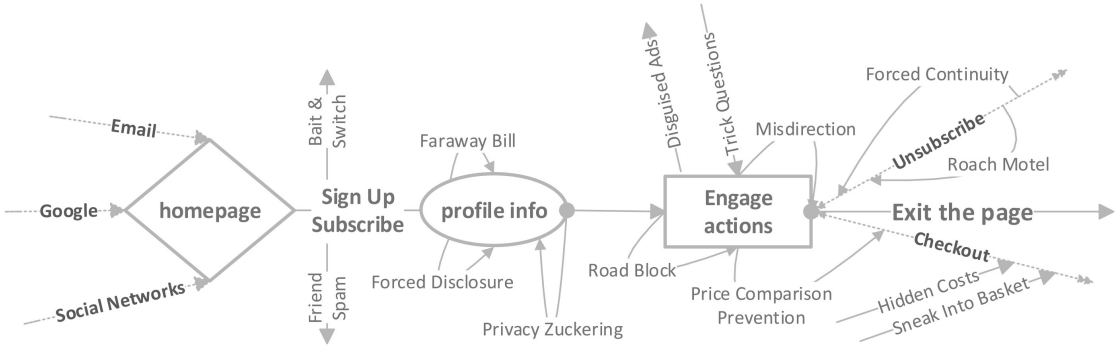


Fig. 1. A journey map taken from a version of darkpatterns.org in 2016. This diagram demonstrates users' interactions when engaging with a website, including experiences of multiple dark patterns.

individual dark pattern types into high-, meso-, and low-level patterns to allow easier access and adaption within and outside this community.

In the past two years, regulators and policymakers have taken action to address issues of technology manipulation and deceptive design practices, resulting in legislative frameworks and regulatory sanctions that aim to protect users from dark patterns' harms. Legislation such as the EU's Digital Service Act (DSA) [6] and California's Privacy Rights Act (CPRA) [2], alongside guidance from governmental bodies such as the Organization for Economic Co-operation and Development (OECD) [30], UK Consumer and Markets Authority (CMA) [4], and the US Federal Trade Commission (FTC) [1] have supported the development of regulatory frameworks related to dark patterns, with the goal of bringing more transparency into digital environments and protecting users' autonomy to make informed decisions. Currently, lawsuits and other sanctions are leveraging these new regulations and, thus, demonstrate the effectiveness of policies where HCI and law work side-by-side to protect end users.

However, existing scholarship often focuses on static dark patterns, driven by sharing screenshots as artifacts as evidence of dark patterns latent in the UI [18]. While contemporary dark patterns scholars often acknowledge aspects of temporal complexity, including feedforward, repetition of actions such as nagging, or actions that are part of a larger sequence, no expert evaluation or automated methods have been proposed that comprehensively support the inspection of an entire user journey. This lack of support for specific methods to support the temporal experience is particularly odd, given that Brignull (the originator of the term "dark patterns" and founder of darkpatterns.org) shared an annotated journey map including the kinds of details mentioned in 2016 for a brief time (Figure 1). As Brignull notes on this archived page:

> *"A journey map is a simple diagram to illustrate users go through in engaging with a webpage, whether it is an online experience, a product, retail, service or any combination. Usually when there are many touchpoints it means the experience is more complex. In this case, we located the Dark Patterns as touchpoints—ideally the map should be clean."*
>
> (From darkpatterns.org, 2016)

We use Brignull's diagram as a source of inspiration and starting point to propose components of a disciplined and rigorous methodology to characterize dark patterns experienced over time. This kind of temporal complexity has been primarily addressed in the dark patterns literature at present through application audits conducted with specific sets

of user goals in mind [15, 29], while other scholarship has focused on automated or semi-automated detection across elements of the user journey [11, 25, 26]. Recent work from Mildner et al. [28, 29], echoing prior work from Gray et al. [20], Luguri and Strahilevitz [24], and guidance from the OECD [30], suggests that not only do dark patterns often occur together in single moments of a user journey, but they can also produce multiplicative or amplified effects both in isolation and across a user journey. We advance this line of research in this short paper, building a foundation for a method of Temporal Analysis of Dark Patterns (TADP) and consider attributes of this method through a case study from the legal literature.

To that end, we make two contributions to the HCI and dark patterns literature. First, we illustrate aspects of temporal complexity that enhance the impact of dark patterns on user behavior through a case study of the Amazon Prime "Iliad Flow," identifying the kinds of user interactions over time that should be considered and characterized by researchers, regulators, and legal scholars. Second, we assess components of a TADP methodology that should be considered when studying the effects dark patterns have on users and identify how these components might be taken up through expert evaluation, automated detection, and human-in-the-loop detection.

2 PROBLEMATIZING DARK PATTERNS EXPERIENCED OVER TIME: A CASE STUDY OF AMAZON PRIME'S "ILIAD FLOW"

A legal complaint filed by the US Federal Trade Commission (FTC) in June 2023 against Amazon is a recent example of enforcement action that includes detailed references to dark patterns [7]. This case follows multiple other cases [5, 12, 13] in the past two years by the FTC and other government bodies that have used the presence of dark patterns as a central form of evidence that user autonomy was not respected. We present the Amazon Prime cancellation process as an explanatory case study [31] to identify how dark patterns are inscribed into the user experience, how these dark patterns relate to each other on specific screens and over time, and what elements of the overall user experience would be useful for scholars to focus on when analyzing other experiences for the presence of dark patterns.

The center of a recent enforcement action by the FTC is Amazon Prime's cancellation process, which gained notoriety for being obstructive to users and led to the adoption of a two-click cancellation option in July 2022—but only for EU consumers [3]. A Norwegian Consumer Council report from 2021 demonstrated how dark patterns were used in Amazon's cancellation process to frustrate consumers, leaving them "[…] faced with a large number of hurdles, including complicated navigation menus, skewed wording, confusing choices, and repeated nudging. Throughout the process, Amazon manipulates users through wording and graphic design, making the process needlessly difficult and frustrating to understand." [22].

In the FTC complaint against Amazon, these same allegations were exposed in further detail, building on evidence that showed how Amazon's design teams were complicit in making this process more difficult than it needed to be:

> *[…] the primary purpose of the Prime cancellation process was not to enable subscribers to cancel, but rather to thwart them. Fittingly, Amazon named that process "Iliad," which refers to Homer's epic about the long, arduous Trojan War. Amazon designed the Iliad cancellation process ("Iliad Flow") to be labyrinthine, and Amazon and its leadership […] slowed or rejected user experience changes that would have made Iliad simpler for consumers because those changes adversely affected Amazon's bottom line.* [7, p. 3]

Notably, legal frameworks such as those used by the FTC or other regulatory bodies rarely require proof of intent in order to produce sanctions. However, in this case, not only were elements of the user experience clearly obstructive, but Amazon's own naming of the user flow indicated their goal of making the process as difficult as possible. The

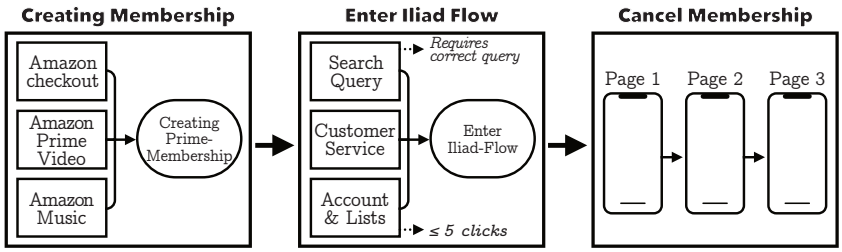


Fig. 2. This flowchart demonstrates the user journeys for becoming an Amazon Prime member, finding the "Iliad Flow", and canceling a subscription.

complaint features an exhaustive description of the user journey, supported by screenshots. Several different aspects of the interactive system are included in the complaint, including the process to subscribe to Amazon Prime, different ways to enter the "Iliad Flow" to cancel Amazon Prime, and the component interactions required to cancel the membership.

2.1 Identifying Dark Patterns in the "Iliad Flow"

The FTC complaint included explicit analysis that demonstrated and named the presence of multiple dark patterns across the "Iliad Flow". To characterize these dark patterns in more detail, we leveraged Mildner et al.'s [29] approach to identify dark patterns in interfaces, using the software Atlas.ti [16] to analyze the complaint through open coding. We used a deductive codebook containing dark patterns from Gray et al.'s [19] ontology and Mildner et al.'s [29] work, thereby analyzing both the text and visual elements of the complaint using a qualitative content analysis approach. One author performed the initial coding work, leveraging dark patterns noted in the complaint and previous sanctions alongside their own expertise from previous studies on dark patterns. A second author who also had prior experience conducting studies on dark patterns confirmed the application of codes. After the document was fully coded, we connected the different interface stages in the form of a journey map including co-occurring and amplification of dark patterns on the one hand and their sequential dependency on the other.

Amazon's "Iliad Flow" describes the user journey leading to the option for cancelling a Prime membership. In our analysis, we not only focused on the "Iliad Flow" but also considered membership creation and the required steps to cancel the service. The cancellation process itself includes three pages, however, there are multiple ways to enter the "Iliad Flow" and even more interactions that terminate the flow without successfully canceling the membership. Figure 2 shows the user journey described in the complaint including membership creation, finding the "Iliad Flow," and three pages users have to successfully navigate to find the option to cancel the membership.

Becoming an Amazon Prime Member. Although not directly a part of the "Iliad Flow," it is noteworthy to demonstrate the ease through which Amazon recruits new members to its Prime program. Options to subscribe Amazon Prime are presented continuously through Amazon's services on both mobile and desktop modalities, including Amazon Music, Amazon Prime Video, and anytime an item is being purchased from Amazon—the service seemingly utilizes every opportunity to offer its Prime membership to users. In doing so, Amazon uses multiple dark patterns that manipulate users' understanding of the choice architecture. Although both Amazon Music and Video offers include stand-alone and cheaper alternatives, the service provider exploits *Interface Interference* (a high-level dark pattern) to promote its Prime membership as a superior subscription, including all of Amazon's premium features but at a higher cost. Consequently,

users are being tricked into more expensive subscriptions through the *Bait and Switch* dark pattern and deploying the *Roach Motel* pattern through the existence of the "Iliad Flow".

Entering the "Iliad Flow". While it is relatively easy to subscribe to Amazon Prime, the complaint describes a complex and labyrinthine procedure to cancel it. In total, Amazon offers users three possible paths to enter the "Iliad Flow". First, customers can use a search function on the website. However, the complaint describes how users have to be highly precise in their choosing of words to be presented with a link to enter the cancellation process. Alternatives refer customers to other settings or help services but do not present quick access. Second, customers can reach out to customer service, which itself requires customers to navigate through multiple options before being able to actually enter a query. Third, customers can enter the "Iliad Flow" by first navigating to Amazon's "Account & Lists," finding the "Manage Membership" option, and selecting the "End Membership" option. Contrary to its name, this option will not end a customer's membership but rather forward them to begin the "Iliad Flow." Together, these three options contain multiple instances of *Interface Interference* and *Obstruction*, for instance, in the form of *Labyrinthine Navigation* or *Misdirection*. The complaint suggests that customers have to take a minimum of two actions to even enter the "Iliad Flow."

Navigating the "Iliad Flow". Once a customer finds themselves in the cancellation flow process, they have to successfully navigate three pages before being able to end their Amazon Prime membership. As Figure 3 demonstrates, the pages repeatedly feature alternative options that, if clicked, remove the user from the "Iliad Flow", exiting the process. Thus, customers have to begin to find and enter the "Iliad Flow" again if they consider one of the alternatives presented on each screen. Each screen of the "Iliad Flow" includes a variety of dark patterns in a labyrinthine interface path deceiving customers in their attempt to cancel their membership. Moreover, the interface emotionally manipulates customers by reminding them about personalized features or contemporary offers that become unavailable once they terminate their membership. Only after customers reach the third page of the "Iliad Flow" are they able to end their subscription immediately. However, Amazon still aims to keep users connected to their service by offering the alternatives to pause a membership or terminate it at a different time. Collectively, customers have to navigate through a plethora of dark patterns in various stages before being able to end their membership.

2.2 Characterizing the Complexity of the "Iliad Flow"

In this section, we describe the findings of our temporal analysis based on Amazon's "Iliad Flow," with a summary of our findings shown in Figure 3. For the sake of brevity, we simplified the "Iliad Flow" in terms of displayed dark patterns and overall complexity to allow a high-level view of both the strategies deployed on individual screens and discrete UI elements and across the entire flow experience. In its original form, the "Iliad Flow" affords users multiple scrolling actions, as options to proceed were otherwise outside the visible frame. Moreover, additional visual and text in the original experience added further *Social Engineering* tactics.

2.2.1 Instances of Dark Patterns. Our temporal analysis of the "Iliad Flow" revealed a plethora of dark patterns customers encounter throughout their attempt to cancel their Amazon Prime membership. In their complaint, the FTC named seven dark patterns specifically: (1) *Forced Action*; (2) *Interface Interference*; (3) *Obstruction*; (4) *Misdirection*; (5) *Sneaking*; and (6) *Confirmshaming*. While our analysis confirms instances of these dark patterns, we extend the FTC's findings by also identifying multiple instances of 22 dark pattern types, including high-level, meso-level, and low-level mapped

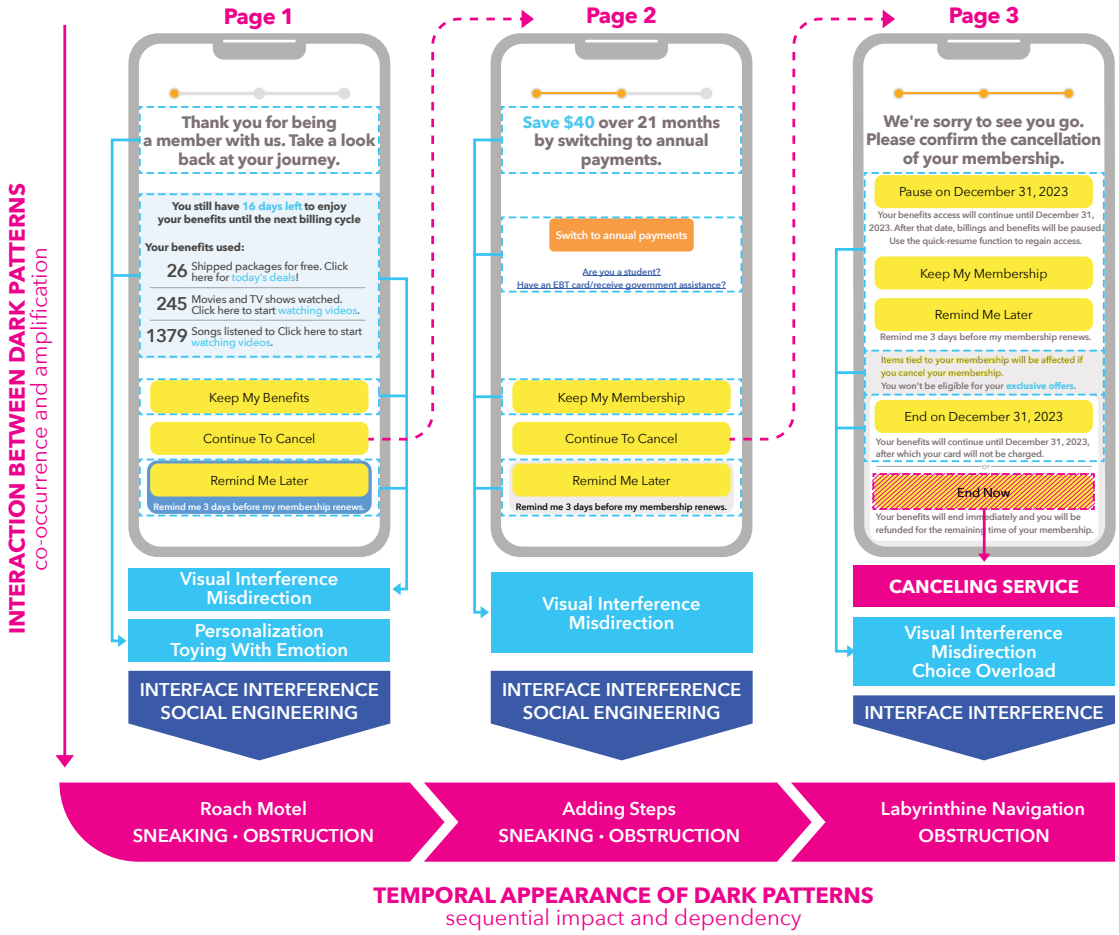


Fig. 3. A summary of our temporal analysis of dark patterns in Amazon's "Iliad Flow." For brevity, we simplified the interface complexity but maintained key options including three screens users have to navigate to be able to cancel their membership. Vertically underneath each page, we summarized co-occurring and amplifying dark patterns. Horizontally, we follow the sequential impacts and dependency of dark patterns. Dark patterns in SMALL CAPS refer to high-level types while lower-case dark patterns refer to meso- and low-level instances (from Gray et al.'s [21] ontology). Descriptions of each dark pattern are included in Table 1.

to Gray et al.'s [21] ontology. Notably, the "Iliad Flow" itself comprises three linked screens on which we counted 70 instances of dark patterns across 22 types.

Most prominently and at the highest level of abstract, we identified *Obstruction* ($n = 25$) and *Interface Interference* ($n = 14$) dark patterns. Other lower-level types that were frequently found in the screens included *Labyrinthine Navigation* ($n = 10$), *Exploiting Errors* ($n = 10$), and *Redirective Condition* ($n = 6$).

2.2.2 Dark Pattern Co-Occurrence & Amplification. Aside from the variety of dark patterns deployed in the "Iliad Flow," our analysis further shows how multiple dark patterns often occur together. As shown in Figure 3, high-level patterns of *Sneaking* and *Obstruction* pervaded the entire interaction sequence, supported by *Social Engineering* in

	Code	N	Definition
1	Aesthetic Manipulation [21]	7	"Any manipulation of the user interface that deals more directly with form than function. This includes design choices that focus the user's attention on one thing to distract them from or convice them of something else."
2	Confirmshaming [10]	7	"Guilting users into opting into something. The option to decline is worded to shame the user into compliance."
3	Confusion [14]	3	"Asking the user questions or providing information that they do not understand. Asking a novice user if they would like to change their default browser, use of double, triple, or quadruple negatives."
4	Decision Uncertainty [29]	1	"This dark pattern confuses users by diminishing their ability to assess situations, leaving them clueless as to what is expected of them or what options are available."
5	Exploiting Errors [14]	10	"Taking advantage of user errors to facilitate the interface designer's goals. E.g. mistyped URL brings up advertisement instead of assistance."
6	Forced Action [21]	5	"This strategy describes dark patterns that require the user to perform a certain action to access (or continue to access) certain functionality."
7	Hard to Cancel [26]	3	"The pattern does not disclose important information upfront to the user that canceling a subscription or membership could not be completed in the same manner they signed up with."
8	Hidden Costs [10]	2	"You get to the last step of the checkout process, only to discover some unexpected charges have appeared."
9	Hidden Information [18]	5	"This dark pattern describes options or actions relevant to the user but not made immediately or readily accessible. It may manifest as options or hidden in fine print, disclosed text, or a product's terms and conditions statement."
10	Interface Interference [18]	14	"This strategy describes dark patterns that manipulate the user interface privileging certain actions over others, thereby confusing the user or limiting discoverability of important action possibilites."
11	Labyrinthine Navigation [29]	10	"This dark pattern describes nested interfaces that are easy to get lost in, disabling users from choosing preferred settings. This pattern is often seen in social media settings menus."
12	Manipulate Navigation [14]	2	"Information architectures and navigation mechanisms that guide the user towards interface designer's goal. E.g. making the free version of an application far more difficult to find than the commercial version on a consumer firewall vendor's website."
13	Misdirection [10]	3	"The design purposefully focuses your attention on one thing in order to distract your attention from another."
14	Nagging [18]	2	"This strategy describes dark patterns that redirect of expected functionality persisting beyond one or more interaction."
15	Obfuscation [14]	4	"Hiding desired information ad interface elements. E.g. reducing contrast of close/stop buttons on video advertisements."
16	Obstruction [21]	25	"This strategy describes dark patterns with intentions of making a process more difficult than it needs to be, with the intent of dissuading certain action(s)."
17	Redirective Condition [29]	6	"Dark patterns of this type contain choice limitations that force users to overcome unnecessary obstacles before being able to achieve their goals."
18	Roach Motel [10]	4	"You get into a situation very easily but getting out is difficult (occurs in subscriptions)."
19	Sneaking [21]	2	"Dark patterns following this strategy attempt to hide, disguise, or delay the divulging of information that is relevant to the user."
20	Social Engineering [21]	8	"Social Engineering is a strategy which presents options or information that causes a user to be more likely to perform a specific action based on their individual and/or social cognitive biases, thereby leveraging a user's desire to follow expected or imposed social norms."
21	Toying With Emotions [18]	2	"[T]his dark pattern includes any use of language, style, color, or other similar elements to evoke an emotion in order to persuade the user into a particular action."
22	Visual Interference [26]	7	"This dark pattern uses style and visual presentation to influence users into making certain choices over others."
	Total =	**70**	

Table 1. Table containing all 22 identified dark patterns from the analysis of the "Iliad Flow". The table includes the number of instances each dark pattern was identified and their definitions.

key decision moments in the first two screens. Additionally, these high level patterns were supported—and even amplified—by numerous lower-level patterns that drew on the higher-level parent types. For instance, all three screens used manipulation of the visual hierarchy (a meso-level pattern) to confuse users about the interactive differences and feedforward between the three options, making options to keep the membership, continue to cancel, or be reminded later appear in parallel. In parallel, *Social Engineering* strategies such as personalization were used to amplify the

interface interference effects by providing specific amounts of media the user might lose access to or provide options on how to choose a different payment plan that would appear more affordable.

Notably, while some patterns are easily traceable to one or more specific UI elements, the interactions among the different types of dark patterns are more nuanced. For instance, the first screen layers choice architecture manipulation and emotional manipulation (*Interface Interference*) and urgency (*Social Engineering*) in a direct way, leaving the roach motel (*Sneaking*) to be realized across the entire user journey. Similarly, the use of labyrinthine navigation (*Obstruction*) applies to the entire user journey as opposed to one discrete UI element or screen.

2.2.3 Sequential Impact & Dependency of Dark Patterns. While co-occurrence between dark pattern types provides insights into the interplay between specific forms of manipulation, deception, and coercion, these types also benefit from each other from a sequential level. To understand their intertwined effects, we considered the dark patterns across the interactions and how they helped maintain deceptive and manipulative pressure on customers. As a customer sets out to end their membership, they constantly face distractions and *Sneaking* strategies to keep them from proceeding. As Figure 3 depicts, each screen contains multiple options deflecting from the goal to end a membership. The screens are visually designed to appear engaging through the *Interface Interference* and *Social Engineering* high-level dark patterns—being both highly visible in their focus and emotionally pressuring. Importantly, engagement with any of the options other than the undifferentiated buttons indicated in the figure instantly exits the customer from the "Iliad Flow" and requires them to begin again. Thus, the combination and sequencing of dark patterns deployed ensures that most consumers will fail at their goal of cancelling the service—particularly the first time they navigate the gauntlet of dark patterns.

3 FOUNDATIONS FOR A TEMPORAL ANALYSIS OF DARK PATTERNS (TADP) METHODOLOGY

Building on the case study we have presented, in this section we outline key characteristics that a Temporal Analysis of Dark Patterns methodology should consider, along with how these characteristics might be supported by expert evaluation, automated analysis, and human-in-the-loop automated analysis.

(1) **Identify which dark patterns are being used, in what combination or sequence, and of what type(s).** This stage requires the use of a standardized source of pattern types and definitions, such as the emergent ontology of dark patterns by Gray and colleagues [19]. Identification of dark patterns should include high, meso, and low-level characterization where possible, although novel dark patterns might only be characterized by high and meso-level, with a low-level characterization leading to the definition of a new potential pattern type. This stage of analysis takes into account: readable text; layout; relative size and positioning of UI elements; use of color, typography, or text decoration; feedforward or other forms of feedback to the user; task flows or other relations between UI elements and screens; and the context or medium of use.

(2) **Identify which UI element(s) are implicated in the use of dark patterns, and how these concentrations of elements within the interface might lead to the user's experience of dark patterns.** This stage requires connections between the presence of a dark pattern and its manifestation in UI or system. This stage of analysis takes into account the relationship between: one or more dark patterns to one or more UI elements; one or more dark patterns to the lack of visible UI elements; or one or more dark patterns to transitions between screens or across the entire user journey. Different levels of dark pattern characterization may allow characterization of high- and meso-level patterns on the screen or journey level that are then inscribed into one or more specific UI elements.

(3) **Describe interactions between dark patterns, co-occurrence of dark patterns types, and/or potential amplification effects.** This stage requires knowledge of which dark pattern types appear and in which combination, both on a specific screen and over time. This stage of analysis takes into account the: combinations of dark patterns that appear in discrete moments of the user journey and over time; the co-occurrence of patterns with shared or differing high- or meso-level parents; the strategies or cognitive biases the patterns exploit; and the causal or other interactive relationship between patterns on a screen or over time.

Based on these proposed stages for a TADP methodology, we can consider which components are best suited for manual expert review, which can be fully automated, and which type of automation may augment expert analysis in a human-in-the-loop system. When detecting dark patterns automatically, several researchers have implicitly recorded temporal interactions with web services in order to reveal the presence of dark patterns; however, the need for temporal detection was not explicitly stated. We list below the most recent advancements in automatic detection of dark patterns in websites and mobile applications, demonstrating where technical approaches might be leveraged in relation to our overall methodology aims.

(1) **Web applications** The foundational work of Mathur et al. in e-commerce websites [27] scaled the detection of dark patterns by (1) automatising the process of product acquisition and capturing HTTP Archive (HAR) [8] files for each crawled page containing HTTP headers and full website response content, (2) detecting visible HTML elements in website content and further automatically clustering them, and (3) using expert analysis to evaluate the occurrence of dark patterns using additional context and surrounding information. This expert analysis included insights from the temporal dimension. For example, researchers found *sneak into basket* instances by noting that no such product was explicitly added earlier thus requiring to observe several steps of the purchase process [26, Fig.3a]; additionally, a *countdown timer* was found on a website where the same offer remained on a day-to-day basis requiring the website to be recorded over several days [26, Fig.4a]. Bouhoula et al. [9] also conducted research on consent banners, automatically detecting dark patterns on these banners using natural language processing (NLP) applied to HTML elements. These researchers also used a temporally grounded two-step process to detect such elements on the first and second layer of the banner, anticipating how a user would interact with these elements in real life. These examples demonstrate a technical foundation that could support automated detection of dark patterns in digital systems by collecting and evaluating HTML element information over time, while also indicating places where expert analysis is needed to characterize what kinds of data are collected, in what time frame(s), and how these data are processed or evaluated.
(2) **Mobile applications** Several scholars have also detected dark patterns in mobile applications, which differ in accessibility as compared to website HTML code. Koch et al. [23] proposed a new solution to download the Android Package (APK) and iOS and iPadOS application archive file (IPA) files to be able to further analyse Android and iOS applications. They also targeted consent banners, extracting app elements that contain visible text, grouping them to detect accept/reject/settings options, and automatically interacting with the options to observe the hidden data flows in each scenario. Chen et al. [11] took a different approach and based their analysis on computer vision and NLP to automatically detect dark patterns in mobile apps, however only using static UI screenshots. These examples demonstrate different technical approaches to identifying and evaluating dark patterns on mobile applications, revealing opportunities for both code-based auditing of APK or IPA files that simulate temporal interaction and scaling up of computer vision or NLP techniques that could be applied to videos of interactions to better characterize temporal characteristics of dark patterns.

We anticipate that future scholarship can productively advance this intersection of automated and expert evaluation techniques to support the temporal analysis of dark patterns, facilitating descriptions of dark patterns on both websites and mobile applications. However, each context presents challenges relating to what level or type(s) of patterns can be detected that future work should consider. In general, low-level patterns *may* be detectable if they can be abstracted in a way that can be supported by web crawlers; however, this detectability is limited by the concreteness of the pattern and the need for human intelligence to detect instances where a pattern is deployed through many different combinations of HTML elements that may require interpretation (as in [26]). For instance, a pattern that manipulates the visual choice architecture might be quite straightforward to detect since this pattern often relates to specific form fields or buttons that can be identified and evaluated in a straightforward manner (as in [9, 23]). However, other patterns—particularly those that involve sneaking or obstruction—will be more difficult to detect in a fully automated manner. In these cases, augmentation technologies may be useful to amplify the abilities of the evaluator, creating an audit trail and also potentially supporting further detection efforts at scale in the future. For instance, an evaluator might manually tag dozens of examples of dark patterns across multiple screens of an interface, indicating where types are present in both static and temporal forms with labels and links to HTML elements or interactive components of the system; these mappings may then be used in combination to train detection systems that can suggest the presence of dark patterns that can then be evaluated and confirmed by an expert. We envision a future TADP methodology that brings together the strengths of both technical detection and expert evaluation, supporting the identification of dark patterns statically and over time in relation to specific UI elements and aspects of the overall user experience.

4 CONCLUSION

In this short paper, we present a case study of the Amazon Prime "Iliad Flow" to characterize the complexity of dark patterns as they are experienced over time. We used this case to demonstrate how dark patterns exist in combination and over time, supporting the foundation for an analysis methodology for Temporal Analysis of Dark Patterns (TADP). We identify key stages that this methodology should include and identification for components that could be automated or augment expert analysis in future work.

ACKNOWLEDGMENTS

This work is funded in part by the National Science Foundation under Grant No. 1909714 and the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR. The research of this work was partially supported by the Klaus Tschira Stiftung gGmbH.

REFERENCES

[1] 2022. *Bringing Dark Patterns to Light Staff Report.* Technical Report. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report%209.14.2022%20-%20FINAL.pdf

[2] 2022. California Privacy Rights Act. https://cppa.ca.gov/meetings/materials/20220608_item3.pdf

[3] 2022. Consumer protection: Amazon Prime changes its cancellation practices to comply with EU consumer rules. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4186 Accessed: 2023-9-12.

[4] 2022. *Evidence review of Online Choice Architecture and consumer and competition harm.* Technical Report. https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers/evidence-review-of-online-choice-architecture-and-consumer-and-competition-harm Accessed: 2022-4-13.

[5] 2022. Fortnite video game maker Epic Games to pay more than half a billion dollars over FTC allegations of privacy violations and unwanted charges. https://www.ftc.gov/news-events/news/press-releases/2022/12/fortnite-video-game-maker-epic-games-pay-more-half-billion-dollars-over-ftc-allegations. https://www.ftc.gov/news-events/news/press-releases/2022/12/fortnite-video-game-maker-epic-games-pay-more-half-billion-dollars-over-ftc-allegations Case number: Docket No. C-4790.

[6] 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[7] 2023. Federal Trade Commission v. Amazon.com, Inc. https://www.ftc.gov/system/files/ftc_gov/pdf/amazon-rosca-public-redacted-complaint-to_be_filed.pdf

[8] 2023. HAR (file format). https://en.wikipedia.org/wiki/HAR_(file_format) Accessed on 13 September 2023..

[9] Ahmed Bouhoula, Karel Kubicek, Amit Zac, Carlos Cotrini, and David Basin. 2023. Automated, Large-Scale Analysis of Cookie Notice Compliance. In *USENIX Security Symposium*.

[10] Harry Brignull. 2018. Deceptive Patterns: User Interfaces Designed to Trick People. http://darkpatterns.org/

[11] Jieshan Chen, Jiamou Sun, Sidong Feng, and Zhenchang Xing. 2023. Unveiling the Tricks: Automated Detection of Dark Patterns in Mobile Applications. *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29-November 1, 2023, San Francisco, CA, USA* (2023). https://doi.org/10.1145/3586183.3606783

[12] Federal Trade Commission. 2021. FTC v. Age of Learning, Inc. https://www.ftc.gov/legal-library/browse/cases-proceedings/172-3186-age-learning-inc-abcmouse

[13] Federal Trade Commission. 2022. FTC v. LendingClub Corporation. https://www.ftc.gov/legal-library/browse/cases-proceedings/162-3088-lendingclub-corporation

[14] Gregory Conti and Edward Sobiesk. 2010. Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, Raleigh, North Carolina, USA, 271. https://doi.org/10.1145/1772690.1772719

[15] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[16] Susanne Friese. 2019. *Qualitative data analysis with ATLAS. ti* (3 ed.). SAGE Publications Ltd, California, United States. 344 pages.

[17] Colin M Gray, Lorena Sánchez Chamorro, Ike Obi, and Ja-Nae Duane. 2023. Mapping the Landscape of Dark Patterns Scholarship: A Systematic Literature Review. In *Designing Interactive Systems Conference (DIS Companion '23)* (Pittsburgh, PA, USA), Vol. 1. Association for Computing Machinery. https://doi.org/10.1145/3563703.3596635

[18] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. dl.acm.org, New York, NY, USA, 534:1–534:14. https://doi.org/10.1145/3173574.3174108

[19] Colin M Gray, Cristiana Santos, and Nataliia Bielova. 2023. Towards a Preliminary Ontology of Dark Patterns Knowledge. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. https://doi.org/10.1145/3544549.3585676

[20] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. 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)*. ACM Press. https://doi.org/10.1145/3411764.3445779

[21] Colin M Gray, Cristiana Santos, Nicole Tong, Thomas Mildner, Arianna Rossi, Johanna Gunawan, and Caroline Sinders. 2023. Dark Patterns and the Emerging Threats of Deceptive Design Practices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. https://doi.org/10.1145/3544549.3583173

[22] Øyvind Kaldestad. 2021. Amazon manipulates customers to stay subscribed. https://www.forbrukerradet.no/news-in-english/amazon-manipulates-customers-to-stay-subscribed/. https://www.forbrukerradet.no/news-in-english/amazon-manipulates-customers-to-stay-subscribed/ Accessed: 2022-3-1.

[23] Simon Koch, Benjamin Altpeter, and Martin Johns. 2023. The OK is not enough: Large Scale Study of Consent Dialogs in Smartphone Applications. In *USENIX Security Symposium*.

[24] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a Light on Dark Patterns. *Journal of Legal Analysis* 13, 1 (March 2021), 43–109. https://doi.org/10.1093/jla/laaa006

[25] S M Hasan Mansur, Sabiha Salma, Damilola Awofisayo, and Kevin Moran. 2023. AidUI: Toward Automated Recognition of Dark Patterns in User Interfaces. (March 2023). arXiv:2303.06782 [cs.SE] http://arxiv.org/abs/2303.06782

[26] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), Article No. 81. https://doi.org/10.1145/3359183

[27] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445610

[28] Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 2362–2374. https://doi.org/10.1145/3563657.3595964

[29] Thomas Mildner, Gian-Luca Savino, Philip R Doyle, Benjamin R Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23, Article 192)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3580695

[30] OECD. 2022. *Dark commercial patterns.* Technical Report. https://doi.org/10.1787/44f5e846-en

[31] Robert K Yin. 2009. *Case study research : design and methods.* Sage Publications, Los Angeles, Calif. 312 pages.

[32] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Foundations of Digital Games.* http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1043332&dswid=1018

PUBLICATION P8

Finding a Way Through the Social Media Labyrinth: Learning From User Perspectives

Authors:
Thomas Mildner, Gian-Luca Savino, Susanne Putze, & Rainer Malaka

The publication contributes to the following angles:

USER

This publication investigates users' expectations of Social Networking Service (SNS) User Interfaces (UIs). Motivated by the lack of discoverability of certain SNS features, the publication includes a card sorting task based on Facebook's UI. To this end, 21 participants were tasked to sort 58 UI features into sensible groups and rate them regarding the importance of individual features and the frequency with which they use them. The findings offer design strategies for optimised groups and inherent structures based on the participants' expectations.

Its contribution to the thesis is to the user angle. Studying SNS users' expectations regarding SNS UI features, the publication garners a better understanding in support of improved SNS interface structures and increased discoverability of individual features.

My contribution to this paper was the design and conduction of the study, data collection, and analysis. I contributed to the interpretation of the data and wrote and revised the manuscript before submitting it.

The contents of this publication are currently under review but pre-published in: Mildner, T., Savino, G.-L., Putze, S., and Malaka, R., "Finding a Way Through the Social Media Labyrinth: Guiding Design Through User Expectations," arXiv:2405.07305 [cs], 2024

Finding a Way Through the Social Media Labyrinth: Guiding Design Through User Expectations

THOMAS MILDNER, University of Bremen, Germany
GIAN-LUCA SAVINO, University of St.Gallen, Swiss
SUSANNE PUTZE, University of Bremen, Germany
RAINER MALAKA, University of Bremen, Germany

Social networking services (SNS) have become integral to modern life to create and maintain meaningful relationships. Nevertheless, their historic growth of features has led to labyrinthine user interfaces (UIs) that often result in frustration among users – for instance, when trying to control privacy-related settings. This paper aims to mitigate labyrinthine UIs by studying users' expectations ($N = 21$) through an online card sorting exercise based on 58 common SNS UI features, teaching us about their expectations regarding the importance of specific UI features and the frequency with which they use them. Our findings offer a valuable understanding of the relationship between the importance and frequency of UI features and provide design considerations for six identified UI feature groups. Through these findings, we inform the design and development of user-centred alternatives to current SNS interfaces that enable users to successfully navigate SNS and feel in control over their data by meeting their expectations.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in interaction design*; Interaction design theory, concepts and paradigms; • **Security and privacy** → *Usability in security and privacy*.

Additional Key Words and Phrases: SNS, social media, deceptive design, dark patterns, ethical user interfaces, ethical design, user experience, user expectation, card sorting, user-centered design

ACM Reference Format:
Thomas Mildner, Gian-Luca Savino, Susanne Putze, and Rainer Malaka. 2024. Finding a Way Through the Social Media Labyrinth: Guiding Design Through User Expectations. In *Proceedings of (Authorversion).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Draft: May 12, 2024

1 INTRODUCTION

Social networking services (SNS) are ubiquitous in many people's everyday lives as both personal and professional drivers for maintaining relationships, retrieving information, and engaging in communities [52]. Yet, in spite of their success, the experience of SNS users is not entirely positive. Misaligned expectations, unfulfilled satisfactions [42, 48], and the feeling of losing control over one's personal data [22, 49] decrease users' satisfaction when using related platforms. To a certain degree, this is the result of difficult-to-navigate user interfaces (UIs), particularly settings menus [30, 43], leading to increasing demands for better control over personal data [5, 38, 57].

Noticeably, two main reasons factor into the difficulty of SNS navigation and, thus, a loss-of-control feeling among their users. Firstly, SNS have matured into complex applications with a wide range of features to engage with others and settings users have to maintain [19, 40, 42]. Many SNS feature multiple feeds or timelines, options for public or personal discourse, and a wide range of controls to customise user experience, such as, but not limited to, control of personal data, advertisement-related data, and notifications from SNS applications. Secondly, research has identified a host of dark patterns in SNS [31, 41, 43, 48] as well as in the design of many of these individual features. In the context of SNS, dark patterns, also referred to as deceptive design patterns [1], are used to steer users' attention [44] or unwillingly increase their engagement while governing their decisions [43]. Building on the existing taxonomy of dark patterns captured and described in related work [26, 29], a recent effort by Mildner et al. [43] identified more than 40 different dark patterns across popular SNS platforms, while Monge Rofarello et al. [44] described design strategies on SNS aimed to capture their users' attention. Furthermore, in a different work, Mildner et al. [41] illustrated SNS users' difficulty in effectively protecting themselves from these strategies, which is in line with similar prior studies [10, 20]. At least partially responsible for users' struggles in this regard, the concept of *Labyrinthine Navigation* [43] describes tangled UI structures that hinder users from successfully and effectively navigating SNS interfaces when controlling personal data within and outside of particular SNS platforms. This combination of increasingly complex interfaces and the prevalence of dark patterns have led to SNS UIs where settings — containing crucial elements to control personal data — are hidden deep within complex and nested interface structures. Although we cannot know whether such design choices fall under malintent, their users experience the consequences through bad usability and a lack of control.

In this research, we take a first step towards improved usability and control for SNS users by understanding their expectations regarding SNS interface design and complexity. To this end, we consider a user-centred design (UCD) approach for investigating the importance of relevant SNS features and the frequency with which they are used. We analysed Facebook's interface as a prototypical example for SNS UIs, as it remains highly relevant after almost two decades in service. Facebook has long been the focus

[1]We chose to use the term "dark pattern" consistent with prior research efforts to describe design strategies that obfuscate or hide consequences of (online) interfaces. Nonetheless, we recognise that the ACM Diversity, Equity, and Inclusion Council has recently [3] identified the term as contentious due to possible negative implications when implying evilness or malintent. Our usage of the term adheres to its initial definition, highlighting the hidden nature of these design strategies with concealed consequences on users [12]. We further acknowledge that this discourse is ongoing and that there is currently no alternative term that fully encapsulates the deceptive, manipulative, obstructive, or coercive nature inherent to the original term.

Authorversion, 2024,

of privacy-concerned research [21, 31, 56] and has been critically reviewed repeatedly from the eyes of public media outlets [1, 51]. Through constantly changing and extending its features, Facebook presents particular challenges for its users to keep personal settings in their desired states and navigate the interface successfully.

In our study, we conducted a card sorting experiment based on Facebook's interface with 21 participants identifying both important and frequently used interface elements as well as seemingly unimportant and less used ones to gain insights about users' expectations about individual and groups of SNS features. Through this study, we aim to answer the following research question:

RQ: What specific design considerations should be taken into account to align SNS UIs with their users' expectations?

Based on our results, we discuss design considerations to structure SNS UI features according to our participants' ratings in terms of the importance and frequency in which a feature is used. Moreover, we identified six groups of SNS UI elements that further capture users' expectations and offer an initial hierarchy within SNS UI features: (1) "User Support", (2) "Legal & Policy Compliance", (3) "Data Security & Privacy", (4) "Profile & Account Management", (5) "Visibility Control", and (6) "User Experience Customization". These groups include common functionalities implicit in SNS but also less used features for privacy, security, and control over users' data. In tandem, these insights offer opportunities to rethink and restructure current SNS to avoid labyrinthine UI structures and instead aid users in navigating them successfully to maintain features and settings according to their preferences. In contrast to current design efforts of commercial SNS applications, which deal with a large number of features and are affected by dark patterns, this paper argues solely from the users' perspective and their expectations. We acknowledge that designing good SNS applications in terms of UI and settings menus is a challenging task. With our design considerations, we contribute a first step towards improving the status quo.

2 RELATED WORK

The main focus of this research lies within traditional UCD concepts in the context of SNS, intending to understand users' perceptions to design optimal UIs that respond to their expectations. However, our study draws from recent efforts in HCI spotlighting unethical, exploitative design strategies that decrease users' ability to make informed decisions through deceptive and manipulative dark patterns. The related work begins by highlighting traditional user-centred and ethical design. Afterwards, we continue with SNS-related studies identifying problematic design and dark patterns that limit user agency.

2.1 From User-Centred Design to Deception

Traditionally, HCI provides designers with the means to develop user-centred interfaces that should be intuitive and easy to navigate. Extending core principles of UCD, the ethically driven school of Value Sensitive Design (VSD) [24] promotes the necessity to uphold users' autonomy and make consequences of interactions transparent to the user. While VSD promotes user autonomy, other interfaces are designed to guide users through complex interactions. In this regard, nudges [53] and persuasive design [23] can be deployed to increase usability in terms of efficiency and engagement, but not necessarily transparency. Although these concepts find many useful applications, especially in health-related contexts [6], they can be exploited to undermine users' ability to make informed decisions [14, 32] leading to deceptive or manipulative interfaces [12, 29]. Offering some insight into how design can accommodate user autonomy to avoid deceptions, Leimstädter [36] build on Hansen and Jespersen's framework [32] to promote reflection through design friction, however, at the cost of user experience and restricted usability. In the scope of persuasive technologies, recent work by Bennett et al. [8] has underlined the general relevance of user agency and autonomy. However, the authors notice certain ambiguities in the terms' usage in related work. In this paper, we follow traditional UCD concepts and Bennett et al.'s terminology suggestion to study users' expectations when interacting with SNS features in terms of importance and frequency [8].

2.2 User Agency in SNS

Similar to autonomy-related work, within the HCI peripheral, research has investigated the effects of SNS on user agency for some time now [34, 48] – particularly regarding users' privacy behaviour [5, 7, 35]. The lack of agency to use SNS as desired may place users in a vulnerable spot. In this regard, an array of studies illustrate the contrast between the positive effects of SNS increasing social connectedness [4, 50] and misuse of SNS, leading to negative consequences on users' well-being [9, 16, 18, 54]. These tensions highlight a continuous need to study how SNS affect their users and what interface features are responsible for potentially problematic outcomes.

This need is further amplified by the constant change and increase of SNS features [19, 40]. Since their advent, SNS have grown into sophisticated platforms that extend their original features when offering users a wide variety of options to engage with content and other users [42]. This upscale of features has led to complex UIs that users may find difficult to navigate [20, 43]. Based on a user study on YouTube's mobile interface, Lukoff et al. [37] noticed design strategies deployed by the platform limiting a sense of agency among its users and found opportune interface mechanisms that could increase their ability to use the platform as preferred. A common approach to enable users in this regard is the implementation of design interventions. Concerned with Facebook's interface, Lyngs et al. [38] developed two interventions that would either remind users about their usage goals or remove Facebook's newsfeed to help users not get distracted. Although their results contain limitations, they demonstrate the benefits of increased control over one's usage behaviour. In a similar vein, Masaki et al. [39] demonstrate how nudges used as design interventions can protect users from exposing personal information unwillingly. Their results are in line with prior results by Wang et al. [55], who confirm that positive impact interface nudges and friction design can have to help users reflect on their decisions before engagement. While related work presents important design interventions as countermeasures to problematic interface design that limit user agency, in this work, we aim to understand SNS users' expectations to inform UIs that avoid the implementation of problematic designs. To this end, we

discuss design considerations for structuring UIs based on users' expectations regarding individual and groupings of SNS UI features.

2.3 SNS Breaking Users' Expectations

Work focusing on design interventions suggests a misalignment of interests between providers of SNS and their users. This discrepancy could stem from commercial incentives [58], leading to unethical design, such as dark patterns, in SNS' UIs [15, 27]. Dark patterns are design strategies that prohibit users from making informed decisions by obfuscating or obstructing informed decision-making [12, 28]. As per their nature, dark patterns are difficult to avoid [20], even when participants were made aware of their existence [10], and SNS are not exempt from this [41]. Consequently, users may be unable to maintain account settings aligned with their preferences depending on the screen modality used to access a service [30, 41] or feel restricted from deleting their accounts alltogether [48].

Arguably, dark patterns restrict users' agency and autonomy to use systems in terms of their beliefs or values and break their expectations. Recent work done by Mildner et al. [43] dismantled popular SNS interfaces identifying a range of dark patterns based on a corpus of 80 types. Moreover, the work described five SNS-specific dark patterns that subscribe to engaging and governing strategies. The engaging strategies fall in line with designs that draw users' attention towards themselves, as described as attention-capturing by Monge Roffarello [44]. The governing strategies, on the other hand, convey interfaces that steer users' interactions while disregarding their goals. Motivated by the *Labyrinthine Navigation* dark pattern falling under this strategy, describing complex and nested interfaces users easily get lost in, we recorded Facebook's interface to study users' perception of its features. Here, our aim was to learn about users' expectations in terms of the importance of features as well as the frequency with which users use them. Based on these criteria, we gained an in-depth understanding of how SNS UI features can be structured to accommodate user preferences.

3 METHOD

To understand SNS users' expectations of SNS UI features, we conducted a card sorting study with 21 participants, including 58 cards representing typical SNS features collected from Facebook's interface. The study was designed to be completed unsupervised and online through the web application Miro [47]. The online study setup was self-contained, meaning the instructions and the task were embedded on the Miro board. Thus, participants did not have to leave the platform throughout the exercise and could concentrate on completing the study.

3.1 Selection of Cards

As a basis for the cards, we analysed the interface of Facebook's mobile application. The decision fell on Facebook as it remains a popular SNS that, in the course of almost two decades, changed its UI and adopted different strategies, growing into the platform it is today. We screen-recorded a walkthrough of the complete Facebook application[2] and identified a total of 102 UI features, including the feed, profile page, and settings menu. In an attempt to limit the scope of cards to suit the purpose of this study (i.e. understanding expectations toward general SNS features), we excluded certain UI features that were specific to Facebook (e.g. marketplace or dating features) or exceeded the scope of this research (e.g. payment methods, management tools for professionals). Finally, this reduction resulted in 58 cards of relevant SNS UI features that participants were asked to sort.

[2]We recorded usage based on Facebook version 397.0 on iOS.

3.2 Card sorting Procedure

Each participant was provided with the necessary information to participate in the study via email, together with a consent form and demographic questionnaire to fill out. After we received their consent and demographic data, participants were given a link to their individual Miro boards where the online card sorting task took place (see Figure 1 for one participant's card sorting results). The study was designed to last about 40 minutes and involved three parts: First, participants were asked to follow a traditional card sorting approach by grouping cards based on similar traits to learn about the relatedness of UI features. Second, we asked them to assign each card an importance score from 1-5 (not important at all - extremely important), following a Likert scale. Lastly, participants colour-coded the cards depending on the frequency with which they would use a particular feature ("frequent" usage in **green**, "moderately" in **yellow**, and "rarely to never" in **red**). We acknowledge that such a 3-point Likert scale is restricting the analysis but follow advice [33] that it can produce interesting insights, especially when a study design demands participants' focus over longer periods of time. The data gained through these additional tasks informed us about the relevance of each feature based on two criteria — importance and frequency — which offered further insights into our participants' expectations for the UI features.

3.3 Participant Demographics

Participants were recruited through various university computer science and HCI programs. Participation was entirely voluntary, and participants were rewarded 10€ for taking part in the study. To qualify for this study, participants had to engage with social media on a weekly basis and needed to be enrolled in HCI-related programs. The latter requirement was chosen as we aimed for some sensibility towards HCI research and technological literacy. In total, 23 participants participated in this card sorting study. However, two participants had to be excluded for incomplete participation. The remaining data from 21 participants was therefore included in the further analysis. Of these 21 participants, nine self-identified as female and twelve as male. At the time of conducting the study, the participants' mean age was 26.52 years (sd=3.56). They were recruited from Germany (n = 10), Netherlands (n = 1), Switzerland (n = 8), and the USA (n = 2). Between participants, their highest education included a high-school diploma (n = 1), bachelor's degree (n = 9), and master's degree (n = 11). They used SNS an average of 6.86 days per week (sd = 0.48).

4 FINDINGS

In this section, we first present the results of the grouping aspect of the card sorting task. We used hierarchical clustering based on

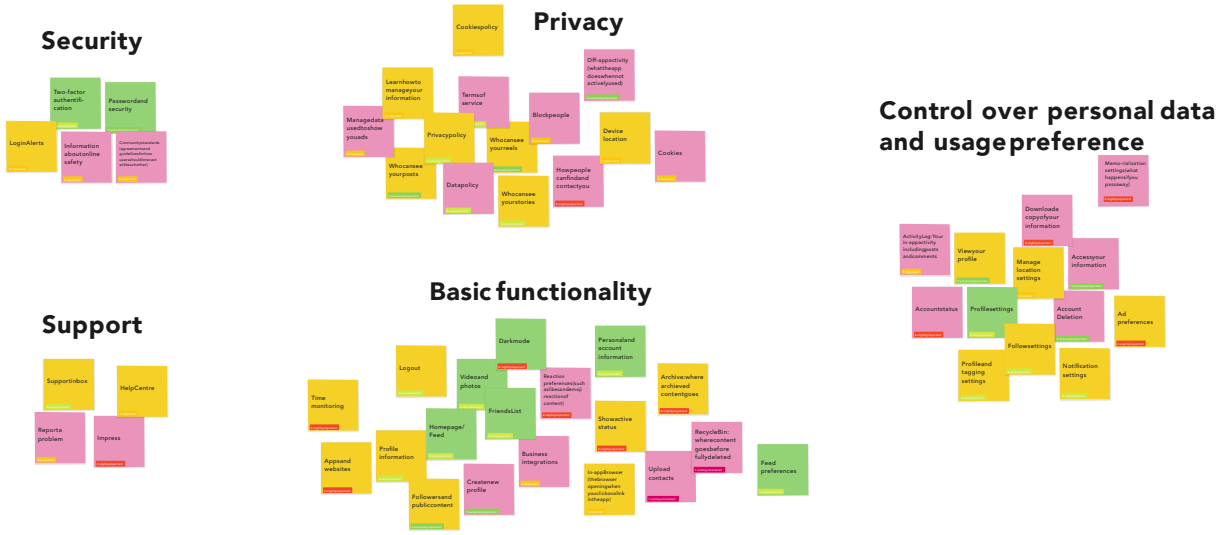


Fig. 1. This figure displays one participant's card sorting results featuring 5 groups of cards with the labels security, privacy, support, basic functionality and control over personal data and usage preference. The cards are both colour-coded and include importance ratings according to the study design.

a similarity matrix generated from each participant's card sorting results. This allows us to assess the relationship between individual cards and to create average groups based on the participants' individual decisions [13]. We then turn to the individual features and report each UI feature's importance and frequency ratings. As the combined data from the 58 cards is too large to be presented in this paper, we focus on the most relevant findings with the complete data included in this paper's supplementary material.

4.1 Groups

The results of the card sorting task offer insights into the collective and individual perspectives of SNS users regarding the sorting of UI features. Moreover, these insights suggest common characteristics shared among SNS UI features, which, in turn, can inform an optimised structure of SNS interfaces. By enhancing the discoverability of individual features in alignment with users' expectations, interface aspects leading to labyrinthine navigation could thus be avoided.

To this end, we began our analysis by transferring the groups created by participants (see Figure 1 for an example) into a similarity matrix, as visualised in Figure 2. This allows us to assess how often participants paired UI features, giving us a first impression of how UI features could be structured. Here, we report noteworthy similarity pairs in percentages based on the number of times the 21 participants paired up individual features. Often used features of SNS were frequently coupled together, such as 'Home feed' and 'Followers and public content' (66.7%). Furthermore, policies were often paired (i.e. 'Data policy' and 'Cookies policy' at 80.9%).

Using the data from the similarity matrix, we proceeded with a hierarchical clustering approach to determine groups with the highest agreement across participants. Figure 3 illustrates the resulting groups. We followed common practice for the hierarchical clustering of our data by using the linkage criterion 'ward' and Euclidean distance. We visually inspected the dendrogram and discussed different cut-off distances to identify meaningful groups among the authors of this work. We chose a cut-off at an Euclidean distance of 40, resulting in six groups (indicated in Figure 3), containing between 3 and 15 features each, with an average of 9 features per group. These results echo similar findings of a related approach by Nawaz [46]. Figure 3 visualises a complete overview of the six groups, their sizes, and their corresponding features.

4.2 Measuring Importance

Alongside our investigation into how participants organised UI features, we also considered their evaluations of the importance of individual features and the frequency with which they use them. These results can inform interface designs to consider perceived importance as a criterion when structuring SNS UIs. We evaluated the data based on the ratings users' gave each card. We used outlier detection and descriptive statistics to identify significantly important or frequently used features. Here, we report the most interesting findings, while we include the full data in the supplementary material of this paper.

Using a Z-score analysis to identify outliers (with a threshold of $Z > 2$), we found two features to be significantly unimportant compared to other UI features: 'Impress' ($Z = -2.08$ and 'Upload contacts' ($Z = -2.36$). Based on our analysis, there were no significantly important features. The average importance rating for all 58 cards was relatively high, with a mean of 3.34 ($sd = 0.74$). Interestingly, the highest ratings were given to 'Password and Security' (mean = 4.70, $sd = 0.56$), 'Account Deletion' (mean = 4.45 $sd = 0.97$), and 'Home Page / Feed' (mean = 4.45, $sd = 0.74$). On the

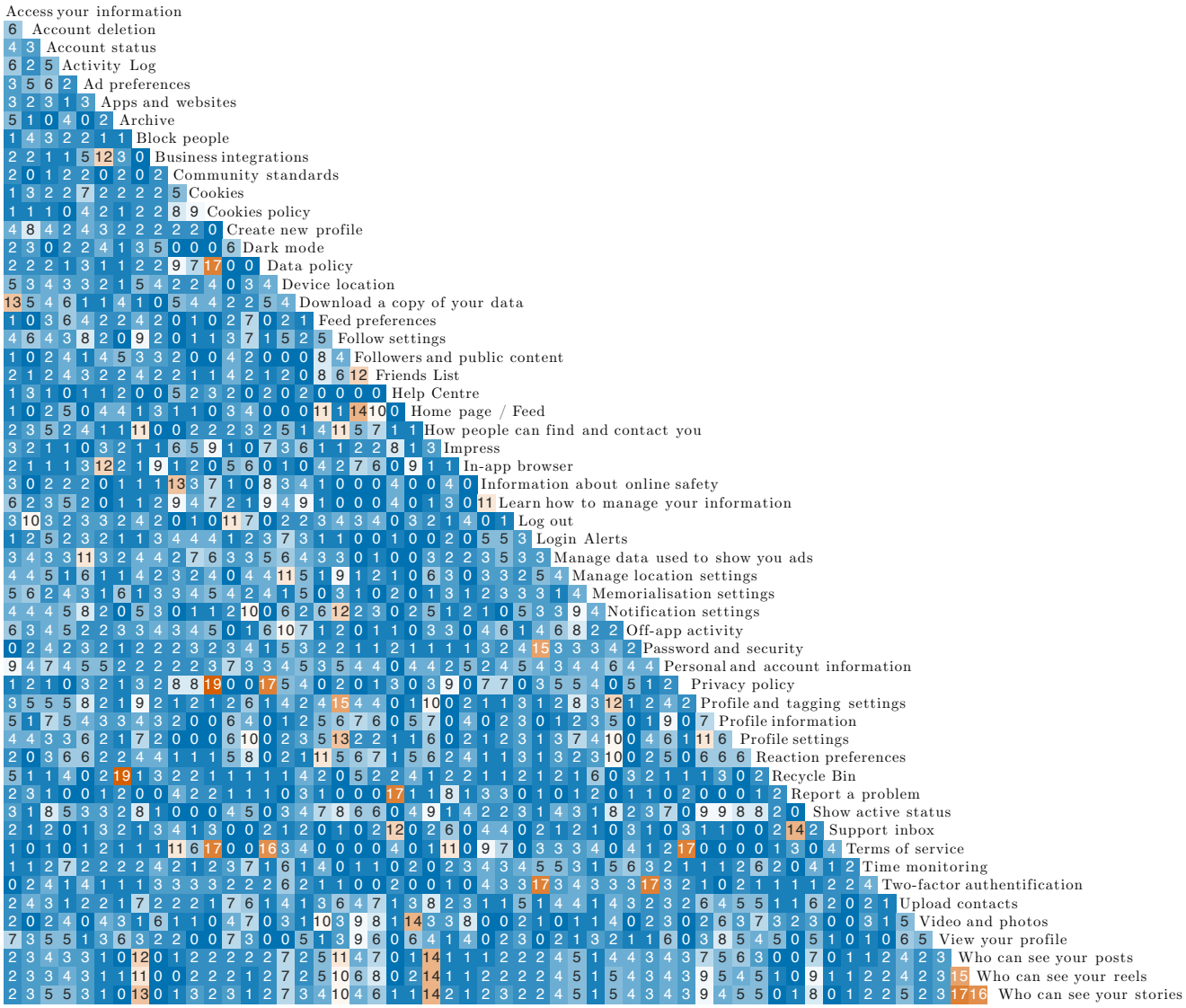


Fig. 2. This figure shows the similarity matrix of all 58 Facebook features based on the 21 card sorting groups.

other end, the UI features with the lowest importance ratings were 'Show Active Status' (mean = 2.10, sd = 0.99), 'Impress' (mean = 1.80, sd = 0.93), and 'Upload Contacts' (mean = 1.60, sd = 1.07) as the least important feature.

4.3 Measuring Frequency

To better understand how relevant participants perceive individual UI features, we asked them to change a card's colour (see Figure 1 for an example) depending on the frequency with which they use it. In this regard, **green** means the feature is often used, **yellow** means the feature is moderately used, and **red** means the feature is rarely used. For our analysis, we mapped the three colours to values from 1 to 3 (1=**red**, 2=**yellow**, 3= **green**) in the form of a 3-point Likert-scale [33]. Similar to the importance ratings, we focus on important findings while we include the complete data in the supplementary material of this paper. We used a Z-score analysis to identify outliers (with a threshold of $Z > 2$). We found two features to be significantly more often used: 'Video and photos' ($Z = 2.16$) and 'Home page / Feed' ($Z = 2.94$). Based on our analysis, there were no significantly rarely used features.

Following descriptive statistics, all 58 cards collectively featured an average score of 1.67 (sd = 0.45) regarding usage frequency. The most frequently used UI features were 'Home Page / Feed' (mean = 3.00, sd = 0, 00), 'Video and Photos' (mean = 2.65, sd = 0.57), and 'View Your Profile' (mean = 2.40, sd = 0.66). In terms of least frequently used features, we find 'Memorialisation Settings'

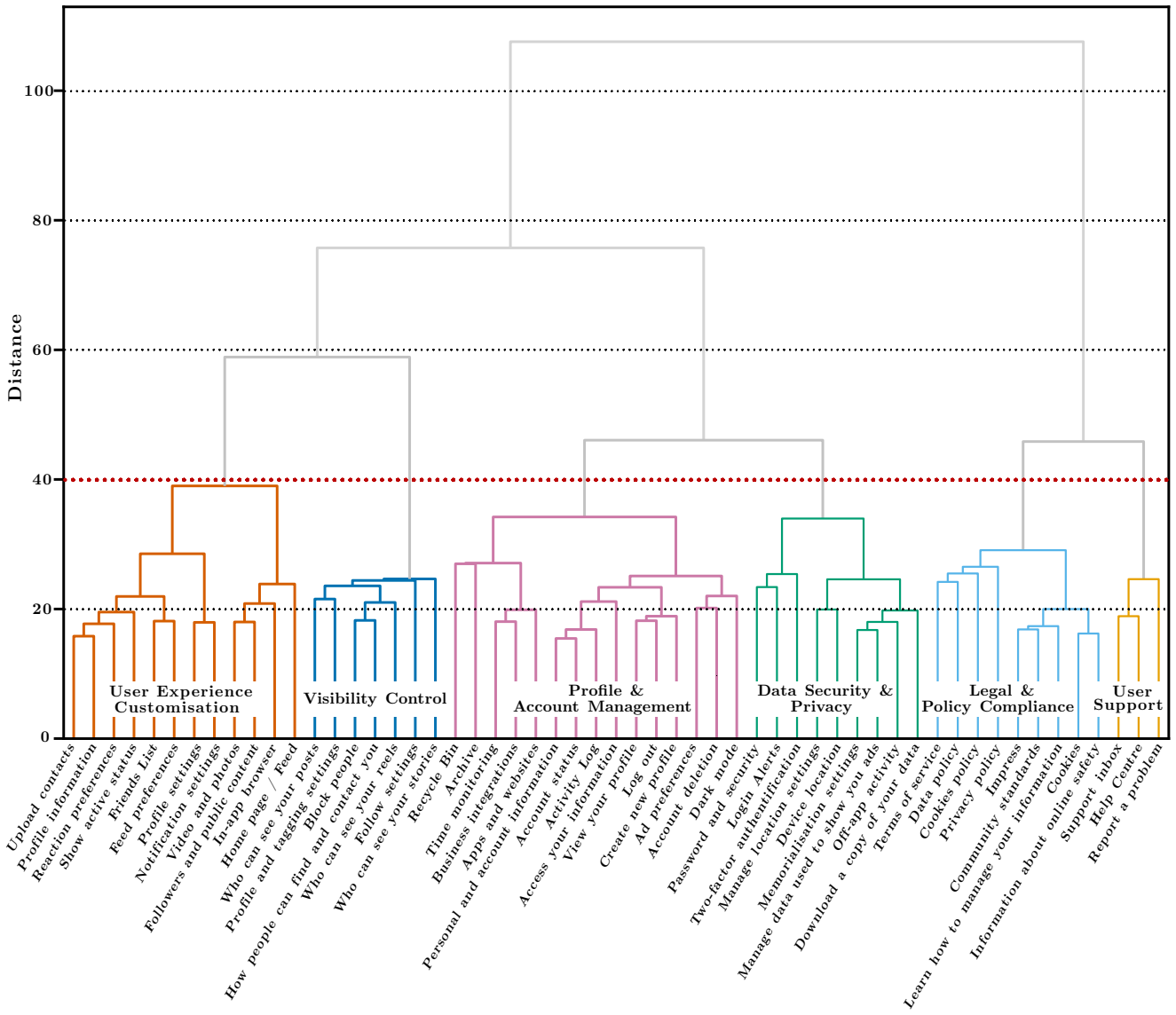


Fig. 3. This hierarchical clustering dendrogram illustrates the optimal number of groups of the card sorting study, highlighted by the dotted red line at 40%.

(mean = 1.05, sd = 0.22), 'Community Standards' (mean = 1.00, sd = 0.00), and 'Terms of Service' (mean = 1.00, sd = 0.00). Notably, three of these UI features were rated with a standard deviation of 0.00, suggesting 100% agreement between participants. To further investigate the results of the card sorting task, we continued by assessing the overall agreement between ratings.

4.4 Agreement

The agreement between participants across UI features in our data indicates similar user expectations regarding how important they find certain UI features and how frequently they use them. To investigate these notions further, we computed the percentage agreement for each feature and across all ratings in terms of importance and frequency. Across all importance scores, we find an average agreement of 44%. Furthermore, the features 'Password and Security' (75.0%), 'Upload Contacts' (70.0%), and 'Account Deletion' (70.0%) have the highest agreement among participants. On the other hand, the features 'Ad Preferences' (30.0%), 'Help Centre' (30.0%), and 'Privacy Policy' (25.0%) have the lowest agreement. For the frequency scores, we noticed an average agreement of 62%. Furthermore, we find that the UI features 'Community Standards', 'Home page / Feed', and 'Terms of service' have an agreement of 100%, while the features 'In-app Browser' (35.0%), 'Time Monitoring' 35.0%, 'Access Your Information' (40.0%) have the lowest agreement among participants.

5 DISCUSSION & FUTURE WORK

Building on prior work [43, 48], this research aims to identify relevant design considerations to bridge otherwise disconnected user expectations with regard to SNS UIs. Offering answers to our research question, the results of our card sorting study reveal sensible groupings of UI features and suggest a hierarchical structuring to afford user expectations. In our discussion, we begin by reiterating how SNS users' expectations are broken in the first place. We then propose design considerations concerning the structure of individual UI features based on the ratings given by our participants. Afterwards, we continue with design considerations focusing on the general grouping of SNS UI features based on our hierarchical analysis.

5.1 Aligning Expectations of SNS Users

It is worth mentioning that we cannot know whether certain UI strategies are deployed with malicious intent; however, users face negative consequences, for instance, in the form of compromised usability and difficult-to-maintain privacy settings. These negative consequences are documented by a series of related work [30, 37, 38, 42], including a study conducted by Schaffner et al. [48], who identified various unethical practices throughout users' attempts to delete their accounts – ultimately limiting their agency over their own account and data. In contrast to decreased agency and studies reporting on SNS deploying dark patterns [43] or on their users misusing related platforms [9, 16, 18, 54], SNS have the opportunity to foster social connectedness and be of great value to maintain meaningful relationships across the globe [4, 50]. It is, therefore, relevant to consider how SNS can be redesigned to offer their users a better experience. By letting users sort SNS UI features, we learned about the possible optimisation of features into sensible groups in terms of restructuring their UIs. Drawing from our participants' card sortings, collecting similar features in closer proximity could improve the discoverability of individual features. Especially with critical UI features, for instance, those related to maintaining personal or ad-related data, a redesign would help users to better control settings according to their preferences [42, 56] To this end, our findings offer guidance and reflections for countering aspects of *Labyrinthine Navigation* dark patterns in today's SNS [41].

5.2 Considerations for Structuring SNS UI Features

The distribution of features across the two dimensions of importance and frequency carries certain insights for the UI design of SNS — some are easier to respect, while others require more attention. For instance, participants rated the often-used 'Home Page / Feed' very important and gave it high frequency, while unpopular features like 'Upload Contacts' were considered less important as they are rarely used. It would be relatively easy to meet our participants' expectations in those regards when restructuring an SNS UI. Briefly, quick access should be granted to frequently used features, while rarely used features can be nested deeper within the UI. Unfortunately, the task becomes more challenging for more complex expectations. The feature 'Account Deletion', for instance, was deemed very important but is, understandably, only rarely used. The feature 'In-app Browser', on the other hand, is quite frequently used but, at the same time, perceived as unimportant. Without certain care for such specific cases, it would be easy to fall back to labyrinthine interface structures that do not meet users' needs and will be difficult to navigate.

The differently perceived importance and frequency ratings of individual UI features demand good design choices to meet users' expectations, especially if a feature is perceived as important but not frequently used, or vice versa. Figure 4 shows a scatter plot of all 58 considered UI features from Facebook. Noticeably, it is divided into four quadrants through the mean importance and mean frequency ratings. The four quadrants suggest four different categories that SNS features can fall into with varying design strategies to address them: (1) High importance and high frequency ratings, (2) low importance and low frequency, (3) high importance and low frequency, (4) low importance and high frequency. For UI design, particularly the two quadrants embedding opposite high and low scores add complexity for finding sensible positions for UI features in an interface. Here, we cover each quadrant independently and propose design considerations based on traditional HCI approaches.

High Importance and High Frequency. The first quadrant is relatively straightforward. UI features that are often used and deemed important should be easy to find and engage with. Here, common practices in HCI, such as the steering-law [2], can help UI designers find a sensible structure for these features. However, designers should be wary of overpopulating interface sections with too many options for users to choose from.

Low Importance and Low Frequency. The second quadrant entails UI features that are neither frequently used nor considered important. These ratings suggest that these features do not require quick access and could be placed further into the background of the interface without breaking users' expectations. Similar to the first quadrant, this is relatively easy to address, even in combination with the first. Again, HCI principles such as the steering law can help structure the UI with respect to features within this quadrant.

High Importance and Low Frequency. These next two quadrants require more attention. UI features that SNS users find important but do not frequently need access to or use often should be positioned in the interface to allow quick access whenever needed — even though they do not necessarily need to be omnipresent. While this may seem difficult at first, HCI has utilities to afford interactions, especially in web and app-based interfaces. UI designers could rely on interface shortcuts [11] to efficiently support users' agency to access otherwise difficult-to-find features. In the same vein, searchbars [45] allow quick and reliable access if the underlying technology can precisely interpret user input in case they are unsure of a feature's name.

Low Importance and High Frequency. Inverse to the former quadrant, UI features that are frequently used but not important to SNS users suggest potential overuse. This further implies that the features of this quadrant should not take space for other, more important features. Instead, especially if excessive or misuse is noticed, the UI structure requires a change to help users better maintain their time and regain agency of their usage behaviour [37, 56]. To this end,

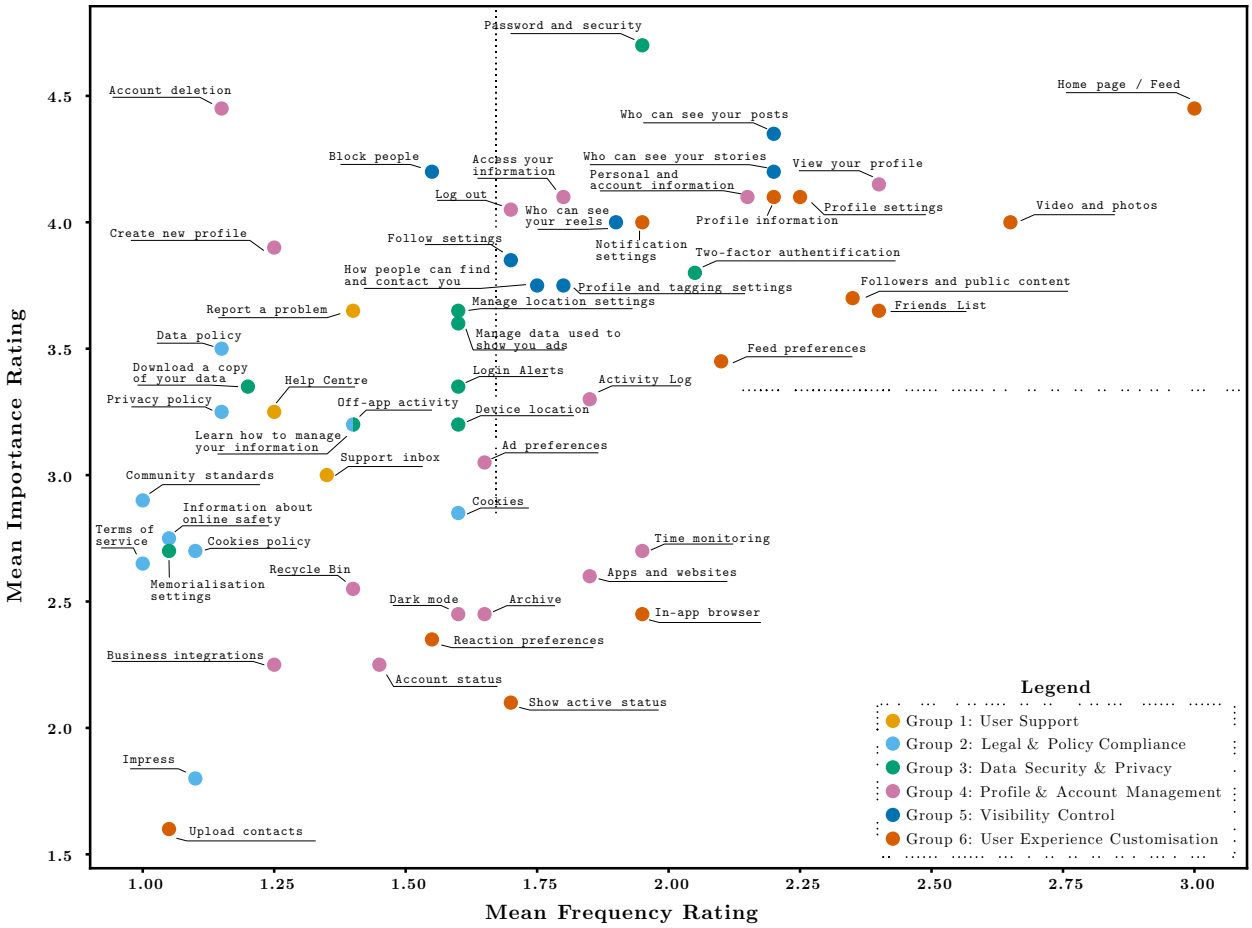


Fig. 4. This figure shows a scatter plot of all 58 Facebook features based on the 21 card sorting groups placed along the two dimensions: importance and frequency. The scatter plot is divided into four quadrants through the overall mean importance and frequency ratings. The four quadrants can be characterised by containing features with low importance and low frequency ratings (lower left quadrant), low importance rating and high frequency rating (lower right quadrant), high importance and low frequency rating (upper left quadrant), and high importance and high frequency rating (upper right quadrant).

design friction is a common design tool to help users make more reflected decisions [17, 36] by hindering impulsive engagement.

5.3 Considerations for Grouping SNS UI Features

Thematic groups of similar UI features help to better arrange their large quantity. Although SNS already structure their UIs based on topics, the lack of discoverability suggested by related work [48], as well as the *Labyrinthine Navigation* dark pattern [43], implies that improvements can be made. Previously, in Section 4.1, we demonstrated optimised groupings of SNS UI features based on participants' card sortings, which we further visualised in Figure 3. Here, we discuss related design considerations per UI feature group. To this end, Figure 5 offers an overview of six individual groups, including convex hulls, to visualise their distribution across the importance and frequency dimensions based on participants' ratings. To describe each group, we used the individual UI features they contained in order to establish overall themes that covered their general scope. In the following paragraphs, we report the contents of each of the six groups independently, as well as discuss the resulting implications in relation to the other groups.

Group 1: "User Support" (3 Features). This group contains three features for user assistance, featuring the 'Help Centre', 'Report a problem', and the 'Support inbox' feature. Such features help users who seek solutions to various issues, ranging from technical problems to policy-related questions. The ratings for both dimensions are mid-ranged compared to other groups, with a mean importance rating of 3.30 (sd = 1.26) and a mean frequency rating of 1.33 (sd = 0.54). These ratings suggest that users do not often require access to these features and while they are not unimportant, they seem not crucial enough to be always present.

Group 2: "Legal & Policy Compliance" (9 Features). Group 2 provides the legal features, including features like 'Community standards', 'Cookies policy', 'Data policy', and 'Terms of service'. These

Fig. 5. This figure shows each feature and their convex hull for each of the six groups as introduced in Figure 4. Each sub-figure visualises the distribution of the contained group across the two dimensions of importance and frequency.

features offer access to legal documents and information. With a mean importance rating of 2.84 (sd = 1.20) and a mean frequency rating of 1.17 (sd = 0.33), this group has the lowest values for importance and frequency among all groups. Thus, it can be argued that these UI features of this group should be situated deeper within interfaces, not to obfuscate other, more relevant groups.

Group 3: "Data Security & Privacy" (9 Features). The next group contains features like 'Device location', 'Download a copy of your data', 'Login alerts', and 'Two-factor authentication'. These UI features offer tools to users they need to secure their own data and remain informed about any attempts to compromise it. With a mean importance rating of 3.51 (sd = 1.09) and a mean frequency rating of 1.56 (Ssd = 0.65), the ratings are higher than those for the Legal & Policy Compliance group but lower than those for the Visibility Control group, putting it quite in the centre in terms of SNS users' expectations.

Group 4: "Profile & Account Management" (15 Features). This group contains features connected to profile and account management and is the largest group with a total of 15 features. They provide options for 'Accessing personal information', 'Account deletion', and 'Ad preferences', among others. The features in this group offer ways in which users can alter how they and their information appear on the platform. Featuring a mean importance rating of 3.22 (sd = 1.11) and a mean frequency rating of 1.67 (sd = 0.69) the group features mid-range values for both importance and frequency ratings. As a consequence of its size, this group spans features of high importance and high frequency (i.e. 'View your profile') as well as low importance and low frequency (i.e. 'Business integrations'). With further attention to the four quadrants between importance and frequency scores, large groups, such as this one, may require further subdivisions into smaller clusters to increase the overall discoverability of UI features.

Group 5: "Visibility Control" (7 Features). Group 5 is focused on features that control online visibility, such as 'Who can see your posts', '-stories', and '-reels'. These features enable users to manage their online visibility, determining how much or how little of their content is visible to different audiences. The mean importance rating for this group is 4.01 (sd = 1.03), which is the highest among all groups, while the mean frequency rating is 1.87 (sd = 0.74). Overall, these ratings suggest that users would expect quick and easy access to the UI features contained in this group.

Group 6: "User Experience Customization" (12 Features). This final group impacts how a user interacts with the platform on a daily basis. Features include 'Feed preferences', 'Notification settings', and controls for 'Show active status' and 'Upload contacts'. These options offer users a level of control over their day-to-day engagement with the platform. For this final group, the mean importance rating is 3.33 (sd = 1.04), and the mean frequency rating is 2.10 (sd = 0.62). Similar to the group "Profile & Account Management", this sixth group contains both low importance and low frequency (i.e. 'Upload

contacts') as well as high importance and high frequency (i.e. 'Home page / Feed') UI features. In fact, this group contains both the lowest and highest-rated features based on our study. As with the fourth group, the size of this group may require further subdivisions to address users' expectations better.

Generally, the overall number of UI features available to SNS users introduces obstacles hindering the discoverability of individual ones — negatively impacting usability. Drawing from our findings, we can rethink the placement of UI features in SNS in consideration of our participants' feedback and expectations to increase the discoverability of those that participants deemed important. Together, Figure 4 and Figure 5 offer a structured overview of users' interface expectations. Moreover, the thematic groups lay some groundwork for possible future research and design directions by focusing on the role that each feature plays within SNS contexts. Importantly, our groups are based on limited SNS UI features and a card sorting task with 21 participants. A future study may allow for more detailed insights into similar groups or even come up with additional or different groups altogether. Nonetheless, our findings offer two general implications that help to better understand how SNS interfaces should be structured to meet users' expectations.

Firstly, our six groups offer sensible themes to organise SNS features that share similar characteristics. Aligned with users' expectations, they present a first impression for restructuring and placing of existing UI features into interface panels, menus, tabs, or any other container that fits the purpose of its application. In this work, we mainly focused on the purpose of UI features and the conceptual structuring thereof. Future work should also consider individual affordances to design the features in line with users' expectations.

Secondly, the individual groups need considerations for the internal structuring of contained features according to importance and frequency ratings. While individual groups show consistency across their features, they often span multiple quadrants as contained UI features vary in importance and frequency. Moreover, the number of included UI features varies considerably, with 'User Support' containing three features compared to 'Profile & Account Management', which contains fifteen. Thus, large-sized groups may require additional design consideration and further analysis, which future work could address. Throughout our analysis of Facebook's interface, we noticed specific menus listing a large number of UI features, often requiring users to scroll in order to find specific settings. In such instances, it may be interesting to recursively apply our considerations and break down clustered settings to increase the discoverability of UI features within them.

Aligning users' expectations with experience, the groups and their themes can positively inform existing usability challenges of large and often difficult-to-manoeuvre SNS interfaces. With this work, we contribute a foundation to rethink and restructure existing SNS interfaces and highlight the relevance of considering the inclusion of the importance and frequency at which UI features are visited in usability studies. In future work, we plan to build on the gained understanding to develop and study alternative SNS UI structures and their effectiveness in avoiding *Labyrinthine Navigation* dark patterns.

6 LIMITATIONS

As with most research, our study has several limitations that we want to disclose here and offer potential avenues for future research. In our selection of UI features, we aimed to stay agnostic to common SNS features to mitigate these limitations. However, the focus is primarily on Facebook's interface, limiting the generalisability of our findings across other SNS platforms. While Facebook can be used as a prototypical example of SNS, this focus could result in insights that are not universally applicable, given the fast-paced evolution of SNS interfaces and their differing scopes for user engagement. However, Mildner et al. [43] have identified *Labyrinthine Navigation* within four SNS: Facebook, Instagram, TikTok, and Twitter, hopefully mitigating this limitation to some extent. In our study, we developed groups and underlying themes based on the card sorting results of 21 participants of similar demographics and knowledge in HCI-related fields. While we particularly opted for these demographics to utilise their technological literacy, this introduces two limitations. First, a higher and more diverse sample could represent a wider user base with different expectations. Second, the groups are accumulated from all participants' card sorting results. Thus, the groups only reflect averaged expectations, removing individual perceptions. In this regard, we noticed further design challenges, as some SNS address particular user needs. For example, Facebook and Instagram have features specifically directed to provide businesses and professionals with useful tools to support their efforts to attract and maintain other users' engagement. Many of these features are only important for a certain user base and, thus, often not relevant for others. While we focused on static interfaces, future work could investigate the feasibility and effectiveness of adaptive or customisable UIs [25] that can be tailored to meet individual user needs

7 CONCLUSION

In the last two decades, social media platforms like Facebook have become ubiquitous companions in many peoples' lives. Based on an analysis of Facebook's interface, this research provides valuable insights for rethinking and restructuring SNS interfaces in alignment with users' expectations. Through a card sorting study involving 21 participants, including considerations of the importance in which UI features are perceived and how frequently users engage with them, we identified six common interface groups allocating UI features into key SNS features that reflect users' expectations. In contrast to current efforts in related work, we do not introduce design interventions to labyrinthine interfaces. Instead, we provide insights based on user-centred design to restructure user interfaces of SNS from the ground up with the goal of increasing the discoverability and usability of individual SNS features.

REFERENCES

[1] 2020. The Social Dilemma - A Netflix Original documentary. https://www.thesocialdilemma.com/
[2] Johnny Accot and Shumin Zhai. 1997. Beyond Fitts' Law: Models for Trajectory-Based HCI Tasks. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '97)*. Association for Computing Machinery, New York, NY, USA, 295–302. https://doi.org/10.1145/258549.258760
[3] ACM. 2023. Words matter: Alternatives for charged terminology in the computing profession. https://www.acm.org/diversity-inclusion/words-matter

[4] Dohyun Ahn and Dong-Hee Shin. 2013. Is the social use of media for seeking connectedness or for avoiding social isolation? Mechanisms underlying media use and subjective well-being. *Computers in Human Behavior* 29, 6 (2013), 2453–2462.

[5] J. Alemany, E. Del Val, and A. García-Fornes. 2023. A Review of Privacy Decision-making Mechanisms in Online Social Networks. *Comput. Surveys* 55, 2 (Feb. 2023), 1–32. https://doi.org/10.1145/3494067

[6] Dmitry Alexandrovsky, Maximilian Achim Friehs, Jendrik Grittner, Susanne Putze, Max V. Birk, Rainer Malaka, and Regan L Mandryk. 2021. Serious Snacking: A Survival Analysis of how Snacking Mechanics Affect Attrition in a Mobile Serious Game. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 113. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445689

[7] Nazanin Andalibi and Justin Buss. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376680

[8] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI Understand Human Agency and Autonomy?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3580651

[9] Ine Beyens, J Loes Pouwels, Irene I van Driel, Loes Keijsers, and Patti M Valkenburg. 2020. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports* 10, 1 (2020), 1–11.

[10] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "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* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 763–776. https://doi.org/10.1145/3461778.3462086

[11] Robert Bridle and Eric McCreath. 2006. Inducing Shortcuts on a Mobile Phone Interface. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (Sydney, Australia) *(IUI '06)*. Association for Computing Machinery, New York, NY, USA, 327–329. https://doi.org/10.1145/1111449.1111526

[12] Harry Brignull. 2023. Deceptive Patterns - Types of Deceptive Pattern. https://www.deceptive.design/types

[13] Miranda G Capra. 2005. Factor Analysis of Card Sort Data: An Alternative to Hierarchical Cluster Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49, 5 (2005), 691–695. https://doi.org/10.1177/154193120504900512

[14] Zhilong Chen, Jinghua Piao, Xiaochong Lan, Hancheng Cao, Chen Gao, Zhicong Lu, and Yong Li. 2022. Practitioners Versus Users: A Value-Sensitive Evaluation of Current Industrial Recommender System Design. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 533 (nov 2022), 32 pages. https://doi.org/10.1145/3555646

[15] Shruthi Sai Chivukula, Ike Obi, Thomas V Carlock, and Colin M. Gray. 2023. Wrangling Ethical Design Complexity: Dilemmas, Tensions, and Situations. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 179–183. https://doi.org/10.1145/3563703.3596632

[16] Dimitri A Christakis. 2010. Internet addiction: a 21 st century epidemic? *BMC medicine* 8, 1 (2010), 1–3.

[17] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[18] Sarah M. Coyne, Adam A. Rogers, Jessica D. Zurcher, Laura Stockdale, and McCall Booth. 2020. Does time spent using social media impact mental health?: An eight year longitudinal study. *Computers in Human Behavior* 104 (2020), 106160. https://doi.org/10.1016/j.chb.2019.106160

[19] Manish Dhingra and Rakesh K. Mudgal. 2019. Historical Evolution of Social Media: An Overview. https://doi.org/10.2139/ssrn.3395665

[20] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[21] Prajwal Eachempati, Laurent Muzellec, and Ashish Kumar Jha. 2022. Examining the Relationship between Privacy Setting Policy, Public Discourse, Business Models and Financial Performance of Facebook (2004–2021). In *Proceedings of the Central and Eastern European EDem and EGov Days* (Budapest, Hungary) *(CEEeGov '22)*. Association for Computing Machinery, New York, NY, USA, 159–168. https://doi.org/10.1145/3551504.3551557

[22] Sindhu Kiranmai Ernala, Moira Burke, Alex Leavitt, and Nicole B. Ellison. 2020. How Well Do People Report Time Spent on Facebook? An Evaluation of Established Survey Questions with Recommendations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376435

[23] BJ Fogg. 2009. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology* (Claremont, California, USA) *(Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, Article 40, 7 pages. https://doi.org/10.1145/1541948.1541999

[24] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman (Eds.). Vol. 16. Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4 Series Title: Philosophy of Engineering and Technology.

[25] Krzysztof Gajos and Daniel S. Weld. 2004. SUPPLE: automatically generating user interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 93–100. https://doi.org/10.1145/964442.964461

[26] Colin M Gray, Nataliia Bielova, Cristiana Santos, and Thomas Mildner. 2024. An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action. *arXiv preprint arXiv:2309.09640* (2024).

[27] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300408

[28] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The dark (patterns) side of UX design. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–14. https://doi.org/10.1145/3173574.3174108

[29] Colin M. Gray, Cristiana Santos, Nataliia Bielova, and Thomas Mildner. [n.d.]. An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action. arXiv:2309.09640 [cs] http://arxiv.org/abs/2309.09640

[30] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2021. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–29. https://doi.org/10.1145/3479521

[31] Hana Habib, Sarah Pearman, Ellie Young, Ishika Saxena, Robert Zhang, and Lorrie FaIth Cranor. 2022. Identifying User Needs for Advertising Controls on Facebook. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–42. https://doi.org/10.1145/3512906

[32] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation* 4, 1 (March 2013), 3–28. https://doi.org/10.1017/S1867299X00002762

[33] Jacob Jacoby and Michael S. Matell. 1971. Three-Point Likert Scales Are Good Enough. *Journal of Marketing Research* 8, 4 (1971), 495–500. http://www.jstor.org/stable/3150242

[34] Se-Hoon Jeong, HyoungJee Kim, Jung-Yoon Yum, and Yoori Hwang. 2016. What type of content are smartphone users addicted to?: SNS vs. games. *Computers in Human Behavior* 54 (2016), 10–17. https://doi.org/10.1016/j.chb.2015.07.035

[35] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security* 64 (Jan. 2017), 122–134. https://doi.org/10.1016/j.cose.2015.07.002

[36] David Leimstädtner, Peter Sörries, and Claudia Müller-Birn. 2023. Investigating Responsible Nudge Design for Informed Decision-Making Enabling Transparent and Reflective Decision-Making. In *Mensch und Computer 2023*. ACM, Rapperswil Switzerland, 220–236. https://doi.org/10.1145/3603555.3603567

[37] Kai Lukoff, Cissy Yu, Julie Kientz, and Alexis Hiniker. 2018. What Makes Smartphone Use Meaningful or Meaningless?, In Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 22, 26 pages. https://doi.org/10.1145/3191754

[38] Ulrik Lyngs, Kai Lukoff, Petr Slovak, William Seymour, Helena Webb, Marina Jirotka, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2020. 'I Just want to Hack Myself to Not Get Distracted': Evaluating Design Interventions for Self-Control on Facebook. In *CHI'20*. ACM, Honolulu, 1–15. https://doi.org/10.1145/3313831.3376672

[39] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–11. https://doi.org/10.1145/3313831.3376666

[40] Karen Elizabeth McIntyre. 2014. The Evolution of Social Media from 1969 to 2013: A Change in Competition and a Trend Toward Complementary, Niche Sites. *The Journal of Social Media in Society* 3, 2 (Dec. 2014). https://thejsms.org/index.php/JSMS/article/view/89 Number: 2.

[41] Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 13. https://doi.org/1010.1145/3563657.3595964

[42] Thomas Mildner and Gian-Luca Savino. 2021. Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–7. https://doi.org/10.1145/3411763.3451659

[43] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 192, 15 pages. https://doi.org/10.1145/3544548.3580695

[44] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. 2023. Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3580729

[45] Dan Morris, Meredith Ringel Morris, and Gina Venolia. 2008. SearchBar: A Search-Centric Web History for Task Resumption and Information Re-Finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1207–1216. https://doi.org/10.1145/1357054.1357242

[46] Ather Nawaz. 2012. A Comparison of Card-sorting Analysis Methods. In *The 10th Asia Pacific Conference on Computer Human Interaction. 2012 - Matsue, Japan*, Vol. 2. ACM, Matsue, Japan, 583–592.

[47] RealtimeBoard, Inc. 2023. Miro | The Visual Collaboration Platform for Every Team. https://miro.com/

[48] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142

[49] Sarita Yardi Schoenebeck. 2014. Giving up Twitter for Lent: how and why we take breaks from social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 773–782. https://doi.org/10.1145/2556288.2556983

[50] Tara J Sinclair and Rachel Grieve. 2017. Facebook as a source of social connectedness in older adults. *Computers in Human Behavior* 66 (2017), 363–369.

[51] Natasha Singer. 2018. Why the F.T.C. Is Taking a New Look at Facebook Privacy. *The New York Times* (Dec. 2018). https://www.nytimes.com/2018/12/22/technology/facebook-consent-decree-details.html

[52] Charles B. Stone, Li Guan, Gabriella LaBarbera, Melissa Ceren, Brandon Garcia, Kelly Huie, Carissa Stump, and Qi Wang. 2022. Why do people share memories online? An examination of the motives and characteristics of social media users. *Memory* 30, 4 (2022), 450–464. https://doi.org/10.1080/09658211.2022.2040534 Publisher: Routledge _eprint: https://doi.org/10.1080/09658211.2022.2040534.

[53] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven. OCLC: ocn181517463.

[54] Jean M. Twenge, Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. 2018. Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science* 6, 1 (2018), 3–17. https://doi.org/10.1177/2167702617723376

[55] Jin-Liang Wang, Linda A. Jackson, James Gaskin, and Hai-Zhen Wang. 2014. The effects of Social Networking Site (SNS) use on college students' friendship and well-being. *Computers in Human Behavior* 37 (Aug. 2014), 229–236. https://doi.org/10.1016/j.chb.2014.04.051

[56] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy nudges for social media: An exploratory facebook study. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web* 01 (2013), 763–770.

[57] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. *SOUPS 2011 - Proceedings of the 7th Symposium on Usable Privacy and Security* (2011). https://doi.org/10.1145/2078827.2078841

[58] Shoshana Zuboff. 2023. The age of surveillance capitalism. In *Social Theory Re-Wired*. Routledge, 203–213.

PUBLICATION P9

Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Deceptive Design Patterns

Authors:
Thomas Mildner, Albert Inkoom, Rainer Malaka, & Jasmin Niess

The publication contributes to the following angles:
GUIDELINE

This publication adds to our understanding of underlying mechanisms that enable dark patterns and unethical design. Based on a focus group study with fifteen experts in dark pattern scholarship and psychology and cognitive science, the publication explores similarities, differences, and facilitators between the concepts of cognitive biases and dark patterns. Discussed in prior work, the notion dark patterns exploiting cognitive biases is illustrated in the "Relationship Model of Cognitive Biases and Dark Patterns". The model follows design from its development into the real world, highlighting potential ethical implications for practitioners to consult when designing interfaces. As counter measures, it further includes pathways for user safeguarding.

Its contribution to the thesis is to the guideline angle. The publication investigates possible exploitations of cognitive biases to harm users through dark patterns and studies similarities, differences, and facilitators. As a result of its findings, the "Relationship Model of Cognitive Biases and Dark Patterns" promotes ethical design by spotlighting ethical implications of designing interfaces.

My contribution to this paper was the design and conduction of the study, the qualitative coding of the data, and the interpretation of our codebook. Through discussions with a co-author, I developed the Relationship Model of Cognitive Biases and Dark Patterns. I drafted the manuscript and revised it before the final publication.

The contents of this publication currently under review but pre-published in: Mildner, T., Inkoom, A., Malaka, R., and Niess, J., "Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Dark Patterns," arXiv:2405.07378 [cs], 2024

Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Dark Patterns

THOMAS MILDNER, University of Bremen, Germany
ALBERT INKOOM, University of Bremen, Germany
RAINER MALAKA, University of Bremen, Germany
JASMIN NIESS, University of Oslo, Norway

Throughout the past decade, research in HCI has identified numerous instances of dark patterns in digital interfaces. These efforts have led to a well-fostered typology describing harmful strategies users struggle to navigate [33]. However, an in-depth understanding of the underlying mechanisms that deceive, coerce, or manipulate users is missing. We explore the interplay between cognitive biases and dark patterns to address this gap. To that end, we conducted four focus groups with experts ($N = 15$) in psychology and dark pattern scholarship, inquiring how they conceptualise the relation between cognitive biases and dark patterns. Based on our results, we constructed the "Relationship Model of Cognitive Biases and Dark Patterns" which illustrates how cognitive bias and deceptive design patterns relate and identifies opportune moments for ethical reconsideration and user protection mechanisms. Our insights contribute to the current discourse by emphasising ethical design decisions and their implications in the field of HCI.

Additional Key Words and Phrases: cognitive bias, deceptive design patterns, dark patterns, design ethics, responsible design

ACM Reference Format:
Thomas Mildner, Albert Inkoom, Rainer Malaka, and Jasmin Niess. 2024. Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Dark Patterns. In . ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Draft: May 12, 2024

1 INTRODUCTION

The last decade of research in Human-Computer Interaction (HCI) has shown a growing interest in unethical design practices — in the midst of it are deceptive design practices and so-called "dark patterns" [1]. Various streams of research have described instances that deceive or manipulate users in domains such as, but not limited to, e-commerce [58], social media [64, 71], and web and mobile interfaces [22, 32, 38]. While this body of work makes significant contributions that inform the protection of people against the harms embedded in these environments, we currently lack a fundamental understanding of the underlying principles that enable the deceptive, coercive, and potentially manipulative characteristics of dark patterns. Previous work by Mathur et al. [58] and Waldman [82] draw tentative connections to the exploitation of cognitive biases in this regard. However, the lack of supportive research leaves a gap for additional, fundamental work to describe this relationship in more detail. Furthermore, it remains unclear how this relationship could manifest itself and how it could be navigated in interaction design.

Generally, exploiting cognitive biases and connected heuristics [5, 50] to manipulate users of any kind poses important ethical questions that require careful evaluation. Incorporating knowledge of human cognition and perception is an integral element of human factors and the design of interfaces. However, Thaler and Sunsteins' 'nudge' theory [76] and the principles of persuasive design [26] illustrate the effectiveness in which cognitive biases can be utilised to alter peoples' choice architecture; and, thus, require responsibility in designers not to hinder informed decision-making as users can be unaware of consequences [16]. Wary of misuse, critique voices ethical concerns against paternalistic implications on agency [43, 60, 70]. Although Thaler and Sunstein independently addressed these concerns [73, 75] by reiterating how nudges were meant to empower people to make individual good decisions, Hansen and Jespersen [40] argued for the necessity of nudging transparently upholding user agency.

This ongoing discourse spotlights the various ethical caveats designers should consider to support the informed decision-making of end-users. While recent work shows positive effects of design friction to assist users in making informed decisions [53, 61], in practice, designers aim for effective user journeys that meet their goals. Online shopping sites, for example, take advantage of the aforementioned principles [16, 58]; they steer users toward effortlessly discovering (recommended) products from where they are quickly manoeuvred to a checkout page [58]. In other cases, design friction is indeed used to hinder certain interactions. This can be sensible to support reflection of an interaction's consequences by interrupting mindless engagement [19, 61]. However, design friction can lead to frustration in other cases. Along similar lines, infamous cookie-consent banners have a history of making selection processes a cumbersome task [34]. They rely on visual interferences and aesthetic manipulation to pre-select choices, effectively pushing users' decisions to their disadvantage [34].

These examples demonstrate the responsibility contained in designers' work. Exploitation of cognitive biases, such as the framing effect or the default effect, have been used to misguide and trick users [58]. Drawing from the importance of this ongoing discourse and the need for a better understanding of the underlying cognitive mechanisms that can be exploited to harm users, this research aims

[1]We opted for the term "dark pattern" in alignment with previous authors when referring to identified harmful design practices. However, we are aware that the ACM Diversity, Equity, and Inclusion Council recently decided to classify the term as problematic, given its association with negative connotations. We use the term in line with its original context of *hidden* consequences for users [11] and acknowledge that there currently exists no alternative terms that convey the full spectrum of deceptive, manipulative, obstructive, or coercive characteristics of the original term.

Authorversion, 2024,

to take the next step by exploring the relationship between cognitive biases and dark patterns and how it should be navigated in interaction design. To this end, we conducted four focus groups with experts in dark pattern scholarship, cognitive science, and psychology research, several of whom also have significant experience in interaction design ($N = 15$). Each focus group included a discussion structured to investigate the following research question:

RQ: How can we conceptualise the dynamic relationship between cognitive biases and deceptive design patterns?

In conclusion, the focus groups facilitated valuable insights by bridging the perspectives of two distinct fields, providing important details to understand the effects of cognitive biases on dark patterns and vice versa. Moreover, our findings offer guidance for future work and ethical design considerations to better protect users. By examining the inherent characteristics of cognitive biases [78] and their utilisation in dark patterns within digital contexts, we have garnered new insights into underlying mechanisms, thereby supporting previous suggestions that linked the harming effects of dark patterns to the exploitation of cognitive biases [58, 82]. Our contribution to this field is the *Relationship Model of Cognitive Biases and Dark Patterns*, which elucidates the connections between design decisions and real-world consequences, emphasizing the necessity of responsible design choices and decision-making to protect end-users.

2 RELATED WORK

This paper synthesises research on autonomy and empowerment, cognitive biases, and dark patterns. First, we revisit contributions about users' autonomy within the periphery of HCI promoting ethical design concepts. Second, we present a background of cognitive bias scholarship that informed this research. Concluding this section, the third part focuses on HCI research on dark patterns and points to an ongoing discourse regarding terminology in this area.

2.1 User Autonomy and Empowerment

As digital technologies have become increasingly ubiquitous, the relationship people share with their daily drivers has been described as complex and emotional [74]. The field of HCI has advanced from focusing on usability issues to designing technologies that foster positive interactions and user well-being, often navigating persuasion and autonomy. As an antagonist to autonomy, persuasive technologies often undermine user agency [15]. While research in health-related environments (e.g. [4]) indicates potential positive applications of the persuasive technology paradigm, it is crucial to remain mindful of the ongoing critiques and ethical considerations surrounding persuasive technology. At the same time, it is important to recognise that participants made an autonomous decision and consented before giving up autonomy to change their behaviour toward more healthy options using this health-care intervention.

Unfortunately, achieving responsible design is not an easy task [37]. While Schneider et al. [72] emphasised the importance of empowerment based on a structured literature review, persuasive design remains an often relied-on strategy for behaviour change interventions [39]. Thereby, the implementation of persuasive design is often well–meant but executed in a problematic manner [15]. To illustrate, in their work, Brynjarsdóttir et al. [12] describe the pitfalls of persuasive sustainability — the attempt to persuade users' behaviour toward environmental sustainability, related to Fogg's Behaviour Model [26]. In sum, Brynjarsdóttir et al. frame a critic concerned with modernist technologies, which they place persuasive sustainability under, and remind about a lack of awareness of a design's impact. Opening future avenues, the authors suggest that, firstly, persuasion and users' reactions have to be better understood before implementation. Secondly, they echo traditional user-centred design (UCD) approaches by reminding practitioners to involve users throughout development stages, and, thirdly, encourage practitioners to move away from the individual and consider larger social environments. Offering both research and industry a reflection tool in this vein, Elsayed-Ali et al. [24] created an online card tool showcasing a variety of critical aspects and considerations for innovative design processes. To that end, the tool fosters the sharing of opinions across hierarchies, accounts for differences among participants, promotes inclusiveness, and gives room to otherwise difficult discussions specific to the participants' work environments.

Work in HCI has long been concerned with conscious control [48], and in that sense user autonomy and agency; although often using these terms interchangeably, as highlighted by a recent literature review [7] considering over 32 years of related work. In this review, Bennet et al. [7] allocate these values as key concepts in various HCI work. In 2014, Grimpe et al. [37] highlighted the difficulties of responsible design, problematizing the challenge into four core issues: *reflexivity*; *responsibility and responsiveness*; *inclusion*; and *anticipation*. A core effort promoting user autonomy when interacting with technologies is value sensitive design (VSD), first proposed by Friedman et al. [27]. Continuing former work [28], the authors encourage interfaces to convey functionalities transparently and truthfully while empowering users. The relevance of VSD is again highlighted by contemporary work from Chen et al. [16]. The authors apply underlying paradigms to often deceptive recommender systems to study notions of disagreement between practitioners and users, illuminating disagreement of associated values. In part, these disagreements may be the result of conflicting ethical stances among practitioners and corporate incentives [17, 31].

Resulting interfaces might restrict users' ability to make informed decisions that are in line with their beliefs or values [26, 84, 85]. In order to return agency to the user, design interventions, design friction, or nudges have been discussed as potential counter-measures to mitigate loss-of-control feelings getting users to reflect on their choices [40, 53, 84]. Focusing on the social medium Facebook, Lyngs et al. [55] compared two design interventions to empower users to reflect on their usage behaviour. One would periodically remind the users about their initial goals; the other removed the newsfeed feature from Facebook to keep users' on track with their goals. In a controlled setup, the study revealed certain shortcomings but highlighted how interventions can help users reflect on their behaviour. Exploring how nudges can be used similarly, Masaki et al. [57] developed nudging interventions deployed to shield users' privacy. These findings gain support by work conducted by Wang et al. [83], demonstrating the positive effects of design interventions, friction, and nudges that can grant users better agency when engaging with social media. In a similar vein, Lukoff et al. [54] conducted

a co-design study based on YouTube's mobile interface to develop alternative mechanisms that redirect a sense of agency to the platform's users. The limited agency of social networking service (SNS) users could be an indicator of decreased well-being. While previous studies promote design interventions as counter-measures to problematic interfaces in this regard, we return to traditional UCD principles aiming to understand users' expectations regarding UI features within and beyond the SNS context. Exploring the possibilities for adaptive interfaces users can customise to meet their needs, Kollnig et al. [51] developed an "app repair framework", after a majority of 85% of their participants demonstrated appreciation for the option to alter elements of their apps. With a focus on audio-related privacy when recording, Dunbar et al. [23] propose design principles to accommodate people's concerns and preferences in this sensitive context. These examples highlight the necessity of understanding when and how in the design process to factor in considerations regarding user agency.

Inspired by this body of work, our objective is to extend our focus beyond specific usage contexts and conceptualise the nuances of design choices practitioners encounter. Furthermore, we intend to incorporate ethical considerations from dark pattern scholarship into the discourse at a more general level. To that end, we connect the fields of cognitive biases and dark patterns to explore their relationships that often lead to negative and harmful effects for users.

2.2 Cognitive Biases

Utilisation of human cognition and perception is an integral part of human factors and HCI [86]. In this vein, Jain and Horowitz et al. [48] discuss opportunities that go beyond interactions requiring users to make conscious decisions. As the work rests its focus on HCI-related work to promote seamless, unconscious interactions that could enhance people's experience with novel technologies, they remind of ethical constraints when designing to alter cognitive processes. While our work is inspired by underlying constructs of cognitive biases, governing our choices unconsciously, the overall history of this field exceeds the means of this paper. Here, we synthesise key contributions as a background of our work.

Tversky and Kahneman's [50] first introduced the term cognitive biases in 1972. Ever since, related work has identified and described a plethora of effects that influence human decision-making. The granularity and diversity of cognitive biases are well reflected in Baron's book 'Thinking and Deciding' [6], highlighting the enormous effort that went into this field of research, reporting that the variety of concepts has been thoroughly catalogued and understood. In an attempt to summarise these past efforts, Hilbert [47] identified eight core biases, provoking the idea that every other cognitive bias can be mapped to one of them. Hilbert placed these core biases alongside additional mathematical definitions to avoid any future ambiguity. His framework can help to understand the logical constraints in cognitive processes and promotes the concept of "The Noisy Memory Channel," demonstrating how noises — confusion and mistakes — affect decision-making.

With the aim to help people make informed decisions, Hertel et al. [45] present a comprehensive overview of cognitive bias modification (CBM) strategies, which encompass procedures designed to prohibit automatic cognitive processes. This includes the alternation of peoples' attention, interpretation of situations, and their memory to affect future decisions. Based on the growing body of work investigating CBM, Jones et al. [49] conducted a systematic literature review. The authors demonstrate the effectiveness of CBM across selected studies, which consistently show that CBM can modify targeted biases in adults but not children. This work underlines the potential of CBM as a tool for influencing automatic decision processes and reducing cognitive biases. Between the lines, these works postulate certain negative connotations and consequences of cognitive biases. Haselton et al. [41] spotlight certain situations where the opposite occurs. Their work on Error Management Theory (EMT) asserts that biased judgments in uncertain conditions can actually result in better decisions compared to unbiased alternatives as part of ongoing research studying the evolution of cognitive biases [42].

While these different perspectives mirror the power with which cognitive biases steer decisions, work illustrates the ease in which they can be exploited, for instance, to personalise recommender systems [77] for increasing purchasing behaviour or govern peoples' search behaviour [5]. To explore possibilities to account for cognitive biases to avoid harmful designs, work by Kahneman and Tversky [78] and Baron [6] informed the design of our focus groups. We relied on the general definitions and selected biases from these works as part of each focus group. We thus base our findings on the fundamentals of cognitive bias and heuristic decision-making theory to learn about their relation to dark patterns.

2.3 Dark Patterns

Critique on nudges [40] spotlights ethical design caveats to obfuscate people's decision-making. In the scope of HCI, these concerns have sparked various streams of research to investigate the negative or harmful effects of technological interactions — in its midst a discourse surrounding dark patterns. Introduced by Brignull in 2010 [11], the concept describes unethical practices that trick users into undesired actions with negative or harmful consequences. Alongside this discourse, research in HCI has fostered a growing typology of unethical graphical user interfaces fitting the definition of dark patterns across digital media. These include, but are not limited to, games [87] and social media [64, 71], as well as context-specific instances where language barriers are exploited [46]. This effort has recently led to the development of a first ontology of related design patterns [30]. *Interface Interference* [32], for instance, promotes certain interface elements, often visually, to gain users' attention with instances identified in e-commerce [58] and social media [63]. Collectively, these works demonstrate the multitude of deceptive and manipulative strategies practitioners use to steer users' decisions against their best interest and toward service providers' goals. To offer an overview, Mathur et al. [59] constructed overarching characteristics by reviewing over 82 types of strategies that include research and regulation-based types. Mildner et al. [64] studied the

effects of 80 empirically studied types in the context of social media, while Gray et al. [30] considered a corpus of over 245 for their ontology.

While studies have frequently proposed a relationship between dark patterns and cognitive biases [58, 82], to our knowledge, we are among the first to explore this relationship in a dedicated study. Allowing interpretation in this vein, various studies convey difficulties among their participants to effectively recognise and avoid the effects of dark patterns [10, 22, 56, 63]. Aiming to create privacy-protecting interventions, work has deployed design strategies as so-called "bright patterns" [36] that invert the mechanisms of dark patterns. The study demonstrates an arguably positive impact on users' decisions. However, the study also shows that users' choice architecture is similarly altered as it would be by dark patterns and, thus, restricting users' autonomy to make informed decisions. This aligns with research conducted by Ahuja and Kumar [2], who demonstrate how dark patterns restrict user autonomy on various levels, bridging the important topics. As the overall goal of this paper is to understand the underlying mechanisms of autonomy-respecting design, we build on their efforts and focus on cognitive and behavioural aspects, particularly nudging and cognitive biases.

2.4 Terminology

Recent voices within the dark pattern community have argued that contemporary terminology — "dark pattern" — should not be used as it poses potential racial misconception [1]. In an attempt to offer an alternative, Brignull [11], who originally coined the concept, suggested using "deceptive design" to describe the unethical design practices. However, voices objected to this change [68], arguing the term "dark" does not insinuate a "bad" interaction but suggests hidden consequences. Furthermore, it dismisses a connection to pattern language [3], which conceptualises reusable design strategies and offers solutions to similar problems. This has led researchers to come up with different terms [66] to describe the same concept. While this discourse still unfolds, we opt to use the term "dark patterns", staying coherent with related work placing its origin in *hidden consequences* of interfaces instead of evilness or mal-intent, but acknowledge the importance of the current discussion problematising the term.

3 METHOD

To answer our research question and conceptualise the dynamic relationship between cognitive biases and dark patterns, we conducted a focus group study with experts in dark patterns and cognitive science/psychology research. Several participants in the study also bring extensive experience in interaction design, enriching our understanding of the practical implications in this field. We opted for expert participants to accommodate the exploratory nature of our research question. Additionally, the engagement of these two relevant perspectives in focused discussions promises novel and interesting insights into the similarities, differences, and facilitators of the two concepts. For each focus group, we invited two expert participants with extensive knowledge of dark pattern scholarship and another two with academic backgrounds in psychology or cognitive science. After agreeing to participate, we sent relevant study information to each participant. At the same time, we gained their consent to record and analyse each session in line with the host university's guidelines. Except for one focus group featuring three participants, each was attended by four experts, resulting in a total of 15 participants. To accommodate their international backgrounds and the different time zones they were living in, the focus groups were held online via the video conference tool Zoom.

3.1 Participants

Participants were recruited from the authors' professional and academic networks or word of mouth. We carefully selected individuals with backgrounds and expertise in dark patterns and cognitive science or psychology with the precondition of having published in esteemed venues of their respective fields. Participation was entirely voluntary and without compensation. Before participating in the focus groups, participants were sufficiently informed about the study's purpose and design, about their rights following GDPR guidelines and then asked to give their informed consent. In total, we recruited 15 participants, six self-identified as female, eight as male, and one as non-binary. The average years of experience participants had in their fields at the time of conducting this study was 6.73 years ($sd = 3.43$). On average, participants were 32.87 years old ($SD = 6.08$). At the time of running this experiment, their professions included (Assistant) Professors (5), Postdocs/Senior Researchers (5), and PhD candidates (5). The demographics of our participants are summarised in Table 1.

3.2 Focus Groups

In the course of 90 minutes, each focus group followed the same procedure: After a brief introduction of all participants, commonly used definitions of the terms *cognitive bias* and *dark pattern* were provided in the form of an online presentation via Zoom. To illustrate previously discussed connections between the concepts, we showed an example illustrating common streaming subscription tiers, including the 'framing effect' and 'decoy effect' to demonstrate their enabling of *aesthetic manipulation*, *visual interference* as well as *price comparison prevention*, *hidden information*, and *sneaking* dark patterns. Figure 1 shows the example given during this introduction.

Initial Discussion. Afterwards, the group was asked three questions to be discussed within ten minutes each: (1) What are the key similarities between cognitive biases and dark patterns? (2) What are the key differences between cognitive biases and dark patterns? (3) What does a cognitive bias facilitate to become a dark pattern, especially with regard to design? Each question affords participants to consider the topic from a different angle. The participants were prompted to use the communication channel of their choice to voice their opinions. They could make use of the chat function of the video conferencing platform or voice their perspective directly via the audio channel.

Card Sorting Task. After the initial discussion, participants were coupled with a disciplinary counterpart to conduct a card sorting task. Using the online collaboration tool Miro, both pairings [2] were

[2]For the group featuring three participants instead of four, we joined all experts within a single group to complete the card sorting task together.

Participant Table

ID	Age	Gender	Country of Residence	Education	Occupation	Years of Experience
P1	47	Male	Ireland	PhD	Senior UX Researcher	3
P2	32	Male	Germany	PhD	Postdoc	8
P3	40	Female	Netherlands	PhD	Assistant Professor	7
P4	25	Male	United States	MSc	PhD Candidate	3
P5	28	Female	United States	MSc	PhD Candidate	5
P6	28	Male	Finland	PhD	Postdoc	5
P7	29	Female	United States	MSc	PhD Candidate	6
P8	29	Female	Germany	Diploma	PhD Candidate	4
P9	39	Non-binary	United States	PhD	Professor	15
P10	31	Male	Ireland	PhD	Postdoc	6
P11	32	Male	Germany	PhD	Postdoc	7
P12	40	Male	Ireland	PhD	Assistant Professor	10
P13	29	Female	Luxembourg	MSc	PhD Candidate	3
P14	29	Female	United States	PhD	Assistant Professor	12
P15	35	Male	Finland	PhD	Assistant Professor	6
Mean	= 32.87				*Mean*	= 6.73
SD	= 6.08				*SD*	= 3.43

Table 1. This table presents an overview of expert participants of the four focus groups.

then tasked to group selected cognitive biases and dark patterns within 20 minutes. To afford the timely constraints of each focus group, we chose to select only prominent cognitive biases and dark patterns based on the citation counts of respective publications. Because of this limitation, however, we did not analyse these results further to avoid a potentially strong selection bias. To give participants more context, cards included definitions taken from original publications. Despite its limitations, we included this exercise as we wanted to engage participants in a transdisciplinary activity to experiment and test previously discussed ideas. Further, our aim was to inspire alternative perspectives and enrich the following discussion. The supplementary material contains a snapshot of the initial Miro board, including the selected cognitive biases (sourced from Tversky's work [78]) and dark pattern types (sourced from the dark pattern ontology by Gray et al. [30]).

Reflective Discussion. Following up on the card sorting task, participants discussed their experiences, difficulties, and ideas behind creating groups. In addition, the experience gained in the previous activity led to new insights with regard to the questions raised in the initial discussion. This allowed a more critical reflection on earlier statements as well as identifying links between cognitive biases and dark patterns that had gone unnoticed before.

4 ANALYSIS

Once we completed the four focus groups, we manually transcribed the discussions and prepared the data for further analysis. To this end, we anonymised any content that could be traced back to any individual participant. For the analysis of the transcribed data, we conducted a thematic analysis and used the card sorting results as an assistive source for a better understanding of participants' perspectives. These steps were concluded using the software Atlas.ti [29].

4.1 Positionality

The authors of this research have mixed backgrounds. One author gained their education in West Africa while the others acquired their education in Central Europe with WEIRD (Western, Educated, Industrialised, Rich, and Democratic) [44] backgrounds. Their research expertise includes design, computer science, and psychology, while their academic work focuses on topics concerned with social justice and user well-being in digital technology contexts. The focus groups were transcribed by two authors and, afterwards, coded by two authors. All participants were recruited through professional networks and by word-of-mouth. Each focus group was conducted by at least two authors, one of which would be responsible for moderating, the other(s) held assistive role(s). As this research aims to describe the relationship between cognitive biases and dark patterns, this lens has guided the structure and analysis of the focus groups. Finally, we acknowledge any possible biases that are the result of our academic, cultural, and personal backgrounds.

4.2 Coding of the Focus Group Transcripts

Prior to the analysis, we manually transcribed and anonymised the focus group discussions. In the first step, two authors coded a representative sample of 50% of the material using open coding in line with Blandford et al. [8]. We then conducted an iterative discussion to establish an initial coding tree. The remaining transcripts were split between the two authors and coded individually. Finally, we conducted a concluding discussion session to finalise the coding tree. This was followed by a thematic analysis to identify emerging dimensions from the material as described by Blandford et al. [8]. The codebook used to analyse the transcripts of the focus groups comprises 52 codes and is included in this paper's supplementary material.

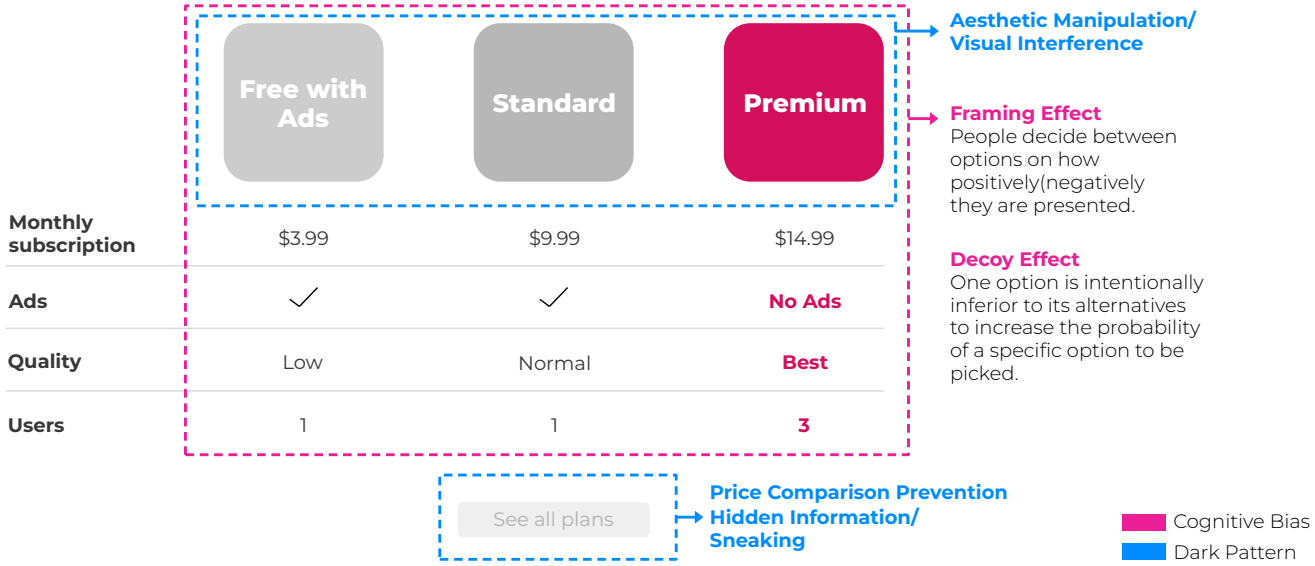


Fig. 1. This Figure was part of the introduction in each focus group and illustrates a common monthly subscription model of service providers with different tiers. The figure highlights in magenta the presence of cognitive biases (framing effect and decoy effect) and in blue dark patterns (aesthetic manipulation/visual interference, price comparison prevention, and hidden information/sneaking).

The main purpose of the card sorting task was to engage participants in an interactive task where dark pattern scholars and psychologists/cognitive scientists worked together to engage with and analyse their ideas from previous discussions. Hence, the set of biases and dark patterns used for the card sorting task was limited and only included subsets of the overall cognitive bias and dark pattern typologies. Thus, results were used to build an understanding of the data collected in the different discussion phases, therefore aiding our analysis of the focus group transcripts and the following identification of relationships.

Using axial coding, two authors began to independently organise all codes systematically into hierarchical structures and groups [8, 20]. This process was extended by connecting individual codes to describe their relationships, noting whether any confirming or contradicting notions exist between them. At this stage, the authors frequently revisited transcripts and card sorting results to ensure they stayed truthful to the data. Afterwards, the authors exchanged their findings in a following discussion. Here, the emergent hierarchical frameworks of the codes were the focus of identifying overarching themes. We thereby aimed to identify phases that followed design practice to real-world consequences in terms of dark patterns. Consequently, based on the analysis outlined above, we constructed a model that encapsulates these phases, effectively bridging the gap between design practices and their tangible impacts in the context of dark patterns.

5 FINDINGS

Based on the thematic analysis, we gained certain insights into the relationship between cognitive biases and dark patterns. However, this relationship appears intricate and, expanding prior suggestions [58, 82], multifaceted. Here, we echo our participants' discussions to provide common and specific characteristics of the two fields. Moreover, we focus on the implications of design, the inscription of functionalities, and real-world applications as discussed by our participants, and promote the "Relationship Model of Cognitive Biases and Dark Patterns". Our presentation of the results as a model is consistent with established practices of presenting results in HCI and Ubicomp (e.g. [23, 25, 48, 62]).

5.1 Similarities, Differences, and Facilitators

Each focus group included granulated discussions about similarities, differences, and facilitation between cognitive biases and dark patterns. Although the participants either had backgrounds in dark pattern scholarship or cognitive science/psychology research, combined with experience in interaction design, providing definitions and a visual example established a common ground for fruitful discourse.

5.1.1 Similarities. When asked about the similarities between the two topics, participants noticed shared attributes regarding the impact on decision-making and autonomy. Arguing from a designer's perspective, P5 discussed this sentiment further when persuading user actions:

> *"I think the key similarity is [...] how the steering happens. [...] You are trying to steer your users from the normative ways of user agency, which is what HCI design mostly preaches." – P14*

By leveraging users' perception and cognition, responsibility plays a pivotal role in the profession of designers. In this regard, P3 noticed another similarity in the shared dangers and harms that occur when not cared for. The same participant went on to discuss the particularities of users' unawareness:

> *"[B]oth can be unconscious. So, in both cases, the user may not be aware of the influence on their decision-making." – P3*

Overall, participants noticed certain synergies between cognitive biases and dark patterns. When triggered, both carry risks of making unfavourable choices as either is difficult to avoid. Certain responsibility was further attributed to designers utilising cognitive biases when used to steer users against their will.

5.1.2 Differences. The second question sought to identify differences between cognitive biases and dark patterns. Across all focus groups, a dominant argument addressed the different natures of either concept. While *"biases are already there [...] and inform our decisions"*, as P9 pointed out, or *"your intuitive brain [...] working in free-flow without really engaging too much in reasoned thought"* (P1), dark patterns, on the other side, are actively created and deployed by practitioners. P1 later continued their argument in consideration of the power dynamic between the designer and user:

> *"[A] designer has complete control over how that pattern takes form, takes shape, and how it's implemented within a system. Whereas a user doesn't have control over their cognitive biases." – P1*

As cognitive biases are intrinsic to our behaviour, dark patterns, like all design patterns, are created and impact our behaviour intrinsically. Importantly, participants noticed that not all dark patterns actually require cognitive biases. This is in line with a statement made by P10:

> *"[O]bfuscation and sneak-into-basket and hidden costs, all of those just seem like outright lying. So, it's not necessarily using a cognitive bias to be the problem." – P10*

5.1.3 Facilitators. During the last question, participants explored possible facilitators between the concepts. As similarities and differences foreshadow related attributes, P9 discusses how dark patterns emerge:

> *"[D]ark patterns are really the design material that allows those cognitive biases to be activated [...] this method is used to take advantage of our cognitive biases." – P9*

A strong sentiment across participants for a facilitator was further noticed in (mal)intent behind deploying dark patterns. In this regard, P3, P12, and P15, thoroughly discussed designers' roles in persuading users' decisions toward commercial goals, limiting their autonomy. Sharing this position, P10 noted:

> *"Something is dark patterns when you use any means whatsoever [to] deprive me of my autonomy or to [...] engage me in practices that violate my privacy." – P9*

Although not all dark patterns rely on cognitive biases, the latter seems to be an effective means to the end for malicious strategies. This reconnects to the previously discussed responsibilities of designers but also highlights additional needs for ensuring user safety. As P3 anecdotally pointed out, however, regulations, such as the GDPR, do not consider intent as a necessity. This would ensure user protection whenever harm is done.

5.2 The Relationship Model of Cognitive Biases and Dark Patterns

To address the echoed implications for responsibility and impact for practitioners, we opted for developing a model that follows design from its creation to real-world implications. In a preliminary version, we modelled how cognitive biases and their exploitation can lead to deceptive design. Our model was constructed based on the findings of our study juxtaposed with relevant previous work. We verified our model by inviting focus group participants to provide feedback on individual and general levels of this preliminary model, helping us iterate and improve it where necessary. To this end, we sent out an online survey to all participants, informing them about the task and gaining their consent before breaking down the model for individual critique. Additionally, we offered participants to reach out to us and share further comments. In total, four participants responded to the survey, helping us to advance the model.

Based on the focus groups, collected feedback, and prior work, we introduce the *Relationship Model of Cognitive Biases and Dark Patterns* as an answer to our research question (visualised in Figure 2): How can we conceptualise the dynamic relationship between cognitive biases and dark patterns? This model comprises three stages that encompass five phases, showcasing a process from design to the real world. More precisely, inspired by ethical considerations of dark patterns [34] and their relationship toward cognitive biases [58, 82], the model follows the implementation of dark patterns (or other deceptive designs) from addressing cognitive biases in design to potential real-world implications, thereby showcasing potential harmful consequences which require the assessment of responsibilities.

Drawing inspiration from Verbeek's theory of technology mediation [80, 81], our model comprises three stages: (1) Inscription / Delegation from a designer's perspective, (2) mediation of the technology or particular interface, and (3) users' interpretation thereof. Adopting Verbeek's theory, the model demonstrates how dark patterns emerge from exploiting cognitive biases. To this end, our model breaks the three stages down into five phases. The first stage describes the designer's perspective to inscribe or delegate functionalities. Happening in phases one and two, **design addresses particular cognitive biases** determining the **balance between autonomy versus coercion**. The second stage – mediation – contains the third phase: the **exploitation of cognitive biases**, which can ultimately lead to deceptive practices and harm. The third stage focuses on the users' point of view, interpreting the design throughout the fourth and fifth phases. First, users **experience the designs' implications**, leading them to **questioning responsibility**. Additionally, we identified crossroads for safeguarding strategies of end-users as well as opportunities for organisations to limit harmful design through exploiting cognitive biases. In the following subsections, we outline each phase in detail. We support the descriptions for each phase through quotes from our participants and connect individual phases to related work where applicable. For improved

readability, we slightly altered some statements, ensuring words and sentiment were maintained.

5.3 From Design to Real World

The aim of the *Relationship Model of Cognitive Biases and Dark Patterns* is to follow the development of (unethical) design from its planning stages until its deployment into the real world. Importantly, as with many things, breaking decisions and their consequences into stages or phases can result in over-simplifying processes. Our model is no exception. To emphasise the responsibility tied to designers' decisions and connect them with the consequences that arise once a design is implemented in the real world, we represented this progression with a continuous arrow beneath the three stages. Research [32, 59] and regulation and legislation [9, 52, 69] suggest that malintent plays a role in the design of online interfaces to deceive or manipulate users. However, many design decisions are not aimed at exploiting users' cognitive biases, and neither are practitioners always able to predict resulting harms. Throughout development, many constraints (e.g. money and time) limit the possibilities in which a design [31, 80] can be tested. Over-simplifying these issues can deflect from the critical aspects when design turns deceptive. Echoing the voices of our participants, the following phases attempt to convey these critical aspects with the goal of respecting the underlying continuum.

5.4 Phase 1: Design Addresses Cognitive Biases

The first phase of this model addresses the utilisation of cognitive biases to control users' attention or afford specific interactions. This is in line with Cross' description of design cognition [21], where it is part of the design process to identify the right problem to be solved: Depending on the situation and the underlying goal, a design may draw the user toward or away from it. Interactions are either better supported or actively obfuscated to guide users' decisions. While there are plenty of reasons for either strategy, dark patterns are used to provoke certain choices. It is important to emphasise here that our data show that specific cognitive biases are addressed both consciously (conscious decision of the designer) and unconsciously (not deliberately addressed by the designer). The *Bad Default* dark pattern [13], for instance, is often used in privacy settings where the system provider is initially allowed to collect personal information until a user decides otherwise. This pattern finds support in the 'default effect' as people often follow existing choices [6]. Whether the deployment of these strategies and usage of supportive cognitive biases is a deliberate choice is questioned by P14:

> *"Are [practitioners] doing this knowingly? I mean, the obvious is that [if] you try to get people to buy a certain thing, [...] how dark is a dark pattern if it's just accidental?" – P14*

Designers can address the *intrinsic* nature of cognitive biases — shared between humans — when creating or using dark patterns to alter peoples' choice architecture. The particular relationship between intrinsic cognitive biases and extrinsic deployment of dark patterns was repeatedly discussed among participants across the focus groups. P15 and P12 populated this idea through the following statements:

> *"[C]ognitive biases are also so embedded in our automatic decision-making process, while dark patterns are our techniques deployed by external entities. [...] One is individual, embedded decision-making, and the other is external to us." – P15*

> *"[C]ognitive biases are the receiver while the [dark pattern] are something from the perpetrator." – P12*

The extrinsic effectiveness of dark patterns thereby hugely benefits from intrinsic cognitive biases, described by some participants as a symbiotic relationship. The knowledge garnered by cognitive scientists and psychologists plays a large role in informing dark patterns and spotlighting opportune moments to leverage cognitive biases. However, designers are not always in possession of this knowledge. P15 elaborated on the potential of this knowledge by saying:

> *"Knowing how we process information and how we heuristically process and think is the way dark patterns become so effective." – P15*

5.5 Phase 2: Balance Between Autonomy vs. Coercion

In the second phase, designers navigate the challenging equilibrium between empowering users by giving them autonomy over their choices and deceptive practices that coerce users' decisions. The (conscious and unconscious) decisions made in the first phase have a direct influence on this balance, where ethical considerations and consequences are introduced. As with the first phase, the balance between providing users with autonomy and coercing their actions can be an active design choice or an unaware consequence [17]. The effect on the user, however, is often the same. Therefore, this phase requires careful attention to support reflective and informed decision-making. P10 argues that it can be dangerous for users when people attribute expertise to themselves in areas they do not actually have and that this can potentially cause a lot of harm, especially in sensitive contexts such as mental health applications:

> *"[S]elf assigned expertise [...] gives [practitioners] that right framing that they could then use dark patterns that exploit cognitive biases to design products without feeling like they are being very unethical."– P10*

Instead of designing responsibly with users' best interests in mind, common or best practices and traditional mindsets within professions may foster excuses for harmful design implications at the user's cost [17, 31]. In this regard, the 'status-quo' bias [6] may offer some insight into why practitioners and designers resort to deploying the same problematic designs. As with persuasive design and nudges, however, their intention might be noble, but the outcome is detrimental to users' well-being. The plethora of deployed strategies that harm users [30] is witness to practitioners' negligence in this regard, reflected by the increased regulatory efforts in place to protect users [52, 69]. However, it has to be noted that this negligence might stem from a culture of oversight or neglect at the organisational level [17]. This second phase concludes the first stage of our model and the perspective of designers who inscribe their ideas into the design.

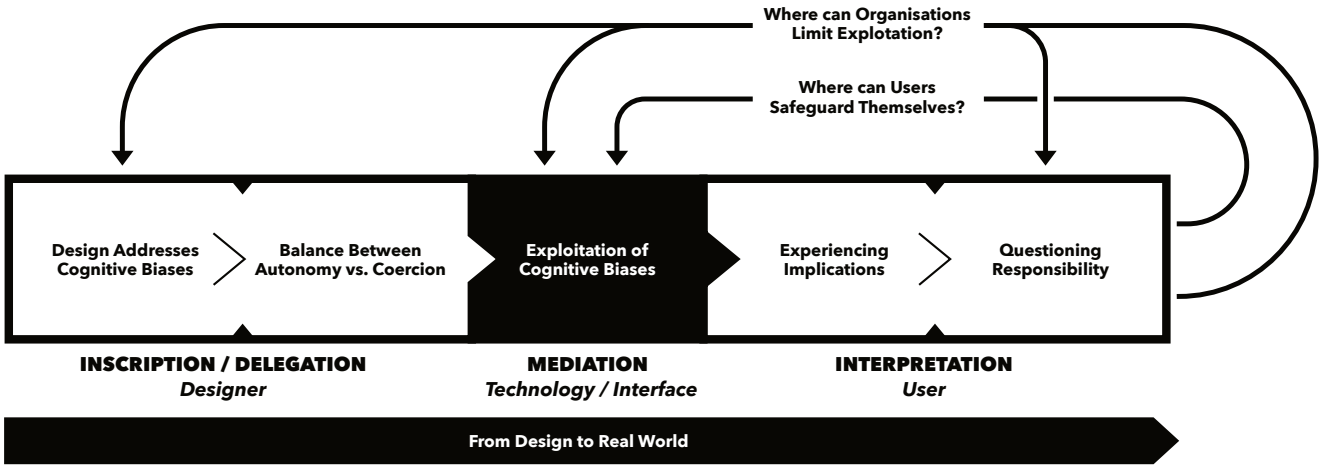


Fig. 2. This Figure presents the *Relationship Model of Cognitive Biases and Dark Patterns*. Following a continuum from (potentially unethical) design to real-world applications, the model comprises three stages spanning five phases. Adopting Verbeek's theory of technology mediation [80, 81], the model follows designers' inscription of functionalities into technology to users' interpretation, leading to the questioning of responsibilities. Depending on the impact and implications of the (unethical) design, end-users may need safeguarding measures, while policy and regulation may be required for their protection.

5.6 Phase 3: Exploitation of Cognitive Biases

Relying on practitioners' decisions in the first stage, the second stage contains the mediation of the design and the third phase where exploitation of cognitive biases manifests — eventually resulting in deceptive practices and harm. It conveys the implications of design that harnesses cognitive biases and controls user behaviour. In other words, this phase is about users' interactions with the design and, thereby, the first phase, where design enters the real world with all its implications. It describes the users' reaction to the design but not yet their evaluation of any interaction. Dark patterns that govern user interactions influence their choice architecture without providing transparent information about their consequences. In this vein, P9 drew the following connection:

> *"[A] cognitive bias is a necessary component as kind of [...] an ingredient to a dark pattern". – P9*

A core element to enable design — whether for good or bad — is to afford specific interactions [67]. To that end, affordances can have many forms and shapes to convey their aims but have to be delivered easily accessible for users to engage. A common choice to create affordances is through nudges [76]. While work demonstrates their effectiveness [14], they limit transparency as P9 detailed:

> *"[N]udges for good fail to respect the autonomy of the person who's using the system if they're not very open and transparent about how the nudges are working and there becomes a power imbalance then between whoever is providing those nudges and the person on the other end of them". – P9*

This power imbalance is a critical implication of dark patterns as the user is rarely in control or aware of any consequences [63]. Thereby, design can easily facilitate its goals without providing sufficient information that would allow users to reflect on their choices. P10 explained how preferences can be exploited in the form of cognitive biases:

> *"[U]sing our tendency to prefer colour or be attention-grabbing so that we look at [the interface] more or our tendency to kind of not want to think more than we have to. [...] [D]ark patterns [are] more of a method and cognitive biases [are] more of an inherent thing that is exploited." – P10*

Our participants also discussed the natural benefits of certain cognitive biases and heuristics to provide effective shortcuts in various situations. In tandem with current design practices and service providers' monetary incentives [16], their mere presence opens the door for serious exploitation, limiting users' ability to formulate reflected decisions and avoid manipulation into unfavourable interactions.

5.7 Phase 4: Experiencing Implications

The fourth phase of the model addresses the real-world implications behind dark patterns and how users experience them. Here, the third and last stage begins, where users interpret the design. At this time, the impacts of the design become noticeable and perceptible, whether originally accounted for or unintentional. Informed by the first two stages of the model, during the fourth phase, practitioners have already deployed the strategies to deliver their design's goal(s). This includes any decisions that affect user agency and autonomy through cognitive biases. In other words, one could say that from a user's perspective, the harm is done. Elaborating on users' expectations, P10 raises important questions from a user's point of view:

> *"[W]hat mental models do I have about these products? And in what ways [do] my mental models of this product align with what's happening before me?" – P10*

Before any interaction, users may have certain expectations of what will happen. The previous phases and underlying choices ultimately lead to the moment when a user interacts with a design for the first time. While certain dark patterns prohibit reflective thinking, others obscure access to specific functionality. Unaware of the practitioners' goal, users rely on a design's attributes to understand required actions. Thereby, they can draw from their mental model and past experience before engaging with it but cannot know for certain what the consequences will be. At this stage, practitioners and users will only be able to assess the truth behind the implications once harm has been done.

5.8 Phase 5: Questioning Responsibility

In response to the aftermath of the prior phase, the fifth phase mirrors past decisions and reflects the question of responsibility behind harmful implications, concluding the third stage of interpretation. However, this question is not simply answered. From a user's perspective, technological illiteracy or unawareness of the presence of dark patterns can lead to seeking responsibility in themselves. Practitioners, on the other end, may displace responsibility behind best practices and common, albeit critically viewed, design paradigms or deflect responsibility altogether, convinced of the noble cause of their design. In any case, the responsibility is difficult to ascribe to a single party. Nonetheless, there are multiple avenues to mitigate negative implications. In this regard, P1 reflected on unintended but harmful interactions, taking the practitioners' perspective:

> *"[T]here is some kind of control over it from a designer's point of view that once [the harm] has been identified and they've observed the problem, that they can go back and fix it then afterwards". – P1*

However, the strategy outlined by P1 would require a thorough understanding of exploited cognitive biases, alternative strategies, and access to necessary resources to change the design post-development. Especially in this phase, it is important to remember this model also as a continuum where decisions are intertwined and causally dependent. Hence, untangling the root causes behind implications and identifying responsibility is difficult. Therefore, practitioners need to acquire a critical view of their own work and reflect on utilised strategies to understand how cognitive biases enable dark patterns that affect end-users' decision-making. This does not mean that practitioners need to become experts in human cognition and perception (although it would not harm). Reflective steps throughout development stages can already help ensure that user autonomy is sufficiently respected to empower informed decision-making.

5.9 Safeguarding and Counter Measures

Collectively, the five phases of the *Relationship Model of Cognitive Biases and Dark Patterns* highlight how dark patterns leverage cognitive biases and how the latter can enable the first. However, counter-measures to existing dark patterns and their harms have not yet been accounted for. We, therefore, expanded the model with two additional arrows that link to opportune moments where users could safeguard themselves and where organisations, such as governmental bodies, could restrict the exploitation of cognitive biases through guidelines and regulations. Suggesting that knowledge can help users mitigate harmful effects, P9 stated the following:

> *"Sometimes, if it's raised to your attention, you can fight back yourself. But in other cases, the pull is so strong that you actually need regulatory bodies or other kinds of policies in place to fight back on your behalf." –P9*

Safeguarding measures for and by end-users are only effective if the problems are understood and can be avoided. In this regard, one arrow leaving the fifth phase carries the question: *Where can users safeguard themselves?* — points to the third phase, where cognitive biases are exploited. Users can learn to mitigate cognitive biases, for instance, through understanding how they manifest [18, 45]. But even then, it may not be possible to always rely on the safeguarding measures as demonstrated by Bongard-Blanchy et al. [10] or Mildner et al. [63]. Importantly, our participants have emphasised that this is not (ever) the user's fault. Hence, our participants envisioned that knowledgeable people would be able to recognise the potential exploitation of cognitive biases in the third phase and could proceed strategically with an appropriate reaction. P9 later continued their argument by discussing the current approach behind dark patterns and proposing legal requirements to assure user protection:

> *"I think a lot of the mainstream dark patterns that seem to be very effective at triggering our cognitive biases and altering our ability to act or understanding what the situation is, do certainly seem like they're illegitimate forms of altering sense-making versus ones that we would describe as normatively acceptable."–P9*

In this regard, P3 points toward already existing regulation, such as the Digital Service Act (DSA) [69] of the European Union (EU), restricting the implementation of dark patterns. Exploitation of cognitive biases, however, is not regulated:

> *"There [is] the word dark patterns in the law now, but cognitive biases are not a term for law." –P3*

In our model, organisational limits to exploitation, such as guidelines or regulations, are alternatives to end-user safeguarding. In this regard, the model incorporates a second arrow leaving the fifth phase asking: *Where can organisations limit exploitation?* The arrow links to the first, third, and fifth phases to highlight opportunities to protect users. In the same direction as pointed out by P3, organisations could request user autonomy to be respected, as done by the EU's DSA [69]. They could also restrict the exploitation of cognitive biases through regulating deceptive design, also done by the DSA as well as other regulations such as the US state of California's CCPA [52] or India's recent issue to prevent dark patterns [65]. Lastly, the arrow loops back into the fifth phase to illustrate the importance of identifying offenders and holding them accountable if users are harmed.

Although we view these measures as more effective compared to users' safeguarding themselves, we recognise that they are also more drastic and finite in the case organisations increase regulations. On the one hand, they could address the exploitation of certain biases directly, for example, by requiring specific user consent or autonomy to prevent harmful, impulsive interactions. This line of defence leaves space for practitioners to choose alternative design

strategies. On the other hand, regulations can restrict the overall available design space, prohibiting the deployment of dark patterns. As with the DSA [69], specific strategies can be directly addressed to ensure user safety.

6 DISCUSSION

Inspired by previous research proposing a connection between cognitive biases and dark patterns [58, 82], this paper offers answers to our research question by problematising the dynamic and intricate constraints within this relationship. To that end, we conducted four focus groups and, based on our results, proposed the *Relationship Model of Cognitive Biases and Dark Patterns*. This model depicts five phases that describe how decisions to deploy cognitive biases can enable exploitation and dark patterns. In this section, we discuss applications of this model, how technology can be devised to preserve user autonomy in such contexts and point toward future avenues HCI could consider to foster the implementation of ethical design and user protection.

6.1 Using the Relationship Model of Cognitive Biases and Dark Patterns

Previous work has described a growing typology of dark patterns [30, 59, 64] users constantly encounter across web and app interfaces. While various patterns have been captured and are relatively well understood, this strand of research lacks in-depth knowledge of the underlying mechanisms. Contributing relevant insights in this regard, the main aim of the *Relationship Model of Cognitive Biases and Dark Patterns* is to support both researchers and practitioners by providing them with reflective phases that can support them in considering the ethical caveats and impacts of their designs. It is not meant to be prescriptive but provides a guiding roadmap reminding about ethical implications throughout the design's lifespan and the interplay of cognitive biases and dark patterns in that regard. While the aim of our model cannot change any malicious objectives of practitioners, it can serve as a reminder about the implicated consequences of utilising cognitive biases and can guide toward potential counter-measures. The model can, therefore, be applied in situations where dark patterns are observed. **Especially in the early development stages of designs, the *Relationship Model of Cognitive Biases and Dark Patterns* can complement decisions made alongside existing, traditional design paradigms that may not always prioritize user agency and autonomy.**

The research community focusing on dark patterns [35] has recently questioned the intent behind deploying strategies that harm users. Throughout their discussions, participants across our focus groups separated practitioners' intent from the harmful impacts of deployed interactions. While some practitioners utilise obvious tactics to manipulate users' choices to their service's advantage, smaller service providers may naively follow existing common practices and deploy dark patterns without realising the consequences. However, in both cases, the end-user may suffer from possible implications, regardless of intent. Based on these insights, we argue that **necessary protective measures should be independent of intentions as they deem the safety of end-users most important.** Recent regulations [52, 69] already try to restrict certain dark patterns. While this offers promising avenues to enhance user protection and safety, we maintain that regulatory measures should be employed judiciously when other protective measures have proven ineffective. The form and shape in which regulation should be enforced pose new, complex challenges. On the one hand, over-regulation may restrict innovation in the early stages. On the other hand, it is difficult to administer regulations that affect domains as large as web and app interactions.

6.2 Preserving User Autonomy

A key aspect of our model is to provide arguments for the importance of supporting and preserving user autonomy.Based on our findings, we argue that it is crucial to allow for and foster informed decisions before interactions happen. Traditional HCI paradigms offer practitioners a range of tools and common practices to support ethical design practice. Although critically reviewed, persuasive design explains how motivation can be directly addressed to steer and alter users' choice architecture and guide them toward a predetermined goal. While behaviour change can be supported through design and technology, it is easy to mean good and do harm [12]. In research, it is customary to behave responsibly, ask for consent, and make users aware that the technologies utilised will potentially affect their choices. Outside research, users are often kept unaware of the consequences behind interactions, often noticed in infamous cookie-consent banners [34, 79]. Knowledge about the enabling effects of cognitive biases toward dark patterns could mitigate certain exploitation and protect users [49]. However, multiple studies [10, 22, 63] investigating the end-users ability to recognise and avoid dark patterns repeatedly demonstrate difficulties among their participants, even if information about dark patterns was provided. This indicates that user safeguarding is limited in its effectiveness, and other means are necessary to uphold their autonomy.

Overall, autonomy and agency-driven design have been a longstanding element in HCI [28]. However, a lack of fine-grained understanding of how to connect these principles with usability [53] suggests more work is needed to take full advantage of existing design paradigms. Our model invites researchers and practitioners to reflect on their responsibilities. We therefore hope that it can support continuing work and foster more responsible designs. Future work could study the effectiveness of the model alongside existing design methods. For example, it would be interesting to learn whether persuasive design or nudges can afford more autonomy if the model is used alongside such practices. Nevertheless, existing work in HCI echoes a balancing act between user autonomy and alternation of choice architecture in every design. **It remains within the designers' responsibility to understand this challenge and follow ethical design principles promoting user autonomy.** A future goal of this work could be the investigation of opportune design strategies that better connect usability, user experience, and autonomy and create better incentives and tangible examples for practitioners to follow such principles.

6.3 Ways Forward

Our proposed model mainly targets its depicted challenges from an HCI perspective. While the model offers novel insights in this

regard and is designed to assist practitioners in their work, it is only one step towards understanding the underlying mechanisms of dark patterns. P15, a participant with a psychology background, speculated into the direction of cognitive processing:

> *"I can come up with the whole information processing pipeline of human psychology here and just identify some [showing that] it's not covering everything." – P15*

Although the discussions of our participants did not come to an overall conclusion or identify specific solutions, they provided rich food for thought about the different angles from which dark patterns could be studied. With more emphasis on human cognition, perhaps future work can work toward models that better support users' perspectives. An already existing angle in this regard is presented in cognitive bias modification (CBM). The work highlighting the effectiveness in which autonomous decisions can be altered to be better reflected [45, 49] spotlights an exciting direction to help users make better decisions when faced with dark patterns. Moreover, research in HCI could adopt CBM to devise technologies that prepare and shape users' expectations in line with possible interactions.

7 LIMITATIONS

Although we were careful in designing and conducting this study, we recognise that this work is prone to certain limitations. Firstly, the last focus group was attended by only three instead of four participants (two experts in dark pattern scholarship and one expert in psychology). While the focus group functioned well in terms of interpersonal dynamics, a fourth member might have changed the dynamic of the discussions. Moreover, this changed the dynamic in which prior card sorting tasks were conducted. Instead of two groups including one expert from the respective fields to discuss and execute the task, we decided to have one group including all three participants. Again, the outcome was comparable to prior focus groups, but we cannot know whether a second expert with expertise in cognitive science or psychology would have impacted the group's discourse and, thus, our findings.

Secondly, the study was designed and administered by three researchers with backgrounds in HCI research. Although one of them has a background in psychology, their current research is situated in the field of HCI. Although the study was informed by relevant work on the concepts of cognitive biases and human behaviour to ensure a levelled discussion in terms of the considered subjects, we acknowledge a certain bias stemming from our expertise. We would welcome any future attempts to reproduce our findings in the fields of human cognition and behaviour to gain different views that may result in additional findings.

Our personal HCI backgrounds also resulted in a third limitation regarding the *Relationship Model of Cognitive Biases and Dark Patterns*. Here, we recognise that our HCI background guided our moderation of the focus groups and informed our interpretation of the data as well as any further analysis. To mitigate these effects, we extensively reviewed material from the field of human cognition, particularly focusing on cognitive bias literature. However, researchers with other backgrounds may interpret the same data differently.

8 CONCLUSION

This research sheds light on the dynamic relationship between cognitive biases and dark patterns. Based on a focus group study with expert participants, we explore the ethical considerations throughout the development of design and describe the *Relationship Model of Cognitive Biases and Dark Patterns* providing an overview of the continuum between design and real-world implications of harmful design.

To that end, we emphasize the critical role of practitioners and researchers in considering the ethical implications of their design decisions, particularly when it comes to user autonomy. By illustrating the dynamic process through which cognitive biases are leveraged to create dark patterns, the model underscores the responsibility designers hold in shaping user experiences. Furthermore, the study recognises the importance of safeguarding strategies for end-users and regulatory measures to protect individuals from the potential harms of dark patterns. This recognition aligns with broader discussions in the HCI community regarding better implementation of ethical design practices and user-centred development of autonomy and agency-enabling technologies.

In summary, this research deepens our understanding of the intricate interplay between cognitive biases and dark patterns but also provides a practical model for practitioners to navigate this complex landscape responsibly. As technology continues to play an increasingly ubiquitous role in our lives, the ethical considerations highlighted in this study become even more pertinent as related work catalogues the variety of unethical, dark patterns. Thus, we hope that this research contributes to the ongoing dialogue on ethical design in HCI and sets the stage for future investigations into creating safer, more user-friendly digital interfaces.

REFERENCES

[1] ACM. 2023. Words matter: Alternatives for charged terminology in the computing profession. https://www.acm.org/diversity-inclusion/words-matter

[2] Sanju Ahuja and Jyoti Kumar. 2022. Conceptualizations of user autonomy within the normative evaluation of dark patterns. *Ethics and Information Technology* 24, 4 (Dec. 2022), 52. https://doi.org/10.1007/s10676-022-09672-9

[3] C. Alexander, S. Ishikawa, and M. Silverstein. 1977. *A Pattern Language: Towns, Buildings, Construction.* OUP USA. https://books.google.ch/books?id=hwAHmktpk5IC

[4] Dmitry Alexandrovsky, Maximilian Achim Friehs, Jendrik Grittner, Susanne Putze, Max V. Birk, Rainer Malaka, and Regan L Mandryk. 2021. Serious Snacking: A Survival Analysis of how Snacking Mechanics Affect Attrition in a Mobile Serious Game. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Number 113. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445689

[5] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21).* Association for Computing Machinery, New York, NY, USA, 27–37. https://doi.org/10.1145/3406522.3446023

[6] Jonathan Baron. 2007. *Thinking and deciding.* Cambridge University Press.

[7] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI Understand Human Agency and Autonomy?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3580651

[8] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. *Synthesis Lectures on Human-Centered Informatics* 9, 1 (April 2016), 1–115. https://doi.org/10.2200/S00706ED1V01Y201602HCI034 Publisher: Morgan & Claypool Publishers.

[9] European Data Protection Board. March, 2022. Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them | European Data Protection Board. https://edpb.europa.eu/our-work-tools/documents/

public-consultations/2022/guidelines-32022-dark-patterns-social-media_en Visited on 2022-03-29.

[10] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. I am Definitely Manipulated, Even When I am Aware of it. It s Ridiculous! – Dark Patterns from the End-User Perspective. *Designing Interactive Systems Conference 2021* (June 2021), 763–776. https://doi.org/10.1145/3461778.3462086 arXiv: 2104.12653.

[11] Harry Brignull, Marc Miquel, Jeremy Rosenberg, and James Offer. 2010. Dark Patterns-User Interfaces Designed to Trick People. http://darkpatterns.org/ (visited on 2021-08-25).

[12] Hronn Brynjarsdottir, Maria Håkansson, James Pierce, Eric Baumer, Carl DiSalvo, and Phoebe Sengers. 2012. Sustainably unpersuaded: how persuasion narrows our vision of sustainability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 947–956. https://doi.org/10.1145/2207676.2208539

[13] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* 2016, 4 (2016), 237–254. https://doi.org/10.1515/popets-2016-0038

[14] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300733

[15] Lorena Sánchez Chamorro, Kerstin Bongard-Blanchy, and Vincent Koenig. 2023. Ethical Tensions in UX Design Practice: Exploring the Fine Line Between Persuasion and Manipulation in Online Interfaces. (July 2023), 61–73. https://doi.org/DOI:10.1145/3563657.3596013

[16] Zhilong Chen, Jinghua Piao, Xiaochong Lan, Hancheng Cao, Chen Gao, Zhicong Lu, and Yong Li. 2022. Practitioners Versus Users: A Value-Sensitive Evaluation of Current Industrial Recommender System Design. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 533 (nov 2022), 32 pages. https://doi.org/10.1145/3555646

[17] Shruthi Sai Chivukula, Ike Obi, Thomas V Carlock, and Colin M. Gray. 2023. Wrangling Ethical Design Complexity: Dilemmas, Tensions, and Situations. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 179–183. https://doi.org/10.1145/3563703.3596632

[18] Andy Cockburn and Carl Gutwin. 2019. Anchoring Effects and Troublesome Asymmetric Transfer in Subjective Ratings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300592

[19] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[20] John W. Creswell. 2009. *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed ed.). Sage Publications, Thousand Oaks, Calif. OCLC: ocn192045753.

[21] Nigel Cross. 2001. Design Cognition: Results From Protocol And Other Empirical Studies Of Design Activity. In *Design knowing and learning: cognition in design education*. Elsevier, Oxford, UK, 29–103. https://oro.open.ac.uk/3285/

[22] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

[23] Julia C. Dunbar, Emily Bascom, Ashley Boone, and Alexis Hiniker. 2021. Is Someone Listening? Audio-Related Privacy Perceptions and Design Recommendations from Guardians, Pragmatists, and Cynics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 98 (sep 2021), 23 pages. https://doi.org/10.1145/3478091

[24] Salma Elsayed-Ali, Sara E Berger, Vagner Figueredo De Santana, and Juana Catalina Becerra Sandoval. 2023. Responsible & Inclusive Cards: An Online Card Tool to Promote Critical Reflection in Technology Industry Work Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–14. https://doi.org/10.1145/3544548.3580771

[25] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. https://doi.org/10.1145/2750858.2804250

[26] BJ Fogg. 2009. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology* (Claremont, California, USA) *(Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, Article 40, 7 pages. https://doi.org/10.1145/1541948.1541999

[27] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman (Eds.). Vol. 16. Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4 Series Title: Philosophy of Engineering and Technology.

[28] Batya Friedman and Helen Nissenbaum. 1997. Software agents and user autonomy. In *Proceedings of the first international conference on Autonomous agents - AGENTS '97*. ACM Press, Marina del Rey, California, United States, 466–469. https://doi.org/10.1145/267658.267772

[29] ATLAS.ti Scientific Software Development GmbH. 2021. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[30] Colin M Gray, Nataliia Bielova, Cristiana Santos, and Thomas Mildner. 2024. An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action. *arXiv preprint arXiv:2309.09640* (2024).

[31] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. https://doi.org/10.1145/3290605.3300408

[32] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The dark (patterns) side of UX design. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–14. https://doi.org/10.1145/3173574.3174108

[33] Colin M. Gray, Cristiana Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns: Foundations, Definitions, and a Structure for Transdisciplinary Action. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 13 pages. https://doi.org/10.1145/123456789

[34] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. *arXiv* (2021). https://doi.org/10.1145/3411764.3445779

[35] Colin M. Gray, Cristiana Teixeira Santos, Nicole Tong, Thomas Mildner, Arianna Rossi, Johanna T. Gunawan, and Caroline Sinders. 2023. Dark Patterns and the Emerging Threats of Deceptive Design Practices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–4. https://doi.org/10.1145/3544549.3583173

[36] Paul Graßl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. 2021. Dark and Bright Patterns in Cookie Consent Requests. *Journal of Digital Social Research* 3, 1 (Feb. 2021), 1–38. https://doi.org/10.33621/jdsr.v3i1.54

[37] Barbara Grimpe, Mark Hartswood, and Marina Jirotka. 2014. Towards a closer dialogue between policy and practice: responsible design in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 2965–2974. https://doi.org/10.1145/2556288.2557364

[38] Johanna Gunawan, Amogh Pradeep, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2021. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–29. https://doi.org/10.1145/3479521

[39] Juho Hamari, Jonna Koivisto, and Tuomas Pakkanen. 2014. Do Persuasive Technologies Persuade? - A Review of Empirical Studies. In *Persuasive Technology*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, Anna Spagnolli, Luca Chittaro, and Luciano Gamberini (Eds.). Vol. 8462. Springer International Publishing, Cham, 118–136. https://doi.org/10.1007/978-3-319-07127-5_11 Series Title: Lecture Notes in Computer Science.

[40] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation* 4, 1 (March 2013), 3–28. https://doi.org/10.1017/S1867299X00002762

[41] Martie G. Haselton and Daniel Nettle. 2006. The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. *Personality and Social Psychology Review* 10, 1 (Feb. 2006), 47–66. https://doi.org/10.1207/s15327957pspr1001_3

[42] Martie G Haselton, Daniel Nettle, and Paul W Andrews. 2015. *The evolution of cognitive bias*. Wiley Online Library. 724–746 pages.

[43] Daniel M. Hausman and Brynn Welch. 2010. Debate: To Nudge or Not to Nudge*. *Journal of Political Philosophy* 18, 1 (March 2010), 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

[44] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (2010), 29–29.

[45] Paula T. Hertel and Andrew Mathews. 2011. Cognitive Bias Modification: Past Perspectives, Current Findings, and Future Applications. *Perspectives on Psychological Science* 6, 6 (Nov. 2011), 521–536. https://doi.org/10.1177/1745691611421205

[46] Shun Hidaka, Sota Kobuki, Mizuki Watanabe, and Katie Seaborn. 2023. Linguistic Dead-Ends and Alphabet Soup: Finding Dark Patterns in Japanese Apps. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

ACM, Hamburg Germany, 1–13. https://doi.org/10.1145/3544548.3580942
[47] Martin Hilbert. 2012. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin* 138, 2 (March 2012), 211–237. https://doi.org/10.1037/a0025940
[48] Abhinandan Jain, Adam Haar Horowitz, Felix Schoeller, Sang-won Leigh, Pattie Maes, and Misha Sra. 2020. Designing Interactions Beyond Conscious Control: A New Model for Wearable Interfaces. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 108 (sep 2020), 23 pages. https://doi.org/10.1145/3411829
[49] Emma B. Jones and Louise Sharpe. 2017. Cognitive bias modification: A review of meta-analyses. *Journal of Affective Disorders* 223 (Dec. 2017), 175–183. https://doi.org/10.1016/j.jad.2017.07.034
[50] Daniel Kahneman. 2003. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist* 58, 9 (2003), 697–720. https://doi.org/10.1037/0003-066X.58.9.697
[51] Konrad Kollnig, Siddhartha Datta, Thomas Serban Von Davier, Max Van Kleek, Reuben Binns, Ulrik Lyngs, and Nigel Shadbolt. 2023. 'We Are Adults and Deserve Control of Our Phones': Examining the Risks and Opportunities of a Right to Repair for Mobile Apps. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 22–34. https://doi.org/10.1145/3593013.3593973
[52] California State Legislature. 2018. CCPA-18 2018. California Consumer Privacy Act of 2018 [1798.100 - 1798.199] (CCPA). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.
[53] David Leimstädtner, Peter Sörries, and Claudia Müller-Birn. 2023. Investigating Responsible Nudge Design for Informed Decision-Making Enabling Transparent and Reflective Decision-Making. In *Mensch und Computer 2023*. ACM, Rapperswil Switzerland, 220–236. https://doi.org/10.1145/3603555.3603567
[54] Kai Lukoff, J Vera Liao, James Choi, Kaiyue Fan, Sean A Munson, and Alexis Hiniker. 2021. How the Design of YouTube Influences User Sense of Agency. In *CHI'21*. ACM, Yokohama. https://doi.org/10.1145/3411764.3445467
[55] Ulrik Lyngs, Kai Lukoff, Petr Slovak, William Seymour, Helena Webb, Marina Jirotka, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2020. 'I Just want to Hack Myself to Not Get Distracted': Evaluating Design Interventions for Self-Control on Facebook. In *CHI'20*. ACM, Honolulu, 1–15. https://doi.org/10.1145/3313831.3376672
[56] Maximilian Maier. 2020. Dark Design Patterns - An End-user Perspective. *Human Technology* 16 (2020), 170–199. https://doi.org/10.17011/ht/urn.202008245641
[57] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376666
[58] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–32. https://doi.org/10.1145/3359183 arXiv: 1907.07032.
[59] Arunesh Mathur, Jonathan Mayer, and Mihir Kshirsagar. 2021. What Makes a Dark Pattern ... Dark ? Design Attributes, Normative Considerations, and Measurement Methods. In *CHI'21*. ACM, New York, NY, USA, Yokohama, 18. https://doi.org/10.1145/3411764.3445610
[60] Christopher McCrudden and Jeff King. 2015. The dark side of nudging: the ethics, political economy, and law of libertarian paternalism. *Choice Architecture in Democracies, Exploring the Legitimacy of Nudging (Oxford/Baden-Baden: Hart and Nomos, 2015), Forthcoming, U of Michigan Public Law Research Paper* 485 (Nov. 2015). https://ssrn.com/abstract=2685832
[61] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. 2019. Design Friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*. ACM, BELFAST United Kingdom, 41–44. https://doi.org/10.1145/3335082.3335106
[62] Thomas Mildner, Orla Cooney, Anna-Maria Meck, Marion Bartl, Gian-Luca Savino, Philip R Doyle, Diego Garaialde, Leigh Clark, John Sloan, Nina Wenig, Rainer Malaka, and Jasmin Niess. 2024. Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users. In *roceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA, Honolulu, HI, USA, 1–18. https://doi.org/10.1145/3613904.3642542
[63] Thomas Mildner, Merle Freye, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 2362–2374. https://doi.org/10.1145/3563657.3595964
[64] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580695
[65] Ministry of Consumer Affairs, Food & Public Distribution. 2023. Central Consumer Protection Authority issues 'Guidelines for Prevention and Regulation of Dark Patterns, 2023' for prevention and regulation of dark patterns listing 13 specified dark patterns. https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1983994 Accessed 21.12.2023.
[66] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. 2023. Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3580729
[67] Donald A. Norman. 2013. *The design of everyday things* (revised and expanded edition ed.). Basic Books, New York, New York.
[68] Ikechukwu Obi, Colin M. Gray, Shruthi Sai Chivukula, Ja-Nae Duane, Janna Johns, Matthew Will, Ziqing Li, and Thomas Carlock. 2022. Let's Talk About Socio-Technical Angst: Tracing the History and Evolution of Dark Patterns on Twitter from 2010-2021. http://arxiv.org/abs/2207.10563 arXiv:2207.10563 [cs].
[69] European Parliament. 2022. Digital Servicess Act***I. European Parliament [A9-0356/2021]. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.html
[70] Douglas Glen Rizzo, Mario J. Whitman. 2009. Little Brother is Watching You: New Paternalism on the Slippery Slopes Themed Issue: Perspectives on the New Regulatory Era. *Arizona Law Review* 51 (2009), 685.
[71] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142
[72] Hanna Schneider, Malin Eiband, Daniel Ullrich, and Andreas Butz. 2018. Empowerment in HCI - A Survey and Framework. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3173818
[73] Cass R. Sunstein. 2015. Nudges, Agency, and Abstraction: A Reply to Critics. *Review of Philosophy and Psychology* 6, 3 (Sept. 2015), 511–529. https://doi.org/10.1007/s13164-015-0266-z
[74] Nađa Terzimehić, Sarah Aragon-Hahner, and Heinrich Hussmann. 2023. The Tale of a Complicated Relationship: Insights from Users' Love/Breakup Letters to Their Smartphones before and during the COVID-19 Pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 28 (mar 2023), 34 pages. https://doi.org/10.1145/3580792
[75] Richard H. Thaler. 2018. Nudge, not sludge. *Science* 361, 6401 (Aug. 2018), 431–431. https://doi.org/10.1126/science.aau9241 Publisher: American Association for the Advancement of Science.
[76] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven. OCLC: ocn181517463.
[77] Georgios Theocharous, Jennifer Healey, Sridhar Mahadevan, and Michele Saad. 2019. Personalizing with Human Cognitive Biases. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. ACM, Larnaca Cyprus, 13–17. https://doi.org/10.1145/3314183.3323453
[78] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. 185 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
[79] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, London United Kingdom, 973–990. https://doi.org/10.1145/3319535.3354212
[80] Peter-Paul Verbeek. 2005. *What things do: philosophical reflections on technology, agency, and design* (1. paperback print ed.). Pennsylvania State Univ. Press, University Park, Pa.
[81] Peter-Paul Verbeek. 2006. Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values* 31, 3 (May 2006), 361–380. https://doi.org/10.1177/0162243905285847 Publisher: SAGE Publications Inc.
[82] Ari Ezra Waldman. 2020. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current Opinion in Psychology* 31 (Feb. 2020), 105–109. https://doi.org/10.1016/j.copsyc.2019.08.025
[83] Jin-Liang Wang, Linda A. Jackson, James Gaskin, and Hai-Zhen Wang. 2014. The effects of Social Networking Site (SNS) use on college students' friendship and well-being. *Computers in Human Behavior* 37 (Aug. 2014), 229–236. https://doi.org/10.1016/j.chb.2014.04.051
[84] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. Association for Computing Machinery, New York, NY, USA, 763–770. https://doi.org/10.1145/2487788.2488038

[85] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. *SOUPS 2011 - Proceedings of the 7th Symposium on Usable Privacy and Security* (2011). https://doi.org/10.1145/2078827.2078841

[86] Christopher D. Wickens, John Lee, Yili D. Liu, and Sallie Gordon-Becker. 2003. *Introduction to Human Factors Engineering (2nd Edition).* Prentice-Hall, Inc., USA.

[87] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. *Foundations of Digital Games 2013* (2013).

Curriculum Vitae

Name: Thomas Mildner
ORCID: 0000-0002-1712-0741
E-Mail: contact@thomasmildner.me
Website: https://www.thomasmildner.me/
Citizenship: German

Education.

- 2021-2024: University of Bremen – PhD candidate in Computer Science, Advisor: Prof. Dr. Rainer Malaka
- 2018-2020: University College Dublin – M.Sc. in Computer Science, Thesis title: Scéalability - Assessing Multi-Agent Storytelling Performances With Amazon's Alexa, Advisor: Prof. Dr. Tony Veale
- 2014-2017: University of Arts, Bremen – B.A. in Digital Media, Thesis title: Anwendung von Style Transfer Algorithmen auf die Generierung geographischer Karten, Advisor: Prof. Peter von Maydell and Prof. Dr. Johannes Schöning

---

Research.

- Researcher at the University of Bremen in the Digital Media Lab led by Prof. Dr. Rainer Malaka
- Researcher at the University College Dublin in the Creative Language Systems group led by Prof. Dr. Tony Veale
- Erasmus+ Internship at the University College Dublin in the Creative Language Systems group led by Prof. Dr. Tony Veale

---

Academic activities.

I was a student volunteer at the following conferences:

- CHI 2022, 2023

I volunteered as Associate Chair at the following conferences:

- MuC 2024

I volunteered as a reviewer for the various HCI conferences and journals, including, but not limited to, ACM CHI, ACM DIS, ACM CUI, ACM CSCW, and IJHCS, receiving 5 special recognitions for outstanding reviews.

I presented my research at the following venues:

- AI IN HEALTH Bremen (2023) — “Auf der Schattenseite von Sozialen Medien - Wie Dark Patterns unsere Entscheidungen beeinflussen”
- Portugal Digital Wellbeing Week (2023) — “How Dark Patterns Influence Our Decision Making”
- M-EPLI Talks (2023) — “Uncovering Dark Patterns In Social Media”

Teaching activities.

I successfully supervised 5 bachelor theses and 1 master thesis.

Teaching at the University of Bremen

- Lecturer for the Seminar “Information Security, Data Protection, and How Dark Patterns Navigate Johnny’s Behaviour”: (2021)

Teaching at the University College Dublin

- TA - Formal Foundations (2019)
- TA - Software Engineering Project (2019)
- TA - Computational Creativity (2018)
- TA - Introduction to Relational Databases and SQL Programming (2018)

Active memberships in scientific societies, fellowships in renowned academies.

- I am an active member of the Association of Computing Machinery (ACM), contributing to conferences associated with ACM and serving as an active reviewer in the ACM community and beyond.
- Moreover, I am an associated researcher of the Early Career Research Academy (ECRA) of the Leibniz Science Campus (LSC) Digital Public Health (DiPH).

Awards.

- ACM Honourable Mention Award (top 5% of papers) - CHI 2023 in Hamburg, Germany for my paper *About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services*

Major scientific achievements

I have contributed to and presented my research at many relevant conferences in HCI, such as ACM CHI, ACM DIS, ACM CUI, and MuC. I have published peer-reviewed full articles as a first author and have contributed to other work as a co-author. Furthermore, I co-organised workshops and special interest groups (SIG) at high-ranking conferences in the field:

Scientific Publications.

1. **Mildner, T.**, Savino, G.-L., Schöning, J., and Malaka, R. (2024). Dark Patterns: Manipulative Designstrategien in digitalen Gesundheitsanwendungen. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz. https://doi.org/10.1007/s00103-024-03840-6

2. Gray, C. M., **Mildner, T.** and Bielova, N. (2023). Temporal Analysis of Dark Patterns: A Case Study of a User's Odyssey to Conquer Prime Membership Cancellation through the 'Iliad Flow' (arXiv:2309.09635). arXiv. http://arxiv.org/abs/2309.09635

3. Gray, C. M., Santos, C., Bielova, N. and **Mildner, T.** (2023). An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building (arXiv:2309.09640). arXiv. http://arxiv.org/abs/2309.09640

4. **Mildner, T.**, Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R. and Niess, J. (2024). Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1-18. https://doi.org/10.1145/3613904.3642542

5. Zargham, N., Alexandrovsky, D., **Mildner, T.**, Porzel, R., and Malaka, R. (2023). "Let's Face It": Investigating User Preferences for Virtual Humanoid Home Assistants". International Conference on Human-Agent Interaction, 246–256. https://doi.org/10.1145/3623809.3623821

6. **Mildner, T.**, Freye, M., Savino, G.-L., Doyle, P. R., Cowan, B. R. and Malaka, R. (2023). Defending Against the Dark Arts: Recognising Dark Patterns in Social Media. Proceedings of the 2023 ACM Designing Interactive Systems Conference, 2362–2374. https://doi.org/10.1145/3563657.3595964

7. **Mildner, T.**, Savino, G.-L., Doyle, P. R., Cowan, B. R. and Malaka, R. (2023). About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3544548.3580695

8. Stefanidi, E., Bentvelzen, M., Woźniak, P. W., Kosch, T., Woźniak, M. P., **Mildner, T.**, Schneegass, S., Müller, H. and Niess, J. (2023). Literature Reviews in HCI: A Review of Reviews. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–24. https://doi.org/10.1145/3544548.3581332

9. Wagener, N., Reicherts, L., Zargham, N., Bartłomiejczyk, N., Scott, A. E., Wang, K., Bentvelzen, M., Stefanidi, E., **Mildner, T.**, Rogers, Y. and Niess, J. (2023). SelVReflect: A Guided VR Experience Fostering Reflection on Personal Challenges. Proceedings

of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3544548.3580763

10. **Mildner, T.**, Doyle, P., Savino, G.-L. and Malaka, R. (2022). Rules Of Engagement: Levelling Up To Combat Unethical CUI Design. 4th Conference on Conversational User Interfaces, 1–5. https://doi.org/10.1145/3543829.3544528

11. **Mildner, T.** and Savino, G.-L. (2021). Ethical User Interfaces: Exploring the Effects of Dark Patterns on Facebook. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3411763.3451659

12. Veale, T., Wicke, P. and **Mildner, T.** (2020). Duets Ex Machina: On The Performative Aspects of " Double Acts " in Computational Creativity.

13. Reinschluessel, A. V., Teuber, J., Herrlich, M., Bissel, J., Van Eikeren, M., Ganser, J., Koeller, F., Kollasch, F., **Mildner, T.**, Raimondo, L., Reisig, L., Ruedel, M., Thieme, D., Vahl, T., Zachmann, G. and Malaka, R. (2017). Virtual Reality for User-Centered Design and Evaluation of Touch-free Interaction Techniques for Navigating Medical Images in the Operating Room. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2001–2009. https://doi.org/10.1145/3027063.3053173

Workshops, Symposia, and SIGs.

- Gray, C. M., Gunawan, J., Schäfer, R., Bielova, N., Chamorro, L. S., Seaborn, K., **Mildner, T.**, and Sandhaus, H. (2024). Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices. Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, 1–6.

- Gray, C. M., Santos, C. T., Tong, N., **Mildner, T.**, Rossi, A., Gunawan, J. T., and Sinders, C. (2023). Dark Patterns and the Emerging Threats of Deceptive Design Practices. Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, 1–4. https://doi.org/10.1145/3544549.3583173

- Avanesi, V., Rockstroh, J., **Mildner, T.**, Zargham, N., Reicherts, L., Friehs, M. A., Kontogiorgos, D., Wenig, N. and Malaka, R. (2023). From C-3PO to HAL: Opening The Discourse About The Dark Side of Multi-Modal Social Agents. Proceedings of the 5th International Conference on Conversational User Interfaces, 1–7. https://doi.org/10.1145/3571884.3597441