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

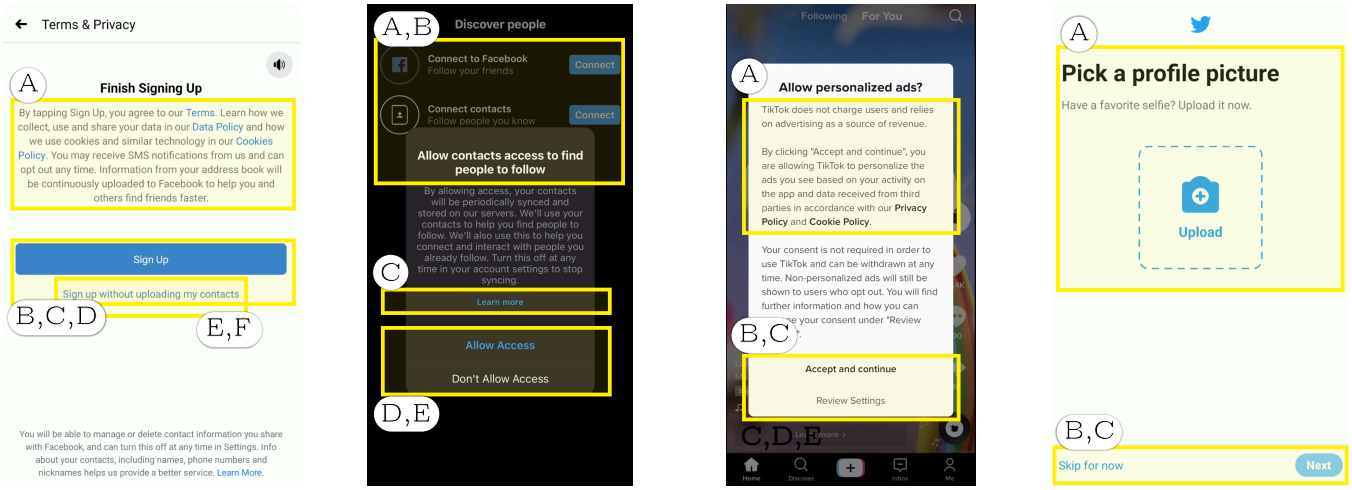
While 44.93% of dark pattern types were recognised during the cognitive walkthrough, the other 55.07% were not. Almost all dark pattern taxonomies contained some dark patterns that were recognised. However, the taxonomy by Zagal et al. [45], being video-game focused, did not contribute any specific dark 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.

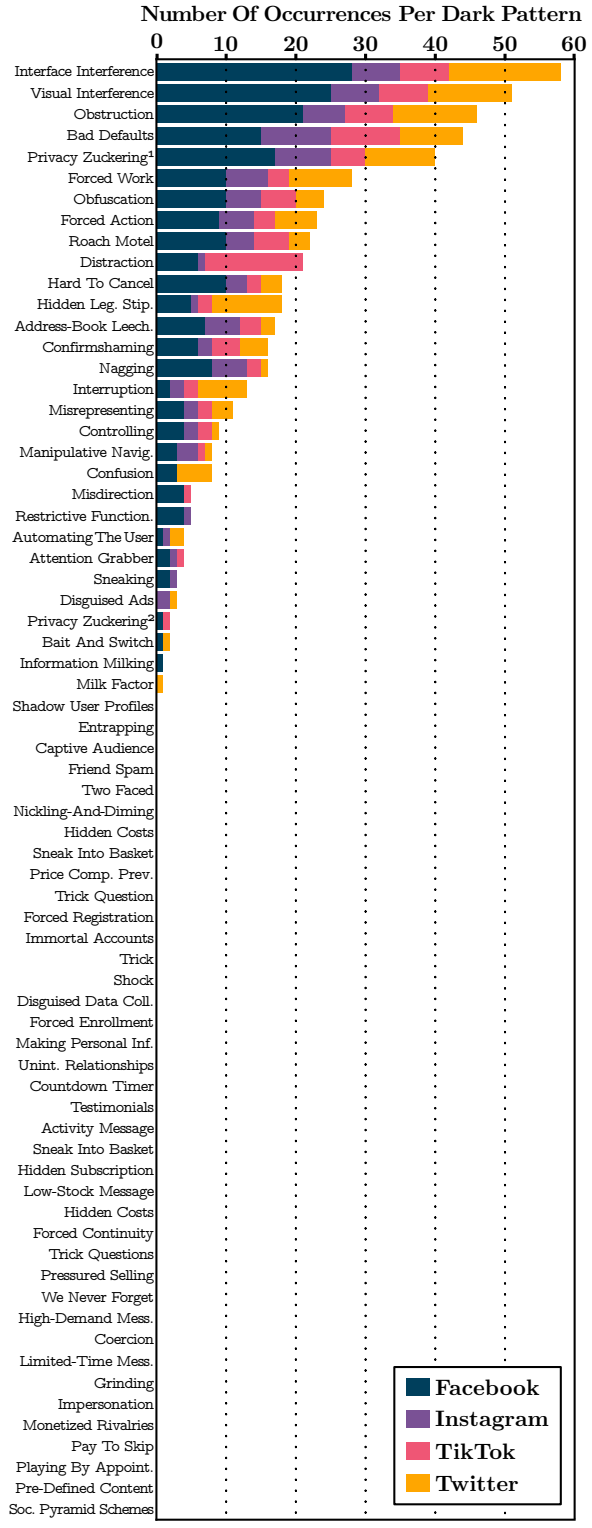


Figure 2: Summary of the occurrences of all 69 considered dark pattern types in four SNSs.Of the 69 types 31 were recognised. *Privacy Zuckering*[1] refers to Brignull's [7] description while *Privacy Zuckering*[2] refers to Bösch et al.'s defintion [5].

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?

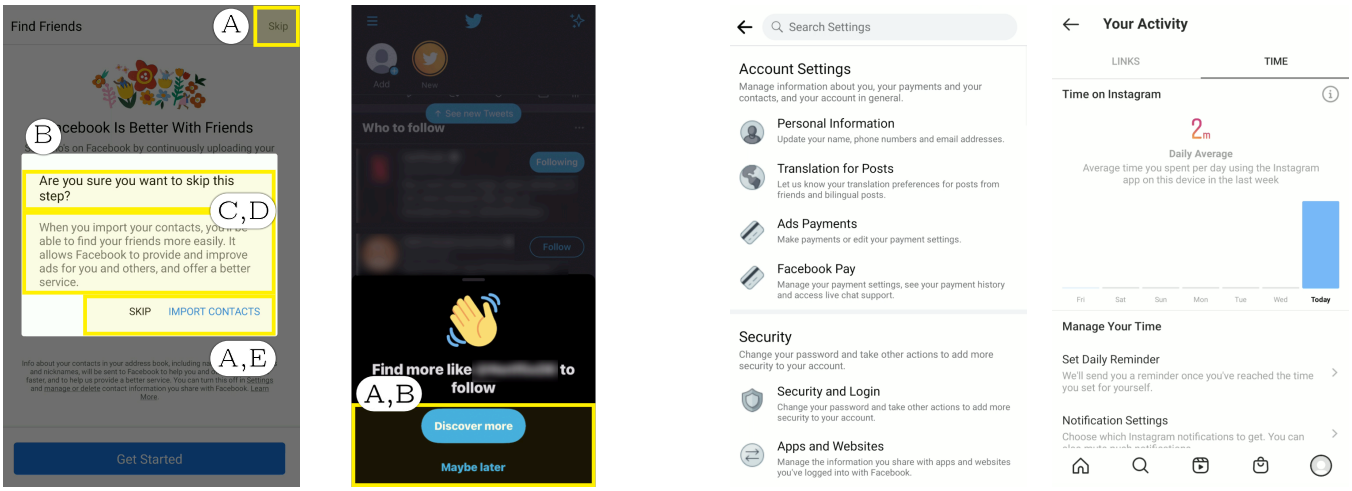
Table 2: This table lists the introductory questions Mathur et al. (2019) [33] gave for each dark pattern characteristic.

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

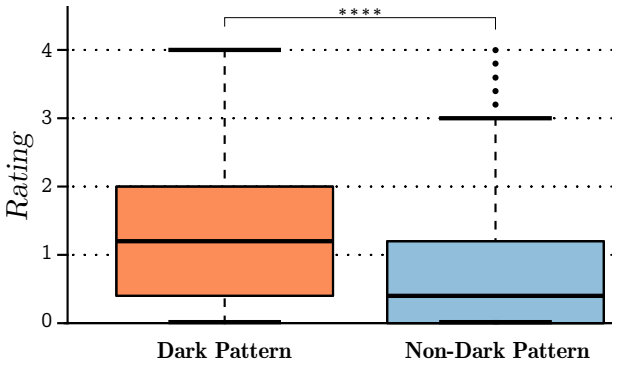


Figure 4: This box plot visualises the differences in which participants, who were provided with a definition for dark patterns, rated the screenshots after being asked if they noticed any malicious designs. The figure shows a significant difference between participants' ratings of screenshots containing 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

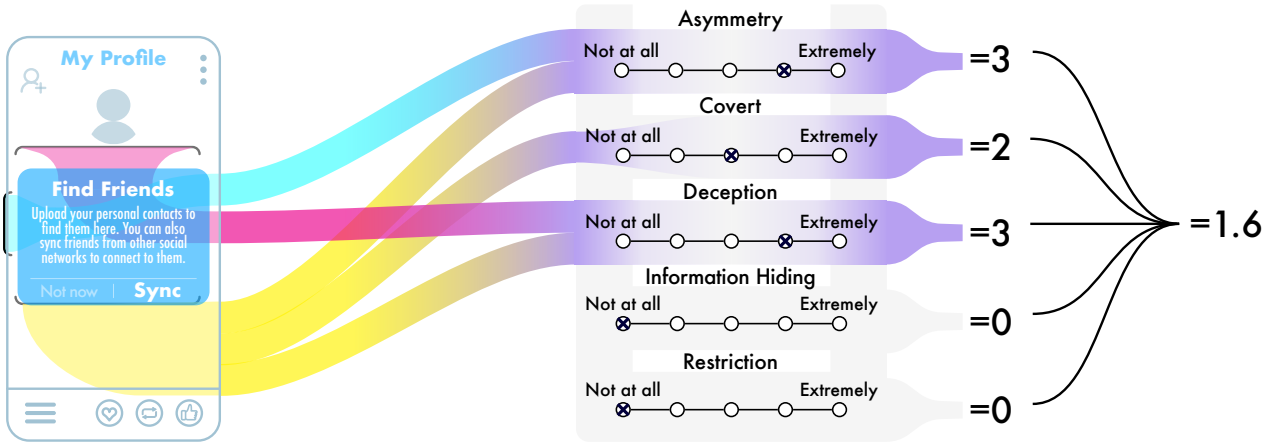


Figure 5: This figure demonstrates the approach to assess malice in interfaces by applying questions based on Mathur et al.'s [33] dark pattern characteristics. First, an interface is selected which is suspected of containing any amount of dark patterns. Using the five questions described in Table 2, 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