2021-12-30

A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically

Ioannis Stavrakakis
Technological University Dublin, ioannis.stavrakakis@tudublin.ie

Andrea Curley
Technological University Dublin, Andrea.F.Curley@TUDublin.ie

Dympna O'Sullivan
Technological University Dublin, dympna.osullivan@tudublin.ie

See next page for additional authors

Recommended Citation

Ioannis Stavrakakis, Andrea Curley, Dympna O'Sullivan, Damian Gordon, Brendan Tierney. A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically. International Journal On Advances in Internet Technology, (2021), Vol 14: 1& 2, DOI: 10.21427/ 20G8-D176.

A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically

Ioannis Stavrakakis, Andrea Curley, Dympna O'Sullivan, Damian Gordon, Brendan Tierney

ASCNet Research Group, School of Computer Science, Technological University Dublin, Dublin, Ireland

Email: Ioannis.Stavrakakis@TUDublin.ie, Andrea.F.Curley@TUDublin.ie, Dympna.OSullivan@TUDublin.ie, Damian.X.Gordon@TUDublin.ie, Brendan.Tierney@TUDublin.ie

Abstract— **This research explores the design and development of a framework for the detection of Dark Patterns, which are a series of user interface tricks that manipulate users into actions that they do not intend to do, for example, share more data than they want to, or spend more money than they plan to. The interface does this using either deception or other psychological nudges. User Interface experts have categorized a number of these tricks that are commonly used and have called them Dark Patterns. They are typically varied in their form and what they do, and the goal of this research is to explore existing research into these patterns, and to design and develop a framework for automated detection of potential instances of web-based dark patterns. To achieve this, we explore each of the many canonical dark patterns and identify whether or not it is technically possible to automatically detect that particular pattern. Some patterns are easier to detect than others, and there are others that are impossible to detect in an automated fashion. For example, some patterns are straightforward and use confusing terminology to flummox the users, e.g. "Click here if you do not wish to opt out of our mailing list", and these are reasonably simple to detect, whereas others, for example, sites that prevent users from doing a price comparison with similar products might not be readily detectable. This paper presents a framework to automatically detect dark patterns. We present and analyze known dark patterns in terms of whether they can be either: (1) detected in an automated way (it can be partially or fully), (2) detected in a manual way (it can be partially or fully) and (3) cannot be detected at all. We present the results of our analysis and outline a proposed software tool to detect dark patterns on websites, social media platforms and mobile applications.**

Keywords-Dark Patterns; User Experience; Digital Ethics; Privacy.

I. INTRODUCTION

Computers and technological applications are now central to many aspects of life and society, from industry and commerce, government, research, education, medicine, communication, and entertainment systems. Computer scientists and professionals from related disciplines who design and develop computer applications have a significant responsibility, as the systems they develop can have wide ranging impacts on society where those impacts can be beneficial but may also at times be negative, thus it cannot be argued that modern technology is value-neutral, as it is clear that it can have both planned and unplanned negative consequences on users.

In this, and previous research [1], we outline and explore the ethical limits of a technology design phenomenon known as "dark patterns". Dark patterns are user interfaces that benefit an online service by leading users into making decisions they might not otherwise make. At best, dark patterns annoy and frustrate users. At worst, they can mislead and deceive users, e.g., by causing financial loss, tricking users into giving up vast amounts of personal data or inducing compulsive and addictive behavior in adults and children. They are an increasingly common occurrence on digital platforms including social media sites, shopping websites, mobile apps, and video games. Although they are gaining more mainstream awareness in the research community, dark patterns are the result of three decades-long trends: one from the world of retail (deceptive practices), one from research and public policy (nudging), and the third from the design community (growth hacking) [2].

The aim of our work is the development of a framework for classifying web-based dark patterns as to which are readily detectable, and which are not. The framework forms the basis of a software tool that can automatically alert users to the presence of dark patterns on websites, social media platforms and mobile applications. In developing the framework we analysed common documented types of data patterns. We present these dark patterns to the reader and classify each dark pattern using the following taxonomy: (1) A pattern that can be detected in an automated way (either partially or fully); (2) A pattern that can be detected in a manual way (either partially or fully); and (3) A pattern that cannot be detected. In this paper we outline the features and functionality of the proposed tool. This research is part of a larger research project (called Ethics4EU) whose goal is develop a repository of teaching and assessment resources to support the teaching of ethics in computer science courses, supported by the Erasmus+ programme [3].

In Section 2, a review of some of the key literature focusing on what dark patterns are, and why they are so successful. Section 3 looks at the specific collection of dark patterns that will be explored in this research. Section 4 presents the initial framework for the detect of dark patterns, looking at which patterns can be detected automatically, which manually, and which cannot be detected at all. Section 5 outlines some other dark patterns that should also be looked at, and finally, Section 6 presents some conclusions and future work about this research.

II. LITERATURE REVIEW

Since the early 1980s computer programmers have used the concept of patterns in software engineering as a useful way of categorizing different types of computer programs. The term dark patterns has been used since 2010 to refer to interface design solutions that intend to deceive users into carrying out undesirable actions [4]. Gray et al. [5] defined dark patterns as "instances where designers use their knowledge of human behavior (e.g., psychology) and the desires of end users to implement deceptive functionality that is not in the user's best interest".

There has been significant research done on dark patterns from the fields of Cognitive Psychology, Usability, Marketing, Behavioural Economics, Design and Digital Media. All this research has led to the abandonment of the rational choice theories for explaining decision making, particularly for matters of privacy [6] and has prompted new examinations that attribute the effectiveness of dark patterns on human cognitive limitations. However, there is still not a universal theoretical explanation of the 'whys' and 'hows' of the effectiveness of dark patterns. For example, Maier [7] argues that manipulation is closely linked to decision making and the latter can be easily influenced through one's emotions and mood leading to decisions lacking rational thought [8].

What is more, according to Kahneman [9] humans are more intuitive than rational thinkers and most of their daily reasoning is performed by their intuition. Below are the main human psychological mechanisms being targeted or exploited by Dark Patterns [10]:

- Nudging, which is based on soft paternalism, positive reinforcement and compliance [11]. Nudging can be and has been used with good intentions in mind and has been proved effective [12][13]. However, because of its proven efficiency, nudging is one of the most common digital manipulation strategies used to mislead users into bad decisions privacy-wise.
- Persuasion techniques built on what Cialdini [14] identifies as the "six basic tendencies of human behaviour" (p. 76). These tendencies namely are: reciprocation, consistency, social validation, liking, authority and scarcity.
- Cognitive biases that fundamentally are information processing limitations of the human mind and are rooted in cognitive heuristic systems [9]. According to Waldman [15] the five most pervasive are: anchoring [16], framing [17], hyperbolic discounting [18][19][20], overchoice [21][22][23] and metacognitive processes such as cognitive scarcity [24] and cognitive absorption [25].
- Cognitive dissonance, an uncomfortable state of mind where one's beliefs and actions are contradictory. Bösch et al. [10] (p. 247) mention "[i]n terms of privacy dark patterns, this process can be exploited by inconspicuously providing justification arguments for sugar-coating user decisions that have negatively affected their privacy".

Although, so far, it appears that the cognitive and psychological factors play a significantly important role on users' failure to protect their privacy when dealing with Dark Patterns, some researchers argue that contextual and social factors are important too. For example, Acquisti et al. [6] claim that incomplete or asymmetric access to information between two agents in a transaction can significantly disadvantage one party leading to problematic decisions. Furthermore, users are not always certain of what they are agreeing to share as the collection of personal data is not always apparent and therefore people remain unaware of what information is collected about them by both private and public organisations [26]. This is usually the norm in digital environments where the user has no control over the design and information processing they are being shown.

On the other hand, research has shown that users, care about their privacy [27], however, the contextual, social and cognitive aspects mentioned earlier lead users to a set of behaviours that are inconsistent to their attitudes towards privacy [15]. Norberg et al. [28] have called this the 'privacy paradox'.

In today's digital environment most digital platforms' provide services seemingly for free. In order for these services to generate revenue they have become dependent on accumulating and processing users' data, oftentimes personal data [29]. According to Zuboff [30] user data is the raw material that produces, what she calls, 'behavioural surplus' which has become a valuable commodity for companies. Behavioural surplus is a powerful tool for predicting user behaviour and many companies use it to influence users into providing more data which leads into a vicious cycle of user data, influence, prediction and so on [31].

Mathur et al. [32] did a meta-analysis of 11,286 shopping websites, and created a taxonomy to try to explain how dark patterns affects user decision-making by exploiting cognitive biases. Their taxonomy has the following characteristics: Asymmetric, Covert, Deceptive, Hides Information, and Restrictive. They found that 11.1% (1254 websites) of the sites had dark patterns, and recommend the development of plug-ins for browsers to help detect these patterns.

Nouwens et al. [33] discuss the growth of Consent Management Platforms (CMPs) which are software systems that manage the interaction between users and the website(s) of an organization, recording (and updating) their privacy preferences, and getting consent for recording interactions with cookies. Crucially these CMPs are compliant with GDPR (the General Data Protection Regulation) however it is still possible for a website to employ Dark Patterns to circumvent GDPR, and almost 90% of the sites with CMPs surveyed were in some way themselves breaching GDPR.

Chromik et al. [34] explore how there is potential for dark patterns to be used in Intelligent Systems. An intelligent system is computer system with an embedded artificial intelligence that can work to solve well-defined tasks, e.g. object recognition, medical diagnosis, language translation. As a consequence of GDPR, these systems must be able to provide some explanation as to how they came to

specific decisions. Some intelligent systems incorporate explanation facilities to support users in understanding decisions. However, this paper discusses the possibility of Intelligent Systems using Dark Patterns in conveying these explanations to get further data from the users. For example, the system could use a Dark Pattern to collect valuable user data under the pretext of explanation. So, the user might be forced to provide additional personal information (e.g., social connections) before receiving personalized explanations. Otherwise, the user would be left off with a generic high-level explanation.

Di Geronimo et al. [35] explore the use of Dark Patterns in mobile apps. They looked at 240 popular mobile apps and explored whether or not these apps included any dark patterns. Their analysis showed that 95% of the apps they reviewed included one or more Dark Patterns, with an average of 7.4 malicious designs per app, with a standard deviation of 5. Almost 10% of the apps included 0, 1, or 2 Dark Patterns (N=33), 37% of the apps contained between 3 to 6 Dark Patterns (N=89), while the remaining 49% included 7 or more (N=118). They also conducted an online experiment with 589 users on how they perceive Dark Patterns in such apps. Overall, the majority of our users did not spot malicious designs in the app containing Dark Patterns (55%), some were unsure (20%), and the remaining found a malicious design in the app (25%). But they found that most users did perform better in recognizing malicious designs if informed on the issue.

Grassl et al. [36] looked at cookie consent requests in the context of Dark Patterns to explore whether or not they undermine principles of EU privacy law. They undertook two online experiments where they investigated the effects of common design nudges on users' consent decisions and their perception of control over their personal data in these situations. In the first experiment (n = 228) they explored the effects of dark patterns to encourage the participants to select the privacy-unfriendly option, and the experiment revealed that most people agreed to all consent requests regardless of dark patterns. The research indicated that the dark patterns made no difference to the participants' behaviour. The first experiment, also showed that despite generally low levels of perceived control, obstructing the privacy-friendly option led to more rather than less perceived control for the participants. In the second experiment (n = 255) the participants we presented with patterns to select the privacy-friendly option (bright patterns). The bright pattern did succeed in swaying people effectively towards the privacy-friendly option. The second experiment also looked at the perceived control of the participants, and it found that it stayed the same compared to Experiment 1. Overall, the researchers concluded about Experiment 1 that whether the participants were presented with a dark pattern or not, they have been conditioned by years of practice to consent, and therefore they concluded that the EU's consent requirement for tracking cookies does not work as intended.

Dark patterns are only just beginning to emerge as a topic in the software development literature. In 2021 Kollnig et al. [37] reported in the development of a functional prototype that allows users to disable dark patterns in apps selectively. This differs from our approach where we are developing a comprehensive framework for identifying dark patterns across a range of platforms, from apps to websites.

Chugh and Jain [38] looked at dark patterns from the perspective of consumer protection as well as their impact on democratic political processes. The researchers distinguish between dark patterns and persuasive advertisements, classifying dark patterns as being manipulative, whereas persuasive advertisements merely attempt to influence people to revise their preferences. They see two major issues with dark patterns, (1) users are typically unaware that they are interacting with dark patterns, and are, therefore, unable to safeguard themselves against the effects of these patterns, and (2) market forces and market competition don't seem to be penalizing organizations for using these patterns. Therefore, they recommend that legislation and regulations are necessary to combat these patterns.

Bongard-Blanchy et al. [39] explored the impact of dark patterns on end-users by surveying 406 individuals. They found that although the participants were aware of the type of manipulative techniques that online services use to impact their online behaviour, they are nonetheless unable to combat their impact. The researchers advocate a multi-faceted approach to addressing these issues, including raising awareness and educating people about the different patterns and how they work, concomitant with this approach, the researchers propose that the users are presented positive information that will encourage them to avoid engaging with new patterns and to cease engaging with existing patterns, e.g. the user could be made aware of how much time they spend engaging with infinite scrolling systems, and they could be reminded that they could be using that time for more enjoyable activities. They also advocate targeting the educational initiatives about patterns based on age-groups and other demographics, and finally they suggest that a combination of strong legal penalties and regulations are needed, as well as new software tools to help detect and highlight the existence of these patterns. However, they do note that some pattens may be more readily detectable in an automated fashion than others.

III. PATTERN DESCRIPTIONS

A vital step in developing the web-based Dark Patterns Framework is to clearly define each pattern and to categorize the patterns into themes. In the research literature previously discussed there is some variance as to the exact meaning of each pattern, therefore below we present definitions that attempt to be as inclusive as possible to the range of definitions for each pattern, but always prioritizing the original canonical definitions developed by the pioneer of dark patterns - user experience designer Harry Brignull [4].

A. Sneaking

- **Sneak into Basket:** When purchasing a product, an additional item is added into the basket, usually the new product is added in because of an obscured opt-out button or checkbox on a previous page. Detection of this pattern is challenging since there may be legitimate reasons for a site to add new items into a shopping basket (e.g. taxes), therefore, automated detection may not be possible, but nonetheless it would still be possible to manually highlight changes in cost, and let the shopper decide if the additional items are valid.

- **Hidden Costs:** When reaching the last step of the checkout process, some unexpected charges have appeared in the basket, e.g. delivery charges, etc. Detection of this pattern is challenging since there may be legitimate reasons for a site to add new items into a shopping basket (e.g. taxes), therefore, automated detection may not be possible, but nonetheless it would still be possible to manually highlight changes in cost, and let the shopper decide if the additional items are valid.

B. Misdiretion

- **Trick Questions:** Often found when registering for a new service. Typically, a series of checkboxes are shown, and the meaning of checkboxes is alternated so that ticking the first one means "opt out" and the second means "opt in". Detection of this pattern is possible at least partially because it is possible to detect pre-ticked checkboxes, and to search for phrases like "opt out" and "opt in".

- **Misdirection**: When the design purposefully focuses users' attention on one thing in order to distract their attention from another, for example, a website may have already undertaken a function and added a cost to it, and the opt out button is small. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.

- **Confirmshaming**: This involves guilting the user into opting into something. The option to decline is worded in such a way as to shame the user into compliance, for example, "No thanks, I don't want to have unlimited free deliveries". Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.

- **Disguised Ads**: Advertisements that are disguised as other kinds of content or navigation, in order to get you to click on them, for example, advertisements that look like a "download" button or a "Next >" button. Detection of this pattern is possible at least partially because it is possible to detect buttons on a webpage. And by using either the ALT tags or OCR to determine the purpose of the button, and then to look at whether it links internally, or to an external site.

C. Obstruction

- **Roach Motel:** When users find it easy to subscribe to a service (for example, a premium service), and find it is hard to get out of it, like trying to cancel a shopping account. Detection of this pattern is possible because it is possible to search for "activate" or "subscribe" links or buttons, that have no reciprocal "deactivate" or "unsubscribe" links or buttons.

D. Forced Action

- **Forced Continuity:** When a user gets a free trial with a service comes to an end and their credit card silently starts getting charged without any warning, and there isn't an easy way to cancel the automatic renewal. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.

E. Variegations

- **Privacy Zuckering:** Tricking users into sharing more information than they intended to, for example, Facebook privacy settings were historically difficult to control. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.

- **Price Comparison Prevention:** The retailer makes it hard for you to compare the price of an item with another item, so you cannot make an informed decision. Retailers typically achieve this by creating different bundles where it is not easy to work out the unit price of the items within the bundles. Detection of this pattern is challenging since it may not be obvious (or clearly labelled) if the products are in different bundles, but it will be possible to manually highlight packaging types, and let the shopper decide if there are any issues.

- **Bait and Switch:** The user sets out to do one thing, but a different, undesirable thing happens instead, for example, Microsoft's strategy to get users to upgrade their computers to Windows 10. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.

- **Friend Spam:** The product asks for users for their email or social media permissions to spam all their contacts. Detection of this pattern is possible since the HTML in the website can be analyzed to determine if the site asked for email or social media permissions.

F. Beyond Brignull

UX researcher Reed Steiner [40] added six patterns:

- **Fake Activity:** On a commercial website, when the page says "three other people are viewing this item right now" this may not be a fully truthful claim. Detection of this pattern is possible at least partially because it is possible to search for phrases such as "other people are viewing this item now" and warn the shopper of this pattern.

- **Fake Reviews:** Research shows that several reviews and testimonials are fake, and exact matches with different customer names can be found on several sites. Detection of this pattern is challenging, but it may be possible to take reviews from the current site, and manually search for them on other similar sites.

- **Fake Countdown:** Some online purchases include countdown timers, in most cases countdown timers only add urgency to a sale. Detection of this pattern is possible at least partially because it is possible to search for phrases such as "offer ends in" or "countdown" and warn the shopper of this pattern.

- **Ambiguous Deadlines:** Some online purchases indicate that a product is only on sale for a limited amount of time, but don't mention a specific deadline. Detection of this pattern is possible at least partially because it is possible to search for phrases such as "for a limited amount of time" and warn the shopper.

- **Low Stock Messages**: Sometimes sites claim that they are low on a particular item. Detection of this pattern is possible at least partially because it is possible to search for phrases such as "only" and "units left" and warn the shopper of this pattern.

- **Deceptive High Demand:** This is similar to the low stock messages. Detection of this pattern is possible at least partially because it is possible to search for phrases such as "in demand" and "in high demand" and warn the shopper of this pattern.

IV. Developing The Framework

With these definitions established, it becomes possible to categorize the patterns into one of three classifications:

(1) A suspected pattern that can be detected in an automated way (partially or fully) based on the text, images or HTML in a webpage or website.
(2) A suspected pattern that can be detected in a manual way (partially or fully) based on the text, images or HTML in a webpage or website.
(3) A suspected pattern that cannot be detected, based on the fact that there is so much variation in either how the pattern is defined or in how the pattern is implemented.

As all of the researchers involved in this project are teaching on an MSc in Data Science, they have knowledge of a wide range of detection techniques, therefore, a Morphological Matrix approach [41] was undertaken, whereby a table was created listing all of the pattern types on the Y-axis, and listing a range of detection techniques on the X-axis (HTML Parsing, Computational Linguistics, Image Processing, Machine Learning, Data Mining, Compiler Design, Regular Expressions) and a series of three online brainstorming sessions were held to identify which patterns might be detectable using which techniques (if any). To help reach a shared understanding of the patterns, not only were definitions of each pattern shared and discussed, but also images from over 100 websites with dark patterns from the Mathur et al. [32] dataset were presented and discussed. Of all patterns discussed, there was general consesus as to which aspects of patterns could be detected, and to what extent that detection was possible. The full framework is presented below in Table 1 where each pattern presented in Section III is classified as to how it can be detected, as well as some detail as to how such a pattern can be detected (if it can) as shown in the *Rationale* column.

Patterns that can be detected automatically will typically have terms in them such as "opt-in", "activate", or "subscribe". These, and other indicators such as the placement or configuration of images, or in the formulation of the HTML tags, allow for the automated detection of dark patterns. In contrast, there are some web-based activities or transactions that cannot, in and of themselves, be automatically detected, but are sufficiently indicative to suggest the presence of a dark pattern. In these cases the framework proposes the development of an ancillary (or appurtenant) window to highlight to the users that there may be something suspicious occurring in the transaction that they are undertaking. Finally, it is worth noting that, there are some patterns that cannot readily be detected, but may be reported using the reporting feature of the system.

The patterns beyond Brignull canon is the only one where it may be possible to do some form of automated detection on all of the patterns (Fake Activity, Fake Reviews, Fake Countdown, Ambiguous Deadlines, Low Stock Messages, Deceptive High Demand). This may be because these patterns focus almost exclusively on text-based enticements to encourage users to purchase content, and because they use text, it is possible to do searches for specific phrases, for example, "offer ends in", "for a limited amount of time" or "in high demand". The one pattern that is slightly different from the others is the Fake Reviews, where instead of searching for a particular phrase on the webpage, we use the entire review to search for that exact same review (or a similar review) on other sites.

TABLE I. DARK PATTERNS DETECTION FRAMEWORK

Category	*Pattern*	*Detection*	*Rationale*
Sneaking	Sneak into Basket	Manual (fully)	Highlight changes in cost
	Hidden Costs	Manual (fully)	Highlight changes in cost
Misdirection	Trick Questions	Automated (partially)	Look for phrases like "opt-in" and "opt-out", as well as pre-ticked checkboxes
	Misdirection	Cannot be detected	There is too much variation in how this pattern is implemented.
	Confirmshaming	Cannot be detected	There is too much variation in how this pattern is implemented.
	Disguised Ads	Automated (partially)	Look for buttons (noting colour and size) and see which ones link to external sites.
Obstruction	Roach Motel	Automated (fully)	Look for sites with "activate" or "subscribe" links or buttons but with no "deactivate" or "unsubscribe"
Forced Action	Forced Continuity	Cannot be detected	There is too much variation in how this pattern is implemented.
Variegations	Privacy Zuckering	Cannot be detected	There is too much variation in how this pattern is implemented.
	Price Comparison Prevention	Manual (fully)	Highlight if products are displayed with different units of the product
	Bait and Switch	Cannot be detected	There is too much variation in how this pattern is implemented.
	Friend Spam	Automated (partially)	Check if the site asks for email or social media permissions, and notify users.
Beyond Brignull	Fake Activity	Automated (partially)	Look for phrases like "other people are viewing this item now".
	Fake Reviews	Manual (partial)	Select the review and search for it on other sites.
	Fake Countdown	Automated (partially)	Look for phrases like "offer ends in" or "countdown"
	Ambiguous Deadlines	Automated (partially)	Look for phrases like "for a limited amount of time"
	Low Stock Messages	Automated (partially)	Look for phrases like "only" and "units left"
	Deceptive High Demand	Automated (partially)	Look for phrases like "in demand" and "in high demand"

Some patterns will have words or images that make them easy to identify ("opt in", "offer ends soon", "in demand", etc.) and therefore we can say that they are automatically detectable (either partially or fully). And, in contrast, some patterns are implemented in such a range of different ways depending on the particular interface (and the definitions of some patterns vary in different research literature), that they are impossible to consistently detect, so we classify these as "Cannot be detected". Other patterns require human judgement, such as determining if using pre-ticked checkboxes is being deceptive, or if the site is asking for security permissions, and so we classify these as being detectable manually (either partially or fully). To help recognise the patterns that can potentially be manually detected, the proposed system will allow the user to display an ancillary window that will help highlight some potential issues of concern on a given webpage or website. The new window can display things like:

- The percentage of the webpage that is visible in the browser window, to ensure the user is aware that there may be instructions or options that are not visible on the current page, but are elsewhere on the page.
- The total number of checkboxes on the page, and the number that are pre-ticked.
- The total number of radio buttons on the page, and the number that are pre-ticked.
- The shopping basket total, that will be zero if there are no items.
- A "fake review detection" tool that allows a user to select the text of a review, and to automatically search for that text elsewhere on the web.
- Highlight the number of links on the page, noting which are from text and which from images (to help detect potential Disguised Ads).

- Highlight which tick boxes or radio buttons are concerned with privacy issues, looking for words such as “privacy” or “GDPR”.
- Indicate if the current webpage or website has already been reported as having a dark pattern.

Further, to help users locate suspected dark patterns on a webpage, the system will provide two modes of operation:

(1) where the system highlights all of the areas on that webpage to show suspected patterns on the page with suitable pointers, and
(2) if the user clicks on a particular type of issue on the auxiliary window, only those areas on the page will be highlighted, for example, if the user selects the “Radio Buttons” section of the panel, then all of the radio buttons on the webpage will be highlighted with pointers.

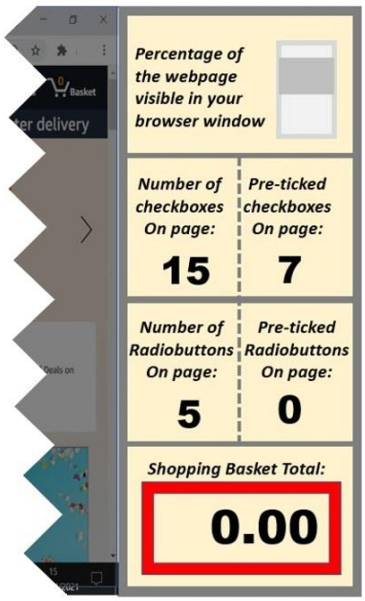

Figure 1. Appurtenant Window with Page Details

Two additional elements of the proposed system are the Reporting and Educational features:

- The *Reporting Feature* is designed to compensate for the fact that some patterns are difficult (or impossible) to detect, and it will allow users to record and report websites and webpages that they suspect have dark patterns. For example, if a user feels that they have been a victim of Forced Continuity, they can report the webpage or website, and indicate which pattern they feel is present.
- The *Educational Feature* which is designed to educate the users on each of the main dark patterns, as well as the variation among different researchers. This feature will help the users appreciate why they are being warned about a particular feature on a website as well as giving them sufficient information to allow them to accurately categorize patterns that they encounter if they wish to report them. It is envisioned that a central part of this feature will consist of a series of videoed micro-lessons.

V. Implementation And Limitations

The goal of this research is to define a collection of dark patterns, and to explore whether or not it is possible to develop a framework to detect these dark patterns - in an automated way, a manual way, or not at all. The detection process not only categorizes whether each pattern is detectable, but it also describes to what extend it is detectable, and suggests some ways it might be detected. The development process of framework was as a result of the brainstorming sessions, and these crucially categorized the patterns into three groupings:

1. Automated Detection ("Disguised Ads", "Friend Spam", "Roach Motel" and "Trick Questions")
2. Manual Detection ("Hidden Costs", "Price Comparison Prevention", "Sneak into Basket")
3. Cannot be Detected ("Bait and Switch", "Confirmshaming", "Forced Continuity", "Misdirection", "Privacy Zuckering")

To help confirm the analysis process, an initial prototype system has been developed using the Python programming language which provides ample software libraries for web crawling and web scraping, specifically the HTMLparser and URLopen libraries were used in this case. The system was developed as a plug-in for the Google Chrome browser and was able to detect four patterns were selected to be implemented, “Trick Questions”, “Roach Motel”, “Friend Spam”, and “Low Stock Messages” were chosen as they are the most straightforward to implement, since that have been classified as “Automated (partial)” and “Automated (fully)” in the above table. These four were implemented, and were tested using over 60 of the dark patterns from the Mathur et al. [28] dataset, and the prototype was able to successfully detect all three of these patterns, each with significant variation. Three key takeaways from the prototype development process were as follows:

1. When testing the prototype system with some users it became evident that the terminology itself was proving to be a barrier to understanding the purpose of the system. Although the participants had experienced the phenomena of being pressured into purchasing goods online, the term “Dark Patterns” was unfamiliar to them, and two of the names of the patterns: “Roach Motel” and “Friend Spam” were equally opaque to the users, proving to be moreso confusing that enlightening. Future development will change some of the terms to more descriptive one, including changing “Dark Patterns Detector” to “Online Shopping Tricks

Detector", changing "Roach Motel" to "Hard to Unsubscribe", and changing "Friend Spam" to "May use your addressbook".

2. A rudimentary Optical Character Recognition (OCR) system was developed to read text off the images on webpages to determine if they have messages that could be considered to be Dark Patterns, for example, text saying "Only a Limited Amount of Stock Left". The implementation proved to be highly effective in terms of reading text from the images, but slowed down the overall detection process significantly, and particularly for websites that had a lot of images on them, it delayed the detection process from being almost instantaneous into taking almost 10 minutes to complete the process.
3. Perhaps one of the most interesting outcomes of the prototyping process was that it allowed the researchers to interrogate their fundamental understanding of the notion of a Dark Pattern. Most websites include some forms advertising, which are not the same as dark patterns, for example, some of the test sites included phrases such as "Customers who bought this product also bought …" which were classified as Dark Patterns by the system, as they are similar to a "Fake Activity" which might say something like "Other Customers are looking at this product". After much discussion it became clear that this is just advertising, and in particular, it is persuasive advertising, which is similar to Dark Patterns, but they differ in that they do not rely on pressuring or confusing the customers.

In terms of the limitations of this research, perhaps the most serious one is the fact that five of the patterns ("Misdirection", "Confirmshaming", "Forced Continuity", "Privacy Zuckering", and "Bait and Switch") have been classified as "Cannot be detected". If these cannot be detected, it significantly limits the efficacy of the final tool, therefore a thorough exploration of the Mathur et al. [32] dataset is planned to determine if there are any implicit characteristics associated with these five patterns that can be used to detect them (either automatically or manually), as well as a number of further brainstorming sessions.

It is also worth noting that that the full implementation of this framework will result in some additional challenges, for example, some sites have a special file called Robots.txt that prohibits the use of web scraping, and it is also the case that some sites use technologies that make them more difficult to parse, for example, frames or webpages implemented in Javascript or CSS.

Finally, another consideration is that many shoppers use mobile applications instead of websites to purchase products and services, and the techniques outlined so far would be ineffective on these applications.

VI. Conclusions And Futre Work

This paper presented a framework for the detection of web-based dark patterns and an accompanying proposed software tool. It begins with a review of some of the key literature in this field, which highlights some of the reasons for the success of dark patterns, as well as their ubiquity. It follows this with an explanation of some of the key dark patterns, and a categorization of the patterns as being in one of the following three classifications:

1. A suspected pattern that can be detected in an automated way (partially or fully), in other words there is some characteristic either in the text, images or HTML of a webpage or website that indicates that it is a dark pattern.
2. A suspected pattern that can be detected in a manual way (partially or fully), in other words there is some characteristic either in the text, images or HTML of a webpage or website that indicates that there is potential for dark pattern on this page or site, but because it cannot be detected definitively, the potential pattern is highlighted to the user.
3. A suspected pattern that cannot be detected, in other words there is so much variation in either how the pattern is defined or in how the pattern is implemented, there is no direct way of detecting it just using web crawling and web scraping techniques.

This classification, in turn, leads to the design of a proposed software tool with the ability to detect patterns from category 1, and to highlight potential instances of patterns from category 2. For those patterns in category 3, even if there is no obvious way to identify them, nonetheless, it is important to deal with them in some way, therefore additional features are required for the system, a *Reporting feature* to address instances of patterns for category 3, as well as an *Educational feature* to create awareness about dark patterns in general.

Future work will focus on full implementation of the software tool and the inclusion of the Reporting and Education features. The Reporting features of the system are envisioned to work either in *stand-alone mode*, or *shared mode*. In stand-alone mode the reporting process is recorded locally on the user's own computer as a series of XML files, whereas in shared mode, the user can share their suspicions about potential dark patterns with other users also using the system, and they can also label and add a description to the suspected pattern.

The Educational features will consist of a series of micro-lessons describing the range of dark patterns. Also, a series of pop-up windows will be developed with simple explanations (and links to examples) of a specific pattern will be developed, to remind the users about the key characteristics of each specific pattern.

Finally, the framework provides a way forward to deal with dark patterns in a comprehensive and comprehensible manner. This has become more and more important as the

number of services that have become available online continues to grow, and in many cases these services are available only exclusively online. It, therefore, becomes a matter of necessity that as many people as possible are aware of these deceitful patterns, and incumbent on IT practitioners to spread the word about these patterns.

ACKNOWLEDGMENT

The authors of this paper and the participants of the Ethics4EU project gratefully acknowledge the support of the Erasmus+ programme of the European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] A. Curley, D. O'Sullivan, D. Gordon, B. Tierney, I. Stavrakakis, "The Design of a Framework for the Detection of Web-Based Dark Patterns". ICDS 2021: The 15th International Conference on Digital Society, Nice, France, 18th - 22nd, July 2021.

[2] A. Narayanan, A. Mathur, M. Chetty, M. Kshirsagar, "Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces". Queue, 18(2), pp. 67-92, 2020.

[3] D. O'Sullivan, D. Gordon, "Check Your Tech – Considering the Provenance of Data Used to Build Digital Products and Services: Case Studies and an Ethical CheckSheet", IFIP WG 9.4 European Conference on the Social Implications of Computers in Developing Countries, 10th–11th June 2020, Salford, UK, 2020.

[4] H. Brignull, "Dark patterns: Deception vs. honesty in UI design". Interaction Design, Usability, 338, 2011.

[5] C. M. Gray, Y. Kou, Y., B. Battles, J. Hoggatt, A. L. Toombs,"The dark (patterns) side of UX design". In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-14, 2018.

[6] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, S. Wilson "Nudges for privacy and security: Understanding and assisting users' choices online", ACM Computing Surveys (CSUR), 50(3), pp. 1-41, 2017.

[7] M. Maier, "Dark patterns: An end user perspective". Master's thesis. Umeå University, 2019.

[8] R. Mehta, R. J. Zhu, "Blue or red? Exploring the effect of color on cognitive task performances", Science (New York, N.Y.), 323(5918), pp. 1226–1229, 2009.

[9] D. Kahneman, "Thinking, Fast and Slow", Penguin Books, 2011.

[10] C. Bösch, B. Erb, F. Kargl, H. Kopp, S. Pfattheicher, "Tales from the dark side: Privacy dark strategies and privacy dark patterns". Proceedings on Privacy Enhancing Technologies, 2016(4), pp. 237-254, 2016.

[11] A. Acquisti, "Nudging privacy: The behavioral economics of personal information". IEEE Security & Privacy, 7(6), pp. 82-85. 2009.

[12] H. Almuhimedi, F. Schaub, N. Sadeh, N.I. Adjerid, A. Acquisti, J. Gluck, Y. Agarwal, "Your location has been shared 5,398 times! A field study on mobile app privacy nudging". In Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 787-796, 2015.

[13] E. Peer, S. Egelman, M. Harbach, N. Malkin, A. Mathur, A. Frik, "Nudge me right: Personalizing online security nudges to people's decision-making styles". Computers in Human Behavior, 109, 106347, 2020.

[14] R. Cialdini, "Influence. The Psychology of Persuasion". New York, NY: William Morrowe Company, 1984.

[15] A. E. Waldman, "Cognitive biases, dark patterns, and the 'privacy paradox". Current opinion in psychology, 31, pp. 105-109, 2020.

[16] D. Ariely, G. Loewenstein, D. Prelec, "Coherent arbitrariness: Stable demand curves without stable preferences". The Quarterly journal of economics, 118(1), pp. 73-106, 2003.

[17] I. Adjerid, A. Acquisti, L. Brandimarte, G. Loewenstein, "Sleights of privacy: Framing, disclosures, and the limits of transparency". In Proceedings of the ninth symposium on usable privacy and security, pp. 1-11. 2013.

[18] A. Acquisti, J. Grossklags "Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior". In 2nd Annual Workshop on Economics and Information Security-WEIS (Vol. 3, pp. 1-27), 2003.

[19] J. Puaschunder, "Towards a utility theory of privacy and information sharing and the introduction of hyper-hyperbolic discounting in the digital big data age". In Handbook of research on social and organizational dynamics in the digital era, pp. 157-200, IGI Global, 2020.

[20] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. Leon, L. F. Cranor, "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, pp. 1-16, 2011.

[21] A. Chernev, U. Böckenholt, J. Goodman, "Choice overload: A conceptual review and meta-analysis". Journal of Consumer Psychology, 25(2), pp. 333-358, 2015.

[22] S. Jilke, G. G. Van Ryzin, G. G. S. Van de Walle, "Responses to decline in marketized public services: An experimental evaluation of choice overload". J. of Public Administration Research & Theory, 26(3), pp. 421-432, 2016.

[23] K. Nagar, P. Gandotra, "Exploring choice overload, internet shopping anxiety, variety seeking and online shopping adoption relationship: Evidence from online fashion stores". Global Business Review, 17(4), pp. 851-869, 2016.

[24] G. A. Veltri, A. Ivchenko, "The impact of different forms of cognitive scarcity on online privacy disclosure". Computers in human behavior, 73, pp. 238-246, 2017.

[25] T. Alashoor, R. Baskerville, "The privacy paradox: The role of cognitive absorption in the social networking activity". In Thirty Sixth International Conference on Information Systems, Fort Worth, Texas, USA, pp. 1–20, 2015.

[26] A. Acquisti, L. Brandimarte, G. Loewenstein, "Privacy and human behavior in the age of information". Science, 347(6221), pp. 509-514, 2015.

[27] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon". Computers & security, 64, pp. 122-134, 2017.

[28] P. A. Norberg, D. R. Horne, D. A. Horne, "The privacy paradox: Personal information disclosure intentions versus behaviors". Journal of consumer affairs, 41(1), pp. 100-126, 2007.

[29] GDPR, EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament, Article 4, 2016, https://gdpr-info.eu/art-4-gdpr/

[30] S. Zuboff, "The Age of Surveillance Capitalism: The Fight for Human Future at the New Frontier of Power". London: Profile Books, ISBN 978-1-7881-6316-3, 2019.

[31] M. Van Otterlo, "Automated experimentation in Walden 3.0.: The next step in profiling, predicting, control and

surveillance". Surveillance & society, 12(2), pp. 255-272, 2014.

[32] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, J. M. Chetty, A. Narayanan, A. "Dark patterns at scale: Findings from a crawl of 11K shopping websites". Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), pp. 1-32, 2019.

[33] M. Nouwens, I. Liccardi, M. Veale, D. Karger, L. Kagal, "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 (pp. 1-13), 2020.

[34] M. Chromik, M. Eiband, S.T. Völkel, D. Buschek, "Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems", IUI Workshops (Vol. 2327), 2019.

[35] L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, A. Bacchelli, "UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, April 2020.

[36] P. Grassl, H. Schraffenberger, F. Borgesius, M. Buijzen, "Dark and bright patterns in cookie consent requests". 10.31234/osf.io/gqs5h, 2020.

[37] K. Kollnig, S. Datta, M. Van Kleek, "I Want My App That Way: Reclaiming Sovereignty Over Personal Devices", 2021, arXiv preprint arXiv:2102.11819.

[38] B. Chugh, P. Jain, "Unpacking Dark Patterns: Understanding Dark Patterns and Their Implications for Consumer Protection in the Digital Economy". RGNUL Student Research Review Journal, 7, 23, 2021.

[39] K. Bongard-Blanchy, A. Rossi, S. Rivas, S. Doublet, V. Koenig, G. Lenzini, "I am Definitely Manipulated, Even When I am Aware of it. It s Ridiculous!--Dark Patterns from the End-User Perspective". arXiv preprint arXiv:2104.12653, 2021.

[40] R. Steiner, "Dark Patterns" . [Online]. Available from: https://www.fyresite.com/dark-patterns-a-new-scientific-look-at-ux-deception/, 2021.06.24

[41] R. G. Weber, S. S. Condoor. "Conceptual design using a synergistically compatible morphological matrix." In FIE'98. 28th Annual Frontiers in Education Conference. Moving from'Teacher-Centered'to'Learner-Centered'Education. Conference Proceedings (Cat. No. 98CH36214), vol. 1, pp. 171-176. IEEE, 1998.