UNIVERSITY OF MODENA AND REGGIO EMILIA
Department of Studies on Language and Culture

Degree Programme: Languages for Communication in International Enterprises and Organizations

Dark Patterns and Cognitive Biases in Digital Marketing: an Empirical Study on Confirmshaming

Supervisor:

Francesca Pancotto

Thesis of:

Martina Fanelli

Co-Supervisor

Enrico Angelo Raffaele D'Ecclesiis

Academic Year 2025 – 2026

Abstract

In the past few decades, the rapid growth of digital marketing has led to a profound change in the relationship between sellers and consumers, as well as in consumers' purchasing decisions. Online interfaces no longer serve merely as a secondary means of communication for companies; rather, they have increasingly developed into real online environments, in which all marketing processes take place. In this context, websites have been designed to guide consumers' behaviour online, as even small modifications to the presentation of information can lead to significant changes in consumers' purchasing decisions. However, this mechanism has increasingly been leveraged by companies and sellers online to maximise profits at the expense of consumers' interests.

The present study aims to provide a detailed description of the so-called dark patterns, namely interface techniques that can be manipulative or deceptive and limit the users' freedom of choice. In addition, it explores how dark patterns are typically effective by drawing on the field of behavioural economics, an interdisciplinary field in which economics, together with psychology, studies people's decision–making processes from the point of view of the cognitive system. Indeed, it has been demonstrated that decisions are influenced by several factors, including cognitive limitations intrinsic to human nature and the emotional sphere.

The thesis is structured as follows: the first chapter provides a theoretical framework on dark patterns, including a taxonomy of the most popular ones, together with real examples; the second chapter illustrates the shift from traditional to behavioural economics, as well as the fundamental concepts of heuristics and cognitive biases and, finally, a taxonomy of biases exploited by the dark patterns presented in the first chapter. The last chapter of this thesis is based on an empirical study focused on the effectiveness of the Confirmshaming pattern, which involves influencing purchasing choices by triggering uncomfortable emotions. In particular, a survey was developed in two versions: the first includes a neutral interface, whereas the second includes an option framed with Confirmshaming. The responses are then compared to observe whether the interface with Confirmshaming reduces the likelihood that participants choose the declining option, and to understand the extent to which the emotional sphere is involved. Results show that the interface with Confirmshaming significantly increases participants' likelihood of choosing the acceptance option. The findings also suggest that this effect occurs with greater guilt elicited by the wording of the declining option, indicating a psychological mechanism consistent with the theoretical framework provided in the first chapters.

Overall, this study contributes to a deeper understanding of the mechanisms underlying online interfaces, in order to raise consumers' awareness of unethical aspects of digital marketing. This connection between research on dark patterns and the field of behavioural economics offers a useful perspective for understanding how intrinsic cognitive processes can be exploited through the interface design. The findings of the empirical study also underline the importance of considering the emotional

dimension of decision-making when evaluating and regulating manipulative interface design. As digital environments constantly evolve, this area of research plays a key role in helping users recognise potentially manipulative or deceptive practices, make more informed choices and navigate the Internet in a more conscious and autonomous way.

Abstract

Negli ultimi decenni la rapida evoluzione del digital marketing ha portato a un cambiamento significativo del rapporto venditore - consumatore e delle decisioni d'acquisto di quest'ultimo. I siti web non servono più solo come canali di comunicazione secondari delle aziende, ma sono gradualmente diventati veri e propri 'ambienti' online, in cui si concentrano tutti i processi di marketing. In questo contesto, i siti sono sempre più progettati per guidare il comportamento dei consumatori online, tanto che anche le più piccole modifiche nel modo in cui le informazioni vengono presentate possono far cambiare notevolmente le decisioni d'acquisto dei consumatori. Tuttavia, tanto le aziende quanto piccoli rivenditori online hanno iniziato a sfruttare questo sistema per massimizzare i profitti a discapito degli interessi dei consumatori.

Questa tesi ha l'obiettivo di presentare nel dettaglio i *dark patterns,* ovvero tecniche manipolative o ingannevoli delle interfacce che limitano la libertà di scelta degli utenti. Studia inoltre come questi siano spesso efficaci basandosi sull'economia comportamentale (*behavioural economics*), un'area di studio interdisciplinare in cui gli economisti e gli psicologi studiano insieme i processi decisionali dal punto di vista del sistema cognitivo. È stato infatti dimostrato che le decisioni sono influenzate da diversi fattori, tra cui i limiti cognitivi intrinseci dell'essere umano e la sfera dell'emotività.

La tesi è organizzata in tre capitoli: il primo capitolo fornisce un quadro teorico sui *dark patterns*, che include una tassonomia delle tecniche più conosciute accompagnate da esempi reali; il secondo capitolo discute dell'evoluzione dall'economia tradizionale all'economia comportamentale, introducendo concetti fondamentali come le euristiche e i bias cognitivi, e una classificazione dei bias sfruttati dai *dark pattern* riportati nel primo capitolo. L'ultima parte si basa su uno studio empirico che ha l'obiettivo di osservare gli effetti del *Confirmshaming*, un pattern specifico che mira a manipolare le scelte dei consumatori toccando la sfera emotiva, in particolare suscitando un disagio emotivo. Ai fini dell'esperimento è stato condiviso un questionario in due versioni diverse: la prima mostra un'interfaccia neutrale, mentre la seconda include un'opzione formulata con *Confirmshaming*. Le risposte dei due gruppi sono state successivamente confrontate per vedere se il *Confirmshaming* riduce la probabilità che i partecipanti scelgano di rifiutare, e per capire quanto la sfera emotiva è coinvolta. I risultati mostrano che l'interfaccia con *Confirmshaming* aumenta in maniera significativa la probabilità statistica che i partecipanti scelgano di accettare. Allo stesso tempo, sembra che i partecipanti siano mossi dal senso di colpa provocato dalla formulazione dell'opzione di rifiuto, mostrando così un meccanismo psicologico in linea con il quadro teorico sui *dark patterns* e sull'economia comportamentale.

In generale, la tesi vuole contribuire a conoscere più a fondo i meccanismi che si nascondono dietro alle interfacce online, in modo da sensibilizzare i consumatori sugli aspetti non etici del marketing digitale. Contestualmente, il collegamento tra la ricerca sui *dark patterns* e la disciplina dell'economia

comportamentale offre una prospettiva utile per capire come i processi cognitivi innati possono essere sfruttati dal design dei siti web. I risultati dello studio empirico contribuiscono a sottolineare l'importanza della dimensione emotiva nei processi decisionali durante la valutazione e regolamentazione delle interfacce. Considerando la continua evoluzione del marketing digitale, quest'area di studio ha un ruolo cruciale per aiutare gli utenti a riconoscere pratiche potenzialmente manipolative o ingannevoli, in modo da fare scelte più informate e navigare su Internet in modo più autonomo e consapevole.

Resumen

En las últimas décadas, la rápida evolución del marketing digital ha provocado un cambio significativo tanto en la relación entre vendedores y consumidores, como en las decisiones de compra de estos últimos. Los sitios web ya no se limitan a ser canales de comunicación secundarios para las empresas, sino que gradualmente se han convertido en verdaderos 'entornos' en línea/online, en los que tienen lugar todas las fases del marketing. En este contexto, los sitios se diseñan cada vez más para guiar el comportamiento de los consumidores en línea, hasta el punto de que cualquier pequeña modificación en la manera de presentar las informaciones puede influenciar significativamente las decisiones de compra de los consumidores. Sin embargo, las empresas y los pequeños negocios en línea han empezado a aprovechar este mecanismo para maximizar sus beneficios en detrimento de los intereses de los consumidores.

Esta tesis tiene el objetivo de presentar en detalle los *dark patterns,* o sea técnicas manipuladoras o engañosas de diseño de las interfaces que limitan la libertad de elección de los usuarios. Además, analiza cómo estas técnicas suelen ser eficaces a partir de la economía conductual, o economía del comportamiento (*behavioural economics),* un ámbito de estudio interdisciplinar en el que los economistas y psicólogos estudian los procesos decisionales desde la perspectiva del sistema cognitivo. De hecho, se ha demostrado que las decisiones están influenciadas por diferentes factores, entre ellos los límites cognitivos inherentes al ser humano, y la esfera emocional.

La tesis se estructura en tres capítulos. El primero ofrece un marco teórico sobre los *dark patterns,* incluyendo una taxonomía de las técnicas más conocidas con ejemplos reales; el segundo trata de la evolución desde la economía tradicional hasta la economía conductual e introduce conceptos fundamentales como las heurísticas y los sesgos cognitivos, además de una clasificación de los sesgos explotados de los *dark patterns* en el primer capítulo. La última parte está basada en un estudio empírico que quiere observar los efectos del *Confirmshaming,* una técnica específica que busca manipular las decisiones de los consumidores afectando a la dimensión emocional, en particular generando malestar emocional. Para el experimento se difundió un cuestionario en dos versiones: la primera incluye un modelo de interfaz neutral, mientras que la segunda muestra una opción formulada con Confirmshaming. Luego, se compararon las respuestas de los dos grupos para comprobar si el *Confirmshaming* reduce la probabilidad de que los participantes decidan rechazar la opción y para determinar en qué medida interviene la esfera emocional. Los datos muestran que la interfaz con *Confirmshaming* aumenta significativamente la probabilidad estadística que los participantes elijan aceptar. Al mismo tiempo, parece que ellos actúen por el sentimiento de culpa provocado por la formulación de la opción de rechazo, revelándose un mecanismo psicológico acorde con el marco teórico sobre los *dark patterns* y la economía conductual.

A la luz de lo anterior, la tesis quiere alcanzar un conocimiento más profundo de los mecanismos ocultos tras las interfaces en línea, con el fin de sensibilizar los consumidores sobre los aspectos poco éticos del marketing digital. Asimismo, la conexión entre la investigación sobre los *dark patterns* y la economía conductual ofrece una perspectiva útil para entender cómo el diseño de las interfaces en línea puede aprovechar los procesos cognitivos innatos. Además, los resultados del estudio empírico subrayan la importancia de la dimensión emocional en los procesos decisionales durante la evaluación y regulación de los sitios web. Teniendo en cuenta la continua evolución del marketing digital, este ámbito de estudio tiene un papel fundamental para ayudar a los usuarios a reconocer prácticas potencialmente manipuladoras o engañosas, así que puedan tomar decisiones más informadas y navegar por Internet de manera más autónoma y consciente.

Table of Contents

Introduction **1**
1. Dark Patterns in Digital Marketing **4**
1.1 The evolution of digital markets 4
1.2 Main definitions of Dark Patterns 8
1.3 Features of Dark Patterns 12
1.3.1 Taxonomy of Dark Patterns *12*
1.4 Beneficial Nudges vs Dark Patterns 25
1.5 Who is vulnerable to dark patterns? 26
2. Behavioural Economics: from Heuristics to Cognitive Biases **28**
2.1 From traditional economics to behavioural economics 28
2.2 The Dual Process Theory, Heuristics and Cognitive Biases 31
2.3 Nudges and Sludges 34
2.4 Cognitive Biases 38
2.4.1 Anchoring bias *38*
2.4.2 Framing bias *40*
2.4.3 Bandwagon effect *42*
2.4.4 Scarcity bias *44*
2.4.5 The Default Effect *45*
2.4.6 The Sunk Cost Fallacy *48*
3. An empirical study on Confirmshaming **52**
3.1 Research questions 53
3.2 Methodology 53
3.2.1 Data collection *53*
3.2.2 Procedure *56*
3.2.3 Limitations of the study *59*
3.3 Findings 60
3.3.1 Data cleaning and acceptance rates *60*
3.3.2 Two-group t-tests on the Likert scales *62*
3.3.3 Levels of attachment to the environmental cause *64*
3.3.4 Robustness check on acceptance rates *64*
3.4 Discussion of the findings 65
3.4.1 Acceptance rates (RQ1, H1) *65*
3.4.2 Guilt for Decision vs Guilt for Wording (RQ2, H2) *66*
3.4.3 Relief and Pride (RQ2, H3) *67*
3.5 Summary 68
Conclusions **69**
References **71**
Appendix A: Surveys (Group A + Group B) **75**
Appendix B: Excel tables and formulas **78**

Introduction

Nowadays people rely on online stores for any kind of purchase: electronics, clothing, home goods and even groceries. The convenience, time efficiency and 24/7 availability of products offered by e-commerce provide consumers with clear and tangible benefits. However, these advantages coexist with potential risks. Users are repeatedly exposed to highlighted limited time offers, countdowns and hidden costs, among other things. Imagine placing an order and realising only later that you paid more than the final cost shown at checkout because the extra fee was not clearly visible, or buying a flight ticket and becoming aware that you also paid for insurance you did not choose. In these cases, to what extent are consumers still willing to risk encountering deceptive practices in order to save time and enjoy the convenience of shopping from home? This thesis approaches the question from a different perspective: rather than trying to understand whether they evaluate risks and benefits, it investigates why they tend not to notice the risks, or fail to resist them when they are clearly visible.

These practices originate from the 'choice architecture' (Thaler & Sunstein, 2008) of online interfaces. Indeed, the rapid advancement of technology has allowed Internet to become central to all human activities, from socialisation to shopping, thus creating an online environment separate from the real world, in which websites are designed to guide users' decisions. Elements of the choice architecture aimed to benefit consumers are considered nudges, while if they aim to deceive them and push them towards choices against their interests, they are considered dark patterns. They are defined as design choices that steer, deceive or pressure users into decisions they would not otherwise make to the benefit of the company (Mathur et al., 2019). Both techniques derive from the same logic: a completely objective and neutral way of presenting information is impossible, therefore any kind of element of the interfaces (layout, colour, font, linguistic choices, etc.) inevitably influences users' behaviour, in an ethical or unethical way.

The main strength of this thesis is the systematic connection between dark patterns' techniques and cognitive biases, moving beyond their mere definition in order to explain why they are typically effective. Building on the dark pattern literature and behavioural economics, this work also aims to shed light on the role of the emotional sphere when people are exposed to manipulative or deceptive practices. The thesis is characterised by two theoretical chapters and a final empirical one. The first chapter provides the context in which dark patterns developed and introduces them by comparing definitions from different authors who focused on different perspectives, such as the mere visual characteristics, the effects on users and the role of the designer, among others. It also highlights that dark patterns can act differently depending on their attributes, for instance, they can be covert, by subtly limiting users' autonomy, or asymmetric, when they intentionally put disproportionate weight on an option at the expense of the alternative.

It is then presented a taxonomy of well-known dark patterns with real examples, followed by the distinction between beneficial nudges and dark patterns and a discussion on consumer vulnerability to these practices, as remaining an open debate.

The second chapter builds on the historical evolution from traditional economics to behavioural economics. The former was based on strict assumptions according to which individuals are rational beings who make decisions solely on the basis of their expected utility; this view was questioned by the more recent collaboration between psychology and economics, that started to consider natural cognitive limitations and emotions as key factors influencing any choice. In this regard, the concepts of Dual Process Theory, heuristics and cognitive biases are explored, distinguishing between cold biases, based on information processing limits, and hot/motivated biases, driven by instincts and emotions. The discussion then links these concepts back to dark patterns by describing the differences between nudges and sludges, as well as between manipulation and deception techniques, always from the perspective of the individual rather than of the design. Finally, a few cognitive biases are classified by connecting them to the examples provided in the first chapter, according to their nature (hot or cold) and to the technique used to exploit them.

The last chapter differs from the previous part of the thesis as it presents an exploratory experiment focused on a specific dark pattern. As one of the objectives is to shed light on the individuals' emotional responses, the chosen technique is Confirmshaming, that consists of framing the options of an interface in a way that elicits uncomfortable feelings such as guilt or fear of being judged, with the aim of steering the consumer towards the choice that most benefits the company. The chapter describes all the steps, from the creation of the surveys to the statistical analyses used to compare acceptance rates between two different groups exposed to different interfaces, and to analyse Likert scales measuring positive and negative emotions. The limitations of this study are also outlined, as it relies on non-probability sampling methods that do not guarantee randomness, and on small sample sizes, that can influence the expected results. The thesis closes with general conclusions and potential directions for future research.

1. Dark Patterns in Digital Marketing

The present chapter aims to define dark patterns by exploring the different definitions of this concept in recent studies and the main characteristics. A first overview of e-commerce and the growth of digital markets will be provided in order to understand the context that allowed for the emergence of dark patterns. Moreover, the most common types will be analysed by illustrating real examples from digital marketing. Finally, the last paragraph of the chapter will address the ongoing debate on consumer vulnerability to dark patterns, thus trying to understand whether everyone is potentially susceptible to such strategies or only a few categories of consumers with specific characteristics.

1.1 The evolution of digital markets

The rapid technological progress and the rise of the Internet have transformed the traditional business model, leading to a significant change in modern consumer behaviour, meaning the shift from physical purchasing to online transactions.

Digital marketplaces emerged in the mid 1990's, they largely expanded between the mid 2000's and the 2010's and nowadays consist of platform-based economies characterised by artificial intelligence-driven personalisation (Wahab & Abdulmajeed, 2025). During the first phase of digital marketing the main function of websites was to provide information to consumers and to enhance sellers' credibility. Indeed, the Internet was primarily used as a secondary channel to promote products and to support offline marketing. Information was merely transferred to the webpage of the company, employing the mass marketing traditional approach, based on only one-way communication and absent seller-consumer interaction. This form of web advertising evolved during the mid-2000's, with the emergence of Web 2.0 together with the increasing interactivity enabled by social media. The name Web 2.0 was coined to identify the second stage of the World Wide Web and to distinguish the main features of Web 1.0 and Web 2.0. While Web 1.0 was based on static and read-only pages with limited interaction, Web 2.0 evolved towards read-and-write pages, interactive social media and dynamic applications. In this online environment all marketing functions started to take place, including not only providing information, but also the so called four Ps[1] of marketing: price, promotion, product and place (Kelly-Holmes, 2015). This development allowed for interactive advertising communication in which consumers started to gain power to spread information and feedback about products to other users, an increasing phenomenon that could not be controlled by companies and sellers.

[1] The four Ps of marketing, also referred to as 'marketing mix', are the key factors at the basis of any successful marketing strategy. The concept was popularized in the 1960's by Neil Borden, an advertising professor at Harvard, and is still used nowadays by companies in marketing. Source: https://www.ama.org/marketing-news/the-four-ps-of-marketing/

Nowadays we are living a further developed phase dominated by artificial intelligence and personalisation, driven by continuous technological advancements and the global diffusion of smartphones, in which data are constantly collected to enable digital markets and social media to provide significantly targeted content and advertising (Kelly-Holmes, 2015). This revolution served as the key driver to the shift from mass market to the current virtual marketplace, described by Kelly-Holmes (2015) as highly segmented, niche-dominated, and global. This means that current digital marketing is no longer addressed to standardised consumer groups, but rather to individual users.

This profound change led to significant advantages for companies, sellers and services providers. In this regard, Fradkin (2018) identified the features shared by major online marketplaces (such as Airbnb and Amazon), that can be considered as key benefits for businesses, also showing that although digital platforms appear highly heterogenous, they actually share certain characteristics. These are:

- **Digital matching**: sellers can easily find consumers through websites, applications or messaging platforms. These tools are able to track the actions of users and, through algorithmic processing of data, to specifically tailor content and interfaces depending on individual preferences and behaviours.
- **Low entry costs**: opening a business online is not expensive since it is not necessary to possess a company or to open a store to sell products. Also, both a company and an individual seller can use the same platform. For instance, Amazon serves as a retail platform by major companies, but also privates can sell second-hand goods. In this regard, consumer-to-consumer platforms such as Vinted and Wallapop are increasingly popular, allowing people to sell both new and second-hand products at low prices, while also promoting circular economy.
- **Ex-post screening**: many digital marketplaces tend to verify products or transactions only after they have been offered for sale, usually through users' feedback, complaints, and ratings, instead of conducting long checks before allowing products to enter the market. While reducing the companies' costs of prior screening, this approach may also lead to a higher risk of selling false or bad-quality products.
- **Non-exclusive and short-run contracts**: consumers can use several platforms and are not obliged to engage in long term contracts. Such flexibility allows sellers to reach different consumers categories by providing products in several different platforms at the same time.
- **Direct transactions**: In digital marketplaces, the money paid by the consumer are directly transferred to the seller, net of fees or additional services, this means that the platform makes the money transfer easier from the consumer to the seller. This is a characteristic specific to e-commerce that distinguishes digital marketplaces from other services such as streaming

platforms and dating applications. In these cases, the platform serves also as service provider, therefore money is transferred to the platform rather than to another seller.

A further strength is the easier internationalisation of markets (Kelly-Holmes, 2015), since products can reach overseas markets at reduced costs, considering that it is not necessary to set up abroad subsidiaries as in physical markets.

From the consumers' perspective, the rapid expansion of e-commerce is fostered by several experienced advantages. Wahab & Abdulmajeed (2025) provide evidence that respondents recognise e-commerce convenience, accessibility, and time efficiency, affirming that you can find wider product ranges, better prices, shorter waiting times and higher availability of products beyond the limitations of traditional store hours. Also, COVID-19 pandemic played a crucial role in this great change, since the forced worldwide lockdown led to consumers' increased demand for alternative shopping methods. Furthermore, as previously anticipated, this marketing revolution has given consumers a new active role, meaning that they are not only active participants of the marketing communication, but also co-creators of value in the whole marketing process itself. In such interactive environment consumers are encouraged to share information and feedback about the product through the so called 'electronic word of mouth' (eWOM), that can influence other consumers' purchasing decisions and can either reinforce or compromise the business activity. Consequently, marketers have no longer information monopoly, and the potential product failure is not under their control (Kelly-Holmes, 2015).

However, this market displacement raises different major issues, since consumers no longer interact directly with sellers, but rather, the relationship seller-buyer is mediated by the online interface. Firstly, online purchasing always involves cybersecurity risks that should be controlled by new regulation in online consumer protection. Together with the growth of e-commerce, the incidence of cybercrime has increased, with cases of identity thefts, data breaches and frauds. In this concern, nowadays new systems of safe transactions have been developed, such as credential verification, fraud detection algorithms and customer service (Fradkin, 2018). Also, policies on consumer protection have further increased. The most comprehensive regulation so far is the General Data Protection Regulation (GDPR), a European Union law put into effect in 2018 that regulates data collection and protects users' privacy. It forces companies and websites to explicitly ask users for their consent before collecting any data, providing clear information about data collection and processing, and giving the possibility to accept or refuse (Kollmer & Eckhardt, 2022). However, scholars and users claim for a more comprehensive law, since this kind of regulations are only applied to consents, whereas users are still exposed to other dangerous and manipulative strategies.

Secondly, although the rise of big data analytics has contributed to the creation of an online environment that is highly controlled and personalised, users are often unaware of the level of personalisation characterising their online experiences. A key practice that allows for this phenomenon is the Online Behavioural Advertising (OBA) (Boerman et al., 2017), namely the practice of tracking people's online behaviour and using the collected information to show people individually personalised advertising. The memorised online behaviour can include web browsing, watched videos, application use data, purchases, etc. Also, websites track personal data such as cookies, small text files that can remember detailed information about a huge number of consumers.

This process is possible thanks to the application of artificial intelligence to marketing, which Lies (2019) refers to as 'marketing intelligence', the process of developing insights from data collection to support marketing planning and decision-making. Traditionally, consumers purchasing decisions were detected mainly through surveys, but this system shifted to automatic collection of marketing decisions thanks to big data analytics. Digital platforms increasingly rely on the tools based on marketing intelligence, such as the recommender systems (RSs), software techniques using algorithms to filter information and to suggest the right items based on the users' specific interests and preferences (Ricci et al., 2010). When the first e-commerce websites started to develop, the RSs were intended to help users to filter an overwhelming number of options that the pages were offering. Therefore, the main function of the recommendation algorithm is to predict the best item that is worth recommending depending on users' previous interactions, preferences and purchasing choices. Consequently, although they help to reduce information overload, the result is an interface with a reduced number of choices specifically tailored for the users, even if they believe to have a higher set of options than in physical stores. At the same time, the online environment is guided by buttons, check boxes, swipe options and strategic organisation of information. In this way, users may perceive their purchasing decisions as completely autonomous, while their actions are actually controlled by the design of the interface. Lies (2019) offers an interesting point of view when analysing such personalisation that deeply influences consumers' decisions. Indeed, he interprets these marketing practices as 'social engineering practices', meaning practices that aim to impact individual behaviour, which in turn leads to a social impact, and thus, a social change.

This overview of digital marketplaces is necessary to understand the environment in which manipulative interface designs have emerged. As previously discussed, there has been a shift from traditional mass marketing into highly personalised virtual marketplaces, creating advantages for both sellers and consumers. On the one hand, companies can reach consumers at global scale at lower costs and can leverage algorithms to show increasingly targeted advertisements. On the other hand, consumers report to benefit from more convenient prices, time efficiency and a higher number of available options.

However, this marketing revolution has led to significant drawbacks. The main concerns regard cybersecurity risks, the need for a more comprehensive regulation for consumer protection and the high personalised online environments. Particularly, big data analysis and algorithms have enabled marketing intelligence practices, online behavioural advertising and recommender systems, techniques that are used by companies online to increase their businesses. These digital marketing tools, that were at first intended to benefit consumers' needs, have been increasingly exploited to maximise companies' profits at the expense of consumers. Their decision-making processes have become more vulnerable to manipulative marketing strategies referred to as 'dark patterns'. This phenomenon is greater explained in the next section.

1.2 Main definitions of Dark Patterns

Digital marketing research provides several definitions describing dark patterns, or deceptive patterns. These will be explored in the present section through both a synchronic and a diachronic analysis, by providing an overview of the development of dark pattern research from the 2010s onwards, while outlining the different perspectives from which dark patterns can be analysed.

The term was first coined in 2010 by Harry Brignull, who identified them as "tricks used in websites and apps that make you do things that you did not mean to, like buying or signing up for something" (Brignull, 2010). This original concept was limited because it only emphasised the visual characteristics of the interface that can negatively influence users, defining them as design "tricks", since according to the author they have been created with great attention to detail with the specific goal of manipulating users. However, research has broadened the concept, taking into account many different aspects of these patterns, that can be divided into four main perspectives.

Firstly, the focus shifted from the interface design to how the design influences consumer decision-making. In this regard, several authors mentioned by Mathur et al. (2021) conceptualised in different ways the effects on consumers, by using verbs such as *subverting, undermining, manipulating, attacking, exploiting*. These verbs carry strong negative connotations and highlight the intention to mislead users in favour of the economic benefits of the other party.

The second interesting perspective moves from the effects on users to the role of the designer of the interface, for instance claiming that "designer abuses their domain-specific knowledge of human behaviour" (Gray et al., 2018), or that "designer intentionally deploys dark patterns to achieve a goal" (Conti & Sobiesk, 2010). These conceptualizations assume that web designers study the user behaviour in detail, are able to predict their reactions to specific design choices and intentionally exploit such knowledge to influence consumers towards the decision that best benefits the company or service provider. This perspective therefore highlights the intentionality and responsibility of the designer.

A further perspective emphasises the opposition between the advantages and the drawbacks generated by dark patterns. Some definitions are focused on benefits for the online businesses, whereas other definitions are more focused on the drawbacks for users. This may be interpreted as a way to emphasise that marketing communication should aim to benefit both the business activity and consumers' needs. However, what actually results from such manipulative strategies is a significant unbalance in favour of sellers.

The last perspective is the most comprehensive one, since the most recent studies integrate all the previous characteristics and define dark patterns in the business context, considering both the characteristics of the interface design and the effects on users' decision-making processes. Firstly, Lukoff et al. (2021) enhanced the very first definition provided by Brignull, studying dark patterns as strategies to maximise consumers' use of digital devices. This concept sheds light on a new aspect, namely these patterns are not always explicit tricks or openly misleading for users; rather, they sometimes seem to be in line with users' intentions. What happens is that they impair users' autonomy, for instance by encouraging to infinite scrolling, autoplay or pull-to-refresh.

The question of the individual autonomy online is fundamental since in such controlled environment the user's autonomy is increasingly vulnerable. Kollmer & Eckhardt (2022) define autonomy as "self-governance that leads to independent choices and the expression of free will among users", whereas Mathur et al. (2021) provide a definition in the normative context, stating that "autonomy is the normative value that users have the right to act on their own reasons when making decisions". Therefore, autonomy does not only mean the freedom to choose among a set of choices, but also that such choices should be conscious and independent from guided interfaces. In this context, dark patterns are a major concern since they are designed to violate users' autonomy, for instance by hiding information or make it difficult for users to find it, and this is why manipulative strategies should be contained.

Building on the impairment of user's autonomy, the last comprehensive definitions useful for the scope of the present study are provided by Mathur et al. (2019) and Kollmer & Eckhardt (2022), who describe dark patterns as follows:

> *user interface design choices that benefit an online service by coercing, steering, or deceiving users into making decisions that, if fully informed and capable of selecting alternatives, they might not make.* (Mathur et al., 2019)

> *user interface design elements that compromise user autonomy by preventing informed choices and that may lead to adverse outcomes for the user, such as invasion of privacy, financial loss, and technology addiction.*
> (Kollmer & Eckhardt, 2022)

In this evolution of the concept, we can observe the shift from general tricks (as originally described by Brignull (2010)) to *coercing, steering*, or *deceiving* practices, including different types of manipulative strategies. In this sense, dark patterns are not exclusively misleading practices, meaning that they are able to influence users' decisions in subtle ways, without necessarily hiding information or provide false information. Also, the last definition highlights the importance of informed choices in users' autonomy and provides possible examples of wrong decisions influenced by dark patterns, with the purpose of raising awareness on serious effects that people may experience, such as personal data breach or technological addiction.

The definitions previously analysed are summarised in Table 1, in order to provide a comparative overview of the different definitions of dark patterns. It should be noted that this table does not provide a complete overview of the existing definitions in the digital marketing literature, but rather, a selection of the most significant ones for the scope of this work from the analysis conducted by Mathur et al. (2021).

Perspectives	Authors	Definitions
Visual characteristics	- Brignull (2010) - Bösch et al. (2016)	"Tricks used in websites and apps that make you do things that you did not mean to" "Misleading interfaces"
Effects on users	- Westin & Chiasson (2019) - DETOUR Act[2] - Conti & Sobiesk (2010)	"Subverting user intent" "Undermining user autonomy" "manipulate", "attack", "exploit" users
Role of the designer	- Gray et al. (2018) - Conti & Sobiesk (2010)	"Designer abusing their domain-specific knowledge of human behaviour" "Designer intentionally deploying dark patterns to achieve a goal"
Benefits and harms from the interface design	- Conti & Sobiesk (2010) - Gray et al. (2018)	"Dark patterns aiming to benefit an online service" "Dark patterns as harms to users"
Impairment of users' autonomy	- Lukoff et al. (2021) - Mathur et al. (2019) - Kollmer & Eckhardt (2022)	"Strategies to maximise consumers' use of digital devices" "[…] 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" "[…] compromise user autonomy by preventing informed choices and that may lead to adverse outcomes for the user […]"

Table 1: Definitions of Dark Patterns

[2] To know more about the DETOUR Act: H.R.8975 - 116th Congress (2019-2020): DETOUR Act | Congress.gov | Library of Congress

1.3 Features of Dark Patterns

Building on the different proposed definitions, Mathur et al. (2019) outlined a set of general attributes characterising dark patterns common in all types of digital platforms, including shopping websites, mobile applications, social media networks, video games, as well as privacy notices. They are as follows:

- **Asymmetric**: dark patterns can impose unequal burdens on the available choices presented to the user in the interface. They are common in the consent interfaces, such as the cookie consents, in which the 'Accept all' button is highlighted, whereas the 'Reject all' button is often less visible or even absent. Indeed, it is often replaced by 'Manage preferences', a button that requires users to make multiple decisions, making it more time-consuming and effortful.
- **Covert:** the dark pattern effect on users' decisions can be covert, when the interface is designed to encourage users to make a specific decision even though they believe their autonomy is not influenced. This is the case of the 'decoy effect', for instance when there are three subscription options, and the second one (the decoy) has been intentionally inserted to make the most expensive option appear the most convenient for the user.
- **Deceptive:** deceptive dark patterns can manipulate users through false or misleading information. For instance, many time-limited offers push users to buy the product as being convenient in those specific hours, but they turn out to be false since at the time of the expiration they often reset.
- **Information hiding:** the interface can be designed to hide or delay necessary information for the users' final decisions, such as when the purchase of a product leads to an automatic subscription without the consent of the user, who becomes aware of this only after payment.
- **Restrictive:** the pattern can be restrictive when the interface offers only a set of options that are actually reduced as compared to the actual number of possibilities. For instance, when users opt-in a specific service but the interface omits the opt-out option or requires a more complex procedure to find it.

1.3.1 Taxonomy of Dark Patterns

In the present subsection real examples of dark patterns will be shown following the results of the study conducted by Mathur et al. (2019), who collected data from 11k shopping websites. Particularly, researchers developed an automatised system able to identify dark patterns on a large scale. They then manually revised the findings to recognise what could be considered dark patterns based on the existing literature. The presented images are partially taken from the website Deceptive.Design, created by Harry Brignull and that provides different categories of dark patterns with explanations

and examples. Other images are personal screenshots, to demonstrate that these techniques are present in applications and websites used in everyday life.

Sneaking:

Sneaking consists of voluntarily hiding information that can significantly influence the consumer's choice, in order to push them to take the action they would not choose if they were exposed to that specific information. This is the typical case of hidden additional costs or unwanted subscriptions; indeed, authors identify a few sub-categories of sneaking dark patterns, as follows:

- **Sneak into basket**: This pattern was widely used roughly ten or fifteen years ago, when online consumer protection regulation was still developing, and it automatically adds products to the shopping cart without consent. One of the most famous companies employing this pattern was Ryanair, that added extra costs such as preselected insurance and other optional services, that substantially increased the cost of the flight ticket, but consumers came to know this only at the checkout (Figure 1). The company was sued for these practices and had to pay a financial penalty, this is why nowadays purchasing procedures are much more controlled. (Source: https://www.lynx2000.it/blog/dark-pattern-il-lato-oscuro-del-design/)
- **Hidden costs**: This pattern involves revealing extra charges only at the checkout. It is similar to the Sneak into basket strategy, but the main difference is that in this case the costs appearing at the checkout are mandatory, such as shipping costs or service fees. Unlike Sneak into Basket, Hidden Costs practices are still used since they involve mandatory charges revealed at the final checkout, rather than optional services automatically added to the cart.
- **Hidden subscription**: also known as Forced Continuity, this pattern aims to hide recurring subscriptions behind one-time offers or free trials. In other words, the company or service provider intentionally omits that the purchase is actually a recurring subscription, thus encouraging the consumer to proceed with the purchase. A famous case regards Adobe, a popular software company that offers many different creative and document tools. In 2024 the Federal Trade Commission (FTC) took action against Adobe for deceptive subscription practices. Particularly, Adobe required a high early termination fee, consisting of the 50 percent of the remaining monthly payments when a consumer cancels within the first year. However, this information was obscured or presented in small-print text, pushing users toward the annual subscription. (Source: https://www.ftc.gov/news-events/news/press-releases/2024/06/ftc-takes-action-against-adobe-executives-hiding-fees-preventing-consumers-easily-cancelling)

The following images present two examples of Sneak into Basket and Hidden Costs, in order to visually compare their use in different online interfaces. Figure 1 shows Ryanair 2010 interface, in which travel insurance was automatically added at the final checkout, and the procedure to deselect it was not intuitive. You had to open the drop-down menu and scroll through a list of countries, and only at the end you could find the option “Don’t Insure me”. Such mechanism may negatively influence consumers, who may not be aware of the deselection option and passively accept the extra cost, or may find the process too difficult or time wasting and decide to pay extra costs anyway.

Figure 2 shows the Deliveroo interface when placing an order. In the first page we can see the selected option discounted to 11,18€, which is the most attractive option since the 25% discount is highlighted. Further information provided are only the free delivery for an order over 10€, together with the advantages with the Deliveroo Plus subscription. However, in the next page of shopping cart the final price is 12,37€ with extra service fee. The problem in this case is not the service fee itself, but the fact that this information was hidden in the first interface, whereas it would be more trustworthy if the information “Service Fee” was added under the delivery costs. The effects on the consumer’s decision may be either to search for another similar delivery platform or to pay more because that specific restaurant is only available on Deliveroo.

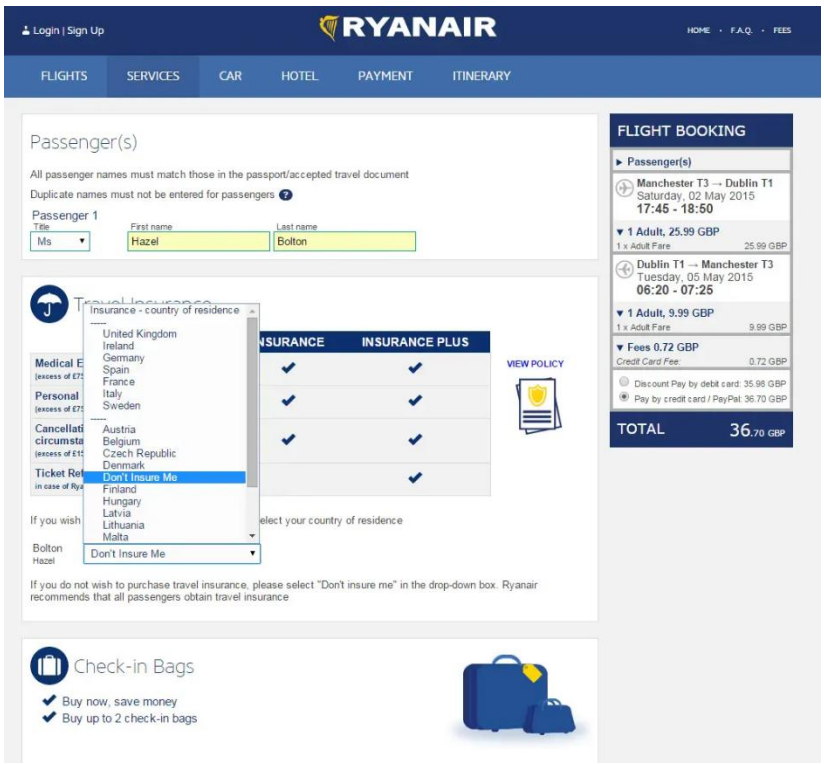


Figure 1: example of Sneak into basket, Ryanair 2010 version.
Source:https://www.lynx2000.it/blog/dark-pattern-il-lato-oscuro-del-design/

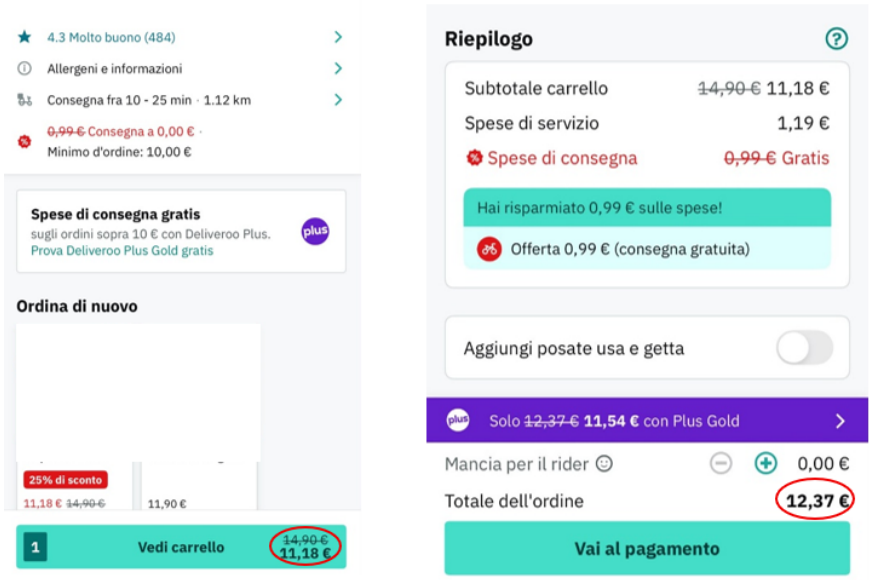


Figure 2: example of Hidden Costs from Deliveroo

Urgency:

Urgency is a category of dark patterns that aim to pressure consumers into purchasing by using fake countdown timers or vague limited time messages. These patterns are widely used nowadays, as demonstrated by Mathur et al. (2019), who found evidence that 437 e-commerce websites use strategies such as "Flash sales ends in X hours, SHOP NOW", even if the majority of discounted products never change, or "Sale ends soon", without stating a specific deadline. Moreover, Harry Brignull on his website presents the case of the app "Hurrify", a plugin used by small e-commerce stores on Shopify to create fake countdown timers. This was demonstrated by the interface for sellers, that provided the default option "Action When Timer ends: Run the campaign all over again" (Source: https://deceptive.design/types/browse/?sel=%2Ftypes%2Ffake-urgency). This plugin was banned in 2023, and nowadays these manipulative techniques are more controlled. However, this kind of design interface is still widely displayed to attract consumers, usually highlighted in red to create a sense of urgency, this is why it is not always easy to distinguish countdown timers reflecting low stock from false ones. For instance, although it is currently a controlled environment, Amazon makes wide use of such strategies, also providing an entire category flash sale, as in Figure 3.

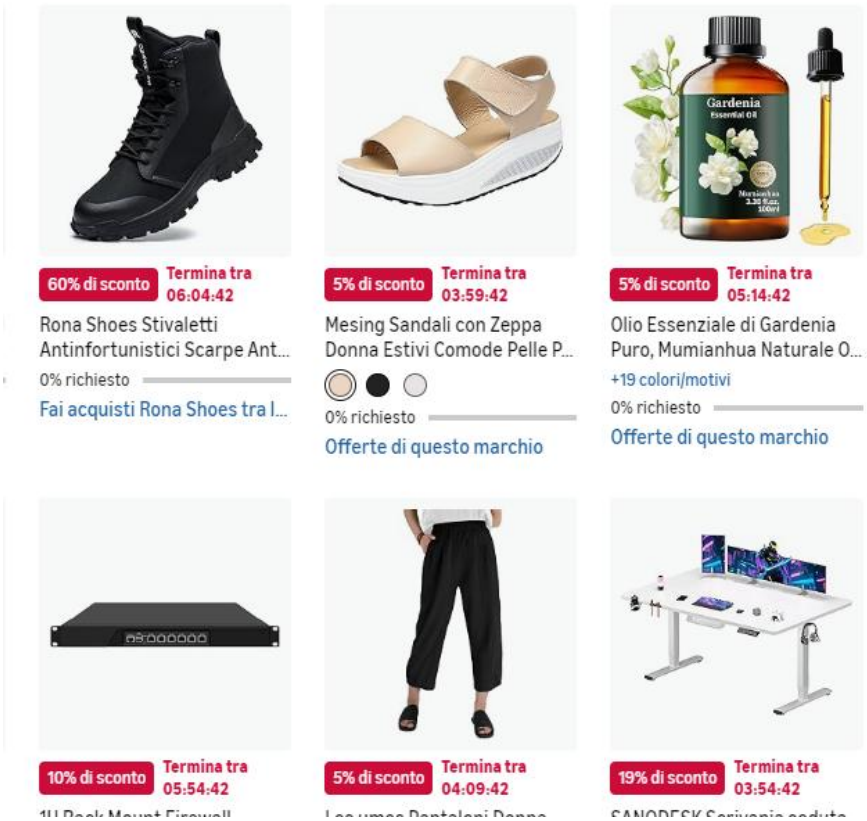


Figure 3: Example of Urgency dark pattern on Amazon

Scarcity:

This pattern is strictly connected to the Urgency category, since both aim to pressure consumers to purchase as soon as possible. However, the previous pattern focuses on time limitation, whereas Scarcity highlights the limited availability of products. In this case the main messages regard low stock (Figure 4) and high demand (Figure 5).

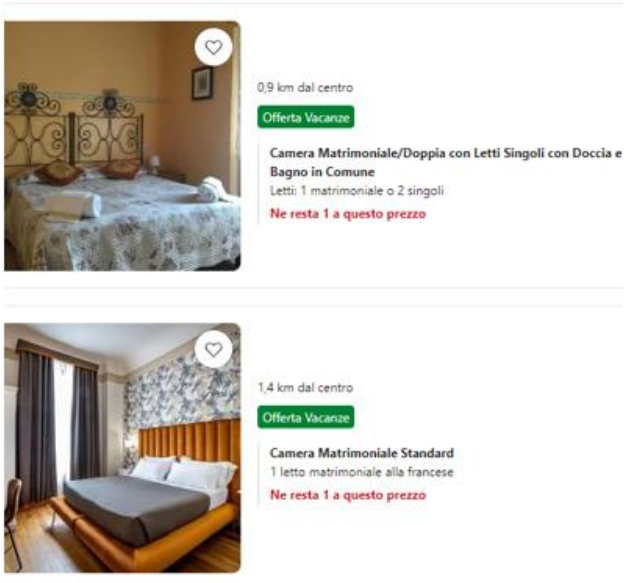


Figure 4: example of Scarcity (low stock) on Booking

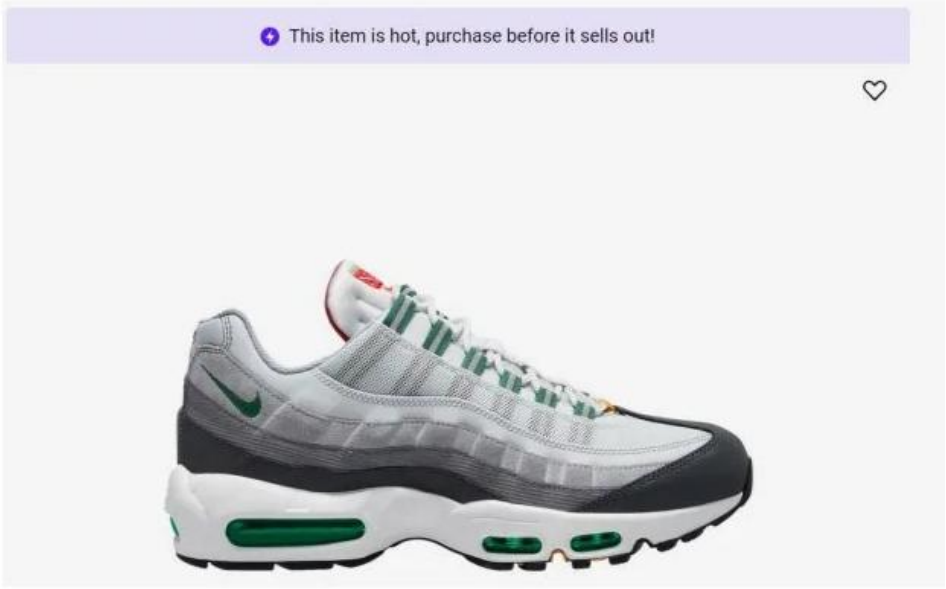


Figure 5: example of Scarcity (high demand) on Foot Locker. Source: https://deceptive.design/articles/foot-locker-creates-false-sense-of-urgency-stating-limited-stock-for-its-products?f_Type=Fake+scarcity

Also, in this case it is not always possible to understand whether the pattern is actually misleading or not, due to urgency perceived by consumers, who are more willing to take into action rather than verifying the patterns trustworthiness. For instance, most products on Booking show the "Only X left" message, and this may appear suspicious due to the huge number of accommodations available. As for the Foot Locker examples, the main problem is the vagueness of the message, that makes consumers intend that the stock is low, but it does not provide precise information neither about the number of purchases nor about the product availability.

Misdirection:

This category includes different types of dark patterns sharing the purpose of steering consumers towards a specific choice through language, visuals and emotions. The main types of dark patterns explored by Mathur et al. (2019) are the following:

- **Confirmshaming:** this pattern uses emotional manipulation to push consumers towards a choice they would not have chosen in another situation. This term was coined by Harry Brignull, who explains that this pattern triggers emotions such as guilt, shame and self-image perception, in order to benefit the service provider. Figure 6 shows two options when booking a flight on American Airlines, that consist of accepting or declining the trip protection. Instead of offering two neutral choices such as "Yes – No", the employed language aims to trigger users' emotions, through the sentences "I'm willing to risk my flight" and "I may be responsible for cancellation fees". Firstly, the use of the first-person pronoun leads to a personal involvement, and secondly, the words "risk" and "responsible" strengthen this effect by carrying negative connotation and increasing feelings of guilt

and fear, as if the cancellation fee was consumer's fault instead of the consequence of the company's policy.



Figure 6: example of Confirmshaming on American Airlines. Source: https://deceptive.design/articles/american-airlines-confirmshaming-users-into-buying-flight-insurance?f_Type=Confirmshaming

- **Visual interference**: also known as false hierarchy or framing, this pattern influences the design interface to make one option visually more attractive than the others, in different ways. For instance, a button may be highlighted with a specific colour that is brighter than the rest of the page, or information may be presented in an ambiguous or overwhelming way across the website. Moreover, it can also be a form of misrepresentation, meaning that a part of the available choices may be hidden, obscured or disguised, and in this case what is omitted is usually the potential negative information, while focusing only on positive aspects. Figure 7 is a typical case that users can find everyday online, in which cookie options are displayed as follows: the 'Accept all' button is highlighted, whereas the other two options are less visible due to the low colour contrast between the text and the button. Above all, the 'Deny all' option is hidden, and it is probably available in the 'Privacy Settings' hyperlink, which is written in small and not visible font. Users are expected to find the most important options (accept and deny options) in a clear and visible way in order to make a fast and free choice, but such interface is more likely to influence their decisions towards 'Accept all' or 'Accept essential only', by hiding the others or making them more effortful and time consuming.

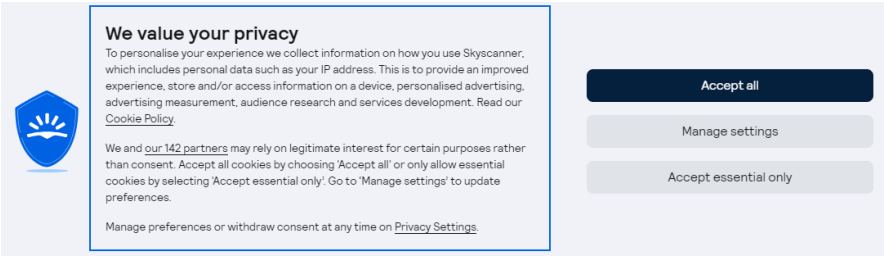


Figure 7: example of Visual Interference from Skyscanner

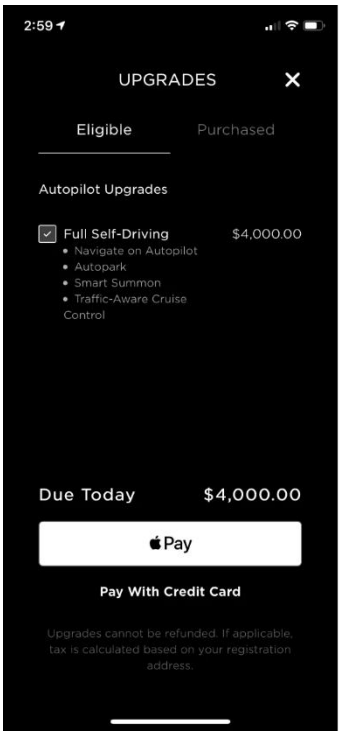


Figure 8: example of Visual Interference. Source:https://deceptive.design/types/browse/?sel=%2Ftypes%2Fvisual-interference

Moreover, Figure 8 shows a reported case of Tesla mobile app, that in 2019 added an eCommerce feature allowing car owners to buy upgrades. The white 'Pay' button strongly contrasts with the black page, whereas one of the most important pieces of information, that is the not refundable payment, is hidden at the end of the page and not visible due to the low contrast text. Consequently, people complained to have purchased by mistake, discovering later of the non-refundable payment. This suggests that while manipulative interfaces do not always cause serious consequences on users, they should be more controlled since in some situations they may damage users' personal data or financial situation.

- **Trick questions**: trick wording uses language to mislead users and guide them to the choice that most benefits the service provider. Particularly, it is mainly used to make users accept promotional and marketing messages and the main strategies consist of writing in a confusing way or providing long and redundant texts. For instance, Figure 7 may be considered as a mix of visual interference and trick questions patterns, since the long and small text is more likely to be skipped by users that tend to prefer the fastest options. Another strategy is the use of double negatives, such as 'Uncheck the box if you prefer not to receive email updates' (Mathur et al., 2019).

This preselected option combines the two negations 'uncheck the box' and 'not to receive', which are confusing since users tend to associate the action to a consensus rather than to the contrary. The most neutral option would be a not selected option such as 'Check the box to receive email updates', but

websites use such strategies to exploit users' tendency to read quickly and make rapid decisions. Finally, a further trick wording pattern may present more ambiguous options providing similar information, in order to push users to accept all options rather than reading and understanding word by word, such as in the following example:

- Recommendations: Trendy or popular features, templates, and other creative contents.
- Products updates: News about new features, product updates or changes, etc.
- Marketing: News about promotions, events, and other campaigns.

(Source: https://deceptive.design/articles/fotor-uses-double-negatives-confusing-users-to-opt-out-of-marketing-advertisements?f_Type=Trick+wording)

In this case, the main confusing features are the repetition of 'new features', 'popular features' and 'news', as if the three options were referring to the same thing, as well as the vague and general 'other campaigns' or 'other creative contents', meaning that users do not exactly know what they are accepting.

- **Pressured selling:** this pattern is highly used both in digital and in offline marketing and consists of persuading consumers to purchase the most expensive product. The two main strategies are cross-selling, a strategy that encourages consumers to buy additional products or services, and up-selling, that intends to push consumers to buy the latest or the most updated version of a product. The first strategy is usually employed by online travel services and retail of electronic devices and household appliances, for instance on Ryanair one or two pages are always included in the purchasing process before the checkout, offering additional services such as taxi, extra luggage or insurance. Also, when buying a mobile phone, they usually advice the phone case or earphones. A particular case regards Apple, that in the past few years is no longer including the power adapter with the iPhone. Retailers therefore tend to exploit the cross-selling strategy by presenting it as an additional accessory together with the main product, even though it is actually necessary to use the product daily. By contrast, the up-selling strategy is mainly used by streaming platforms that offer different subscription options, but the prices are usually strategically decided and displayed in a way that the most expensive option serves as an 'anchor' to push consumers towards the mid-option, instead of the cheapest option (see anchoring bias in 2.4.1).

Social proof:

This pattern induces consumers to believe that a product is the most popular among a variety of different products, through users' reviews and testimonials that sometimes may be false and randomly generated. The main employed strategies are notifications or popups stating that a high number of users are viewing the product at the same time, or testimonials and feedback that should always be verified to consider them trustworthy.

As a result, consumers may feel pressured to act quickly before the product is sold out, instead of taking time to think whether they really need it or to check if there are similar options available online. E-commerce websites often employ this strategy, such as Amazon (Figure 9), that shows the number of people buying the product to demonstrate its quality, credibility and to persuade the consumer to conform to the behaviour of others, that is actually the intrinsic meaning of social proof.



Figure 9: example of Social Proof from Amazon

While these proofs may be true since Amazon is strictly controlled, as previously anticipated social proof patterns are not always truthful but also difficult to be verified. Figure 10 illustrates a fake case of Social Proof on One Travel reported by a user who checked the HTML code related to '38 people are looking at this flight'. The element "view_notification_random" demonstrates that it was a random generated number, rather than true statement based on the previous purchases. It is also worth noting that this interface combines Social Proof pattern and Urgency pattern, by including the timer. Combining patterns is a common practice online, mainly when it comes to similar patterns related to scarcity and urgency that are used together to create a higher perception of low availability and a greater pressure on consumers.



Figure 10: example of fake Social Proof.
Source: https://deceptive.design/hall-of-shame/browse/?f_Type=Fake+social+proof

Obstruction:

This category of dark patterns aims to make a certain action harder for the consumer in order to discourage them from taking it. The main strategy is referred to as Hard to Cancel or Reach Motel, that is mainly employed by subscription services and consists of easing the sign-up process but making it harder to cancel the subscription, particularly in terms of interface design. A well-known case of exploitation of this manipulative pattern regards Amazon, that was sued by Federal Trade Commission in 2023, for subscribing users in Amazon Prime without consent and making the cancellation process excessively difficult. There was a significant difference between the enrolling process and the cancellation before introducing new changes: one or two clicks to subscribe, whereas four pages, six clicks and fifteen options to cancel. In these pages Amazon provided alternative options such as keeping the subscription at a discounted price, opting out the automatic renewal and denying the cancellation (Source: https://www.ftc.gov/news-events/news/press-releases/2023/06/ftc-takes-action-against-amazon-enrolling-consumers-amazon-prime-without-consent-sabotaging-their).

While Prime subscriptions are now more controlled, the cancellation process still requires many passages as illustrated in Figure 11. What stands out in these screenshots is not only the number of passages required, but also the interface designed to persuade the consumer to keep the subscription. In page one, three and four 'Join Prime' texts are central and highlighted in bright blue, as well as the different services included (such as Prime Music, Prime Video, etc.) and repeated reminders of the Prime benefits, such as 'Discover all Prime benefits'. By contrast, the buttons concerning the cancellation are in small text at the end of the page. These design strategies are not immediately intuitive and may dissuade users from unsubscribing, whereas offering a cancellation procedure as quick and easy as the sign-up process would be a more neutral approach.

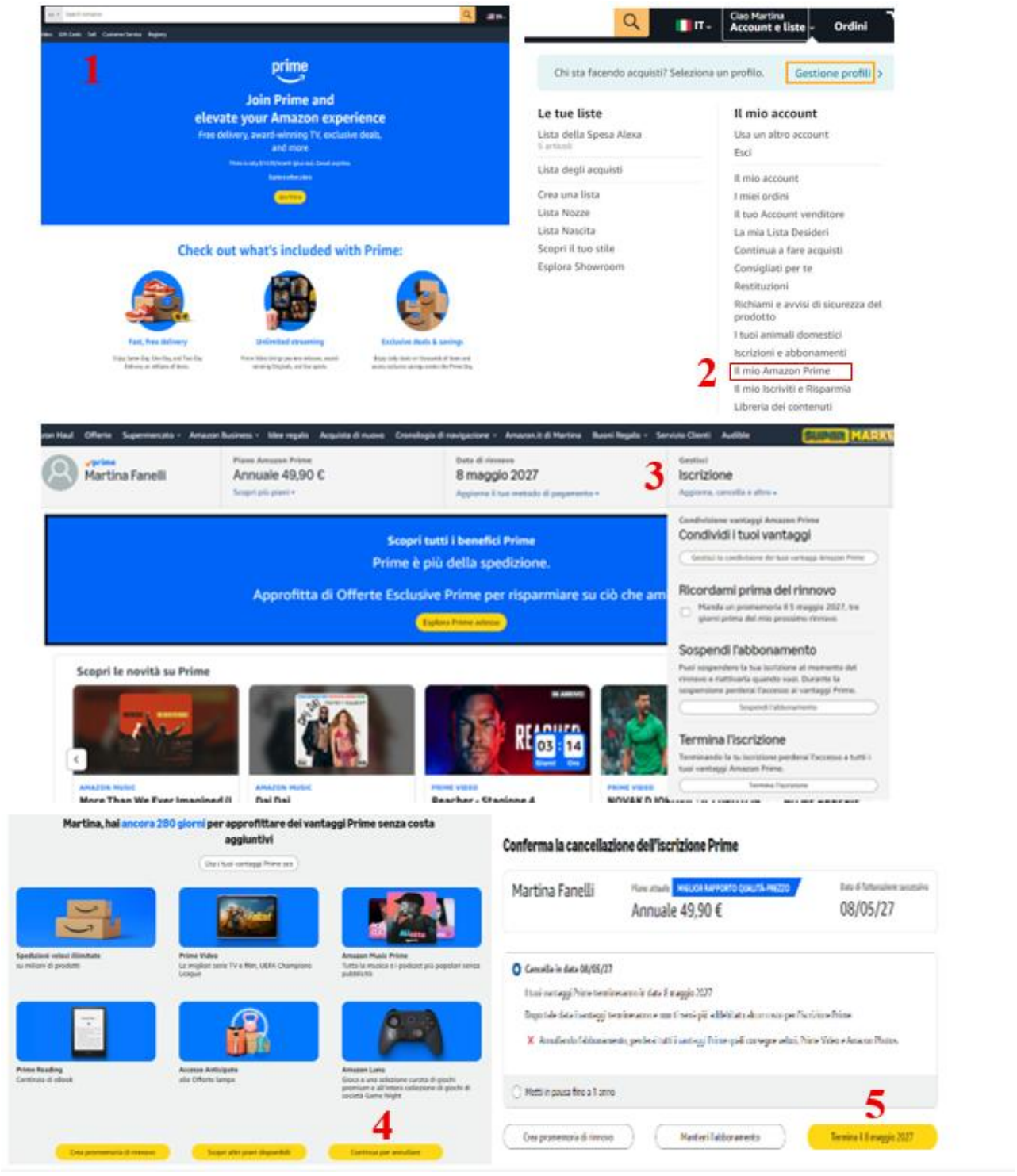


Figure 11: example of Hard to Cancel on Amazon

Forced action

This is the last category of dark patterns proposed by the taxonomy of Mathur et al. (2019) and includes manipulative interfaces that require users to perform a specific action in order to access a functionality. The main patterns are Forced Enrolment and Nagging, the latter of which was introduced by Gray et al. (2018).

On the one hand, forced enrolment compels users to provide personal information or to create an account in order to complete the desired activity, such as purchasing a product, otherwise the website prevents them from purchasing. On the other hand, nagging consists of numerous interruptions while browsing the website requiring users to perform an action that they can apparently refuse. In this case there is not a mandatory action to be taken in order to conclude a purchase, but the pressure caused by insistent interruptions is likely to persuade consumers to finally accept. For instance, Figure 12 shows the two pop-ups appearing repeatedly on Instagram website if you try to see posts and navigate without signing in. If you click the close button, you can still navigate for a while before another similar pop-up appears.

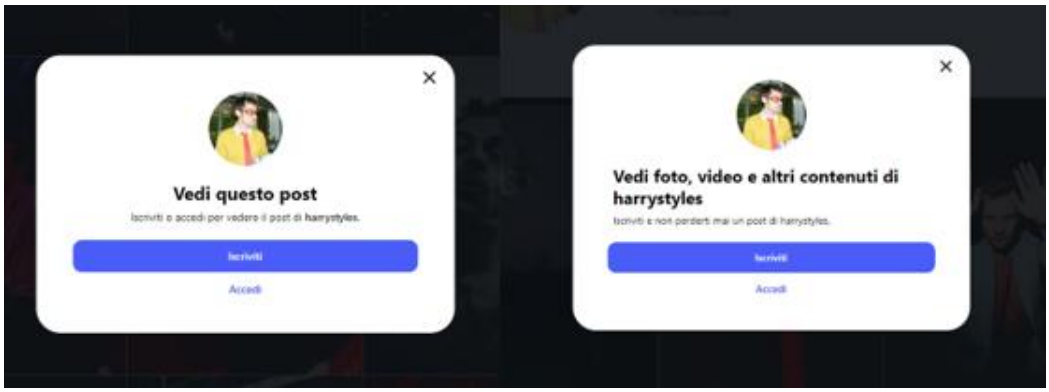


Figure 12: example of Nagging on Instagram

Overall, the taxonomy provided above is intended to make a systematic mapping of dark patterns, although literature offers more categories and provides further contributions to explore unethical practices. Also, this is not a fixed taxonomy since companies and online service providers used to combining different patterns, as explained in the case of the social proof pattern, combined with urgency in Figure 10. Real examples of dark patterns were provided to show typical strategies to which users are exposed, and to raise awareness about what is thought for their interests and what is not. In the next chapter these examples will be considered from the point of view of the behavioural and psychological functioning guiding consumers decision-making processes.

1.4 Beneficial Nudges vs Dark Patterns

Generally speaking, both digital and offline marketing are based on strategies aimed at increasing profits while satisfying consumers' needs. This is why it is fundamental to make a distinction between dark patterns and other strategies that are beneficial to consumers, also known as nudges. According to Zac et al. (2025) two main criteria should be considered to distinguish dark patterns and beneficial nudges, namely the purpose of the pattern and the principal beneficiary. On the one hand, beneficial nudges aim to guide users towards choices that are expected to enhance their long-term welfare and to align with their interests. On the other hand, the main purpose of dark patterns is to deliberately complicate users' decision-making processes in order to serve commercial objectives, even though the final choice may also disadvantage consumers. While the first criterion regards the underlying purpose of the interface design, the second criterion focuses on the final beneficiary of that specific design. Therefore, beneficial nudges act in the consumers' interests, whereas dark patterns are designed to advantage the seller or the provider of the online service.

A similar distinction is proposed by Kollmer & Eckhardt (2022) who compare 'digital bright nudging' and 'digital dark nudging' from an ethical perspective. Digital nudges are considered bright nudges when they follow ethical principles, such as when complex information is presented in a simplified way that is familiar to users, or when they respect users' expectations and present truthful information. For instance, grocery shopping websites often summarise long and complex nutritional information using symbols, colours and simple language, such as the Nutri-Score, a 5-Colour nutrition label widely used to easily compare nutritional value of food products.

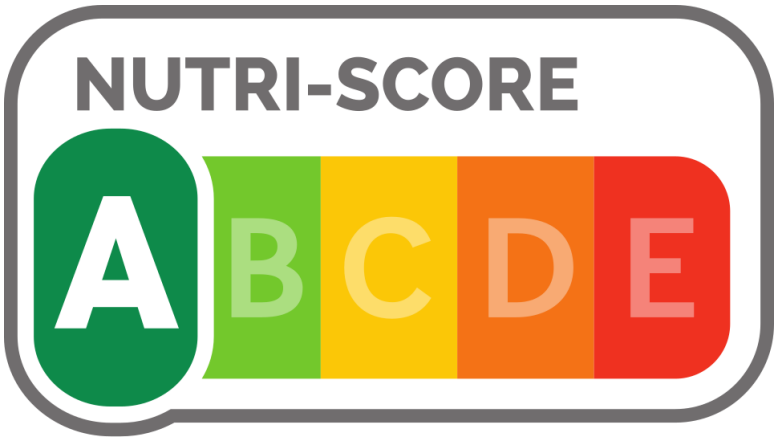


Figure 13: Nutri-Score. Source: Wikimedia Commons

By contrast, nudges that violate these principles and therefore compromise users' decisions are considered digital dark nudges, for instance when presenting complex interfaces that include unnecessary information in order to mislead users' choices.

Finally, Luguri & Strahilevitz (2021) explore this theme stating that dark patterns and beneficial nudges may use similar techniques, but the latter foster prosocial behaviour, such as organ donation or retirement saving. This is a further point of view that highlights that the distinction between dark patterns and beneficial nudges do not necessarily depend on the employed techniques, but mainly on the intentions and on the interests they favour. Overall, it is fundamental to raise awareness on this distinction among consumers since attractive strategies are specific to marketing communication, while it becomes problematic when the purposes do not respect the consumers' interests and free decision-making.

1.5 Who is vulnerable to dark patterns?

Research on dark patterns highlights the importance of consumers' protection from misleading strategies, thus leading to another important question concerning consumers' vulnerability: are all users equally susceptible to dark patterns, or are specific categories more vulnerable than others?

The traditional approach, also known as 'victim approach' (Cole, 2016) states that the effects of dark patterns' harmful manipulation depend on observable and concrete features such as age, income and education. Particularly, first studies claimed that specific consumers groups, for instance elder people or consumers having financial difficulties, may be more vulnerable to dark patterns since these characteristics automatically imply that they are weaker or not completely able to protect their own interests and autonomy online. Consequently, a part of the regulatory debate was based on the idea that these groups needed specific or added protection.

However, over the last few years this approach has been challenged by current studies focused on individual characteristics and psychological aspects that will be explored in detail in the next chapter. For instance, Zac et al. (2025) empirically tested the vulnerability of different groups of users to dark patterns, by conducting an experiment in which participants were exposed to a realistic website specifically designed to include manipulating strategies. As a result, they found that susceptibility to dark patterns is widely uniform among all users regardless of income, age, and education, meaning that everyone may potentially be influenced by dark patterns in their decision-making processes. Moreover, they found a reduction in the effectiveness of these strategies when more than one action was required, whereas a 'single click' or a single option was more likely to encourage users to complete the purchase. Nowadays this is the prevailing shared view among scholars, together with an international consensus about the need to create regulations that control these manipulation techniques and protect consumers.

A further work adopting an individual perspective rather than considering objective factors (e.g. age and income) is Pflügner et al. (2021). They state that individual personality traits influence users' vulnerability to 'technostress', meaning stress caused by the exposure to information and communication technologies, that may lead to effects such as burnout, exhaustion and reduced performance, among others. Moreover, these traits do not act singularly, but rather, they interact to constitute personality profiles that determine stress levels during the interaction with online interfaces. Although this topic will not be explored in detail in the present work, it can be considered an important contribution to literature that studies dark patterns from an individual and psychological perspective. Indeed, in light of what stated above, it may be inferred that users with a higher susceptibility to technostress may also be more vulnerable to manipulative design strategies. In this regard, Kollmer & Eckhardt (2022) recommend further investigation into dark patterns susceptibility depending on users' personality profiles. As for the scope of the present study, this literary overview is useful to understand the shift from an objective and superficial approach to a more individual and psychological approach to dark patterns.

The analysed research has illustrated that consumers vulnerability is highly influenced by two factors that are strictly intertwined. On the one hand, the online environment that is specifically created as a controlled environment, although it seems an open environment made of different options among which consumers can freely decide. On the other hand, marketing strategies that are originally thought as nudges beneficial to consumers have increasingly become manipulation techniques, with the main purpose of maximising the seller's profits at the expense of consumers' interests. In the following chapter we will analyse which cognitive mechanisms are exploited by dark patterns, in order to better understand why these practices are effective.

2. Behavioural Economics: from Heuristics to Cognitive Biases

The present chapter sheds light on the psychological processes intrinsic to human beings that lead most consumers to be vulnerable to dark patterns online. It will first illustrate the transition from the traditional economic model to behavioural economics, which constitutes a fundamental passage to explain such susceptibility. Indeed, economic studies of the 20th century expected people to make rational choices based on their needs, this is why scholars could not explain why real-world choices were actually irrational, meaning far from the expected rational behaviour. The later collaboration between economics and psychology raised awareness about the fact that decision-making processes are often influenced by factors other than rational thinking, such as emotions and cognitive limitations. Building on these studies, the fundamental concepts of heuristics, cognitive biases and the Dual Process Theory will be introduced to explain how they impact individuals' decision-making processes. Finally, these theoretical foundations will be connected to dark patterns, by explaining how they exploit cognitive biases, going back to the examples shown in the first chapter.

2.1 From traditional economics to behavioural economics

The main economic models prevailing in the 20th century, also referred to as 'neoclassical models' (Congiu & Moscati, 2022), considered humans as rational beings who make logical economic decisions based on the Expected Utility Theory. The mathematical formalisation of this theory was developed by John von Neumann and Oskar Morgenstern, even though the theory itself dates back to the 18th century, and aims to demonstrate that a rational decision maker wants to achieve the largest value of an expected utility in a situation of uncertainty. In other words, the rational individual when making decisions ideally evaluates the known and objective risks as well as the likelihood of such risks to occur, taking into account that the results of the choice can be influenced by external factors that cannot be controlled. Building on this idea, not only does the decision maker choose the option with the highest expected utility, but also, they are expected to have coherent preferences and to make choices that are stable over time and independent from the contexts. Moreover, elaboration of information is ideally always optimal, and the final choice is always the best option aimed to maximise their well-being. Thaler & Sunstein (2008) define these strictly rational beings as 'Econs', as compared to 'Humans', real individuals characterised by limited rationality. This distinction reflects the shift from the strict assumptions of the traditional economic model to the more realistic assumptions based on psychological studies.

In the 1980s economists started to question the theories on rational human behaviour, since empirical evidence identified several deviations of human behaviour from neoclassical theories. Such evidence gave rise to new psychological assumptions that started to consider human beings as having bounded rationality and influenced by emotions as well as other unconscious mechanisms. In the 1990s they were named 'behavioural models' (Congiu & Moscati, 2022), according to which consumers always try to satisfy their needs and interests but in a more realistic way, admitting that preferences can change over time and in different contexts, and that information processing is not as easy and optimal as neoclassical theories stated. This transition led to the development of behavioural economics, the new interdisciplinary field in which psychology and economics work together to study the economic decision-making processes of consumers, building on the assumption that people tend to make irrational decisions. Nowadays this field of study is widely accepted not only among academic scholars, but also by the public and the policymakers. In this concern, in the early 2000s behavioural economists realised that economic policies based on the rationality model could not work and opened the debate on the necessity to create new policies taking into consideration that not every economic decision is taken to maximise their welfare. This specific field is referred to as 'behavioural welfare economics', but for the purposes of this study, the chapter will focus on the cognitive aspects of decision-making processes rather than on their application to economic policies.

A parallel field of study based on the behavioural approach is neuroscience, whose application in marketing is called consumer neuroscience (in academic research) or neuromarketing (in industry practice). The main difference between behavioural economics and consumer neuroscience is that the former aims to explain how specific cognitive mechanisms influence consumers' decisions, whereas the latter is a subfield of a science that studies how the brain works from a physical point of view, in terms of development, structure and functions of the nervous system. Particularly, neuroscience is originally a branch of biology, but it has increasingly become an interdisciplinary science collaborating with psychology, linguistics, computer science, among others. This field also evolved from a first traditional phase in which the human organism was defined as a 'black box' (Hubert & Kenning, 2008), meaning that it was not possible to directly access it and gain deeper knowledge. As a consequence, researchers made abstract assumptions based on external factors (such as the expected utility) to explain human behaviour and decisions, but the internal processes were not observable. In recent years advanced technologies and new methodologies allowed for a direct observation of the brain, thereby exploring unconscious and automatic functions of the brain that deeply influence human behaviour. The most interesting techniques are the Functional Magnetic Resonance Imaging (fMRI), that tracks changes in blood flow in the brain, and the Electroencephalography (EEG), that measures electrical activity in the brain and detects immediate reactions (Hubert & Kenning, 2008).

They are significantly useful when applied in marketing, when it comes to studying which parts of an advertisement can be more attractive, creating products' packaging, predicting consumers' reactions to prices or also measuring the unconscious connections that people may create with certain brand messages.

Behavioural economics and consumer neuroscience shed light on two main findings regarding the processes underlying consumers' choices, significantly useful to study digital marketing strategies. Firstly, as previously anticipated, human rationality is inherently limited, therefore decision-making does not always follow a fully rational process, but may rely on simplified thinking patterns. For instance, if consumers were to make fully rational decisions, when faced with the choice between a 200€ Android and a 1000€ iPhone, according to the neoclassical model they would always opt for the cheapest option. However, in real-life thousands of people choose iPhone every day, also who do not actually need it (for work, for example). Several reasons can lead to this choice, for instance the attractive and unique design through which Apple products are presented, the constant exposure to information and advertisements regarding iPhone, as well as the need to belong to a socially privileged group owning an iPhone. As a result, it can be considered an irrational choice since it conflicts with the consumer's economic interest.

Secondly, this type of behaviour is not an exception, but rather, it represents a common feature of human decision-making, as people share similar cognitive skills and limitations that can lead to similar patterns of decision-making. In some cases, the outcomes may be advantageous, while in other cases they may lead to suboptimal choices. Returning to the previous example, these mechanisms help explain why a high number of people worldwide prefer purchasing iPhone rather than cheaper mobile phones, even though the cheaper alternative would be more consistent with the assumptions of the neoclassical model. As bounded rationality is intrinsic to humans, the behavioural economist Dan Ariely (2008) introduced the concept of predictable irrationality, meaning that irrational behaviour, conceived as deviations from neoclassical economic models, is systematic rather than exception in real-life decision-making, thus playing a crucial role in predicting consumers' choices.

This new economic perspective is inconsistent with the neoclassical model, which tried to predict consumers' behaviour assuming that they always make rational decisions. However, observable evidence actually showed deviations from this theory, and demonstrated that irrational behaviour also follows predictable patterns. As explored in the next paragraph, such predictable behaviour became a fundamental basis for developing successful marketing strategies but also started to be exploited to create manipulative patterns with the aim of maximising business profits at the expense of the consumer welfare.

2.2 The Dual Process Theory, Heuristics and Cognitive Biases

In order to understand how individuals make decisions that sometimes result in systematic errors, it is necessary to build on a foundational theory in cognitive psychology and behavioural economics, namely the Dual Process Theory by Tversky and Kahneman (1974). The psychologists demonstrated through several experiments that people usually rely on mental shortcuts, known as heuristics, to make fast decisions in situations of uncertainty, instead of engaging in deliberate thinking and making statistical and probabilistic calculations. One of the main experiments consisted of evaluating whether people consider the real base rate frequency of an event, meaning the probability of an event occurring, when making judgments in an uncertain situation. Participants were provided with personality profiles of a few individuals, who were part of a group of one hundred professionals formed by engineers and lawyers, and they had to decide whether the profile belonged to an engineer or to a lawyer. While most profiles provided details identifying an engineer or a lawyer, Tversky and Kahneman created a personality profile completely neutral and with no useful information to decide the job. The profile stated:

> *Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.* (Tversky & Kahneman, 1974)

Since no relevant information was provided in this description, deciding the probability of Dick being an engineer or a lawyer should have been based only on the proportion of the two groups. In this concern, the researchers divided the participants into who was informed that the sample was formed by 70 engineers and 30 lawyers, and who was informed that the sample was formed by 30 engineers and 70 lawyers. The findings showed a violation of probability laws: both groups estimated roughly a 50% probability of Dick being an engineer, ignoring the proportions in the sample. Tversky and Kahneman concluded that people rely on probability laws when they have no other information, while when dealing with descriptions, although irrelevant, they ignore prior probabilities and focus their decision on the profile similarity to the stereotype represented by a certain group. In other words, according to participants a skilled and successful young man is more likely to be engineer even though statistical evidence suggests otherwise. This experiment empirically explains what will be later called Representativeness heuristic, namely the mental shortcut used to decide whether an individual belongs to a specific group basing this evaluation on the extent to which the subject is representative of that group, meaning how close they are to the group stereotype.

This and other experiments conducted by the two researchers constitute the foundation of the Dual Process Theory, that was developed later in 2011 by Daniel Kahneman. The theory suggests that the brain is divided into two cognitive systems guiding the decision-making processes. System 1 is fast, automatic, intuitive, emotional and effortless; it activates in situations of uncertainty and emergency, when there is no time for deliberate reasoning. For instance, System 1 enables us to slam the brakes when a car suddenly cuts us off: this action reflects a quick and instinctive reaction since there is no time to analytically think about what to do. By contrast, System 2 is slower, effortful, deliberate and logical; it activates in situations that allow to take time and make a logical choice. System 1 is therefore related to irrational decisions, whereas System 2 leads to rational decisions. As for the experiment of Tversky and Kahneman (1974), it can be concluded that participants activate System 2 when there are no available descriptions and have to make statistical calculations, whereas they activate System 1 when relying on descriptions that represent a particular stereotype, since it is the fastest and more intuitive option. It is important to highlight that both cognitive systems are necessary for humans' survival. Indeed, it would be unrealistic to expect people to deliberately think in every kind of choice they have to make. The brain always tries to minimise cognitive effort mainly in situations of uncertainty, and this is an important skill in everyday life that allows people to make rapid decisions without processing all available information.

Building on this theory, System 1 is the cognitive system that relies on heuristics, namely mental shortcuts employed to make fast and irrational decisions in order to reduce cognitive effort. While sometimes they are useful, in other cases they can lead to disadvantageous choices for the individual. These errors are known as cognitive biases and are defined as systematic and predictable errors resulting from heuristics. Congiu and Moscati (2022) also define biases as 'rationality failures', referring to systematic deviations from the strict neoclassical rationality models. Literature has identified several types of cognitive biases that will be explored at the end of this chapter. However, it is important to distinguish the two major categories that include further subtypes.

Cognitive biases can be divided into cold biases and hot (or motivated) biases. On the one hand, cold biases are caused by intrinsic limitations in human information processing, meaning that we tend to use heuristics due to the limited capacity of the cognitive system to handle complex information, but there is no influence of personal feelings. This is the case of the Representativeness heuristic, that leads the brain to opt for the option that seems to be the most representative of a particular group, avoiding a more complex and deliberate thinking or calculations. On the other hand, hot or motivated biases reflect feelings and instincts. In this case there is not a limitation in information processing, but rather a choice influenced by the individual's needs, desires or personal motives.

For instance, consumers may purchase an expensive product over a cheaper but similar alternative because they are motivated by positive feelings, such as esteem or a sense of connection, towards the person promoting the product in an advertisement. This categorisation derives from a theory developed in the 1960's by the psychologist Robert Abelson, who coined the term 'hot cognition' for the first time to explain that thinking is deeply influenced by emotional states, wishes and social attitudes rather than cold logic. This strong distinction was later challenged by the more recent approach of Ziva Kunda (1990), who stated that motivational factors and rational thinking are not isolated processes, but rather, personal motivations can influence rational cognitive processes. In other words, emotions do not interrupt or replace rational thinking, but can influence information processing and interpretation. This is a fundamental point since, although biases can be classified depending on whether they are more influenced by emotions or by information processing, these categories are not mutually exclusive. Cognitive biases do not follow strict rules and may also be neutral, without a clear cognitive or emotional influence.

Overall, the following schematisation is useful to summarise the concepts discussed in this section:

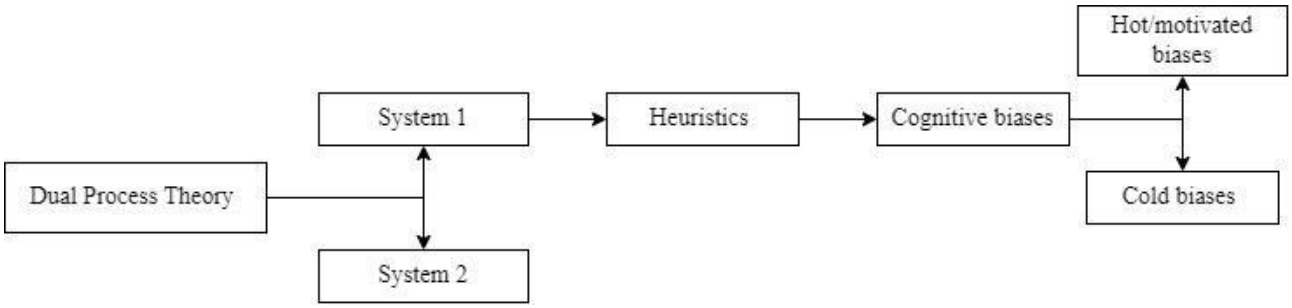


Figure 14: from Dual Process Theory to cognitive biases

In the next paragraph we will reintroduce dark patterns from the perspective of cognitive processes starting from the difference between nudges and sludges, to understand how marketing strategies originally thought to help consumers can become misleading techniques.

2.3 Nudges and Sludges

The first chapter introduced the opposition between beneficial nudges and dark patterns, highlighting that the aim of marketing strategies is not necessarily to harm consumers, but depends on the purpose and on the beneficiary of the employed strategy. In order to explore this distinction in terms of their effects on decision-making, it is important to illustrate the original definition of nudge in behavioural economics, its application to digital marketplace and the opposition between nudging and sludging.

The original definition of nudge was provided by Thaler & Sunstein (2008), thus giving a significant contribution to the field of behavioural economics. The nudge is described as "any aspect in the choice architecture that alters people's behaviour in a predictable way without forbidding any option or significantly changing their economic incentives" (Thaler & Sunstein, 2008). The main interesting points are the reference to the choice architecture, namely the structure and presentation of choices, and the predictability element. In other words, the way in which information is presented, and any slight modification can significantly impact people's choices. Moreover, such choice architecture is not casual, but rather, it is based on previsions and expectations from consumer behaviour. This predictability in marketing strategies is possible thanks to behavioural economics' studies and the transition from rational to irrational economic model, discussed in 2.1. The studies on cognitive processes in consumer decision-making paved the way for the creation of techniques that aim to activate System 1, so that people make fast and effortless decisions, encouraged by several factors that generate a stimulus which is not completely rational but guided by emotions or cognitive limitations.

This process has increasingly become the basis of the digital marketplace. In this regard, Weinmann et al. (2016) contextualised nudges in the online environment, redefining 'digital nudging' as "the use of user-interface design elements to guide people's behaviour in digital choice environments". The key points addressed in this research are the absence of neutrality and the exploitation of human bounded rationality. Firstly, since human decisions only partially depend on rational thinking, but are mostly influenced by how information is presented, this implies that there is not a universal and objective way to present information. Any kind of choice architecture in the interface design will inevitably influence consumers, voluntarily or unintentionally. Moreover, digital nudging is based on the concept originally introduced by behavioural economics according to which people act guided by bounded rationality, they therefore rely on cognitive System 1 by using heuristics that can influence positively or negatively the decision-making. This idea was not originally conceived to undermine consumers interests, and this is explicitly stated by Thaler & Sunstein (2008) when saying that nudges alter behaviour "without forbidding any option or significantly changing their economic incentives".

Therefore, nudges are thought to help consumers make beneficial choices, at the same time respecting their autonomy and taking into consideration the unconscious cognitive processes.

However, service providers and companies online started to exploit these acknowledgments in psychological studies to employ unethical nudges. Weinmann et al. (2016) presented a set of nudging principles based on Thaler et al. (2010), that can be applied in the choice architecture, and that can take on a positive or negative connotation, depending on the purpose and on the beneficiary ('digital bright nudging' or 'digital dark nudging', see 1.4). For the scope of the present work only a few will be cited in order to understand that overall nudges exploit heuristics and cognitive biases but they become negative practices when the main objective is to benefit the company at the expense of the consumer's interests. For instance, the *incentive* principle consists of making incentives more visible and salient. Indeed, when benefits or drawbacks of a certain choice are obscured to the user, this may lead to suboptimal decisions. By contrast, if incentives are made explicit, the user can make more conscious choices. This principle can have a positive connotation when aimed to help users, but also a negative connotation when fake incentives are shown. For instance, displaying the real-time cost of a phone call is a beneficial nudge, whereas hiding costs or using a fake countdown timer on a product on sale is a dark pattern. Displayed costs advantage the user because it shows real costs and therefore, they can decide either to continue the call being aware of the costs or keep the call short. On the other hand, fake timers benefit the company since they encourage consumers to buy quickly based on fake information. From the point of view of cognitive processes both options engage System 1. In the first case, the cognitive effort of making calculations of the call's final cost is bypassed in favour of displaying the costs in real time. This enables the user to make an aware choice based on whether to continue or end the call, potentially allowing them to reduce the final cost. By contrast, in the second case the same cognitive bias is triggered but with the objective of encouraging people to purchase quickly, maybe leading them to act differently from their original intentions.

Moreover, the *understanding mapping* principle involves presenting originally complex data in a more intuitive or familiar way to help users understand what is actually useful for them. It is beneficial for consumers when it aims to help them make informed decisions, but it can become a dark pattern when information is intentionally presented in a way that an option seems to be more convenient than what it actually is, and at the same time when really useful information is hidden. For instance, when advertising a digital camera, mapping megapixels (the complex information) to maximum printable size (familiar information) can be considered a beneficial nudge since it makes information clearer for the user. By contrast, when making an annual subscription, if price is displayed as "0,99€ per day" instead of "annual subscription: 360€", the salient information becomes the low price per day instead of the real cost that the consumer will bear, which may be hidden or written in small font.

The first case activates System 1 by avoiding the cognitive effort of trying to process complex information, whereas the second case still activates System 1, but it is more likely to trigger the anchoring effect on the low initial price (see 2.4.1), thus potentially leading the consumer to base their purchase decisions on a misleading information.

In more recent studies, Thaler (2018) denounced negative nudges in the broader context of public policies and coined the new term 'sludges'. Sunstein (2019) provided a more precise definition, describing them as

> *nudges that induce excessive or unjustified friction and cost time or money; make life difficult to navigate; may be frustrating, stigmatizing or humiliating; and that might end up depriving people of access to important goods, opportunities and services.*
> (Sunstein, 2019)

In this case the idea of creating obstacles complicating people's life is related to all bureaucratic processes that are required to access a certain benefit, to which citizens should be entitled (such as welfare benefits, healthcare, etc.). Therefore, the sludge may have a double effect: it requires excessive cognitive effort, potentially causing psychological stress and frustration, while also increasing the likelihood that people will give up on obtaining the benefit.

Kollmer & Eckhardt (2022) adapted this approach to the digital context, redefining the digital sludges as an excessive friction or obstacle that hinders people from performing an action or activity online. They introduced the opposition volitional user agency – non-volitional user agency to explain the psychological mechanisms exploited by dark patterns to manipulate consumers. Volitional user agency refers to the clear, intentional and conscious intention of the user to complete a certain action, such as to cancel a subscription. This dimension is impeded by digital sludges that intentionally insert unnecessary passages such as repeated pop-up menus or extra waiting times. For instance, this is the case of Amazon Prime that requires several non-intuitive passages to cancel a subscription (see 1.3.1). Sludges do not deny or prevent users from taking an action, but rather, they intentionally exploit the excessive cognitive effort required to discourage users from proceeding.

By contrast, non-volitional user agency refers to the prior phase and in which the user is still evaluating the available options and has not decided yet. In this case digital nudges do not impact an already established intention, but rather influence the decision-making process itself. For instance, countdown timers affect this dimension: when the consumer has not yet decided whether to purchase a product or not, the timer may trigger loss aversion (see 2.4.4), thus creating a sense of urgency that leads to a fast decision before engaging in deliberate thinking. The following table summarises the main differences between nudges and sludges:

	Nudges	Sludges
Original definitions	"Any aspect in the choice architecture that alters people's behaviour in a predictable way without forbidding any option or significantly changing their economic incentives" (Thaler & Sunstein, 2008)	"Nudges that induce excessive or unjustified friction and cost time or money; make life difficult to navigate; may be frustrating, stigmatizing or humiliating; and that might end up depriving people of access to important goods, opportunities and services." (Sunstein, 2019)
Digital nudges vs digital sludges	"the use of user-interface design elements to guide people's behaviour in digital choice environments" (Weinmann et al., 2016)	Excessive friction or obstacle that hinders people from performing an action or activity online. (Kollmer & Eckhardt, 2022)
Type of user's agency (Kollmer & Eckhardt, 2022)	Exploitation of non-volitional user agency by influencing the decision-making process itself.	Exploitation of volitional user agency by impeding the performance of an action that has already been decided.
Examples	Countdown timers on limited offers.	Complex process of cancelling a subscription (Hard to Cancel).

Table 2: differences between nudges and sludges

These two concepts provide a fundamental conceptual basis for understanding dark patterns and cognitive processes they exploit. The next section will be dedicated to a series of cognitive biases documented in behavioural economics literature. They will be explained in great detail and will be linked to the examples of dark patterns presented in the first chapter. Also, they will be categorised based on the following features:

- Cold bias, hot/motivated biases or neutral biases
- Biases exploited by manipulation techniques or by deception techniques (Kollmer & Eckhardt, 2022)
- Attributes of dark patterns (Mathur et al., 2019) (see 1.3).

2.4 Cognitive Biases

As previously discussed, cognitive biases are systematic and predictable errors caused by humans' bounded rationality, and result from heuristics, namely mental shortcuts that allow people to make rapid and effortless decisions. This section will explore the cognitive biases that are exploited by the dark patterns described in Chapter 1 (see 1.3.1), by re-analysing the images from the point of view of the exploited cognitive process that may be activated when exposed to that particular interface. In order to make an ordered categorisation of cognitive biases they will be described taking into account the distinction between cold biases, hot/motivated biases and neutral biases (see 2.2), the taxonomy of attributes of dark patterns (see 1.3) and the difference between manipulation and deception techniques employed by dark patterns (Kollmer & Eckhardt, 2022). The main difference between the two techniques is based on the nature of provided information. Manipulation techniques present realistic information but manipulate the way in which they are presented to exploit human cognitive limitations, whereas deception techniques present false information or omit information that may be fundamental for the consumer's final choice. This analysis will show that biases do not follow a fixed categorisation, but each dark pattern is able to exploit multiple biases. The taxonomy of cognitive biases follows the one proposed by Mathur et al. (2019), to which further concepts will be added.

2.4.1 Anchoring bias

Anchoring bias involves people's tendency to place disproportionate weight on the first piece of information available when making choices. This initial value to which people focus attention is called *anchor,* and becomes the reference point according to which further information is evaluated. Instead of activating System 2 through which all available information can be evaluated accurately, it is more likely that System 1 operates to make a choice that is influenced by the anchor. This occurs because the anchor shapes early and fast hypotheses and limits the integration of information that comes after, thereby reducing objectivity of the final choice.

This concept was first introduced by Tversky & Kahneman (1974), who proposed the anchor-and-adjust hypothesis: individuals make estimates or predictions starting from an initial value (the anchor), then they adjust from this point when incorporating new information. The problem is that adjustments are typically insufficient, leading to judgements or decisions too close to the anchor.

Anchoring bias can be strategically leveraged in digital marketing since it is particularly effective on pricing, namely the way in which one or more prices are presented, that can impact consumers' perception. For instance, if consumers first encounter a product at its original and non-discounted price, it becomes the anchor and any subsequent discount will be perceived as particularly attractive. By contrast, if the initial exposure is to a discounted price, the standard price will be judged as unreasonably high.

This mechanism is strictly connected to a similar effect, known as 'decoy effect' (Huber et al., 1982), that is frequently exploited by subscription services such as Netflix, Spotify, etc., when proposing their plans. They usually present three options as follows:

- Basic Plan: $5.99/month (Limited features)
- Standard Plan: $11.99/month (Popular choice)
- Premium Plan: $18.99/month (Extra features but costly)

When exposed to this kind of interface, the most expensive plan usually serves as the decoy to encourage the consumer to opt for the standard plan. Companies do not expect them to choose the premium plan since it is not convenient, but it is intended to serve as the evaluation measure to finally choose the second plan. On the contrary, if only the basic and standard plans were proposed, the latter would be the decoy and, consequently, the consumer may opt for the cheapest one, thus resulting as not beneficial for the company. The main difference between the anchoring effect and the decoy effect is that the anchor influences judgement by using a single value before being exposed to subsequent values, whereas the decoy impacts the final choice between a set of options proposed together, and one of them serves as reference to guide the choice towards the mid-tier option. Building on the taxonomy of dark patterns in the first chapter, the following dark patterns exploit this bias:

- **Hidden costs (Figure 2):** in the Deliveroo example, the first value to which consumers are exposed is the discounted price compared to the original price (11,18€ instead of 14,90€). It serves as the anchor to attract the highest number of consumers at the expense of other restaurants offering similar products, and to evaluate the subsequent final price adding extra costs. When exposed to the final price, the consumer has already spent time and invested mental effort into the process of finding a restaurant, choosing the product, etc. The evaluation therefore focuses on the extent to which the price has changed from the anchor, rather than evaluating the price in the absence of the anchor. In other words, in the absence of an anchor the consumer would evaluate whether 12.37€ is worth the product, whereas when taking the anchor as the reference value they are more likely to evaluate if it is worth changing the restaurant considering only a slight difference between the anchor and the final price. This effect is typically effective since the brain always tends to reduce cognitive effort.
- **Pressured selling:** considering the example of the purchase of an iPhone in which the service offers additional items, the initial value is the price of the product that serves as anchor from which the consumer evaluates the final cost at the checkout. For instance, if the anchor is 1099€ and the consumer is offered a 30€ charger, they will tend to consider the charger's price as a low additional price compared to the value of the mobile phone. By contrast, if the purchase of the charger alone was evaluated, the price would be much

more relevant to the purchasing choice. This is why pressured selling patterns can leverage the anchoring effect: in the context of high-priced purchase, the price of additional products appears less costly and therefore more acceptable to consumers.

Overall, the anchoring bias can be characterised according to the following features:

- **Nature of the bias**: the anchoring bias can be considered a cold bias since it involves judgement and information processing, while the emotional sphere is not involved.
- **Dark patterns' techniques**: dark patterns exploiting the anchoring bias are mainly manipulative as the prices are real, but are framed in a way so as to manipulate the consumer towards the choice that is most beneficial for the company.
- **Dark patterns' attributes** (Mathur et al., 2019): the patterns exploiting this bias are generally covert, since the options are subtly presented to intentionally steer consumers towards a specific choice even though they believe their autonomy is not influenced.

2.4.2 Framing bias

The framing bias represents the way in which people make different decisions building on the same information but depending on how options are presented. This is a broader concept since information may be presented with positive or negative connotations, but also showing a piece of information as more salient than another one. The original conceptualisation of the framing effect was provided by Tversky and Kahneman in their *Asian disease problem* (1981), in which they observed that people change their risk perception depending on how information is presented. People present risk aversion towards what is presented as gain, whereas they are willing to take risks when information is presented as loss. The experiment consisted of providing participants with different programmes aimed at saving people from an unusual disease that would kill 600 people. This programme was exposed in two different ways to the two groups. The first group received the positive frame (gain approach), that included two options: saving 200 people for sure, versus a one-third probability of saving everyone and a two-thirds probability of saving none. In this case most participants chose the first option because it is the certain option. The second group received the negative frame (loss approach), divided into two options: 400 people would die for sure, versus one-third probability to save everyone and a two-thirds that everyone would die. In probabilistic terms the two frames are equal, but in the negative approach participants preferred the risky option since in this case the certainty implied losses. Therefore, Tversky and Kahneman demonstrated that people's decision-making process in uncertain situations is not typically based on deliberate thinking or on probabilistic calculations, but rather on their perception of information depending on whether it is framed in a positive or negative way.

This also suggests that people are influenced by ambiguity aversion, meaning they prefer known or positive risk over unknown or negative risk, and reflects the brain's function of reducing cognitive effort by avoiding deliberate reasoning under excessive uncertainty.

The framing effect is widely leveraged in digital marketing; indeed, marketers' strategies consist of presenting information by reducing uncertainty, thus making products more appealing to hesitant consumers, and of highlighting positive aspects avoiding negative ones.

The main dark patterns exploiting the framing bias are the following:

- **Confirmshaming (Figure 6)**: it involves intentionally choosing specific words to steer the consumer towards a certain choice. This technique is generally effective since it triggers irrational processes intrinsic to human behaviour. For instance, in Figure 6 the service decides to add 'I'm willing to risk my flight' rather than leaving the neutral 'No', thus probably eliciting emotions of shame, fear of losing the flight and personal responsibility. The pattern is exploiting the affect heuristic, namely people's reliance on feelings or instincts rather than concrete information when making decisions. This also happened in the experiment by Tversky and Kahneman (1981): their specific choice of wording guided the perception of the participants even though the outcomes were the same for both programmes. This is explained by the fact that people typically think of something positive as less risky and more beneficial, and something negative as more risky and less beneficial, but this evaluation is only based on their perception activated by System 1, rather than on System 2 critical reasoning.
- **Visual interference (Figure 7):** this pattern includes different techniques aimed at guiding consumers' attention towards a specific choice, either through visual elements such as contrasting colours or font styles, or through linguistic choices aimed at presenting only positive information. The framing bias is exploited by the latter type of visual interference, whereas the former one will be discussed when exploring the default bias (see 2.4.5). Figure 7 shows a cookie banner that provides only positive information, ranging from the title 'We value your privacy', to expressions such as 'This is to provide an improved experience' to justify personal data collection. The positive tone of the message is conveyed by a statement of care towards the user, but what is actually stated is that in certain cases partners are allowed to collect personal data even though you did not give consent. The mechanism is similar to the one exploited by confirmshaming: this kind of message can be perceived as positive and therefore less risky to accept all cookies. Rather, if the tone of the message was more neutral or negative, by using for instance the title 'Data collection and cookie preferences' or by stating that "if you do not 'Accept all', your data

may still be shared…", the user would be more likely to perceive the option negatively and would evaluate the 'Accept all' option as riskier.

In light of what discussed above, the framing bias can be characterised according to the following features:

- **Nature of the bias**: the framing effect itself may be considered a cold bias since it is closely related to the way in which information is presented, which in turn influences information processing. However, in the case of Confirmshaming pattern, the affect heuristic can lead to a motivated bias as the emotional sphere plays a key role in shaping the purchasing choice.
- **Dark patterns' techniques**: these patterns are mainly manipulative since the provided information is real, but is manipulated to highlight the positive aspects or less risky aspects, while negative information can be inferred only by activating cognitive System 2.
- **Dark patterns' attributes**: Visual Interference may be considered covert, since the user may not be aware of such manipulation of information and may believe to be exposed to neutral information. At the same time both Visual Interference and Confirmshaming are considered asymmetric, since the emphasis is imbalanced towards one option at the expense of the alternative one. By contrast, Confirmshaming is not covert because the wording is visible (Mathur et al., 2019), although users may not always recognise the manipulative intention behind it.

2.4.3 Bandwagon effect

The Bandwagon effect, or Herd effect, refers to people's tendency to follow the crowd, meaning to conform to the judgement of the majority. Although their own choice is different from the choice of the majority, they tend to question their decision for this reason. This phenomenon is widely studied in psychology, philosophy and behavioural economics and the main causes underlying this behaviour are considered to be social education, fear and uncertainty, among others. Keynes studied this bias even before the advent of rational expectations theory, particularly in the context of investments and financial instability (Keynes, 1936). He explained that financial instability in stock markets is the outcome of the sociological and psychological forces that dominate in uncertain times. These forces are not only the propensity to consume or the desire to hold money, but also waves of optimism and pessimism among investors. As for sociological forces, the economist noticed that investors tend to believe what others believe and to do what others do. He compared this mutual influence to the example of a newspaper beauty contest where readers have to pick the six prettiest faces among one hundred images, but the winning rule is based on selections that most reflect the average preferences of all readers. In other words, they base their decisions on what they believe others expect, rather than on their own preferences. This conceptual analogy refers to investors' tendency to buy stocks by

trying to guess what others will buy, rather than buying assets with stronger economic value. Consequently, stock prices can fluctuate due to human unconscious cognitive processes rather than to real-world performance.

In the digital marketing context, this behaviour takes the form of mutual influence among consumers who make purchasing choices following others' decisions. Feedback about products play a crucial role in guiding consumers, since the greater the number of consumers choosing and positively reviewing certain products, the more likely other consumers are to follow the same choice. The dark pattern that most exploits this bias is **Social Proof (Figure 9-10)**. Small text labels such as 'X people bought in the past month' or 'X people are looking at this flight' are intentionally inserted to trigger different heuristics. Firstly, the consumer may decide to purchase the product because many people have previously chosen it, and this decision may be influenced by the belief that the more consumers buy the product, the more likely the product is trustworthy and of high quality. Another related process is the Veblen effect, namely people's need to belong to a group or community and to create a social identity related to a product. The economist Thorstein Veblen studied this effect in the context of the luxury market, and it can explain the choice shared by a huge number of people to purchase an iPhone instead of a cheaper Android (see 2.1).

As for Figure 9, the consumer may decide to buy the same laptop that other 300 people had already bought due to an unconscious desire to belong to that group of gamers. By contrast, Figure 10 shows a combination of dark patterns that exploit other kinds of heuristics related to scarcity and urgency which will be explained in greater detail when describing Scarcity bias (see 2.4.4). However, the indication of the number of people looking at the flight may also be interpreted as a sign of popularity of that specific airline or of the destination, thus encouraging the consumer to purchase the ticket.

Overall, the Bandwagon Effect can be characterised according to the following features:

- **Nature of the bias**: this bias can be considered a hot/motivated bias, because people's tendency to follow the majority stems from the need to belong to a group, which is an inner and emotionally driven motivation rather than a rational evaluation.
- **Dark patterns' techniques**: the patterns exploiting the bandwagon effect may be both manipulative and deceptive, depending on whether the provided information is real or false. They are manipulative when the reported number is accurate but is intentionally shown to activate cognitive System 1, therefore the decision is influenced by unconscious and not rational processes rather than deliberate evaluation of the product considering the characteristics, quality, etc. By contrast, they are deceptive when the number is randomly generated and therefore the decision is based on fake information, thus creating a false belief about the product's popularity.

- **Dark patterns' attributes**: these patterns can be considered covert, as consumers are generally unaware that the way information is presented is intentionally aimed to leverage this bias.

2.4.4 Scarcity bias

Scarcity bias refers to people's tendency to place higher value on a product when they perceive limited availability, and it is one of the most exploited biases in digital marketing. This effect is closely related to other heuristics that can be triggered at the same time when exposed to certain marketing strategies or dark patterns. First, people typically feel compelled to act quickly not only when a product has low availability, but also without a solid rationale. Indeed, it is intrinsic to human beings to favour action over inaction even though there is no evidence that the action will lead to better outcomes, since inaction usually makes us feel uncomfortable. This has been identified as an evolutionary instinct, also known as Action Bias (Patt & Zeckhauser, 2000), caused by the need to feel safe (i.e. doing something to protect yourself) as well as to feel socially rewarded. Indeed, in society people tend to perceive others more positively when taking actions than when remaining inactive, which may be interpreted as a sign of laziness or carelessness.

Moreover, this irrational behaviour is significantly emphasised by people's typical loss aversion or fear of missed opportunity. In other words, people are uncomfortable with the feeling of regret and the thought that something should have been done. In this case the fear is usually related to missing the opportunity to purchase the discounted product. At the same time, being aware of the low availability of a product may also increase the desire to purchase it because scarcity may be interpreted as an indicator of good quality. In digital marketing, highlighted scarcity messages, timers and limited offers are all techniques aimed at encouraging immediate action by triggering the fear of missing out. This is the fast and effortless decision made by cognitive System 1, whereas System 2 would involve a slower evaluation of other similar options rather than an immediate purchase.

As previously anticipated, the dark pattern shown in Figure 10 clearly exploits the scarcity bias related to the action bias and the loss aversion. Not only do they show the timer in big and highlighted font, but also, the sentence 'Book now before tickets run out!' contributes to creating a higher sense of urgency that may lead the consumer to buy without rationally evaluating the choice. Further examples showing urgency and scarcity dark patterns can be observed in Figures 3, 4, and 5. In particular, in Figure 3 the strategy is focused on time limitation, whereas Figures 4 and 5 are focused on limited product availability. In both cases scarcity bias can be triggered, but the slight difference may also tend to activate different unconscious responses. While limited-time scarcity tends to generate a stronger sense of urgency to act, limited-quantity scarcity may also trigger the bandwagon effect, since desirability towards a product may be increased by the fact that in that specific moment

other people are purchasing it. Overall, the scarcity bias together with action bias and loss aversion can be characterised as follows:

- **Nature of the biases:** these biases can be considered mainly hot biases as the sense of urgency perceived by the consumers is an instinctive response guided by the emotional sphere, even though, as with all other biases, this also reflects the broader need of the brain to reduce mental effort.
- **Dark patterns' techniques:** dark patterns exploiting these biases may employ either manipulation techniques or deceptive techniques. As in previous cases, they can be manipulative when the provided information is truthful, but it is deliberately designed to encourage consumers to act quickly. For instance, real low stock is shown in a manipulative way when highlighted in red or accompanied by messages emphasising scarcity. By contrast, they can become deceptive when information is false and randomly generated, as in the case of a fake countdown timer, or when a low-availability message does not represent the actual stock level. This practice may be revealed by refreshing the page once the timer has expired: if the same offer does not switch to the original non-discounted price, it is likely that the time limitation was not real.
- **Dark patterns' attributes:** these patterns can be classified as mainly covert since they are designed to subtly limit the user's autonomy, but also deceptive when information is misleading.

2.4.5 The Default Effect

The default effect, also known as status quo bias, indicates people's tendency to passively accept options that are preselected by default due to inertia (Mathur et al., 2019), rather than actively evaluating other options. This concept was first introduced by Thaler et al. (2010) when presenting the nudging principles at the basis of the choice architecture to guide users towards advantageous purchasing decisions. Later it was further developed in behavioural economics literature providing real-life examples. For instance, putting healthy food at eye level in a cafeteria is a basic default technique to steer the client towards choosing that product. Indeed, default techniques do not only consist of preselecting an option, but overall aim to reduce cognitive effort by providing the most 'comfortable' option. In the cafeteria example, the default option is the first one to capture the client's attention and may discourage them from looking for other less visible alternatives. Moreover, it may also allow for reduced physical effort: putting healthy food at eye level and unhealthy food in less comfortable positions may lead clients to choose healthy food since it is easier to reach.

Another famous example involves organ donation consents (Johnson & Goldstein, 2003). They compared organ donation consent rates in different countries and demonstrated that the differences are not related to cultural, religious or economic factors, but rather to which option is preselected as the default option. They identified the 'Opt-in' System, in which the person is non-donor by default, and the 'Opt-out' System, in which the person is donor by default. In the former case, people are required to actively decide to be organ donors, meaning they have to fill in a form and also intentionally think of accidental death scenarios. From a psychological perspective, this involves a significant cognitive effort as well as emotional discomfort. In the latter case, the individual is automatically considered as an organ donor and is required to fill in a form if they choose to refuse. Consequently, opting-in is easier than opting-out, which implies higher cognitive effort. This study revealed that countries with the 'Opt-in' System have low consent rates (e.g. Germany and the United Kingdom, average 15% consent rate), whereas countries with the 'Opt-out' System have much higher consent rates (e.g. Austria and France, average 98% consent rate).

In the digital marketing context, the default effect is typically exploited by interfaces that provide a set of options among which one of them is preselected or highlighted, so that users should actively deselect and explore in detail the other options. The following dark patterns illustrated in Chapter 1 typically exploit the default effect:

- **Sneak into basket (Figure 1):** in the old Ryanair website version the flight insurance was automatically added to cart by default and the consumer had to follow a series of passages to remove it, including opening the drop-down menu and scrolling through the list of countries to find the 'Don't insure me' option. This is comparable to the 'opt-out' system: since changing the option is more effortful (in terms of too many passages or too complex to remove it), the consumer may decide to keep the default option.
- **Visual interference (Figure 7-8):** This pattern has been already introduced in 2.4.2, analysing Figure 7 as an example of a pattern exploiting the framing bias. In this case this dark pattern is explored from the point of view of elements that immediately catch the eye thanks to their visual characteristics. In Figure 8, the 'Pay' button is the first element attracting attention since the white colour contrasts the black background, it may therefore lead users to click without reading other information that may be potentially important for the purchase, such as the small hidden text 'upgrades cannot be refunded'. The higher visibility creates the same effect of putting healthy food at eye level in a cafeteria: the user tends to follow the easiest and less effortful choice. Another hypothesis in Figure 8 is that the 'Full Self-Driving' option had already been preselected when going on 'Eligible'. In this case the default bias may be triggered by two characteristics of the interface at the same time, namely the preselected option and the most attractive button.

As for Figure 7, the unconscious mechanism is the same when focusing on the three buttons about cookies consent: although the 'Accept all' button is not preselected, it is highlighted, while the other options require higher cognitive effort, such as when it comes to looking for the 'Deny all' button that may be hidden in 'Manage settings'. This mechanism is closely related to the salience bias, a concept introduced by the psychologist Fritz Heider in the 1970s and later developed in behavioural economics, that refers to people's tendency to focus attention on items or information that are more prominent from the visual point of view.

- **Trick questions:** this pattern was described in Chapter 1 citing the example of the double negatives 'Uncheck the box if you prefer not to receive email updates' (Mathur et al., 2019). This pattern exploits both the default effect and the bounded rationality intrinsic to humans. As observed in the other biases explored so far, all dark patterns exploit the brain's tendency to save cognitive resources, but in this case this pattern is deliberately aimed to increase the cognitive effort required to understand what you are choosing, by using the double negatives. At the same time, what already appears as complex information is presented as the default option that the user should deselect, which requires further effort since people tend to associate action to consensus.

Overall, the default effect can be characterised as follows:

- **Nature of the bias:** the status quo bias together with the salience bias are mainly cold biases as they are strictly connected to the way in which information is presented, and therefore how it is processed and evaluated by consumers.
- **Dark patterns' techniques:** these patterns are more manipulative than deceptive because it is not a matter of providing fake information, but rather, of manipulating the way in which information is presented, by preselecting an option or making it more salient.
- **Dark patterns' attributes:** dark patterns exploiting the default effect can be considered asymmetric and covert. In the case of Sneak into Basket, the design of the interface is clearly asymmetric, as the cognitive effort required to make a decision between the options is unbalanced. In other words, the choice of adding flight insurance is much easier since it is automatically added to the cart, whereas deciding to remove it involves deliberate thinking about how to deselect. By contrast, a more neutral design would present an 'Add flight insurance' button to click if the consumer wished to purchase the insurance. Moreover, dark patterns exploiting the default effect are covert since the default option is often preselected or highlighted to benefit the website or the company rather than to guide consumers towards the best choice for their interests. Moreover, Visual Interference (Figure 7) is also information hiding due to the intentional choice of hiding the 'Deny all' option in 'Manage settings', instead of presenting all options equally.

2.4.6 The Sunk Cost Fallacy

The sunk cost fallacy is people's tendency to continue investing in an activity even when it would be more logical to stop it, justifying this choice on the basis of time, money or mental energy they have already spent on it. This bias is studied not only in behavioural economics, but it is present in several aspects of people's life, from social relationships to career choices. For instance, people tend to keep staying in a long-term relationship that is no longer healthy because they focus on the time and the emotional energy they have already invested, rather than focusing on the potential future advantages that the choice of ending the relationship may provide. Moreover, the sunk cost effect may deeply influence the professional career, since people may feel stuck in a job that is no longer satisfying or aligned with their personal interests. However, changing job may be a difficult decision since the effort invested in education, training and integration into the professional environment may outweigh the uncertain future benefits related to a new career path (Examples' source: https://thedecisionlab.com/biases/the-sunk-cost-fallacy).

In the 1980s Thaler explained this mechanism by showing that people are more likely to use a product or service if they pay for it and the cognitive effort required to accept the loss of the product, or the ending of an activity, leads people to continue even if it is not the best decision for their interests. This concept contrasts with the earlier traditional economic theory according to which people were expected to consider only future benefits when making decisions. By contrast, System 1 tends to make the easiest choice, namely staying in the same situation in which the individual had already invested. This bias is strictly associated with other unconscious processes such as loss aversion or emotional involvement, depending on the specific case.

In the context of digital marketing, this bias is typically exploited by leveraging time and money spent by consumers in a purchasing process or service, to discourage them from abandoning it although it would be the most rational choice. Among the dark patterns described in Chapter 1, the following exploit the sunk cost fallacy.

- **Hard to cancel (Figure 11):** in the case of the Amazon Prime subscription, the sunk cost fallacy is the secondary mechanism that can be triggered by that type of interface. As shown in the first chapter, the primary mechanism exploited is the cognitive effort required to go through all the passages involved to cancel the subscription. At the same time, the user may be influenced by the personal investment in this platform, in terms of previously purchased products, the annual fee, as well as the established habit of using the service. Although paying for the service may be disadvantageous in the long term (for instance when the use of the service is gradually reduced), deciding to cancel the subscription may require higher cognitive effort than continuing to pay for it.

- **Hidden costs (Figure 2):** as previously discussed, this pattern mainly exploits the anchoring bias when it comes to evaluating the final cost of a product after adding extra fees. In the Deliveroo example, the sunk cost fallacy may be triggered in the final phase, following the anchoring effect. Users may be less likely to abandon the cart after seeing that the final cost is unexpectedly higher than the original one, because they are influenced by the time and effort already spent to find that restaurant, choose the meal, and enter required information to place the order (such as personal data, address, etc.). In this case the low price of the purchase is another reason to not restart the process for a cheaper alternative. However, what is relevant here is the psychological mechanism, that may be activated also when it comes to expensive purchases.

Overall, this bias can be characterised as follows:

- **Nature of the bias:** this bias can be considered neutral since, depending on the situation, it may be driven either by cognitive limitations or by the emotional sphere. In the specific context of digital marketing, it tends to be a cold bias as the choice is mainly guided by the tendency to reduce cognitive effort. By contrast, in the context of ending a relationship or changing the professional career the bias is more driven by emotional attachment.
- **Dark patterns' techniques:** these patterns can be considered partially deceptive, since although information is accurate, hiding certain options may create a false belief and therefore mislead users in favour of the company's interests.
- **Dark patterns' attributes:** the dark patterns exploiting this bias are mainly information hiding, since in the first case the cancellation option is difficult to find, whereas in the second case the extra costs are hidden until the checkout. In both cases users are not provided with complete information that would allow them to evaluate in a more rational way.

The following table is useful to summarise cognitive biases and related dark patterns explored in this chapter.

Cognitive bias	Nature of the bias	Dark patterns exploiting the bias (chapter 1)	Dark patterns' techniques (Kollmer & Eckhardt, 2022)	Dark patterns' attributes (Mathur et al., 2019)
Anchoring bias	Cold bias	Hidden costs (Fig. 2) Pressured selling	Manipulation	Covert
Framing bias **Affect heuristic**	Cold bias Motivated bias	Confirmshaming (Fig. 6) Visual interference (Fig. 7)	Manipulation	Asymmetric Covert/Asymmetric
Bandwagon Effect **Veblen Effect**	Motivated biases	Social Proof (Fig. 9-10)	Manipulation/ Deception	Covert
Scarcity bias **Action bias** **Loss aversion**	Motivated biases	Urgency/Scarcity (Fig. 3, 4, 5, 10)	Manipulation/ Deception	Covert Deceptive
Default effect **Salience bias**	Cold bias	Sneak into basket (Fig. 1) Visual interference (Fig. 7, 8)	Manipulation	Asymmetric Covert Information hiding
Sunk Cost fallacy	Neutral bias	Hard to cancel (Fig. 11) Hidden costs (Fig. 2)	Partially deceptive	Information hiding
None		Forced action (Fig. 12) Hidden subscription (Adobe)		Restrictive (all-or-nothing) Information hiding

Table 3: classification of Cognitive Biases

As observed in Table 3 the two exceptions that are not related to any bias illustrated above are Forced action and Hidden subscription, since they do not actually exploit any cognitive bias to steer users towards a specific choice.

Firstly, following Mathur et al. (2019) attributes of dark patterns, Forced Action can be classified as a restrictive dark pattern since only one option is available to navigate into a website or access a particular functionality, such as creating an account, signing up or accepting cookies. Similarly, the

nagging pattern consists of repeated pop-ups requiring a specific action before continuing to navigate, thereby making the browsing activity more complex and limited. In these cases, there is no covert mechanism through which the user's autonomy is subtly limited, without their awareness. By contrast, they are totally aware of the mandatory action that is required since the alternative is absent and there is no possibility other than doing the action or abandon the website, this is why Mathur et al. (2019) define this pattern as an 'all-or nothing proposition'.

Secondly, Hidden Subscription can be classified as information hiding pattern, since when the user pays for a service, they are unaware of the fact that the payment is for a subscription. Similarly, as in the case of Adobe, many users paid for temporary use without knowing that they would have to pay a significant termination fee within the first year, as this information was hidden or obscured. Again, no specific cognitive bias is exploited because users act based on the information they have, and in this case the necessary elements to objectively evaluate the choice are lacking. Overall, these exceptions show that not all dark patterns act in a deceptive way but simply limit the user's autonomy by explicitly interrupting their action or not giving alternatives.

The present chapter provided the theoretical foundation of behavioural economics to understand how dark patterns influence consumers' purchasing decisions online. The first part illustrated the passage from the neoclassical economic model to the rise of behavioural economics, highlighting that in the first-place humans were expected to be rational beings able to make economic decisions based on the expected utility. In more recent studies, thanks to the collaboration of psychologists and economists, individuals started to be seen as characterised by unconscious cognitive mechanisms that limit this rationality, and most influenced by emotions and instincts when making decisions. This explains why dark patterns are sometimes effective: these deviations from rationality, that can result in cognitive biases, are systematic and predictable since they are intrinsic to human behaviour. Building on this key point, other linked concepts were explored, such as the Dual Process Theory, in order to explain how dark patterns tend to trigger System 1, encouraging consumers to make fast and effortless decisions instead of deliberate thinking. Finally, the topics in the first chapter were revisited by analysing in great detail the differences between nudges and sludges, and then by presenting a taxonomy of cognitive biases exploited by the dark patterns described earlier. In the last chapter an exploratory experiment will be conducted by taking under analysis the Confirmshaming pattern to see in practice its influence on people who are required to make a fast choice.

3. An empirical study on Confirmshaming

The first two chapters of this thesis illustrated in detail dark patterns and cognitive biases from a theoretical perspective, by describing different concepts and providing several examples. By contrast, the present chapter is based on a small empirical study of a specific case previously mentioned in the taxonomy of dark patterns, namely Confirmshaming. In particular, this experiment aims to observe the effects of a message framed with Confirmshaming on participants, in terms of decision-making influenced by the emotional sphere.

As explained in 1.3.1, Confirmshaming is part of the broader category of Misdirection dark patterns and consists of manipulating consumers' decisions through specific wording that aims to trigger uncomfortable emotions such as guilt. This activation of the emotional sphere is due to the so-called affect heuristic, a mental shortcut that leads to a fast decision affected by emotions felt at that moment. When exploring cognitive biases, this effect was classified under the framing bias (see 2.4.2) since the related dark pattern involves influencing the decision-making process through specific and intentional wording. It was also categorised as a hot/motivated bias since such wording aims to appeal to the emotional sphere of the individual. The proposed experiment will try to understand which emotions are triggered when it comes to accepting or declining an option framed with Confirmshaming. The feeling of guilt will not be the only emotion considered, as one emotion may trigger other closely related emotions depending on the decision being faced. For instance, the user may feel guilty for declining an option or proud of themselves for accepting, which in turn may be influenced by an underlying fear of being judged. Therefore, positive or negative feelings are also related to the self-image that individuals tend to build in a positive way. This concept will play a key role in the experiment by considering not only the sense of guilt, but also the feelings of relief and pride connected to a specific choice.

This experiment builds on existing research on dark patterns, and the closest study is conducted by Luguri & Strahilevitz (2021), who compared the acceptance rates of participants subject to different conditions: neutral (no dark patterns), mild and aggressive, with Confirmshaming being part of the mild condition. Similarly, this experiment will compare the effects caused by a neutral interface to those caused by an interface framed with Confirmshaming. Additionally, the study will assess emotional responses related to guilt, relief and pride, while Luguri & Strahilevitz (2021) measured participants' mood on a single item ranging from relaxed to irritated.

The present chapter is structured as follows: the next section presents the research questions together with the related hypotheses; section 3.2 describes the research methodology, including the sampling methods, the description of the surveys and the limitations of the study; section 3.3 presents

the findings in quantitative terms, followed by the final section, which discusses the findings from a qualitative point of view, and a final summary of the outcomes.

3.1 Research questions

This experiment aims to answer the following questions:

- **RQ1**: To what extent does a Confirmshaming increase the likelihood that participants accept a donation request, compared to a neutral framing?
- **RQ2**: To what extent does Confirmshaming elicit different emotional responses (pride and relief when accepting, guilt when declining), compared to a neutral framing?

Building on these questions, the hypotheses are:

- **H1**: The acceptance rate is higher in the group exposed to the interface framed with Confirmshaming, than in the group exposed to the neutral interface. This is expected due to the nature of this dark pattern, which is intended to exploit the emotional sphere of the individual.
- **H2**: The feeling of guilt is higher in the Confirmshaming group, among participants who declined the donation request. The underlying factors eliciting this emotion may be the wording of the declining option, but also the personal attachment to the environmental cause.
- **H3**: The feelings of relief and pride are higher in the Confirmshaming group, among participants who accepted the donation request. In this case participants may be influenced by a sense of relief associated with the need to avoid negative judgement, or a sense of pride derived from accepting the option, thereby building a positive self-image.

3.2 Methodology

This section provides information about the creation of the surveys, starting from the webpage mockups to the sampling methods, and how data were collected with the questionnaires on Google Forms. Moreover, a few limitations of the study will be discussed and, finally, the procedure of data analysis will be described, including chi-squared tests and t tests.

3.2.1 Data collection

The first step of this experiment consists of creating the two interface mockups that will later be the subject of the two surveys. They were built on Canva, a useful platform for graphic design and visual communication through which it is possible to create any kind of material, such as social media graphics, marketing materials, websites, presentations, etc. The two interfaces are examples of checkout pages, in which the price of the articles as well as the final value including shipping fee are shown.

Between this information and the ‘Pay’ button, a further option was inserted, asking users to donate 1€ to fund reforestation projects. In the two surveys the mockups are identical, except for the declining option, which is a neutral ‘No, thanks’ in Figure 15 and ‘No, I don’t care about the planet’s future’ in Figure 16.

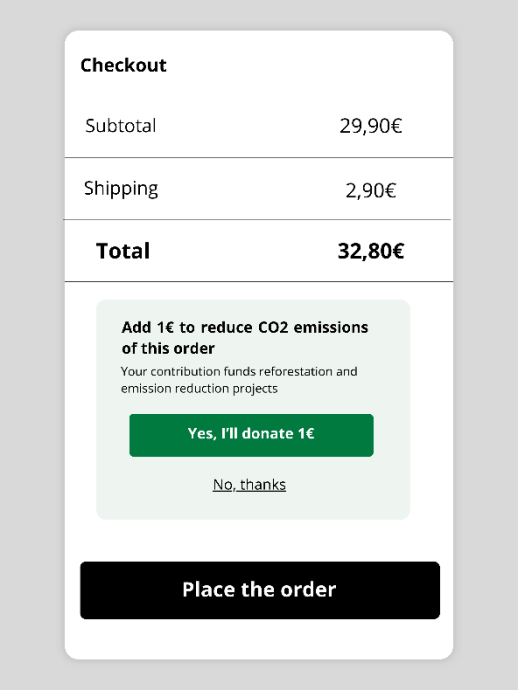


Figure 15: Checkout page with neutral framing.

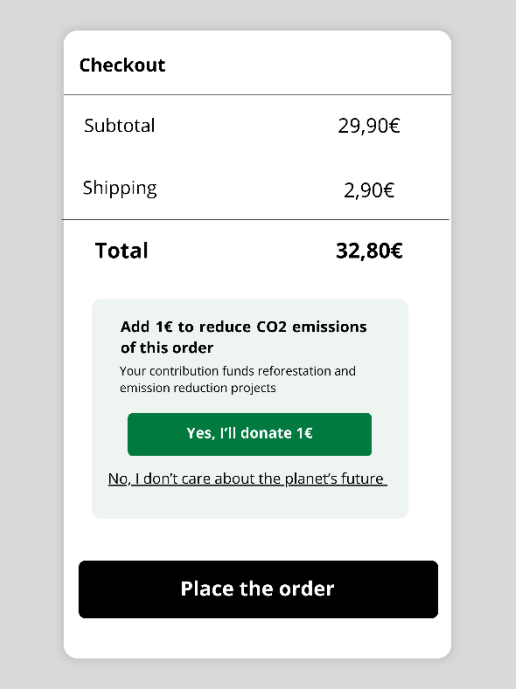


Figure 16: Checkout page framed with Confirmshaming.

The latter message is framed with Confirmshaming and is expected to create a different effect on participants’ choices compared to the neutral message. The choice of an environmental cause as the motivation for donating is justified by its relevance as being a social salient issue, which may be more meaningful to some participants than to others. This may be an influential factor for the results of the tests, as will be explained when describing the surveys. It is also worth observing that other design techniques were included in order to make the interfaces as realistic as possible, such as highlighting the accepting option while using a small font for the declining option, which can be considered an example of Visual Interference (see 1.3.1).

The second step involves recruiting participants and illustrating the related sampling methods. The two surveys are intended to be presented to two separate groups of participants: Group A will answer the questionnaire including the Neutral checkout page, while Group B will answer the questionnaire including the page framed with Confirmshaming. Participants are mainly recruited by sharing the surveys on the messaging application WhatsApp, ensuring that they are distributed to different people so as to obtain two distinct samples.

This is done both through WhatsApp groups and by directly contacting people known to the researcher, who are then asked to forward the surveys to other potential participants.

In the context of the sampling methods (Saunders et al., 2023) this type of participant recruitment is considered a non-probability method, particularly convenience sampling. The main difference between probability and non-probability sampling methods is based on the degree of randomness. In the first case, the survey should guarantee that everyone in the target population has an equal opportunity to participate, therefore, participation should solely depend on chance. The more this requirement is respected, the more representative the results will be of the target population. However, this method requires a significant amount of time and financial resources, and it is usually almost impossible to reach total randomness, this is why sampling errors and limitations are a fundamental part of empirical research. By contrast, non-probability methods do not aim to reach representativeness, but the main focus is on the number of responses. This is a faster and more accessible methodology, but it cannot guarantee randomness because not everyone has an equal possibility to participate. The sampling method used for this test is classified as convenience sampling due to the easier accessibility and availability of the target population. Also, the samples are relatively limited and include people known to the researcher, although none of them is influenced by the topic of this study. The recruitment phase resulted in 42 participants per group, all aged between 20 and 30, in order to limit sample heterogeneity. However, participants' exact age was not collected, as it was only used as a criterion when recruiting people.

Moreover, the two surveys were created on Google Forms, a useful service that allows users to create surveys or quizzes and offers several types of custom questions, such as multiple-choice questions, text boxes or drop-down menus. It also automatically saves and organises results so that the researcher can later download them in Excel or Google Sheets.

The surveys share an identical structure and, as previously anticipated, the only difference is the wording of the refusal option in the two checkout pages (see Figures 15 and 16). The full text of the surveys is reported in Appendix A. They are organised in five sections, as follows:

- **Section 1**: this is the introduction page that provides information about what participants are required to do, also indicating the completion time and highlighting that the questionnaire is completely anonymous (no demographic data were collected) and that they can interrupt at any time. The aim of the survey and the description of the scenario are deliberately vague to avoid influencing participants' choices throughout the questionnaire.
- **Section 2**: it shows the checkout page mockup, which is the starting point of the questionnaire, and the type of framing of the refusal option is the independent variable of the research hypotheses.

Participants are asked to imagine being in a real situation in which they have to complete an online purchase and to decide whether to donate 1€ for a reforestation project. The image is followed by the accept/decline question, and the answers constitute the dependent variable of RQ1 and H1 (see 3.1). Moreover, this question is followed by a recall check, meaning that participants are required to remember the exact words of the refusal option. This type of question is usually included in questionnaires since participants' inattention is a common source of sampling error. However, in this case it will be interpreted with caution since participants are allowed to go back and check the options, and this constitutes a limitation of the study, as will be explained in 3.2.3.

- **Section 3-4:** these sections are organised depending on what participants answered in section 2. In particular, participants who accepted to donate are directed to a specific set of questions, while those who declined are directed to a different set of questions. Depending on their choice, they are asked to rate their emotional responses in terms of guilt or relief and pride on 5-point Likert scales, and the answers are the variables for RQ2, H2 and H3.
- **Section 5**: this final section aims to measure the participants' attachment to the environmental cause. This question was intentionally included since the wording of the refusal option is not the only motivation behind the consumer's choice, but there may be several reasons, including the attachment to the cause. In this regard, a further section was included in the questionnaires, aiming to thank participants for completing the survey and to clarify that the questions are not intended to judge their values or choices, but solely to observe the effects of this specific dark pattern for academic research purposes. It was intentionally inserted as the final section in order not to influence participants' choices.

3.2.2 Procedure

After receiving the answers, data were analysed following precise and ordered steps. Firstly, the results were downloaded and combined in a single Excel file, and a table was created providing the following information:

- **Participants**: identified by numbers
- **Conditions**: Neutral - Confirmshaming
- **Accepted – Declined**: this column presents the descriptive results, namely "Yes, I'll donate 1€", "No, thanks" and "No, I don't care about the planet's future"
- **Accepted (0) – Declined (1):** in this column the results in words are converted into 0 or 1, in order to ease the following calculations.
- **Recall check**: this column allows to verify whether all participants remembered the exact words of the donation request options.

- The next four columns show the **Likert scales scores** for guilt related to the decision, guilt related to the wording of the message, or, in case of acceptance, for relief and pride.
- **Env. Cause**: the last column indicates the scores related to the attachment to the environmental cause, which will be the last set of data to be analysed.

Secondly, data were cleaned by excluding what may interfere with the analysis. The recall check results and potential missing responses were the two main aspects to verify in order to obtain a dataset as clean as possible. In the first case, wrong answers indicate that the participant may not have read carefully and therefore may not have been exposed to the intentional manipulation. Consequently, the choice to accept or decline cannot be related to the way in which the message was framed, and including the scores of participants with wrong answers in the analysis risks mixing those who received the manipulation with those who did not. In the latter case, the presence of missing responses may alter the findings in different ways, considering that a blank cell is not always automatically ignored by Excel formulas. Depending on the formula, it may be either excluded or interpreted as 0, especially in arithmetic operations, which would itself indicate a value and therefore compromise the validity of the results. An example of missing responses in this case concerns the Environmental Cause variable, that caused a reduction of Group B compared to the original sample. While the questionnaire was still open for responses, a technical issue occurred due to a small modification in the conditional structure of the survey. It was corrected as soon as it was noticed, but during that time slot a few participants did not receive the question.

Once the cleaned data were obtained, the analysis was organised by separating the acceptance rates and the Likert scales, since they address different research questions. Firstly, acceptance rates were compared between the two conditions (Neutral – Confirmshaming) through the chi-squared test of independence. This type of test was chosen over a t-test since, although both tests produce a p-value and lead to similar conclusions, the chi-squared test is specific to comparisons of categorical variables (such as 0 = Accepted vs 1 = Declined), whereas t-tests compare the means of continuous variables. However, the t-test p-value was calculated to check whether both tests lead to similar conclusions. This part of the analysis aims to answer RQ1 by assuming as the null hypothesis (H0) that the two conditions do not produce any effect on the acceptance rates, meaning that the observed differences are due to chance rather than to a real effect of the conditions. Through the chi-squared test it is possible to verify whether the observed data provide statistically significant evidence to reject the H0 and to state that the conditions actually produce a significant difference in the acceptance rates.

Moreover, Likert scales measuring guilt, pride and relief are analysed through the two-group t-test, as they are treated as continuous variables.

This test is useful to compare two independent samples and to check whether they differ from each other from a statistical perspective. In this context, it helps answer RQ2 by assuming as H0 that the Neutral and Confirmshaming conditions do not cause any significant difference between the two groups, and this null hypothesis is valid for all the rated emotions. For instance, in the case of guilt for the decision itself, the H0 will be 'Group B does not feel guiltier than Group A because of the Confirmshaming condition, but the difference is due to chance'. If the observed results are statistically significant, the H0 is rejected and it can be affirmed that the degree of guilt of the two groups depends on the condition they faced. Otherwise, non-significant results are not sufficient to reject the H0.

As for the analysis of the Likert scales, it is important to outline a few points. Each scale was analysed taking the corresponding subgroup as the sample, rather than the whole groups. As explained in Sections 3-4 (see 3.2.1), the questionnaires are structured in a way that do not allow participants to answer all the questions; therefore, the scales for Guilt are related to the subgroup that declined, while the scales for Relief and Pride are related to the subgroup that accepted.

Moreover, throughout the analysis with two-group t-tests, a significance level of $\alpha = 0.05$ is conventionally adopted. In other words, it is assumed that if the value of the H0 falls within the 95% confidence interval, the H0 cannot be rejected, while if it falls outside this interval (the remaining 5% of probability), it is a statistically significant result that allows the H0 to be rejected (Quirk & Rhiney, 2021). In our data analysis both the limits of the confidence interval and the p-value are calculated as a robustness check. The limits are useful to verify whether the value assumed under $H0 = 0$ (as the null hypothesis assumes that there is no significant difference between the means of the two groups) falls within the confidence interval. By contrast, the p-value provides a probability measure to be compared directly to the significance level: if the p-value is < 0.05, the H0 is rejected as it corresponds to the value falling outside the confidence interval, whereas if the p-value is > 0.05, there is no statistically significant evidence to reject the H0.

Finally, it is important to clarify the function of the degrees of freedom (df) and how they are calculated. This value is necessary for calculating the confidence interval of the two groups separately, as well as the t-test value, and indicates the number of independent values that are free to vary before the remaining values are fixed by a set of rules or calculations. Quirk & Rhiney (2021) provide a table with degrees of freedom already calculated for sample sizes ≥ 10, while for smaller samples they were calculated as $df = n-1$ (the sample size of the smaller group minus one), following the logic of the table.

Before concluding the data analysis, a robustness check was conducted by re-calculating the chi-squared value on acceptance rates of the whole sample, including those who did not pass the recall check, in order to verify whether their exclusion affected the final conclusions.

The last step of the data analysis consisted of examining the relationship between the personal attachment to the environmental cause and the acceptance rates. A Pivot Table was built on the sample including Group A and Group B, cross-checking condition, attachment level (High – Medium/Low), and acceptance rates. However, this comparison is solely descriptive and suggests a potential correlation between the two factors, given the reduction of the sample size resulting from the missing responses previously described.

3.2.3 Limitations of the study

This subsection outlines a few limitations fundamental for a correct interpretation of the findings. Firstly, this is an exploratory study and for this reason the results are indicative rather than real conclusive. This exploratory nature is due to the non-probability sampling method, which does not allow the results to be generalised to the population. People do not have the same opportunity to access the survey since it is not randomly distributed and, consequently, the samples cannot be representative. Moreover, being convenience samples, they are limited in size, and small samples tend to reduce the likelihood that tests will find statistically significant differences. In other words, estimates are quite unstable with a low number of participants, and only large differences can be significant, whereas smaller differences risk being not statistically significant. This does not mean that the effect does not exist, but small sample sizes may reduce the statistical power of the tests. This research may be replicated in future studies with larger and more representative samples in order to increase the statistical power of the tests.

Secondly, it is worth outlining that the surveys provide a hypothetical scenario in which people are required to act as they would in a real situation, which is significantly different from making a real donation. Consequently, the acceptance rate may not reflect the one observed in a real life. In this regard, interesting future research may be based on a real donation or, in general, on data collected from a real website using dark patterns.

Moreover, the recall check can constitute another limitation. As previously anticipated, it is likely that people do not have the same level of attention throughout the surveys. For instance, they may immediately focus on the two options when starting the survey, but some of them may also take a quick look at the image and directly move to the questions. The recall check in large samples is useful to exclude participants who could not remember the exact words, thereby reducing the sampling error. However, this is effective when the platform does not allow to go back to previous answers, but Google Forms has no setting preventing respondents from doing so. In addition, the recall check question was placed in the same section as the image, although not directly visible, and they could easily scroll up to recheck the options. Although the introduction asked participants not to change

previous answers, this depends only on their good faith and does not ensure their compliance. Therefore, including those who did not read the options carefully risks overestimating their attention.

Finally, the motivation behind the choice is another limitation to be outlined. Participants may be motivated by the way in which the options are worded, but also by their personal attachment to the cause or by the need to build a positive self-image. Indeed, the attachment to the environmental cause is socially perceived in a positive way, and consequently participants may feel more motivated to accept, even though in real life they may not show such interest in the cause. The final question in Section 5 was included to try to understand this matter, but it should be interpreted with caution since the answer may be influenced by their previous choices.

The next section will illustrate the quantitative findings of this research, following the procedure described in 3.2.2.

3.3 Findings

The results of this study are provided by first presenting the final sample obtained after data cleaning, followed by the acceptance rates observed in the two conditions. The results of the chi-squared test are then compared with the t-test p-value as a robustness check, followed by the results of the t-tests on the Likert scales. Then, the comparison between acceptance rates and the attachment to the environmental cause is presented and, finally, the acceptance rates are then re-examined on the whole sample. Complete calculations with formulas are presented in Appendix B.

3.3.1 Data cleaning and acceptance rates

Firstly, the original sample of N= 84 was reduced to N= 57 (Neutral= 32, Confirmshaming= 25) because 27 participants gave wrong recall check answers (see Figure 17).

	A	B	C	D
1	Participants	Condition	A Accepted / B Declined	Recall check
2	2	Neutral	B: No, thanks	B: I'll pass on this one
3	3	Neutral	A: Yes, I'll donate 1€	B: I'll pass on this one
4	42	Neutral	B: No, thanks	C: No, I'm not interested
5	36	Neutral	B: No, thanks	D: I don't remember
6	34	Neutral	B: No, thanks	C: No, I'm not interested
7	25	Neutral	B: No, thanks	B: I'll pass on this one
8	7	Neutral	B: No, thanks	C: No, I'm not interested
9	13	Neutral	B: No, thanks	B: I'll pass on this one
10	15	Neutral	B: No, thanks	B: I'll pass on this one
11	17	Neutral	A: Yes, I'll donate 1€	D: I don't remember
12	84	Confirmshaming	B: No, I don't care about the planet's future	A: No, thanks
13	81	Confirmshaming	A: Yes, I'll donate 1€	C: No, I'm not interested
14	75	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
15	70	Confirmshaming	A: Yes, I'll donate 1€	C: No, I'm not interested
16	71	Confirmshaming	A: Yes, I'll donate 1€	D: I don't remember
17	72	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
18	68	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
19	63	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
20	64	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
21	65	Confirmshaming	B: No, I don't care about the planet's future	A: No, thanks
22	60	Confirmshaming	B: No, I don't care about the planet's future	A: No, thanks
23	58	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
24	51	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
25	52	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
26	49	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
27	43	Confirmshaming	B: No, I don't care about the planet's future	C: No, I'm not interested
28	44	Confirmshaming	B: No, I don't care about the planet's future	A: No, thanks

Figure 17: Excel Table showing wrong Recall Check answers

Secondly, a Pivot Table was built on the remaining sample including conditions and the counts of the given answers expressed as percentages, as in Figure 18.

Conteggio di 0 Accepted / 1 Declined	**0 Accepted / 1 Declined**		
Condition	**0: Accepted**	**1:Declined**	**Totale complessivo**
Confirmshaming	64%	36%	100%
Neutral	16%	84%	100%
Totale complessivo	**37%**	**63%**	**100%**

Figure 18: Acceptance Rates in both conditions

A chi-squared test was necessary to verify whether such differences are significant (see Figure 19). The H0 assumed that any observed difference is not related to Neutral or Confirmshaming conditions but is due to chance. After calculating the expected values (if H0 were true), the chi-squared value was computed as

$$X^2 = \frac{\sum(O - E)^2}{E} = 14.12$$

Observed (O)	Accepted (0)	Declined (1)	Total
Neutral	5	27	32
Confirmshaming	16	9	25
Total	21	36	57

Expected (E)	Accepted (0)	Declined (1)
Neutral	11,79	20,21
Confirmshaming	9,21	15,79

(O-E)^2/E	
Neutral (0)	3,91
Neutral (1)	2,28
Confirmshaming (0)	5,00
Confirmshaming (1)	2,92
X^2=	14,12
Critical value	3,84
df	1
p-value	0,00017 <0,05

Figure 19: chi-squared test on sample N=57

With df = 1 and a critical value of 3.84, H0 is rejected as 14.12 > 3.84. At the same time, we obtained a t-test p-value of p = 0.00017 < 0.05, that confirms the previous result. In other words, the observed differences are statistically significant and not attributable to chance.

3.3.2 Two-group t-tests on the Likert scales

The next step consisted of calculating the two-group t-tests for the following subgroups: Guilt for wording of the message, Guilt for decision, Relief and Pride. In the present section, the specific calculations and related values will be shown only for the first case, since the procedure is the same for each scale. By contrast, only the main results of the remaining three subgroups will be summarised in a single table not to make the section repetitive (all calculations and formulas are available in Appendix B).

As for the subgroup Guilt for Wording, we calculated for each condition the mean, standard deviation and standard error, which are necessary to obtain the limits of the confidence interval and the t-test value. The latter was calculated as

$$t = \frac{\bar{M}1 - \bar{M}2}{SE_{\bar{M}1 - \bar{M}2}}$$

where M1 and M2 are the means of Group A and Group B, and SE is the standard error. It was assumed as H0 that the wording of the message does not produce any difference effect in the degree of guilt between Group A and Group B, and with a df = 8 and a critical value of 2.306, we obtained an absolute t value of 3.22 > 2.306, meaning that the observed results are statistically significant and H0 is rejected. Also, we obtained a p-value of 0.009 < 0.05, which confirms the result of t-test (see Figure 20).

Neutral	Confirms.
Guilt for wording	
1	5
1	4
2	1
1	4
3	4
2	5
3	4
2	2
4	2
1	
1	
2	
2	
1	
1	
1	
3	
1	
1	
3	
3	
2	
3	
2	
1	
1	
1	

	Neutral	Confirms.		
N	27	9		
∑	49	31		
Mean	1,81	3,44		
ST.DEV. (SD)	0,92	1,42		
ST.ERROR (SE)	0,18	0,47		
Upper limit	2,18	4,54		
Lower limit	1,45	2,35		
SD^2/N (Neutral)	0,03		Critical t	2,306
SD^2/N (Confirms)	0,23			
sum	0,26			
SE	0,51			
t value	-3,22		pvalue	0,009 < 0,05
|t| > critical t	3,22 > 2,306			

Figure 20: two-group t test on guilt related to the wording of the message.

The same procedure was repeated for the remaining subgroups, obtaining the results in Table 4[3].

Condition	Neutral		Confirmshaming					
Variable	**Mean**	**ST.DEV.**	**Mean**	**ST.DEV.**	**|t|**	**df**	**Critical t**	**p-value (α = .05)**
Guilt for Decision	1,89	1,01	2,89	1,36	2,02	8	2,306	0,068
Relief	3,6	1,52	3,5	0,89	0,14	4	2,776	0,89
Pride	3,4	1,14	3,69	1,01	0,5	4	2,776	0,63

Table 4: Values of two group t-tests on Guilt for Decision, Relief and Pride (Neutral vs Confirmshaming)

Unlike the first subgroup, these results are not statistically significant, as the t-value is always smaller than the critical value and the p-value is always higher than 0.05. Therefore, there is not sufficient evidence to reject H0, meaning that the observed differences may be due to chance.

[3] Critical t values follow the approximation df = n-1 of the smaller group (Quirk & Rhiney, 2021), whereas p-values were calculated with the Excel T.TEST function.

3.3.3 Levels of attachment to the environmental cause

This part of the analysis differs from the previous ones since the question about the attachment to the environmental cause was included to observe a potential correlation between the level of attachment and acceptance rate, rather than to test a formal hypothesis. Indeed, it should be interpreted with caution as Group B was reduced from 25 to 18 participants due to a technical issue (see 3.2.2), thereby resulting in a total sample of N = 50 instead of N = 57.

The following table (Figure 21) was built to cross-check levels of attachment (High = 5, the maximum score on the scale; Medium/Low = 1 – 4) and acceptance rates. The interesting point is that in Neutral condition a high level of attachment seems to be correlated with the choice of accepting to donate, while this correlation disappears in the Confirmshaming condition, which shows identical acceptance rates, meaning that participants seem to accept or decline at similar rates, regardless of their level of attachment.

Conteggio di Participants		0 Accepted / 1 Declined		
Condition	Attachment	0	1	Totale complessivo
Confirmshaming	High	66,7%	33,3%	100%
	Medium/Low	66,7%	33,3%	100%
Neutral	High	21,4%	78,6%	100%
	Medium/Low	11,1%	88,9%	100%

Figure 21: Pivot Table of acceptance rates related to the levels of attachment to the environmental cause.

Moreover, the t-test p-value was calculated to verify that the results of the previous table do not reflect a pre-existing imbalance between the two groups in terms of attachment to the environmental cause. We obtained a p-value of $p = 0.51 > 0.05$ (see Appendix B), suggesting that no statistically significant difference was detected in their levels of attachment, and thereby supporting that the groups were comparable with respect to this variable. In other words, if a disproportionate number of people highly attached to the environmental cause were randomly assigned to the Confirmshaming group, this imbalance would likely have produced a significant difference between the two means. By contrast, the small difference observed (M1 = 4.34 vs M2 = 4.50) and $p = 0.51$ suggest that there is no pre-existing imbalance, and that these results represent normal random variation, meaning that the groups can be considered comparable.

3.3.4 Robustness check on acceptance rates

The final step of data analysis consisted of repeating the chi-squared test on the acceptance rates for the whole sample of N = 84, including participants who gave wrong answers to the recall check, in order to verify whether the conclusions obtained on a sample of N = 57 are similar or differ because of the sample reduction. The following table shows the obtained values (Figure 22).

Observed (O)	Accepted (0)	Declined (1)	Total
Neutral	7	35	42
Confirmshaming	19	23	42
Total	26	58	84

Expected (E)	Accepted (0)	Declined (1)
Neutral	13,00	29,00
Confirmshaming	13,00	29,00

(O-E)^2/E	
Neutral (0)	2,77
Neutral (1)	1,24
Confirmshaming (0)	2,77
Confirmshaming (1)	1,24

X^2	8,02	>3,84
Critical value	3,84	
df	1	
p-value	0,0043	<0,05

Figure 22: results of the chi-squared test on the whole sample of N=84

The result remained statistically significant, as the chi-squared value is higher than the critical value (8.02 > 3.84). Following the prior procedure (see 3.3.1), this is further confirmed by p = 0.0043 < 0.05, thereby rejecting the H0 according to which the presence of the dark pattern does not produce any significant difference in the choice of accepting to donate.

In the next section the results presented up to now will be interpreted from a qualitative perspective, with the aim to answer research questions and accept or reject the alternative hypotheses.

3.4 Discussion of the findings

This section is organised following the research questions in 3.1 and the related research hypotheses, which will be supported or rejected depending on the results presented in the previous section. A few concepts from dark pattern literature will then be revisited in order to provide a qualitative explanation of the obtained results.

3.4.1 Acceptance rates (RQ1, H1)

The first research question was:

> **RQ1**: To what extent does Confirmshaming increase the likelihood that participants accept a donation request, compared to a neutral framing?

The H0 assumed that Neutral and Confirmshaming conditions produce no significant effect on the choice of accepting or declining the donation request, while the alternative hypothesis (H1) stated that the acceptance rate is higher in Group B (Confirmshaming), than in Group A (Neutral interface). The results in 3.3.1 allowed the H0 to be rejected in favour of the H1, as they showed a statistically significant difference between the two groups. RQ1 can therefore be answered by stating that the acceptance rate of participants receiving the donation request framed with Confirmshaming is significantly higher than that of the group receiving a neutral interface (64% vs 16%) and this was further confirmed on the whole sample of N = 84 (see 3.3.4). These findings can be compared with Luguri & Strahilevitz (2021), who divided their sample into a group exposed to a neutral interface (the control group), and a group exposed to gradually higher levels of manipulation. They found acceptance rates of 11.3% in the control group,

25.8% in the mild dark pattern condition (which included also Confirmshaming) and 41.9% in the aggressive condition (Luguri & Strahilevitz, 2021, p. 65, Table 2). The main difference is that, in this case, the wording of 'No, I don't care about the planet's future' is more aggressive, and it is therefore a stronger version of Confirmshaming rather than a mild one. Although the sample sizes are highly different and the experiments are also structured differently, the results obtained show a similar pattern. Building on what has been explored in the previous chapters, this is consistent with Mathur et al.'s (2019) definition of Confirmshaming as an intentional and explicit way of wording a message aimed at steering consumers towards a certain choice by triggering an uncomfortable emotional state. An exploratory analysis was also conducted comparing acceptance rates with the levels of attachment to the environmental cause to understand whether the acceptance rates actually depend on the framing of the message or on a pre-existing attachment to the cause. Although this was not included as a formal hypothesis, being solely a descriptive analysis, the results suggest no significant imbalance between the two groups in terms of attachment to the cause. In other words, the results indicate that the specific wording, rather than a pre-existing inclination to donate to the cause, is more likely to account for the observed difference in acceptance rates.

3.4.2 Guilt for Decision vs Guilt for Wording (RQ2, H2)

The second research question was:

> **RQ2**: To what extent does Confirmshaming elicit different emotional responses (pride and relief when accepting, guilt when declining), compared to a neutral framing?

As for the participants who declined (in both the Neutral and the Confirmshaming conditions), two Likert scales were built to distinguish those who felt guilty because of the wording of the message from those who felt guilty for the decision itself not to donate. This distinction aimed to highlight that guilt can arise from different motivations, which in this case can be either a general sense of moral responsibility toward the environmental cause, or a reaction intentionally triggered by the wording of the message. H0 assumed that Group B do not experience greater guilt than Group A, either because of their decision not to donate or because of the wording of the message, and that any observed difference is likely due to chance. By contrast, H2 assumed that Group B feel guiltier than Group A. Results show that guilt for the decision itself is higher in Group B, but the difference is not statistically significant, while guilt for the wording of the message differs significantly between the two groups (see 3.3.2). H0 can therefore be rejected in favour of H2 as regards guilt caused by the wording of the message, and it can be concluded that Group B feels significantly guiltier than Group A, which instead received a neutral framing.

Overall, the values obtained suggest that the prevalent mechanism affecting participants' choices is the emotional discomfort triggered when reading 'No, I don't care about the planet's future', rather than a broader sense of guilt related to the decision itself. This is in line with the classification of Confirmshaming as manipulative dark pattern rather than deceptive (see 2.4.2), as it acts explicitly through the wording of the option instead of hiding information. It also supports the classification of this dark pattern as exploiting hot or motivated biases (such as the affect heuristic) (see 2.4.2), since the guilt observed in the experiment seems to result from an immediate affective reaction to the wording, rather than from a deliberate evaluation of the decision.

3.4.3 Relief and Pride (RQ2, H3)

Building on RQ2, this part of the findings regards participants who accepted to donate in both conditions, and the aim was to observe, through Likert scales, the extent to which they felt relieved or proud of themselves after accepting to donate. The H0 follows the same logic, assuming that the choice of accepting produces no significant emotional response, and the observed differences between Group A and Group B are due to chance. By contrast, H3 assumed that Group B feels more relieved and/or proud of themselves after donating than Group A. Both emotional states arise from the individuals' need to build a positive image of themselves, and this concept was included as the counterpart to the uncomfortable emotions discussed in relation to H2.

However, the results obtained by analysing the Likert scales were not statistically significant for either variable (see 3.3.2): relief scores are almost identical in the two groups, while Pride scores are only slightly higher in Group B, but these results do not provide sufficient evidence to reject H0 and to support H3. These results may be caused by the small sample sizes that substantially reduce statistical power of the tests, even though this does not explain the almost identical means under Relief variable. One possible explanation it that with a donation of 1€ in a hypothetical scenario, the emotional reward of accepting may be minimal. This is also in line with Luguri & Strahilevitz (2021), according to which "when dark patterns are effective at leading people to a certain answer, there is no affective backlash". Indeed, they found no significant mood differences among participants who accepted, but only among those who declined. Moreover, this is consistent with Mathur et al. (2019), who describe Confirmshaming as specifically targeted to those who decline, without any equivalent mechanism of reward for those who accept. This is why it is considered asymmetric rather than covert: the framing of the message is visible and, at the same time, it places disproportionate emotional burden on the declining option compared to the accepting one. However, we offer a further interpretation according to which Confirmshaming may also be considered covert because, although the wording is explicit, the user may not be aware of the manipulative intent.

3.5 Summary

The following table (Table 5) summarises the research questions, hypotheses and outcomes obtained from the data analysis.

Research Questions	Hypotheses	Outcomes
RQ1: Acceptance rates (Neutral vs Confirmshaming)	H1: higher acceptance in Confirmshaming group (B)	Supported
RQ2: Guilt for the wording of the message	H2: Higher sense of guilt in Group B	Supported
RQ2: Guilt for the decision itself	H2: Higher sense of guilt in Group B	Not Supported
RQ2: Relief for accepting to donate	H3: Higher sense of relief in Group B	Not Supported
RQ2: Pride for accepting to donate	H3: Higher sense of pride in Group B	Not Supported

Table 5: Summary of Research Questions, Hypotheses and Outcomes

As outlined above, the Confirmshaming condition significantly increases the acceptance rates compared to the Neutral condition, and this result is further confirmed by the robustness check on the whole sample. Moreover, the hypothesis that greater guilt is elicited in Group B than in Group A by the framing of the declining option is also supported by statistically significant evidence. By contrast, hypotheses related to guilt for a general sense of responsibility toward the environmental cause, as well as to relief and pride within the subgroup that accepted, are not supported by sufficient evidence, which may be due to the small sample sizes that limit statistical power. Overall, this empirical study is consistent with the dark pattern literature and behavioural economics, which have studied the way in which the wording of marketing messages triggers consumers' emotional response, in line with the affect heuristic and the framing bias. Given the exploratory nature of this study, the results should be interpreted as indicative rather than conclusive.

Conclusions

Overall, the present thesis aimed to study how dark patterns influence consumers' decisions, and particularly why consumers are usually unaware of them or fail to resist them when they are clearly visible. The main strength of the work is the analysis of the relationship between dark patterns and cognitive biases, to explain the effects of these techniques on individuals' cognitive system, integrating an empirical test focused on the emotional sphere. The first chapter defined dark patterns, highlighting the difference between beneficial nudges aimed at helping consumers and dark patterns aimed at maximising companies' profits, and provided a taxonomy of the most common techniques with real examples. The second chapter introduced the Dual Process Theory, heuristics and cognitive biases, showing that individuals have intrinsic cognitive limitations which, although often result in useful skills, can also lead to not fully rational choices. Biases are related either to information processing or to gut feelings and emotional states, and the last chapter focused on the latter, presenting an exploratory empirical study of the effects of an interface framed with Confirmshaming. In particular, two different groups were exposed to a checkout page mock-up with a request to donate for environmental projects: the first group received a neutral interface, while the second received the same interface framed with Confirmshaming. The experiment revealed a statistically significant difference in acceptance rates, with the group exposed to Confirmshaming accepting more often than the neutral group. Moreover, positive and negative emotional states were analysed through Likert scales, revealing that participants who declined the request framed with Confirmshaming felt significantly guiltier because of the wording of the message than those who declined in the neutral condition. By contrast, the hypotheses related to guilt elicited by the decision to decline itself, as well as relief and pride for accepting, were not supported by sufficient evidence. The significant findings are consistent with the theoretical framework on cognitive biases, as higher acceptance rate aligns with the framing bias exploited by dark patterns such as Confirmshaming and Visual Interference, which steer users through the way in which information is presented, either through visual elements or linguistic choices. Moreover, the results are also consistent with the affect heuristic, the mental shortcut that leads to fast decisions influenced by the emotional state in the moment. Indeed, it was assumed that Confirmshaming would influence the choice by eliciting guilt, and results go in this direction. However, positive feelings related to the individual's self-image did not differ significantly between the groups.

This may be due to the limitations of the study, but it may also indicate that Confirmshaming acts mainly on uncomfortable emotions, without a corresponding reward mechanism. The main limitations make the experiment an exploratory study providing indicative results rather than real demonstrations. They include the small sample sizes, which may have contributed to some non-significant results, as well as the sampling method employed, that did not guarantee total randomness or representativeness of the target population.

Building on the initial question of this thesis, Confirmshaming belongs to the category of dark patterns that are clearly visible, but consumers tend to fail to resist them. This suggests that unethical practices are effective not only through deceptive techniques that hide or omit information, but also through manipulative techniques that elicit particular emotional states through the framing of the message, or the way visual elements are presented.

Given the exploratory nature of this study, future research could replicate it, building on the same theoretical framework with larger and more representative samples. Demographic data such as age, education and digital competence could be collected to measure variability, so that the issue of vulnerability introduced in the first chapter could be addressed as a formal research question. Also, participants could be exposed to a realistic website, in which the donation request is presented as real and, consequently, a real risk of fraud. In this way, it could be verified to what extent awareness of the hypothetical nature of the experiment affects the final choice and related emotions.

The topic addressed by this work has implications for different actors. For consumers and users online, it raises awareness of their vulnerability to unethical practices, as the more informed they are about these practices when browsing the Internet, the more they can protect their freedom of choice and their economic interests. Moreover, companies and designers should use marketing strategies ethically, balancing their profits with consumers' interests. Persuasion can be considered ethical when high-quality products or real advantageous offers are promoted, but it becomes unethical manipulation when consumers are induced to make disadvantageous choices presented as the best options. Finally, this knowledge is useful for policymakers to further develop consumer protection law online and to limit unethical practices. As the online environment constantly evolves, more comprehensive regulation over unethical practices are necessary to ensure that the choice architecture supports consumers rather than manipulating them.

References

Ariely, D. (2008). Predictably irrational: The hidden forces that shape our decisions. HarperCollins Publishers.

Boerman, S. C., Kruikemeier, S. & Zuiderveen Borgesius, F. J. (2017). Online Behavioral Advertising: A Literature Review and Research Agenda. In *Journal of Advertising*, 46(3), pp. 363-376.

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the dark side: Privacy dark strategies and privacy dark patterns. In *Proceedings on privacy enhancing technologies.*

Brignull, H. (2010). *Deceptive Patterns.* https://deceptive.design/

Brignull, H. (2010). Dark Patterns: Dirty Tricks Designers Use to Make People Do Stuff. Available at: https://deceptive.design/articles/dark-patterns-dirty-tricks-designers-use-to-make-people-do-stuff/

Cole, A. (2016). All of us are vulnerable, but some are more vulnerable than others: the political ambiguity of vulnerability studies, an ambivalent critique. In *Critical Horizons*, 17(2), pp. 260–277.

Congiu, L. & Moscati, I. (2022). A review of nudges: Definitions, justifications, effectiveness. In *Journal of Economic Surveys*, 36, pp. 188–213. https://doi.org/10.1111/joes.12453

Conti, G. & Sobiesk, E. (2010). Malicious interface design: exploiting the user. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. pp. 271–280. https://doi.org/10.1145/1772690.1772719

DETOUR Act, H.R. 8975, 116th Cong. (2020). https://www.congress.gov/bill/116th-congress/house-bill/8975

Federal Trade Commission. (2023). *FTC takes action against Amazon for enrolling consumers in Amazon Prime without consent and sabotaging their attempts to cancel* [Press release] https://www.ftc.gov/news-events/news/press-releases/2023/06/ftc-takes-action-against-amazon-enrolling-consumers-amazon-prime-without-consent-sabotaging-their

Federal Trade Commission. (2024). *FTC takes action against Adobe and executives for hiding fees, preventing consumers from easily cancelling software subscriptions* [Press release]. https://www.ftc.gov/news-events/news/press-releases/2024/06/ftc-takes-action-against-adobe-executives-hiding-fees-preventing-consumers-easily-cancelling

Fradkin, A. (2018). Digital marketplaces. In *The new palgrave dictionary of economics,* pp. 2885-2898. Palgrave Macmillan, London.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, L. A. (2018). The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. 534, pp. 1–14. https://doi.org/10.1145/3173574.3174108

Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. In *Journal of Consumer Research*, 9(1), pp. 90–98. https://doi.org/10.1086/208899

Hubert, M. & Kenning, P. (2008). A current overview of consumer neuroscience. In *Journal of Consumer Behaviour*, 7(4-5):272-292. https://doi.org/10.1002/CB.251

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? In *Science,* 302(5649), pp. 1338-1339.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kelly-Holmes, H. (2015). Digital Advertising. In *The Routledge Handbook of Language and Digital Communication*.

Keynes, J. M. (1936). The General Theory of Employment, Interest and Money. Wordsworth Classics of World Literature. Herts, UK: Wordsworth Editions.

Kollmer, T. & Eckhardt, A. (2022). Dark Patterns: Conceptualization and Future Research Directions. In *Business & Information Systems Engineering,* 65(3). https://doi.org/10.1007/s12599-022-00783-7

Kunda, Z. (1990). The case for motivated reasoning. In *Psychological Bulletin*, 108(3), pp. 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lies, J. (2019). Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing. In *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(5), pp. 134-144.

Luguri, J. & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. In *The Journal of Legal Analysis,* 13(1), pp. 43-109.

Lukoff, K., Hiniker, A., Gray, C. M., Mathur, A., Chivukula, S. S. (2021). What can CHI do about dark patterns? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems,* ACM. https://doi.org/10.1145/3411763.3441360

Lynx. (2025). *Dark Pattern: il lato oscuro del design*. https://www.lynx2000.it/blog/dark-pattern-il-lato-oscuro-del-design/

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. In *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp. 1-32. https://doi.org/10.1145/3359183

Mathur, A., Mayer, J. & Kshirsagar, M. (2021). What Makes a Dark Pattern... Dark?: Design Attributes, Normative Considerations, and Measurement Methods. In *CHI Conference on Human Factors in Computing Systems*, pp. 1-18. https://doi.org/10.1145/3411764.3445610

Patt, A., & Zeckhauser, R. (2000). Action bias and environmental decisions. In *Journal of Risk and Uncertainty*, 21(1), pp. 45-72.

Pflügner, K., Maier, C. & Weitzel, T. (2021). The direct and indirect influence of mindfulness on techno-stressors and job burnout: A quantitative study of white-collar workers. In *Computers in Human Behavior*, 115. https://doi.org/10.1016/j.chb.2020.106566.

Quirk, T. J., & Rhiney, E. (2021). Excel 2019 for marketing statistics: A guide to solving practical problems. Springer.

Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In *Recommender systems handbook* pp. 1-35. Boston, MA: springer US.

Saunders, Mark & Lewis, P. & Thornhill, A. (2023). Research Methods for Business Students.

Sunstein, C. R. (2019). Sludge and Ordeals. In *Duke Law Journal,* 68, pp. 1843-1883. Available at: https://scholarship.law.duke.edu/dlj/vol68/iss8/6

Thaler, R. H. (2018). Nudge, not sludge. In *Science*, 361(6401), p. 431.

Thaler, R. H. & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2010). Choice Architecture. In *SSRN Electronic Journal*. Available at: https://ssrn.com/abstract=1583509

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. In *Science*, 185(4157), pp. 1124-1131.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. In *Science*, 211(4481), pp. 453–458. https://doi.org/10.1126/science.7455683

Wahab, F. K. & Abdulmajeed, T. I. (2025). The Rise of E-Commerce and Digital Marketplaces: Implications for Consumer Engagement and Market Efficiency. In *Journal of Business & Entrepreneurship*, 35(1), Article 4. https://repository.ulm.edu/jbe/vol35/iss1/4

Weinmann, M., Schneider, C., & vom Brocke, J. (2016). Digital Nudging. In *Business & Information Systems Engineering*, 58(6), pp. 433 - 436.

Westin, F., & Chiasson, S. (2019). Opt out of Privacy or "Go Home": Understanding Reluctant Privacy Behaviours through the FoMO-Centric Design Paradigm. In *Proceedings of the New Security Paradigms Workshop (NSPW '19)*, ACM, pp. 57–67. https://doi.org/10.1145/3368860.3368865

Wikimedia Commons. (2019). *Nutri-score-A light background logo.svg*. https://commons.wikimedia.org/wiki/File:Nutri-score-A_light_background_logo.svg#filehistory

Zac, A., Huang, Y. C., Von Moltke, A., Decker, C. & Ezrachi, A. (2025). Dark patterns and consumer vulnerability. In *Behavioural Public Policy*, pp. 1-50. https://doi.org/10.1017/bpp.2024.49

Appendix A: Surveys (Group A + Group B)

In this section only one survey was reported as the two versions are identical, except for the checkout mockups, which are both shown in Figures 24 and 25.

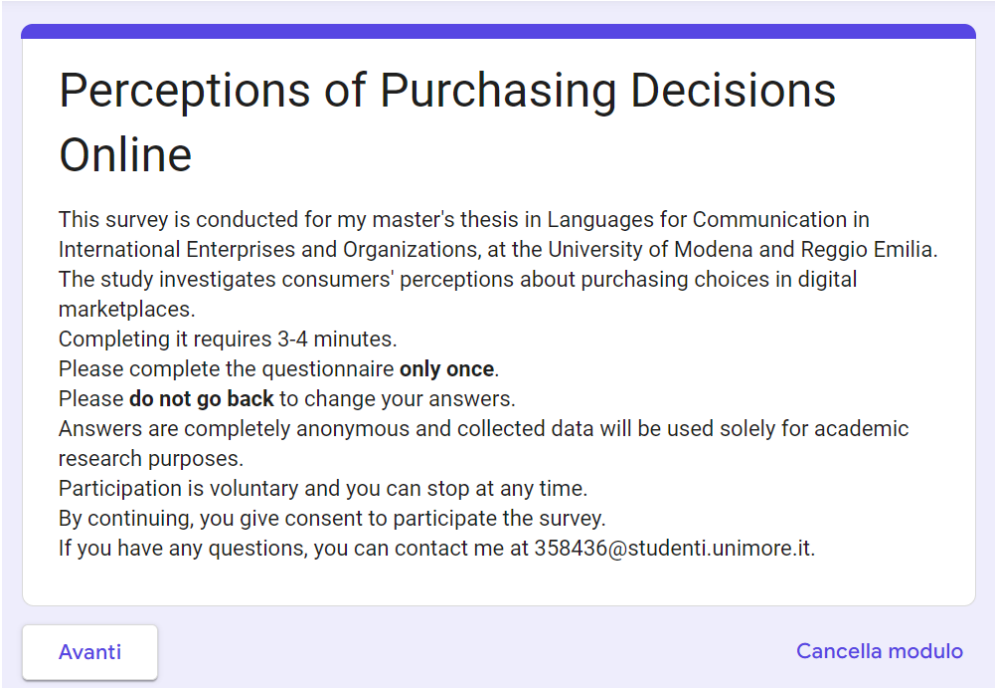


Figure 23: Section 1 of the survey

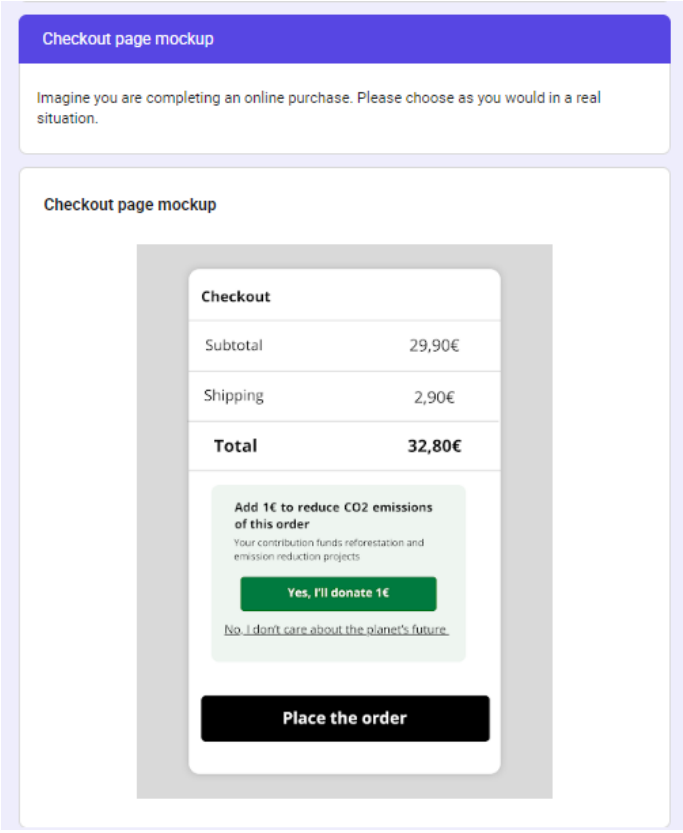


Figure 24: Section 2 of the survey (Group A – Neutral)

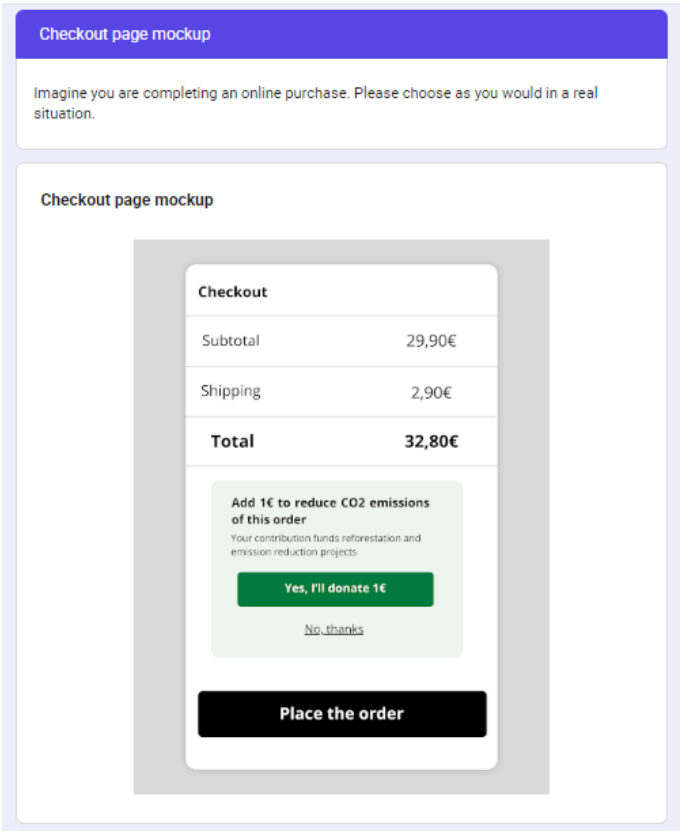


Figure 25: Section 2 of the survey (Group B – Confirmshaming)

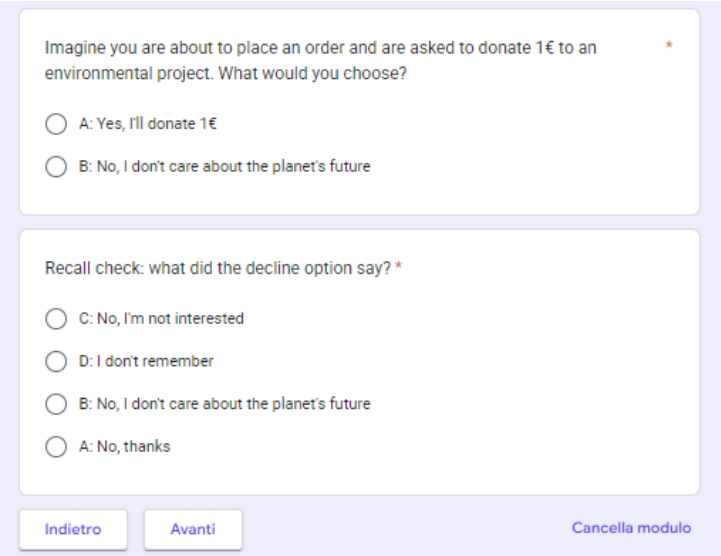


Figure 26: Section 2 of the survey

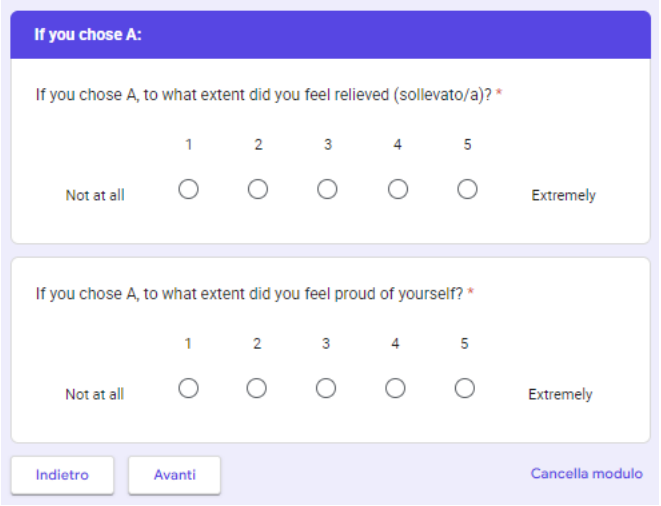


Figure 27: Section 3 of the survey

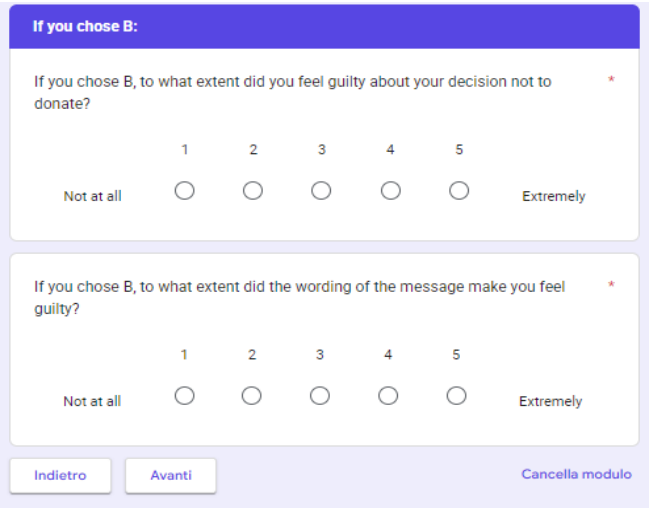


Figure 28: Section 4 of the survey

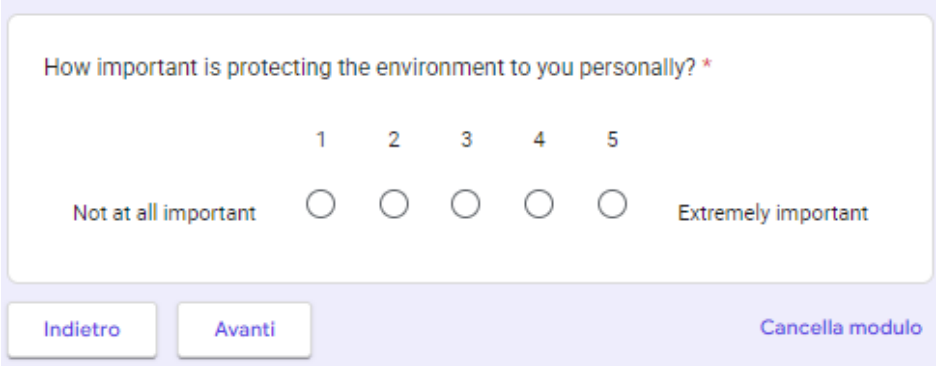


Figure 29: Section 5 of the survey

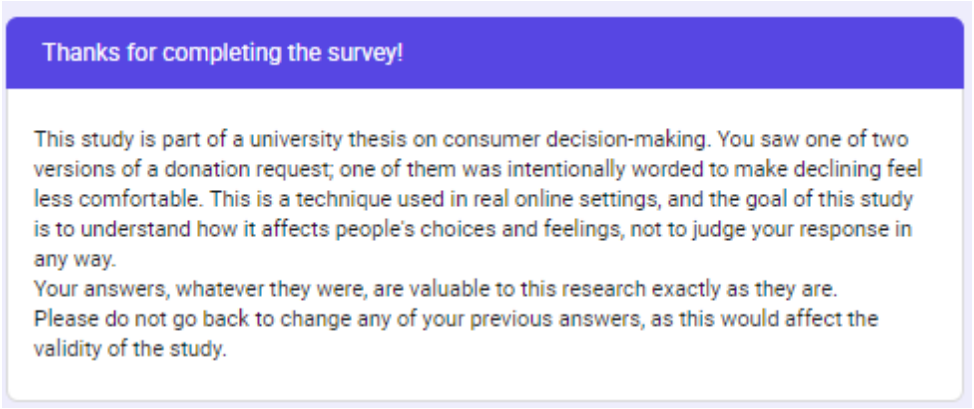


Figure 30: final section of the survey

Appendix B: Excel tables and formulas

	A	B	C	D	E	F	G	H	I	J
1	Participants	Condition	Accepted (A) / Declined (B)	Accepted (0) / Declined (1)	Recall check	Relief	Pride	Guilt for decision	Guilt for wording	Environ. Cause
2	1	Neutral	B: No, thanks	1	A: No, thanks			1	1	4
3	4	Neutral	B: No, thanks	1	A: No, thanks			1	1	5
4	5	Neutral	B: No, thanks	1	A: No, thanks			3	2	4
5	6	Neutral	B: No, thanks	1	A: No, thanks			1	1	4
6	8	Neutral	B: No, thanks	1	A: No, thanks			3	3	5
7	9	Neutral	B: No, thanks	1	A: No, thanks			1	2	5
8	10	Neutral	B: No, thanks	1	A: No, thanks			2	3	3
9	11	Neutral	B: No, thanks	1	A: No, thanks			3	2	4
10	12	Neutral	A: Yes, I'll donate 1€	0	A: No, thanks	5	4			5
11	14	Neutral	B: No, thanks	1	A: No, thanks			4	4	4
12	16	Neutral	B: No, thanks	1	A: No, thanks			2	1	5
13	18	Neutral	B: No, thanks	1	A: No, thanks			1	1	4
14	19	Neutral	B: No, thanks	1	A: No, thanks			1	2	4
15	20	Neutral	B: No, thanks	1	A: No, thanks			2	2	4
16	21	Neutral	B: No, thanks	1	A: No, thanks			1	1	3
17	22	Neutral	B: No, thanks	1	A: No, thanks			1	1	5
18	23	Neutral	B: No, thanks	1	A: No, thanks			4	1	5
19	24	Neutral	B: No, thanks	1	A: No, thanks			2	3	5
20	26	Neutral	B: No, thanks	1	A: No, thanks			1	1	4
21	27	Neutral	B: No, thanks	1	A: No, thanks			1	1	4
22	28	Neutral	B: No, thanks	1	A: No, thanks			1	3	5
23	29	Neutral	A: Yes, I'll donate 1€	0	A: No, thanks	2	3			5
24	30	Neutral	B: No, thanks	1	A: No, thanks			3	3	3
25	31	Neutral	B: No, thanks	1	A: No, thanks			2	2	5
26	32	Neutral	B: No, thanks	1	A: No, thanks			3	3	5

Figure 31: Cleaned data (Group A)

	A	B	C	D	E	F	G	H	I	J
1	Participants	Condition	Accepted (A) / Declined (B)	Accepted (0) / Declined (1)	Recall check	Relief	Pride	Guilt for decision	Guilt for wording	Environ. Cause
34	45	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	2	2			
35	46	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			4	5	
36	47	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			2	4	
37	48	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			3	1	
38	50	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	2	2			
39	53	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	4	5			
40	54	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	4	3			
41	55	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			3	4	3
42	56	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	3	4			4
43	57	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			3	4	5
44	59	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	4	5			5
45	61	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	4	5			5
46	62	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	3	3			5
47	66	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	4	4			5
48	67	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	5	4			5
49	69	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	3	3			4
50	73	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	4	4			4
51	74	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	3	3			5
52	76	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	3	3			5
53	77	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			5	5	5
54	78	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	5	5			5
55	79	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			4	4	5
56	80	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			1	2	2
57	82	Confirmshaming	A: Yes, I'll donate 1€	0	B: No, I don't care about the planet's future	3	4			4
58	83	Confirmshaming	B: No, I don't care about the planet's future	1	B: No, I don't care about the planet's future			1	2	5

Figure 32: Cleaned data (Group B)

F79 | =H74*F76/H76

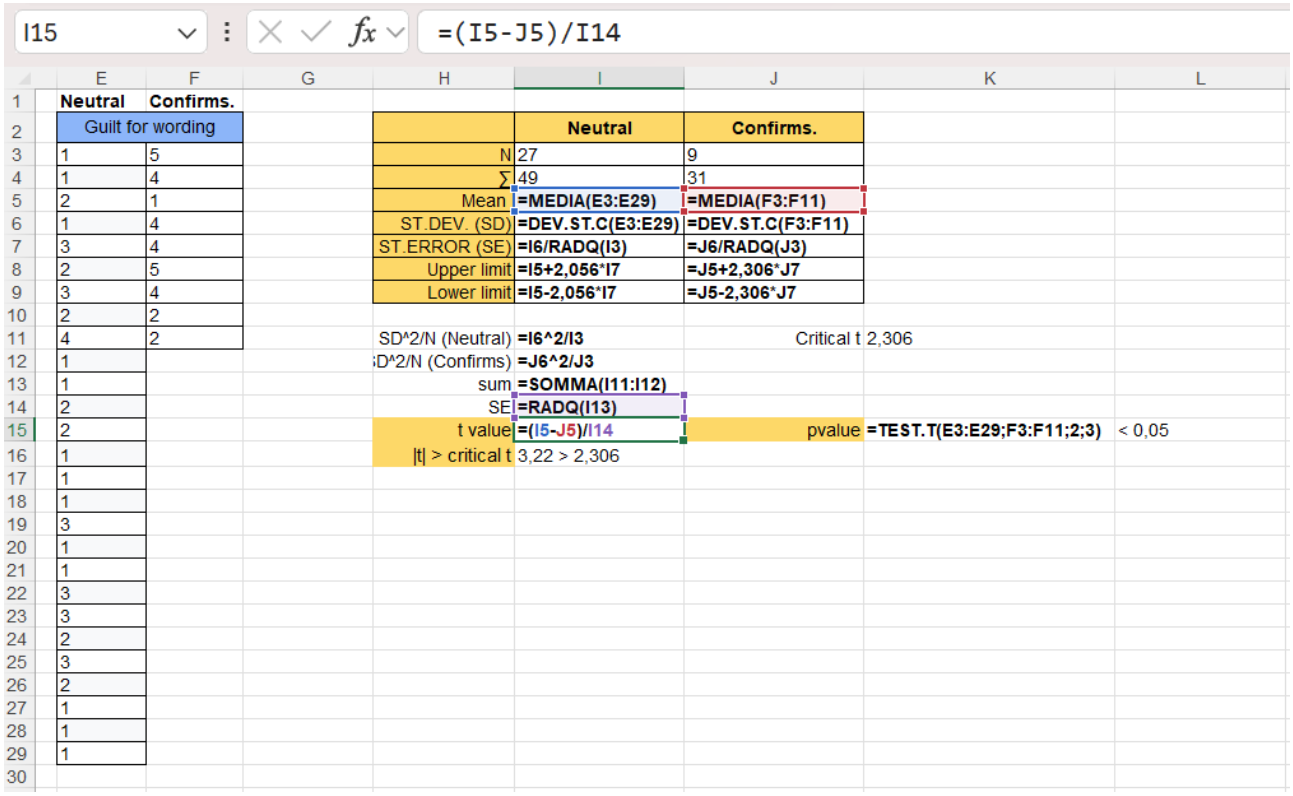
	B	C	D	E	F	G	H
72	Neutral	Confirms.					
73	1	0		**Observed (O)**	Accepted (0)	Declined (1)	Total
74	1	1		Neutral	5	27	32
75	1	1		Confirmshaming	16	9	25
76	1	1		Total	21	36	57
77	1	0					
78	1	0		**Expected (E)**	Accepted (0)	Declined (1)	
79	1	0		Neutral	=H74*F76/H76	=H74*G76/H76	
80	1	1		Confirmshaming	=H75*F76/H76	=H75*G76/H76	
81	0	0					
82	1	1					
83	1	0		(O-E)^2/E			
84	1	0		Neutral (0)	=(F74-F79)^2/F79		
85	1	0		Neutral (1)	=(G74-G79)^2/G79		
86	1	0		Confirmshaming (0)	=(F75-F80)^2/F80		
87	1	0		Confirmshaming (1)	=(G75-G80)^2/G80		
88	1	0					
89	1	0		X^2=	=SOMMA(F84:F87)		
90	1	0					
91	1	0		Critical value	3,84		
92	1	1		df	1		
93	1	0		p-value	=TEST.T(B72:B104;C72:C97;2;3)	<0,05	
94	0	1					

Figure 33: Acceptance Rates on sample N=57

M13 | =(M3-M8)^2/M8

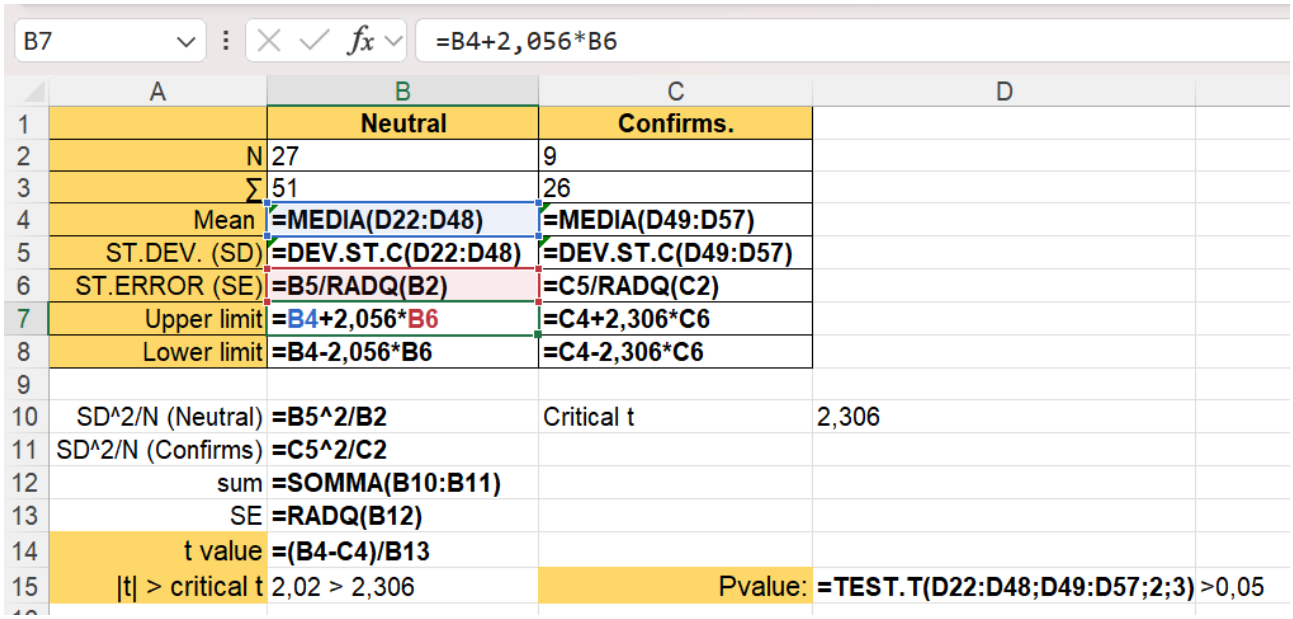
	K	L	M	N	O
1					
2		**Observed (O)**	Accepted (0)	Declined (1)	Total
3		Neutral	7	35	=M3+N3
4		Confirmshaming	19	23	=M4+N4
5		Total	=M3+M4	=N3+N4	84
6					
7		**Expected (E)**	Accepted (0)	Declined (1)	
8		Neutral	=O3*M5/O5	=O3*N5/O5	
9		Confirmshaming	=O4*M5/O5	=O4*N5/O5	
10					
11					
12		(O-E)^2/E			
13		Neutral (0)	=(M3-M8)^2/M8		
14		Neutral (1)	=(N3-N8)^2/N8		
15		Confirmshaming (0)	=(M4-M9)^2/M9		
16		Confirmshaming (1)	=(N4-N9)^2/N9		
17					
18		X^2	=SOMMA(M13:M16)	>3,84	
19					
20		Critical value	3,84		
21		df	1		
22		p-value	=TEST.T(H2:H43;I2:I43;2;3)	<0,05	

Figure 34: Acceptance rates on sample N=84



	E	F	G	H	I	J	K	L
1	Neutral	Confirms.						
2	Guilt for wording				Neutral	Confirms.		
3	1	5		N	27	9		
4	1	4		∑	49	31		
5	2	1		Mean	=MEDIA(E3:E29)	=MEDIA(F3:F11)		
6	1	4		ST.DEV. (SD)	=DEV.ST.C(E3:E29)	=DEV.ST.C(F3:F11)		
7	3	4		ST.ERROR (SE)	=I6/RADQ(I3)	=J6/RADQ(J3)		
8	2	5		Upper limit	=I5+2,056*I7	=J5+2,306*J7		
9	3	4		Lower limit	=I5-2,056*I7	=J5-2,306*J7		
10	2	2						
11	4	2		SD^2/N (Neutral)	=I6^2/I3	Critical t	2,306	
12	1			SD^2/N (Confirms)	=J6^2/J3			
13	1			sum	=SOMMA(I11:I12)			
14	2			SE	=RADQ(I13)			
15	2			t value	=(I5-J5)/I14	pvalue	=TEST.T(E3:E29;F3:F11;2;3)	< 0,05
16	1			|t| > critical t	3,22 > 2,306			
17	1							
18	1							
19	3							
20	1							
21	1							
22	3							
23	3							
24	2							
25	3							
26	2							
27	1							
28	1							
29	1							
30								

Figure 35: t-test on Guilt for Wording



	A	B	C	D
1		Neutral	Confirms.	
2	N	27	9	
3	∑	51	26	
4	Mean	=MEDIA(D22:D48)	=MEDIA(D49:D57)	
5	ST.DEV. (SD)	=DEV.ST.C(D22:D48)	=DEV.ST.C(D49:D57)	
6	ST.ERROR (SE)	=B5/RADQ(B2)	=C5/RADQ(C2)	
7	Upper limit	=B4+2,056*B6	=C4+2,306*C6	
8	Lower limit	=B4-2,056*B6	=C4-2,306*C6	
9				
10	SD^2/N (Neutral)	=B5^2/B2	Critical t	2,306
11	SD^2/N (Confirms)	=C5^2/C2		
12	sum	=SOMMA(B10:B11)		
13	SE	=RADQ(B12)		
14	t value	=(B4-C4)/B13		
15	|t| > critical t	2,02 > 2,306	Pvalue:	=TEST.T(D22:D48;D49:D57;2;3) >0,05

Figure 36: t-test on Guilt for Decision

B10 =C5^2/C2

	A	B	C	D	
1		**Neutral**	**Confirms.**		
2	N	5	16		
3	∑	17	59		
4	Mean	**=MEDIA(J2:J6)**	**=MEDIA(J7:J22)**		
5	ST.DEV. (SD)	**=DEV.ST.C(J2:J6)**	**=DEV.ST.C(J7:J22)**		
6	ST.ERROR (SE)	**=B5/RADQ(B2)**	**=C5/RADQ(C2)**		
7	Upper limit	**=B4+2,776*B6**	**=C4+2,131*C6**		
8	Lower limit	**=B4-2,776*B6**	**=C4-2,131*C6**		
9	SD^2/N (Neutral)	**=B5^2/B2**			
10	SD^2/N (Confirms)	**=C5^2/C2**			
11	sum	**=SOMMA(B9:B10)**			
12	SE	**=RADQ(B11)**			
13	t value	**=(B4-C4)/B12**	Critical t	2,776	
14	|t| > critical t	0,50 < 2,776	pvalue	**=TEST.T(J2:J6;J7:J22;2;3)**	>0,05

Figure 37: t-test on Pride

B7 =B4+2,776*B6

	A	B	C	D	
1		**Neutral**	**Confirms.**		
2	N	5	16		
3	∑	18	56		
4	Mean	**=MEDIA(I2:I6)**	**=MEDIA(I7:I22)**		
5	ST.DEV. (SD)	**=DEV.ST.C(I2:I6)**	**=DEV.ST.C(I7:I22)**		
6	ST.ERROR (SE)	**=B5/RADQ(B2)**	**=C5/RADQ(C2)**		
7	Upper limit	**=B4+2,776*B6**	**=C4+2,131*C6**		
8	Lower limit	**=B4-2,776*B6**	**=C4-2,131*C6**		
9	SD^2/N (Neutral)	**=B5^2/B2**			
10	SD^2/N (Confirms)	**=C5^2/C2**			
11	sum	**=SOMMA(B9:B10)**			
12	SE	**=RADQ(B11)**			
13	t value	**=(B4-C4)/B12**	critical t	2,776	
14	|t| > critical t	0,14 < 2,776	pvalue:	**=TEST.T(I2:I6;I7:I22;2;3)**	>0,05

Figure 38: t-test on Relief

K3 =TEST.T(G2:G19;H2:H33;2;3)

	A	B	C	D	E	F	G	H	I	J	K
1	Participants	Condition	0 Accepted / 1 Declined	Environ. Cause	Level of attachment		Environ. Cause Confirmshaming	Environ. Cause Neutral			
2	57	Confirmshaming	1	5	High		5	5			
3	59	Confirmshaming	0	5	High		5	5		P value:	0,51
4	61	Confirmshaming	0	5	High		5	5			
5	62	Confirmshaming	0	5	High		5	5			
6	66	Confirmshaming	0	5	High		5	5			
7	67	Confirmshaming	0	5	High		5	5			
8	74	Confirmshaming	0	5	High		5	5			
9	76	Confirmshaming	0	5	High		5	5			
10	77	Confirmshaming	1	5	High		5	5			
11	78	Confirmshaming	0	5	High		5	5			
12	79	Confirmshaming	1	5	High		5	5			
13	83	Confirmshaming	1	5	High		5	5			
14	56	Confirmshaming	0	4	Medium/Low		4	5			
15	69	Confirmshaming	0	4	Medium/Low		4	5			
16	73	Confirmshaming	0	4	Medium/Low		4	4			
17	82	Confirmshaming	0	4	Medium/Low		4	4			
18	55	Confirmshaming	1	3	Medium/Low		3	4			
19	80	Confirmshaming	1	2	Medium/Low		2	4			
20	4	Neutral	1	5	High			4			
21	8	Neutral	1	5	High			4			

Figure 39: Levels of attachment to the environmental cause