

# Testing for Manipulation: Experimental Evidence on Dark Patterns

FRANCESCO BOGLIACINO, LEONARDO PEJSACHOWICZ, GIOVANNI LIVA, FRANCISCO LUPÍÁÑEZ-VILLANUEVA<sup>1</sup>

---

<sup>1</sup> Bogliacino: *Corresponding author, Dipartimento di Scienze Economiche, Università degli Studi di Bergamo, Via dei Caniana 2, 24127, Bergamo, Italy, [francesco.bogliacino@unibg.it](mailto:francesco.bogliacino@unibg.it)*, Pejsachowicz: Université de Paris I La Sorbonne, Liva: Open Evidence, Lupiáñez-Villanueva: Universitat Oberta de Catalunya. Funding: *EU Consumer Programme (2014-2020) under a service contract with the European Innovation Council and SMEs Executive Agency (EISMEA) acting under the mandate from the European Commission. Behavioural Study on Unfair Commercial Practices in the Digital Environment (European Commission - Framework Contract with reopening of competition - Behavioural studies) and Agence Nationale de la Recherche (programme d'Investissements d'avenir EUR)*. We are grateful to Egelyn Braun, Hannah Nohlen, and Silvia Pella for comments during the project. A special thanks to Marcello Negrini for commenting on a previous draft. We thank Cristiano Codagnone, Teresa Rodríguez de las Heras Ballell and Lucie Lechardoy for help and discussion during the writing of the proposal and the development of the project. A special thanks to Alba Boluda for excellent research assistance throughout the project. We thank Setecem for programming the online experiment and Schlesinger group for data collection. We thank the lab at PSE and in particular Maxim Frolov for assistance for the second experiment. We are grateful for comments received at various seminars and conferences, including BEBES in Bogotá, the 4<sup>th</sup> Winter Workshop in Bergamo, and a seminar at PSE-Paris I. All the remaining errors are ours. The information and views set out in this article are those of the author(s) and do not necessarily reflect the official opinion of the Commission/Executive Agency. The Commission/Executive Agency do not guarantee the accuracy of the data included in this study. Neither the Commission/Executive Agency nor any person acting on the Commission's/Executive Agency's behalf may be held responsible for the use which may be made of the information contained therein.

## Abstract

Several countries and supranational authorities are debating whether to regulate or ban dark patterns, deceptive users' interfaces. A key empirical component to this debate is how to assess manipulation. In this study, we develop a transaction test which measure to what extent the dark patterns lead to decisions inconsistent with elicited preferences.

We conducted a large preregistered online study (N=7430) with a representative population of six countries to identify both the effect of dark patterns on consumers' choice consistency and the potential counteracting effects of protective measures. Our treatments include three dark patterns - hiding information on the product, toying-with-emotions, and the use of psychological profiling to personalize the display for the consumer – and two versions of a protective measure that discloses information and requires subject to confirm the selection.

Participants are assigned to either a motivated delay or incentive compatible time pressure environment, allowing to identify the impact of treatments on consumers paying enough attention and on situationally vulnerable consumer.

Dark patterns do manipulate consumers, showing remarkable effects on both average and vulnerable consumers. The cool down intervention has a null effect.

We stress test the transaction test in a controlled experiment, where the preference elicitation is incentive compatible, we collect repeated measurement of choices among lotteries and we manipulate the extent of the mistake.

In this additional experiment, the TWE treatment resulted in greater inconsistency compared to the control group, particularly in lotteries where the point of indifference was less likely to be located at the boundaries of the MPL grid. While subjects learned to be consistent through multiple rounds of choice and with decision problems further from their area of indifference, the learning effect is less pronounced under the TWE treatment.

**Keywords:** dark patterns; manipulation; inconsistency; time-pressure

**JEL Classification:** C92; D18; D91; L13

## 1. Introduction

Online platforms employ website customization techniques that subtly influence users' choices in a manner that can be frustrating and appear misleading. For instance, they may make it difficult to unsubscribe, add items to the shopping cart without clear consent, or dynamically disclose prices. For all these instances of online choice architectures, Brignull (2010) introduced the concept of dark patterns: user interfaces that intentionally confuse, coerce, deceive the consumers, leading to decisions inconsistent with individual preferences. Dark patterns differ from standard sales techniques whose aim is to persuade consumers. Online platforms continuously introduce new versions of deceptive design and deploy them at scale (Bösch et al., 2016; Cara, 2019; Luguri & Strahilevitz, 2021; Mathur et al., 2019). Dark patterns are typically restricted to online transactions, where the *consent of the parties*, sovereign in the contractual law, no longer represents the gold standard. In fact, consent represents the gold standard (i) when it is sparsely expressed, (ii) when the harms are easy to envisage, and (iii) when there are the right incentives to consent (Richards & Hartzog, 2019). Each of the three criteria is less likely to be satisfied in online transactions.

Several countries and supranational bodies have been undertaking initiatives to regulate dark patterns, sparking heated discussions (BEUC, 2022; Calo, 2014; Guidance on the Interpretation and Application of Directive 2005/29/EC of the European Parliament and of the Council Concerning Unfair Business-to-Consumer Commercial Practices in the Internal Market, 2021; Hartzog, 2018; Netherlands Authority for Consumers & Markets, 2020; Norwegian Consumer Council, 2018; Richards & Hartzog, 2019; SERNAC, 2021, 2022; UK Competition and Market Authority, 2022). Despite the widespread utilization of dark patterns in user interfaces and the presumably extensive undisclosed internal research conducted by platforms, there is a scarcity of empirical evidence examining the adverse impact of these patterns on consumer choices. Furthermore, we have limited knowledge regarding the potential effectiveness of protective measures against such practices. A plausible rationale behind this lack of understanding lies in the absence of a consensus regarding a definitive empirical examination. Dark patterns are not indicted based on their efficacy alone (as marketing strategies can also be effective), but rather due to their manipulative nature.

In this article, we provide a controlled test of manipulation by dark patterns. Participants are presented with a choice between two goods, A and B, which are presented as entertainment packages provided by a platform (no specific labels are used in the experiment). B is identical to A but requires sharing personal data with a third party. Although B is priced lower, the individual price is intentionally set below the participants' stated willingness to exchange money for data protection. By choosing this option, participants violate the Weak Axiom of Revealed Preferences (Varian, 2006). This transaction test allows us to induce value for consistency, as participants are incentivized to choose in line with their revealed preferences. Additionally, it enables us to examine how various dark patterns, such as hidden information, toying with emotions (TWE), and personalization, increase the likelihood of inconsistent choices when targeting alternative B. One notable advantage of our approach is its portability and adaptability to different dark patterns and the associated protective measures.

Data for our assessment comes from an online experiment in six countries, our sample is high powered (around 1200 observations per country) and representative (it matches the population in each country according to several observable characteristics).

Our task tests for manipulation in a controlled environment in a way that agrees with the legal discipline in Europe and the US. The discipline of dark patterns in Europe falls in the domain of the Unfair Commercial Practice Directive, which regulates “any act, omission, course of conduct or representation, commercial communication including marketing, by a trader, directly connected with the promotion, sale, or supply of a product to consumers” (*DIRECTIVE 2005/29/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL “Unfair Commercial Practice Directive,”* 2005). The Directive states that practices are unfair if they breach professional diligence, are deceptive, or coercive, and if they cause or be likely to cause an average consumer to take a transactional decision that they would not have taken otherwise. The US discipline states the following: an unfair trade practice is one that causes or is likely to cause a financial loss to consumers, not compensated by the benefits, and is not reasonably avoidable by a consumer (*FTC Policy Statement on Unfairness,* 1984). Our transaction test aligns with the criteria for unfairness outlined in both the Directive and the US discipline. In the EU context, our test serves as a reduced form evaluation of the deceptive nature and effectiveness of the practice. In the US interpretation, we directly apply the criterion by measuring preferences, enabling us to quantify the net outcomes in terms of benefits and financial costs.

We manipulate two factors in a between subject design. Factor one is the interface. To assess the impact of dark patterns and protective measure, we randomly expose a subset of participants to one of the following three dark patterns. Hidden information is a dark pattern obfuscating some information that can be accessed through further actions, in our case, clicking over three dots. TWE is the delivery of messages that recommend the purchase by leveraging regret aversion. Personalization is the label used for specialized offers built upon previously collected navigation data, obtained through the platform. Building on previous results in psychological targeting (Matz et al., 2017), we use the self-reported profiling on an extroversion-introversion scale to choose a picture that is likely to increase the appeal of the product and we add it to the TWE treatment.

To assess a protective measure, we test two versions of a “cool down”. The cool down is implemented as a confirmation prompt, grounded in the principle that autonomous decision-making necessitates a thoughtful evaluation and potential reconsideration of actions (Dworking, 1988). Participants are exposed to the dark pattern and subsequently presented with a summary of their selection, requiring them to confirm their choice. This constitutes the first version of the cool down, which is customized differently for each dark pattern. The second version of the cool down is specific to personalization, wherein the confirmation prompt is accompanied by a disclosure statement informing participants that their information has been utilized to tailor the offer.

To summarize, factor one has eight conditions: three dark patterns, four treatments with cool down, and the control.

Factor two has two levels. We randomly assign participants to a motivated delay or an incentive compatible time pressure environment (Alós-Ferrer & Garagnani, 2020). The existing legislation values the manipulative nature of the practice from the perspective of an “average” consumer, which pays sufficient attention and processes the relevant information. Essentially, consumers are expected to exercise caution and not blindly accept the assertions made by sellers, unless they are deemed vulnerable due to structural or situational factors. The European *average consumer* (well informed and careful) and the American *consumer reasonably attentive* are close to our motivated delay condition. Similarly, the incentivized time pressure condition is close to the vulnerable consumer.

We found that the three dark patterns increase the likelihood of inconsistent choices for an average consumer, with an estimated effect of 25.5% (hidden information), 12.5% (toying with emotions), and 20.5% (toying with emotions and personalization) of a standard deviation (computed for the average consumer in the control). Consumers under time pressure become much more vulnerable, increasing their likelihood of inconsistency up to 45%. The effect of dark patterns continues to be positive and significant, but smaller in size: 11.8% (hidden information), 8.4% (toying with emotions), and 11.2% (toying with emotions and personalization) of a standard deviation (computed for the vulnerable consumer in the control). Our protective measure does not provide a solution. The cool down reduces the likelihood of inconsistency but results are not statistically significant at conventional levels.

There is an alternative way to interpret our results. One reason to assume consistency of choice is that when consistency is violated, the consumer can be money pumped (Angner, 2012; Echenique et al., 2013). A person choosing Package B, when stating a willingness to exchange the data protection for  $x$  euros would go through the following transactions: owning the A would shift to B in exchange for a  $y(<x)$  euros discount (based on the choice she made), and then would be willing to exchange  $x$  euros for data protection (via the stated preferences). At the end of the transaction, she would get back the original A package but having lost  $|x-y|$  euros.<sup>2</sup> In our experiment, the incentives for consistent choices are ten times those for inconsistency. Think at this 90% as the size of the money pump. A dark pattern which increases the odds of selecting package B by one percentage point (pp) would induce an expected loss of 0.9%. Given our results, the potential loss for the average consumer from hidden information amounts to 18.45%, from TWE to 11.25%, and from TWE with personalization to 18.45%. For the vulnerable consumer, the potential loss from hidden information amounts to 10.62%, from TWE to 6.72%, and from TWE with personalization to 10.08%. Given the scale of operation of large platforms, these potential consumer losses project into sizable, extracted rents.

To test the robustness of our findings, we run an additional experiment with a traditional lab subject pool. In an MPL task, the participants reveal their point of indifference between a set of four mixed lotteries and four alternatives where there is a trade-off between outcome and probabilities. Subsequently, the subjects perform a set of dichotomic choices that are randomly

---

<sup>2</sup> Notice that we explicitly frame our task using the willingness to exchange to avoid falling into the WTA-WTP gap. Even assuming the latter though, inconsistency of choice is still present because the gap cannot account for a difference in price around 50% of the expenditure (Chapman et al., 2021). And clearly it is incompatible with our incentive system where the cost of inconsistency is tenfold.

rows from the MPLs, under two dark pattern conditions, where the design promotes inconsistency, and a control.

Our experiment complements the online in several ways: it elicits preferences ensuring incentive compatibility, administers the treatment within subjects with repeated measures allowing for learning, and introduces variation in the degree of inconsistency as the choices vary according to their distance from the point of indifference. Finally, it provides symmetry between the two options and allows dark patterns targeting both the left and the right lottery, whereas the online task was designed to always focus on the option without data protection.

The setup with lotteries provides us with a stress test, as the choice between abstract objects deprives the settings from certain social expectations that reinforce the role of dark patterns. When facing a seller in a brick-and-mortar store or via an online platform, there exist shared beliefs regarding environmental cues and design features. Default settings are not assumed to be deceptive because this manipulation would breach a code of conduct. This makes the consumer more likely to fall for dark pattern. In our setting, the lack of framing introduces the consumer to a less familiar environment where social expectations are less entrenched. Additionally, dark patterns in the lab are obviously much milder.

A TWE treatment induces more inconsistency than the control, especially for those lotteries where the point of indifference was less likely to be placed at the boundaries of the MPL grid. While subjects become less inconsistent as the round or the distance (from the switching point) increase, the behavior stabilizes under dark patterns. In other words, we found that the dark pattern counteracts learning. A simple dark pattern where we highlight one of the options does not replicate the same results. This finding is informative for policy as simple nudging of a choice does not systematically alter the behavior.

Our findings contribute to the discussion on regulating dark patterns. The literature has not reached a consensus on how to assess dark patterns and their manipulative nature. Luguri and Strahilevitz (2021) run two large, ecologically valid experiments, where survey respondents are given the choice to unsubscribe from an online identity protection service with, depending on treatment and experiment, either increasingly higher levels or different types of dark patterns. They find dark patterns to be highly effective, but measure this by looking at the comparative take-up rate against control. Moreover, the study uses deception: survey participants are convinced by ruse that they have been subscribed to a service. Thus, their experimental method has a drawback as a general paradigm for the evaluation of dark patterns. Esposito et al. (2017) measure purchase of movies but somehow arbitrarily define some choices as compatible and some as incompatible. The problem is that in the incentive system, “compatibility” is not salient, and thus violates induced value.

Very compelling is the evidence that comes from *field* experiments, although limited to whether they alter behavior and not whether they manipulate. The consumer protection authority in Chile manipulates its own webpage to test different version of cookies consent notice (SERNAC, 2022). They conclude that a *bright pattern* (a Nudge) is the most effective, whereby the default is the most favorable option for the consumer (for a similar design, and results, see Grassl et al., 2021).

We contribute to this literature by providing a very flexible approach that can be adapted to the assessment of a several dark patterns and other forms of business practices. Moreover, since the literature has completely overlooked the difference between vulnerable and average consumers, we provide an adaption of an established method that can be easily integrated into alternative approaches.

In recent years, the popularity of behavioral economics has led to various studies assessing the impact of selling techniques to confuse consumers. Kalayci & Potters (2011) and Kalayci (2016) use “confusion” manipulations. In this series of experiments firms can obfuscate either quality or price of an alternative by increasing the number of parts the consumer needs to add up to recover the value in question. This fits well the study of a particular practice, add-on pricing, but is not adaptable to other types of dark patterns. A similar literature looks at strategic obfuscation (Gu & Wenzel, 2015, 2020). A few papers test strategic display of price (drip pricing) but mostly look at standard impact on efficiency and not on manipulation (Huck & Wallace, 2015; Rasch et al., 2020). Evidence on drip pricing comes also from two field experiments, both reporting large effects on purchase (again lacking a direct measure of manipulation) (Blake et al., 2021; Dertwinkel-Kalt et al., 2020). We could not find experimental economics papers on hidden information, toying with emotions, personalization.

This article also enhances our overall understanding of choice architectures. In contrast to the traditional research that aligns with libertarian paternalism (Thaler & Sunstein, 2008), dark patterns represent a *Nudge for Bad*. Nevertheless, it has been argued on multiple occasion that “choice architects do not always have the best interests of the people they are influencing in mind” (Akerlof & Shiller, 2015; Thaler, 2018; Thaler et al., 2012; Thaler & Sunstein, 2003), thus serving as a justification for behavioral interventions. In our contribution to this literature, we delve into the discussion of countermeasures aimed at mitigating manipulative strategies employed by companies.

The assumption of rationality as consistency of choices has a long story in economics and its empirical content is well known (Echenique et al., 2013; Varian, 2006). Reviewing the literature on the causal determinants of rationality violations is far beyond the scope of this insight, but we are not aware of any previous application to dark patterns. By the same token, this article contributes to the vast literature on persuasion and manipulation (Della Vigna & Gentzkow, 2010). Scholars studied persuasion and manipulation of consumers, donors, voters, and investors. The contribution of this paper is methodological, as it provides a transaction test for manipulation, and it is one of the very few controlled evidence of dark patterns manipulation.

## **2. Experimental Design**

To test the impact of dark patterns and protective measures, we designed a choice task where we induced value on consistency. Participants are offered two entertainment packages, one more expensive with data protection and one cheaper with data sold to third parties. Hereafter, we define Basic the former and Premium the latter (we do not use any label in the experiment). The difference in price is calibrated to be lower than the stated monetary willingness to exchange data protection. As a result, choosing the cheaper package violates consistency. Preferences are gathered in an instrumental task. To identify the effect of practices for vulnerable or average

consumers, participants are either paid to explain their choice or incentivized to make faster choices (Alós-Ferrer & Garagnani, 2020).

This is a between-subject design with 8X2 conditions. Factor 1 is the choice architecture, with the following conditions: control, three separate conditions with dark patterns (hidden information, toying-with-emotions, TWE with personalization) and four conditions with dark pattern and a protective measure (a cool down with a recall of the choice to be confirmed, for each of the three dark patterns, and a cool down with a disclosure on the use of data, for personalization). Dark patterns are always targeting the Premium package.

Factor 2 is a manipulation of the mode of deliberation, with two levels: deliberative versus intuitive. In the former, participants are paid to explain their choice, in the latter, participants are assigned to an incentivized time delay, where each second reduces the total amount of incentives from an initial endowment.

### *2.1 Tasks*

Participants go through the following experimental sequence.

In the first section, participants are described a scenario where they own an entertainment package provided by a platform. They are asked their willingness to accept (WTA) for five features of the package: data protection, interruption by commercials, access to movie pre-release, unsubscribe in one click, and access to a companion service (music platform). The WTA is stated in an eleven points Likert scale from 0 to 5 EUR or more, where each point corresponds to 50 cents (or equivalent in the local currency) discount from the monthly payment. The target question concerns data protection, but we included the other four as distractors to eliminate concerns for experimenter's effect. Questions are in random order. To ensure non-random answers, we follow a standard procedure and we incentivize participants to repeat the answer to one randomly chosen question (Diaz et al., 2021).

After stating their preferences, participants reveal their gender and personality profile (big five) in a standardized set of questions and then move to the explanation of the incentive systems for the main task, where they answer a comprehension question with feedback.

The main task is the choice between the Basic and Premium package. Experimental conditions are assigned between subjects. Factor one has the following conditions:

1. Control: the participant should choose between the two packages. The two packages are shown aside. Description includes price, additional features, and condition on data. Data are not shared in the Basic and shared with third parties in the Premium. The price is calibrated from the question on the WTA for data protection to be 2.5 euros lower than the required discount.
2. Hidden information: all as in the Control, but the premium package has three dots on the data protection feature, the information is visible by clicking on it.
3. TWE: all as in the Control, but the Premium comes with a message in a blue label stating "Don't waste time looking for what to watch. Choose this offer and we will give you personalized suggestions to new content you'll love. We have prepared this personalized offer just for you". (Notice that the statement is true given the calibration).



4. TWE with personalization: all as in TWE but the Premium has a picture within the description. The picture is personalized according to the answer to the gender and extroversion questions in the second module (Matz et al., 2017). If the participant is classified as extrovert, we show a picture with a group of friends having fun watching TV. If classified as introvert and self recognizes himself as male, we use a picture of a male relaxing in front of the TV. Similarly for the female. All pictures were purchased from Getty®.
5. Hidden information and Cool down: all as in condition two above, but after making their choice, a box appears stating “This is a summary of the offer you selected”, showing the picture corresponding to the selected package, and asking whether the participant wants to confirm or change selection.
6. TWE and Cool down: all as in condition three above, but with the same cool down explained in the previous condition.
7. TWE with personalization and cool down: all as in condition four above, but with the same cool down explained in condition five.
8. TWE with personalization and cool down with transparency: all as in the previous condition, but the cool down box includes “Your data has been collected and used to prepare this personalized offer for you”.

Factor two has two conditions:

1. Time Pressure: The main task appears with a clock on the right-hand side, counting down from 30 seconds to zero and then becoming red. The 30 seconds correspond to a monetary amount which is reduced for every second spent performing the task.
2. Motivated Delay: The main task appears with a text box on the right-hand side. The participant receives a monetary amount by providing a meaningful explanation of the choice.

Figure 1 presents the experimental stimuli for each dark pattern and the cool down.

In the main task, we further randomize three elements. The first is the left/right position in the task. The second is how the price is shown: either the left package has the price for one year and the right for a quarter or the other way around. Third, since the two packages have features besides price and data protection that we cannot vary, we randomize the left/right order of two descriptions of the same features. These additional features concern the presence of ads, the access to content, the access to pre-release, the fees for additional services. This is done to ensure ecological validity, as the packages look like real ones, but also create some information overload that is externally valid to the choice architecture where dark patterns are implemented in real platforms. We cannot vary these other features as the stated WTA for data protection is *ceteris paribus*.

The post experimental questionnaire includes socio demographics, three questions on preferences towards risk, patience, and trust (Falk et al., 2018), and a short questionnaire with six statements and a Likert scale of agreement (Kaptein et al., 2012), measuring the sensitivity towards the six principles of persuasion (Cialdini, 2009). The six principles of persuasions are commitment, reciprocity, scarcity, likeability, authority, and social pressure.

## 2.2 Procedure

We sent an invitation to an online panel in Europe that matches the population in each country according to gender, age, and region of residence. We define equal quotas by country, gender, and age group (18-24, 25-54, 66-65). The target sample was computed at 7200 observations. Countries included are Spain, Italy, Germany, Poland, Bulgaria, and Sweden. A native speaker translates the protocol in each language, further revised by separate experts.

In conducting the power analysis, we estimate an average inconsistency in the baseline of 20% and an overall 40% standard deviation. Considering significance at 10% (one side) and 80% power, we wanted to identify a minimum sizeable effect of around 15% of a standard deviation. This gives at least 400 observations per cell. Given available budget, we raised it to 450.

Each participant received a fixed fee for completing the questionnaire plus the incentives from the main task and from the WTA task. Average variable incentives are in line with the mean payment on other online platforms. The time of completion was around ten minutes. The final sample includes 7430 observations, and descriptive statistics are reported in Table 1 below.

All the incentives are computed in points that are exchangeable in the local currency, to ensure replicability. One point is equal to half a euro. Participants receive one point if they correctly recollect the stated WTA in one randomly chosen question and 10 points if they choose consistently (1 point if they fail to). In the motivated delay, they receive 1 point if they write at least 40 characters to explain their choice; in the incentivized time pressure, they are endowed with 30 seconds (corresponding to 1.5 points) to make their choice. If they spend more than 30 seconds, they are not prevented from making their choice but do not receive additional incentive.

The full experimental protocol, with images of all the treatment is available in the Supplementary Online Appendix. The companion report of the study includes additional information and analysis (European Commission et al., 2022). Ethical Committee at UOC granted the IRB approval on the 26<sup>th</sup> of October 2021. The hypotheses and the analysis plan were pre-registered on Aspredicted #84620 on the 9th of January 2022 after the pilot.

## 2.3 Analysis plan

Our outcome variable  $Y_i$  is a dummy equal to one if the participant  $i$  chose the targeted package. This choice violates consistency. The outcome per participant can be written in the following form:

$$Y_i = \alpha + \sum_j \sum_k d_i^{j,k} \beta^{j,k} + \varepsilon_i$$

Where  $i$  is the participant,  $j=1,\dots, \delta$  is the condition with respect to factor one,  $k=1,2$  is the condition with respect to factor two,  $d_i^{j,k}$  is the treatment dummy (equal to one if the participant  $i$  is in the condition  $j,k$  and zero otherwise) and  $\varepsilon_i$  is any unobservable variable affecting the outcome. Control and Motivated Delay is the omitted category.  $\beta^{j,k}$  tells us how many pp more likely to be inconsistent a participant is, when in experimental condition  $j,k$ , with respect to the omitted category. All tests are one sided, given that we have a prediction on the direction of the effect.

The main hypotheses to be tested are four: (H1) Dark patterns induce more inconsistency in average consumers; (H2) Dark patterns induce more inconsistency in vulnerable consumers; (H3) protective measures reduce inconsistency with respect to dark pattern in average consumers; (H4) protective measures reduce inconsistency with respect to dark pattern in vulnerable consumers. These general hypotheses can be further broken down per dark pattern and per protective measure. We expect hidden information, TWE, and TWE & personalization to increase inconsistency. We expect cool down and cool down with transparency information condition to reduce inconsistency with respect to the dark pattern condition.

The equation above can be estimated by Ordinary Least Squares. Although the outcome is a dummy, the focus of analysis is causal inference from the comparison between groups, making linear model both suitable and more transparent (Angrist & Pischke, 2009; Gomila, 2021).



In the estimation, we include three additional dummies, which refer to technical elements of the design that we randomized to reduce experimenter demand: whether basic was presented on the left or right (v1), (v2) whether basic (premium) was annual (quarterly) or quarterly (annual), and (v3) whether we use the first (second) or the second (first) description for the additional features for the basic (premium) package. These additional dummies are randomly assigned, and their inclusion increases the precision of the estimates.

Figure 1 The main task: the dark patterns

### Panel A: Hidden Information

Question 8

Please select your preferred option



 <b>12 month Subscription Plan</b>	 <b>3 months Subscription Plan</b>
<p>Package 1</p> <p>Total price per year: 60.60</p> <p>Movies and TV shows, without any limits</p> <p>Access to pre-releases at no cost</p> <p>No fee for additional features and services</p> <p>No third party ads</p> <p>Your data will not be shared with any third parties</p> <p><input type="radio"/></p>	<p>Package 2</p> <p>Total price per three months: 10.65</p> <p>Movies and TV shows, without any limits</p> <p>Free access to pre-releases</p> <p>You will receive free additional bonus features and services</p> <p>Your page will not show any advertisement from third party</p> <p>...</p> <p><input type="radio"/></p>

Next

### Panel B: Toying with Emotions

Question 8

Please select your preferred option

 <b>12 month Subscription Plan</b>	 <b>3 months Subscription Plan</b>
<p>Don't waste your time looking for what to watch. Choose this offer and we will give you personalized suggestions to new content you'll love</p>	<p>Package 1</p> <p>Total price per three months: 15.15</p> <p>Movies and TV shows, without any limits</p> <p>Free access to pre-releases</p> <p>You will receive free additional bonus features and services</p> <p>Your page will not show any advertisement from third party</p> <p>Your data will not be shared with any third parties</p> <p><input type="radio"/></p>
<p>Package 2</p> <p>Total price per year: 42.60</p> <p>Movies and TV shows, without any limits</p> <p>Access to pre-releases at no cost</p> <p>No fee for additional features and services</p> <p>No third party ads</p> <p>Your data will be shared with third parties</p> <p><input type="radio"/></p>	

Next

### Panel C: Toying with Emotions and Personalization

### Panel D: Control

Question 8



Please select your preferred option

 <p><b>3 months Subscription Plan</b></p>	 <p><b>12 month Subscription Plan</b></p>
	<p>Package 1 Total price per year: 60.60 Movies and TV shows, without any limits Free access to pre-releases You will receive free additional bonus features and services Your page will not show any advertisement from third party Your data will not be shared with any third parties</p> <input type="radio"/>
<p>Don't waste your time looking for what to watch. Choose this offer and we will give you personalized suggestions to new content you'll love We have prepared this personalised offer just for you.</p>	
<p>Package 2 Total price per three months: 10.65 Movies and TV shows, without any limits Access to pre-releases at no cost No fee for additional features and services No third party ads Your data will be shared with third parties</p> <input type="radio"/>	

Next

Question 8

Please select your preferred option

 <p><b>3 months Subscription Plan</b></p>	 <p><b>12 month Subscription Plan</b></p>
<p>Package 2 Total price per three months: 10.65 Movies and TV shows, without any limits Access to pre-releases at no cost No fee for additional features and services No third party ads Your data will be shared with third parties</p> <input type="radio"/>	<p>Package 1 Total price per year: 60.60 Movies and TV shows, without any limits Free access to pre-releases You will receive free additional bonus features and services Your page will not show any advertisement from third party Your data will not be shared with any third parties</p> <input type="radio"/>

Next

### 3. Results

#### 3.1 Main Results

7430 participants took part to the online experiment. The descriptive statistics are reported in Table 1.

The average consumer who is not exposed to dark patterns (control, motivated delay) shows a 37.80% likelihood of inconsistent preferences. Hidden information increases this likelihood by 12.25 pp ( $t=3.90$ ,  $p<0.001$ , all tests are one-sided), TWE increases the likelihood by 6.02 pp ( $t=1.90$ ,  $p=0.028$ ), and TWE & personalisation increases the likelihood by 9.84 pp ( $t=3.08$ ,  $p=0.001$ ). Personalisation does not reveal an added value with respect to TWE alone (3.81 pp,  $t=1.17$ ,  $p=0.123$ ).

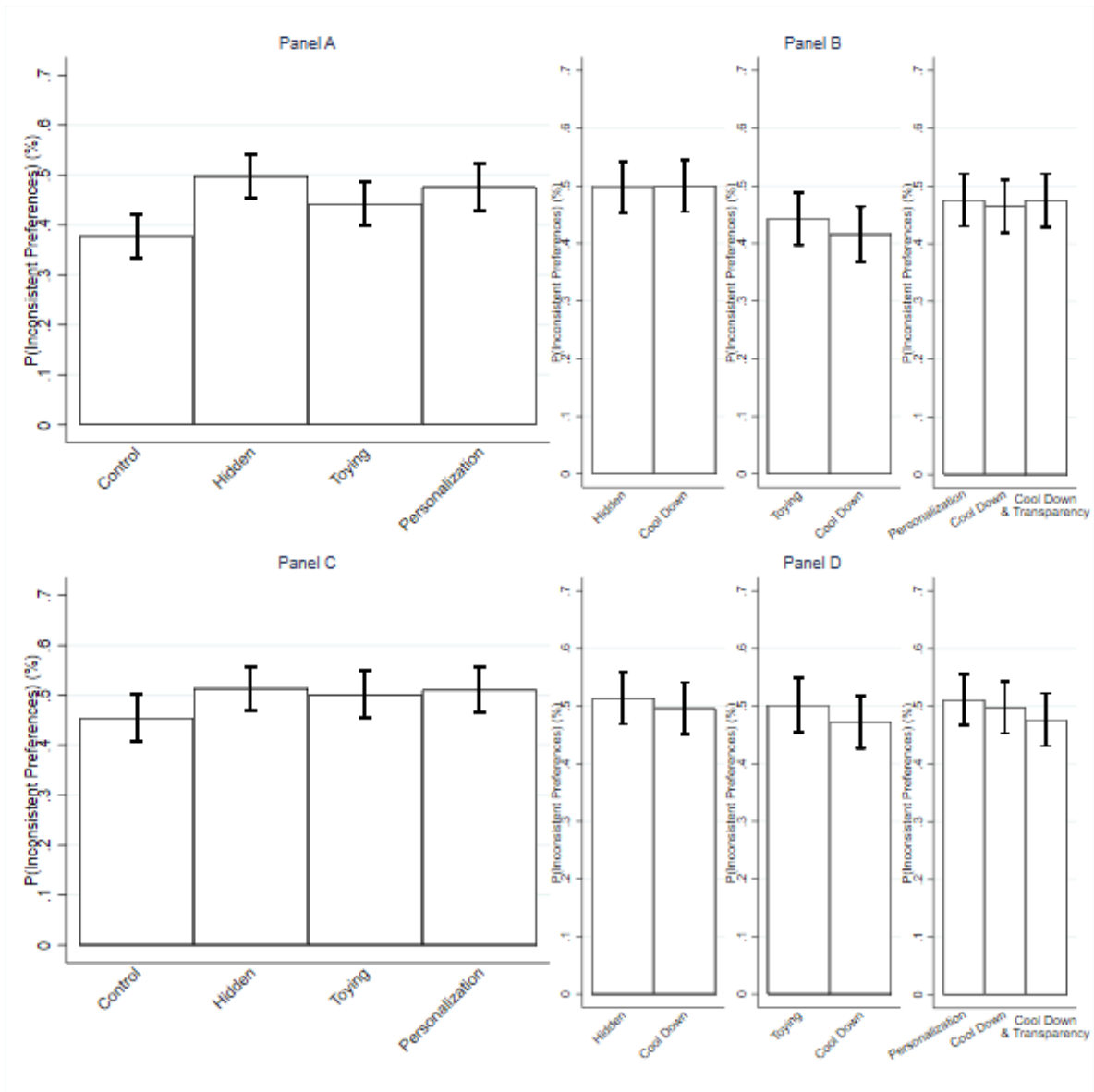
Adding protective measures reduces the outcome but the effect is negligible and not significant at conventional levels. Cool down with transaction information decreases the likelihood of inconsistent preferences with respect to hidden information by 0.29 pp ( $t=0.09$ ,  $p=0.463$ ), with respect to TWE by 2.53 pp ( $t=0.76$ ,  $p=0.224$ ), and with respect to TWE & personalisation by 1.05 pp ( $t=0.01$ ,  $p=0.375$ ). Cool down with transaction and targeting information decreases the likelihood of inconsistent preferences with respect to TWE & personalisation by 0.11 pp ( $t=0.03$ ,  $p=0.486$ ). Overall, we cannot reject the null hypothesis for the cool down protective measure.

The outcome variable is significantly higher under incentive compatible time pressure than motivated delay: vulnerable consumers have a greater likelihood to be inconsistent than average consumers (45% versus 37%). Hidden information increases this likelihood by 5.80 pp ( $t=1.80$ ,  $p=0.036$ ), TWE increases the likelihood by 4.16 pp ( $t=1.26$ ,  $p=0.103$ ), and TWE & personalisation increases the likelihood by 5.50 pp ( $t=1.71$ ,  $p=0.043$ ). We can test whether TWE with personalization outperforms TWE: this is not the case ( $t=0.41$ ,  $p=0.340$ ).

Protective measures show a negligible size effect and are not statistically significant. The cool down decreases the likelihood of inconsistent preferences with respect to hidden information by 1.67 pp ( $t=0.52$ ,  $p=0.299$ ), with respect to TWE by 2.24 pp ( $t=0.69$ ,  $p=0.246$ ), and with respect to TWE & personalisation by 0.68 pp ( $t=0.21$ ,  $p=0.415$ ). Transparency and cool down decreases the likelihood of inconsistent preferences with respect to TWE & personalisation by 3.52 pp ( $t=1.09$ ,  $p=0.137$ ). Therefore, this study did not detect a statistically significant effect of introducing a cool down period or a transparency message.

These results are plotted in Figure 2: Panel A shows the average likelihood of inconsistent choices (and 95% confidence interval) in control and dark patterns conditions and Panel B the effect of protective measures with respect to dark patterns, in both cases for the average consumer. Panel C and D shows the same measures for the vulnerable consumer. The OLS regression supporting the graphs and the results is reported in SOM, Section III: Table 1.

Figure 2 The impact of Dark Patterns and Protective Measures on the Average and Vulnerable Consumers



Note: Panel A and B only include participants in the motivated delay condition. Panel C and D only include participants in the incentivized time pressure condition.

Table 1 Descriptive statistics

<b>Total observations</b>	7430
<b>Gender</b>	
Male	48.63%
Female	51.02%
Other	0.35%
<b>Age</b>	43 (sd 15)
<b>Education</b>	
Primary education	6.72%
Secondary education	40.05%
At least some tertiary education	11.33%
Completed tertiary education	29.30%
<b>Marital status</b>	
Single	37.04%
Married/civil union	53.38%
Divorced/widowed	9.58%
<b>Household yearly income</b>	
9.999 Euro or below	18.28%
10.000 Euro – 29.999 Euro	41.71%
30.000 Euro – 49.999 Euro	23.18%
50.000 Euro – 149.999 Euro	15.90%
150.000 Euro or above	0.94%
<b>Labour market status</b>	
Employed	64.41%
In search of job	9.34%
Students/retired/housekeeper	21.61%
Other labour market status	4.63%
<b>Country</b>	
Bulgaria	16.51%
Germany	17.07%
Italy	16.57%
Poland	16.68%
Spain	16.62%
Sweden	16.55%

Note: share of participants.

### 3.2 Robustness check and validity of the design

In this subsection, we report the statistical analysis of several elements of the design: balancing of covariates, assessment of the time manipulation, plausibility of the preferences elicitation, assessment of the correct recall of the stated WTA, and plausibility of the measurement of personality trait.

We perform a multinomial logit regression to assess whether the socio-demographics (gender, age, marital status, employment status, education, income level) and the attitudes towards persuasion predict assignment to treatment cell. Out of the 180 tests of hypothesis, only six are significantly different from zero and no covariate appears significant more than once, which is consistent with random assignment.



We can test whether the time manipulation worked. The seconds spent making the decision under time pressure (mean 24.05, sd 29.61) is lower than the time spent in the motivated delay (mean 144.49, sd 245.78) condition ( $t=29.54$ ,  $p<0.001$ ).

Before moving to the main task, participants stated their willingness to accept (WTA) for the following features of the platform package: 1) sharing their data with third parties; 2) interruption by commercials; 3) access to other content; 4) unsubscribe in one click; 5) access to music platform.

The responses were collected on a discrete grid, which goes from zero to five euros or more (per month). We explore whether the aggregate distribution of WTA for sharing data with third parties is like those for other features, to see whether it is consistent with the measurement of preferences, which should be heterogeneous across domains. We compute the Kolmogorov-Smirnov test statistics. As it turns out, the distribution of the WTA for data sharing is statistically different from that of interruption by commercials (K-S=0.022,  $p=0.046$ ), of access to other content (K-S=0.033,  $p<0.001$ ), of unsubscribe in one click (K-S=0.033,  $p<0.001$ ) and of access to music platform (K-S=0.035,  $p<0.001$ ). This dissimilarity is driven by dissimilarity of individual preferences across features, consistent with participants answering truthfully to the questions. We perform a t-test with matched sample: the mean WTA is not statistically different from that for interruption by commercials ( $t=1.02$ , two sided  $p=0.30$ ), is statistically different from that for access to other content ( $t=4.79$ ,  $p<0.001$ ), is statistically different from that for unsubscribe in one click ( $t=4.10$ ,  $p<0.001$ ) and is statistically different from that for access to music platform ( $t=5.59$ ,  $p<0.001$ ).

SOM, Section IV, Figure 1, Panel A shows the corresponding histogram. The modal response is 5 (24.1%), followed by 2 (14.79%). It is interesting to see whether the choice of the grid conditions the response, by looking at the right tail of the distribution. 24.01% of the respondents answer the rightmost category. This is larger than the closest options, as a result, a finer grid would have changed the shape of the distribution. This may imply that we are underestimating the effect of dark patterns.

At the end of the first instrumental task, respondents were incentivised to correctly recall the stated WTA. 100% of participants recall the value for the WTA, defining correct a one-euro difference from the original stated value (the price discount in the choice task was 2.5 euros). This suggests the absence of a potential issue of internal validity of the results.

Finally, we analyze the answer to the questions on personality traits, elicited using a standardized instrument to measure the Big Five: openness, conscientiousness, extraversion, agreeableness, and neuroticism (Woods & Hampson, 2005). Since the personalization requires separately identifying those that score low and high on this dimension, we can check whether the distribution has most of participants in the middle and a fuzzy separation between extroversion and introversion or rather has significant tails indicating the presence of two types. Had that the distribution be approximately normal, (1) most of the response would be around the mean and (2) a small level of measurement error will shift the type of a significant portion of people, classifying a lot of slightly introvert as extravert and vice versa. Under this scenario, the personalization would not change the behavior, but because of mistargeting (sending the wrong message) and not because of ineffectiveness (the message does not work).

SOM, Section IV, Figure 1, Panel B reports the empirical distribution for the response to the extraversion item, together with a normal distribution. It is shown that the distribution has much more mass on the tails, contrary to a normal distribution. This weights in favor of the instrument. We can infer that the limited value added of personalization may be due to the choice of image or to the absence of treatment effect but is not due to measurement error in personality traits.

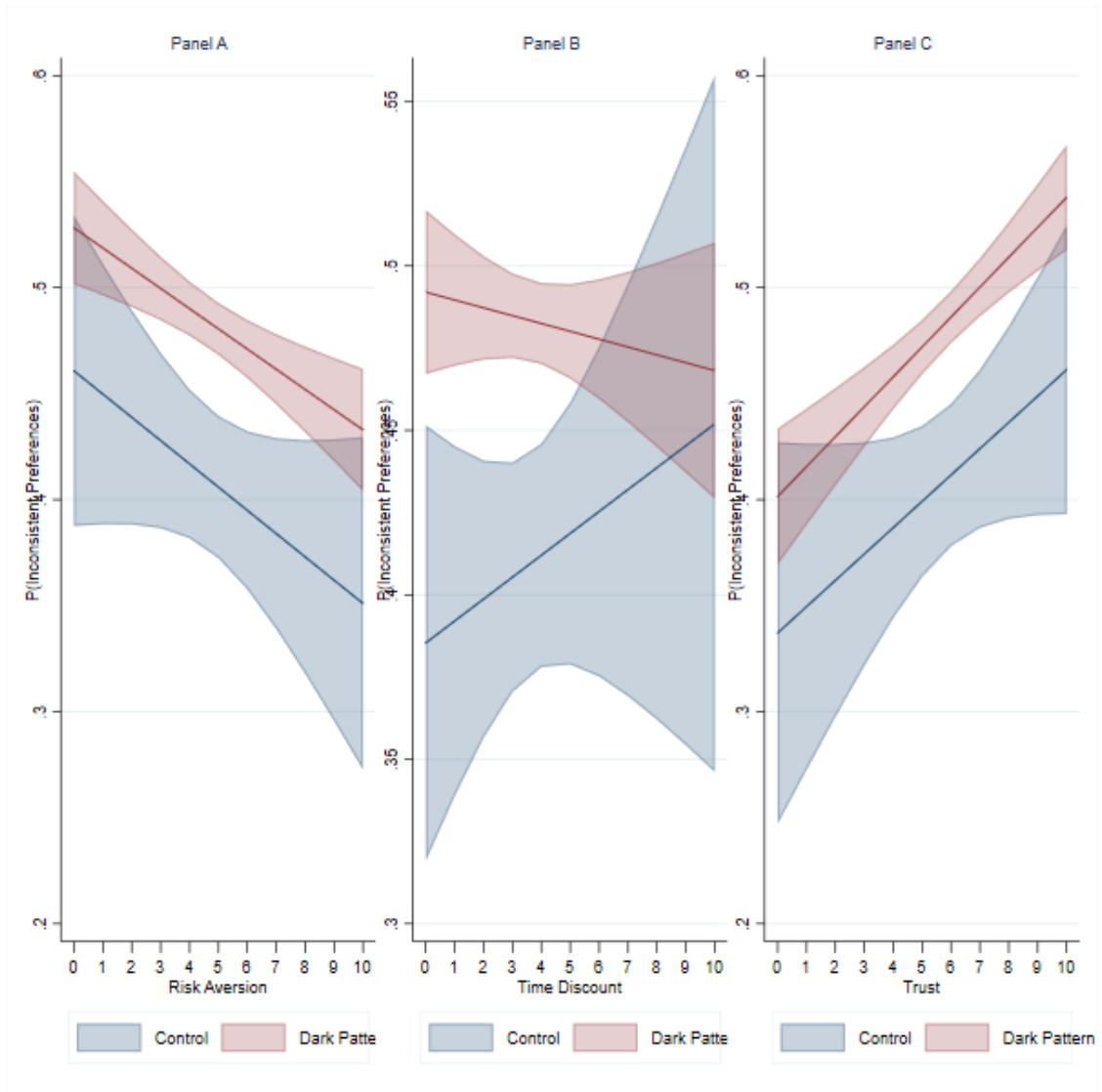
### *3.3 Heterogeneity*

Since the post experimental questionnaire collects information on risk aversion, time discount, and trust, we analyze heterogeneity along these dimensions. The questions belong to the Global Preference Survey (Falk et al., 2018). The answers are collected on a Likert scale from zero to elevens, we rescale them to make sure that they are increasing in the level of risk aversion, time discount, and trust.

The following results should be interpreted as correlations and are subject to low statistical power. In fact, preferences are not randomly assigned. Additionally, differences in subgroups are smaller than the overall main effect, and the variability is higher. The combination of the two entails a much larger sample than what is necessary to estimate a main effect. This part of the analysis does not distinguish across dark patterns. We control for protective measure but given the lack of main effects, the analysis overlooks differences across the three dark patterns. In three separate regressions, the dummy for dark patterns interacts with the mediating variable. We conduct a mediation analysis following Spiller et al. (2013) and compute Johnson-Neyman points, following the same procedure as in Alós-Ferrer & Garagnani (2020).

The three graphs of the interactions are plotted in Figure 4, Panel A for risk aversion, Panel B for time discount, and Panel C for trust. The first mediating variable explored, risk aversion, does not mediate dark patterns. The effect of dark patterns is statistically significant (at the 10% threshold) for any level of risk aversion. Alternatively, time discount (the extent to which someone favors the present vs the future) mediates the effect of dark patterns. We found two Johnson-Neyman points, at time discount zero and six (recall that the variable goes from zero to ten). The effect of dark patterns on choice inconsistency is stronger for lower values of present bias. Trust mildly mediates dark patterns, there is a Johnson-Neyman points at level of trust of two. This suggests that the effect of dark patterns on choice inconsistency vanishes only for the lowest levels of trust.

Figure 3 Heterogeneity analysis and Johnson-Neyman point



## 4. Experiment with lotteries and elicited preferences

### 4.1 Objectives

We perform an additional experiment to evaluate the reliability of our findings regarding the effects of dark patterns.

In this companion experiment, subjects choose among lotteries. Lotteries provide several advantages. In fact, preference elicitation among lotteries can be fully incentivized. We do it via a Multiple Price List task, where we elicit the point of indifference between a mixed lottery and one that trades-off outcomes and probabilities.

Moreover, preferences can be visually represented: we chose a tree representation. This representation opens the possibility to arbitrarily switch between simple and composite lotteries, introducing elements of complexity that are deemed crucial for the surreptitious introduction of deceptive features. Simple and composite lotteries are shown in Figure 4 below (Panels A and B respectively) for the case of a generic outcome  $x$  to be determined via a an MPL.

Furthermore, by construction, MPL comprises many dichotomic choices that differ in their distance from the point of indifference. Subsequently, by randomly selecting rows from the MPL, we can evaluate whether dark patterns remain equally effective.

Finally, we opt to assign treatments within subjects to stress-test whether learning mitigates the impact of dark patterns.

On the negative side, lotteries present certain drawbacks due to their abstract nature. Their lack of framing disrupts established social norms and expectations at the point of sale. In conventional transactions, such as purchasing from a physical store or an online platform, there exists a shared understanding among customers regarding certain environmental cues. For instance, if a supermarket designates a specific aisle with the word “Pasta”, a customer will reasonably assume that all pasta varieties are located there, not solely the priciest brands. Similarly, the presence of a nudge button would typically be interpreted as opting for the “conventional” choice rather than an inferior one. However, these ingrained social expectations may not extend to abstract objects like lotteries. Consequently, the absence of such shared beliefs regarding lotteries can lead to reduced trustworthiness among individuals, thereby undermining the efficacy of dark patterns.

The other shortcoming is that not all treatments can be reproduced in this setting. We cannot implement hidden information in a meaningful way within this setting. Due to the visual representation of lotteries, the act of omitting information is rendered futile as it is immediately noticeable or becomes counterproductive to the extent that it heightens the prominence of the alternative.

The main treatment is a blue colored button to nudge towards an option, with the sentence “I want to choose this” on top (IWE). We incorporate an additional simple *highlight* treatment, limited to the blue colored button only, a widely used web design element. The comparison between the two provides a test of the added value of the sentence. Figure 5 presents the experimental stimuli for each dark pattern. In Panel D-E of Figure 4, we represent respectively

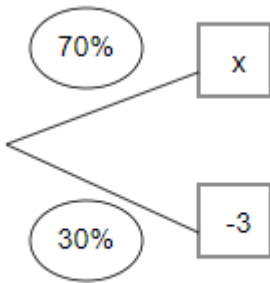
the highlight, TWE and control conditions. Each dichotomous choice consists of two lotteries, represented in tree form.<sup>3</sup>

---

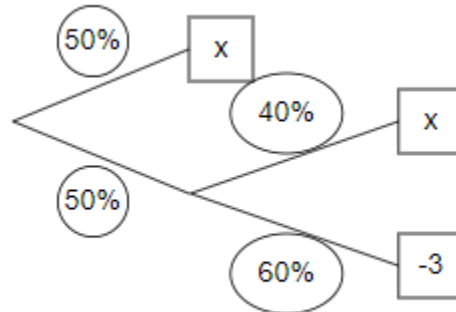
<sup>3</sup> In our original plan, we intended to implement a personalization treatment by utilizing the responses from the risk elicitation task to craft a message along the lines of “since in Task I you systematically preferred the risky/riskless option, why don’t you choose this one?”. However, we ultimately decided to omit this approach due to concerns that in an abstract task with no framing of a “seller”, it is unclear how the subject would interpret the suggestion and whether it represents an experimenter demand effect.

Figure 4 Tree representation of lotteries

A. Simple representation

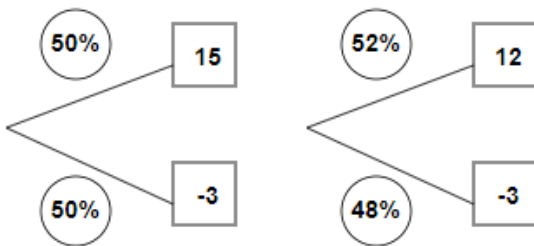


B. Composite representation



C. Treatment Highlight

Veillez choisir entre l'option A et l'option B.

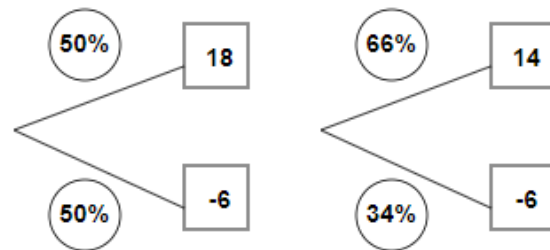


Option A

Option B

D. Treatment TWE

Veillez choisir entre l'option A et l'option B.

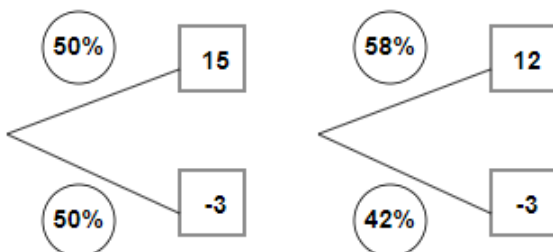


Option A

Option B  
Je veux choisir cette option

E. Control

Veillez choisir entre l'option A et l'option B.



Option A

Option B

#### 4.2 Experimental Design, Procedures and Analysis Plan

This is a within subject design with two treatments and a control. We manipulate the choice architecture, with the following conditions: control, highlight, TWE. Dark patterns are always targeting the choice opposite to the one taken in the preference elicitation task.<sup>4</sup>

Participants go through the following experimental sequence. The introductory phase included instructions on the tree-representation of lotteries (two comprehension questions with feedback are posed at the end) and on the MPL with single switching point. In the first section, participants report their risk aversion, loss aversion and aversion to compounded lotteries, using MPL. This part is common to all treatments.

Participants answer four MPLs. They choose their switching point for two different mixed lotteries (50% 15, 50% -3 and 50% 18, 50% -6). For each mixed lottery, participants report the  $x$  such that (70%  $x$ , 50% loss) is equivalent to the original lottery and the  $z$  such that ((50+ $z$ )% gain, (50- $z$ )% loss) is equivalent to the original lottery. In the formula, *gain* is either 12 or 14 and *loss* either -3 or -6 according to the original lottery. Then they face 16 dichotomous choices that are randomly selected rows from the MPLs. The 16 dichotomous choices satisfy two constrains: each lottery is shown at least four times and each lottery is shown at least once with each treatment.

In the main task, we further randomize the format of representation of lotteries between compounded or simple. If the respondent sees the MPL task as simple, it sees the dichotomous choice as compounded, and vice versa. Orders of MPLs and of lotteries is randomized. The order of sections cannot be randomized because the programming of the DP is sequential (they require the responses from Section II).

The post experimental questionnaire includes socio demographics, a Cognitive Reflection Task (Frederick, 2005), the set of GPS questions and attitude to persuasion from Experiment I, and one question on the frequency of eCommerce during the previous twelve months.

The procedure is standard. We sent an invitation to a random sample from PSE-Paris I lab subject pool. A native speaker translates the protocol in French.

Each participant received a fixed fee for completing the experiment plus the incentives from Section I task (risk aversion, loss aversion, and aversion to compound lottery) and from the main task (either the consistency or one random choice among a set that included both the MPL rows and the dichotomous choices). Average variable incentives are in line with the mean payment in the lab for online experiments: participants receive on average 11.21 euros, with 4.21 standard deviation for a 22' experiment (with 20' standard deviation). We preregistered at least 100 subjects.

All the incentives are computed in euro exchangeable points as in the previous experiment (the exchange rate is one EUR for two points). The show up fee is three euros.

The experiment has been programmed in Lab.js (Henninger et al., 2022). The full experimental protocol in English is available in the SOM Section III. IRB approval has been

---

<sup>4</sup> For a companion methodological paper, we also run a treatment with stated preferences and pay for consistency in lottery, but we will not discuss the results here.

granted by PSE on the 25<sup>th</sup> of April 2023. The hypotheses and the analysis plan were pre-registered on OSF (<https://osf.io/m8hxt>) on the 24<sup>th</sup> of May 2023.

We test the following main hypothesis: there is no effect of a dark pattern on consistency ( $\Delta\mu^{DP} = \mu^{DP} - \mu^C = 0$ ). The alternative hypothesis is that  $\Delta\mu^{DP} > 0$ . We can test this with a one-sided paired t-test. This hypothesis is separately tested for the two dark patterns.

### 4.3 Results

#### 4.3.1 Main result

We collected 120 observations. Half participants are female, around 60% are students, the average age is 25 (7 sd) and the two thirds of participants have an annual income below 30k euros. The rest of the descriptive statistics is presented in Table 2 below.

The main outcome is an indicator variable equal to one if the choice is inconsistent. The average outcome in the control is 36.82%. TWE increases the likelihood by 5.73 pp (t=2.12, p=0.018 all tests are one sided unless otherwise specified). Highlight alone does not significantly increase the likelihood (1.89 pp, t=0.71, p=0.23). TWE presents an added value with respect to highlight alone (3.84 pp, t=1.42, p=0.079). Overall, data support our main hypothesis. The supporting OLS regression is reported in Column (1) in Table 3 below. These results are plotted in Figure 5, which shows the average likelihood of inconsistent choices (and 95% confidence interval) in control and dark patterns conditions.

The standard OLS regression includes whether the main task was performed under composite or simple representation and a dummy for left-right order of presentation (the latter varies within subject).

Additionally, we performed a standard battery of robustness check to assess the main results. The effect of TWE on inconsistency is maintained when we include the round of decision (column (2) in Table 3 below). Fixed effect estimators are reported in columns (3) and (4), but results do not vary. Column (5) restricts the estimation to those that correctly answered the questions before receiving the feedback, but the results stay the same. Column (6) exclude those that employs more than 42' to complete the experiment (i.e. more than one standard deviation above the mean). The results are robust.

One feature of our design suggests that the impact of dark patterns may be underestimated. Surprisingly, approximately one fourth of the participants in the MPLs opted for extreme switches on the MPL grid when it came to the Outcome lotteries. This occurrence facilitates participants in accurately recalling their choices, thereby making it more challenging for dark patterns to display their effects. In Section 4 of the SOM, Figure 9 portrays the histograms depicting the switching points for the four lotteries. It is notable that for the two Outcome Lotteries, the distribution is more likely to be censored than for the two Probability Lotteries. This may reflect lower levels of risk aversion than expected. To further analyze this, we conducted OLS and FE regressions on the subsamples of Outcome and Probability Lotteries in Table 4 below. We report the statistical tests from the FE columns.



The average inconsistency for outcome lottery, in the control, reaches 37.68%. In the TWE, it reaches 40%. The difference is not statistically significant ( $t=0.75$ ,  $p=0.227$  one sided). In the highlight, the outcome is on average 38.09% ( $t=0.13$ ,  $p=0.448$ ). On the contrary, for probability lotteries, the baseline is 35.93%, whereas in the TWE is more than 9 pp larger (9.06,  $t=2.29$ ,  $p=0.011$ ). The highlight shows an average inconsistency of 39.26% ( $t=0.90$ ,  $p=0.186$ ).

#### *4.3.2 Learning, discriminability, and the effect of dark patterns*

Taking advantage of our design, we construct a variable that quantifies the distance, in terms of grid points, between the dichotomic choice in Task III and the switching point for the corresponding lottery in the MPL task. As we move along the grid towards the switching point, the discriminability between the left and right options gradually diminishes. This is the crucial juncture where the deceptive role of dark pattern is confounded with the role of a “decision device”.

To provide a visualization, we arbitrarily split the level of discriminability using the threshold of two grid points. The outcome by conditions is plotted in Figure 7, in the two top panels.

When the choices are close to the switching point, inconsistency increases, as one would expect, reaching 45.69% in the control. TWE raises the level of inconsistency by 6.04 pp ( $t=1.25$ ,  $p=0.106$  one sided). The simple highlight generates the same level of inconsistency (44.94%) as the control. The supporting regression is in the SOM, Section IV, Table 2, column (1).

When we randomly draw a dichotomic choice far from the switching point, the average level of inconsistency drops to 33.26% in the control. The effect of TWE remains flat at 5.46 pp ( $t=1.67$ ,  $p=0.048$ ). In this case, under the highlight condition, inconsistency reaches 35.89% (the difference is not statistically significant, column (2) in the Table). In column (3), we also estimate a model with the interaction with a continuous measure of the distance. Notice that choosing one point of the grid farther from the indifference decreases inconsistency by 2 pp in the control. As a result, even a mild TWE dark pattern has a 5.29 pp effect ( $t=1.23$ ,  $p=0.110$ ) that stays constant across the grid, representing a remarkable finding.

Another intriguing aspect to consider is the measurement of learning. In the main experiment, the treatment was administered between subjects, and participants were only required to perform the task once. However, in this setting, participants were tasked with making 16 choices, allowing us to explore the potential effects of repeated exposure on learning.

In Figure 7, the bottom two panels depict the level of inconsistency by treatment within the first and the second halves of the task. Upon closer examination of the control group, the average inconsistency drops remarkably from 41.04% to 32.61%. This represents a striking learning process, as consistency improves by around 0.84 pp per round, as reported by a simple OLS regression of inconsistency over the round (Column (1) of SOM, Section IV, Table 3).

In contrast, learning progresses at a comparable slower rate in the dark pattern condition. Specifically, the average inconsistency for TWE stands at 43.76% during the initial eight rounds and decreases to 41.19% in the final eight rounds. The pattern observed in the highlight condition is even more striking as the average inconsistency increases from 35.82% to 41.31%. The impact of TWE in the first part is not statistically significant, displaying a 2.75 pp difference

( $t=0.64$ ,  $p=0.261$ ) but emerges in the second part, reaching 8.63 pp ( $t=2.32$ ,  $p=0.010$ ). It is worth mentioning that the highlight condition also exhibits a similar effect in the second part, with an 8.59 pp difference ( $t=2.16$ ,  $p=0.016$ ). The regression analyses for both sets of choices are presented in Columns (2) and (3) of the SOM, Section IV, Table 3, while Column (4) incorporates a model with interaction.

The impact of Dark Patterns on learning becomes even more evident when we focus on the sample of Probability Lotteries, considering the earlier finding regarding the switching point. These results are visually represented in Figure 8 below. In the first half of the experiment, the average inconsistency in the baseline stands at 41.46%, whereas it decreases to 30% in the second half. The reduction in inconsistency is significantly lower in the TWE condition, where the inconsistency is initially at 47.27% and subsequently reduces to 42.58%. (In the highlight condition, the inconsistency remains relatively stable at around 39%.) Comparing the control and TWE conditions, an already significant 6 pp difference in the first half expands to a 13 pp difference in the second half.

Finally, we look at decision time, recorded automatically in Lab-js (Figure 7). In the control condition, the average choice takes 10.57 seconds. Decision time does not change in the highlight condition (10.51 second), whereas it increases by almost two seconds in the TWE (12.50). We report OLS and Fixed effect regressions in SOM, Section IV, Table 4. The results are not statistically significant, but the coefficients are stable across specifications. The analysis of decision time was pre-registered because we wanted to assess to what extent dark patterns operates as decision devices. The evidence seems to contradict that hypothesis although the current analysis is very preliminary. The two seconds may reflect the reading time of the additional text on the buttons or the abstract environment. Whereas, in online platforms, certain social expectations drive behavior of the buyer, these may be different than in the lab.

#### *4.3.3 Heterogeneity Analysis*

Within the sample, 17.50% are risk averse, 36.66% are risk neutral (their switching point is either 6/6 or 7/6 of the expected value of the lottery). Around 60% are loss averse and 47% are compound lottery averse. 35.83% of the participants do not incur into any mistake in the CRT.

The likelihood to be inconsistent is poorly predicted by individual traits. Among GPS variables (risk, trust and time discount), (elicited) risk aversion, loss aversion, attitude to compound lottery, CRT, gender (female), the dummy for student, and the attitude to persuasion (the average of the six questions) the only ones that predict the average inconsistency are CRT ( $\rho=-0.23$ ,  $p=0.011$ ) and the attitude to compound lotteries ( $\rho=0.16$ ,  $p=0.0732$ ).

In the SOM, Section IV, we conduct a mediation analysis mimicking the one in Section 3.3. We report the estimated outcome for TWE and Control, at different levels of the mediating variable, with the confidence intervals. The regression controls for the usual variables. Figures 2-8 plot the output for self-reported risk aversion, self-reported time discount, self-reported trust, CRT, elicited risk aversion, attitude to compound lottery, and loss aversion. None mediates the effect.

Table 2 Descriptive statistics

<b>Total observations</b>	120
<b>Gender</b>	
Male	52.50%
Female	46.67%
Other	0.83%
<b>Age</b>	25 (sd 7)
<b>Student</b>	58.33%
<b>Marital status</b>	
Single	37.04%
Married/civil union	53.38%
Divorced/widowed	9.58%
<b>Household yearly income</b>	
9.999 Euro or below	35.83%
10.000 Euro – 29.999 Euro	34.17%
30.000 Euro – 49.999 Euro	20.00%
50.000 Euro – 149.999 Euro	10.00%
150.000 Euro or above	0%
<b>Frequency of eCommerce (previous year)</b>	
Once or twice	17.50%
3-5 times	32.50%
6-10 times	20.00%
More than 10 times	30.00%

Note: share of participants.

Figure 5 The impact of Dark Patterns on choice among lotteries

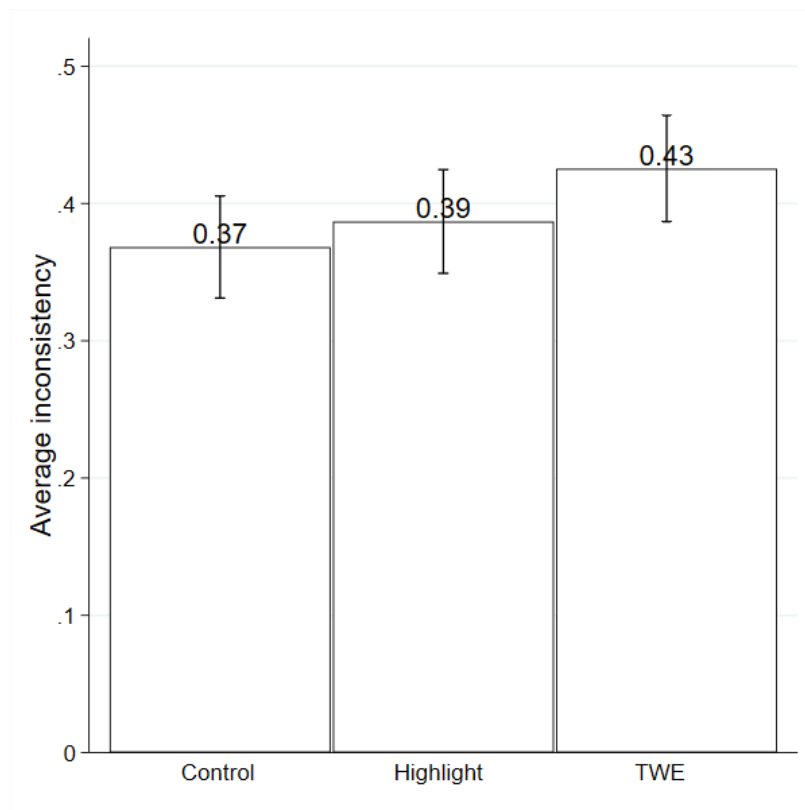


Figure 6. Learning, inconsistency and the effect of Dark Patterns.

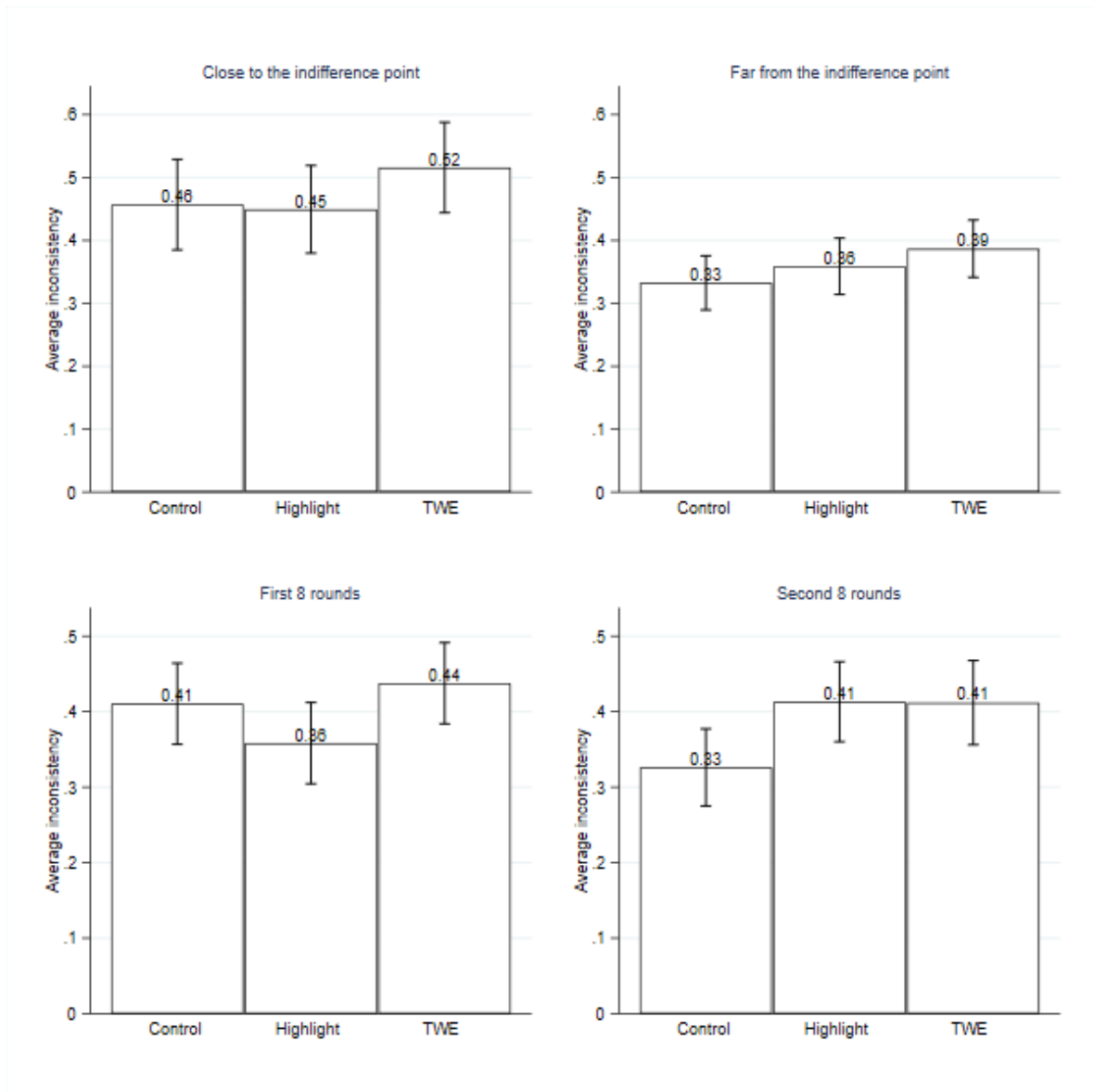


Figure 7 Decision times across conditions

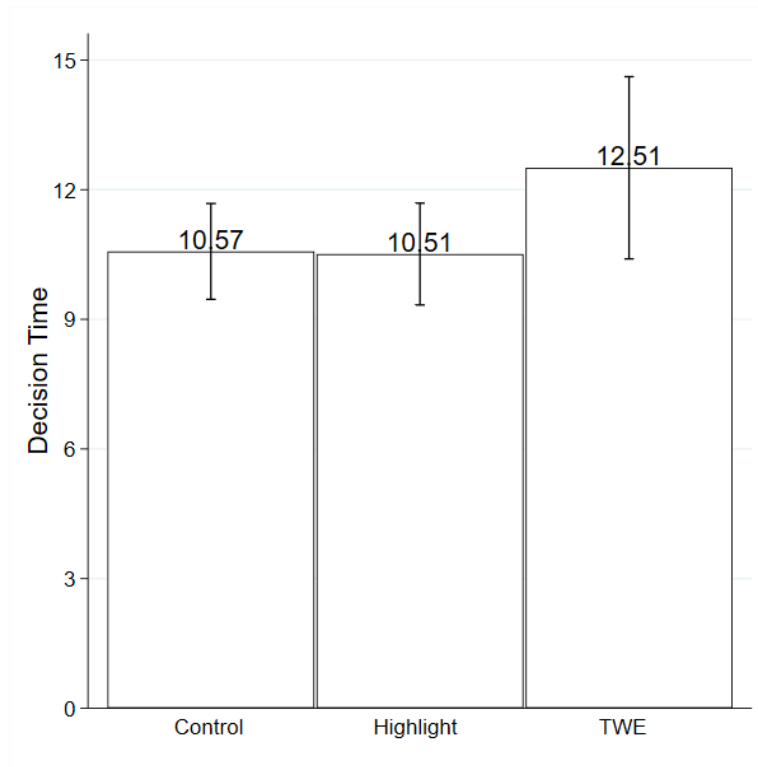


Figure 8 Learning, inconsistency and the effect of Dark Patterns. Results from probability lotteries

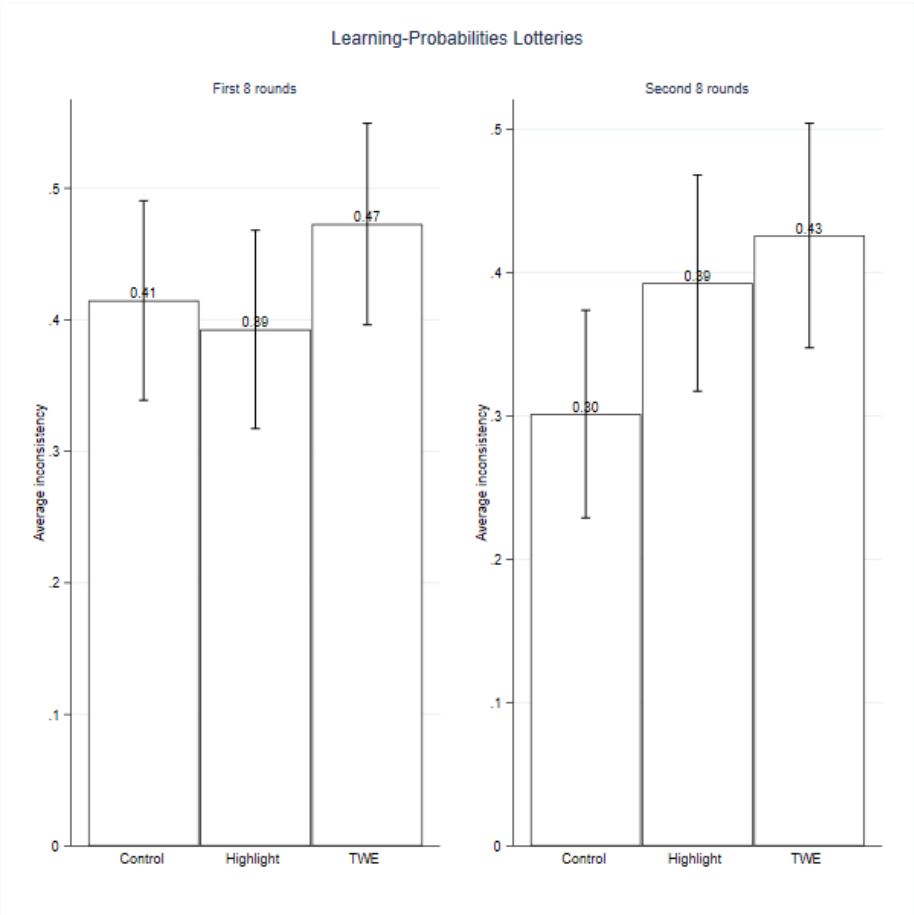


Table 3 The effect of DarkPattern on inconsistent choices

	(1)	(2)	(3)	(4)	(5)	(6)
	Inconsistent Choice (%)	Inconsistent Choice (%)	Inconsistent Choice (%)	Inconsistent Choice (%)	Inconsistent Choice (%)	Inconsistent Choice (%)
Highlight	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)	0.01 (0.03)	0.01 (0.03)
TWE	0.06** (0.03)	0.06** (0.03)	0.06** (0.03)	0.06** (0.03)	0.06* (0.03)	0.06** (0.03)
Constant	0.36*** (0.03)	0.36*** (0.04)	0.37*** (0.02)	0.37*** (0.02)	0.37*** (0.03)	0.38*** (0.02)
Observations	1,920	1,920	1,920	1,920	1,552	1,856
R-squared	0.00	0.00	0.00	0.00	0.00	0.00
Composite	Yes	Yes	Yes	Yes	No	No
Reversed	Yes	Yes	Yes	Yes	No	No
Round	No	Yes	No	Yes	Yes	Yes
Estimator	OLS	OLS	FE	FE	FE	FE
Sample	All	All	All	All	NoErrors	Within 42'
Number of iid			120	120	97	116

Note: Clustered (id) std err in parenthesis. \*p<0.1 \*\*p<0.05 \*\*\*p<0.01

Table 4 The effect of DarkPattern on inconsistent choices, separating Outcome and Probability Lotteries

	(1)	(2)	(3)	(4)
	Inconsistent Choice (%)	Inconsistent Choice (%)	Inconsistent Choice (%)	Inconsistent Choice (%)
Highlight	0.00 (0.04)	0.04 (0.04)	0.00 (0.04)	0.03 (0.04)
TWE	0.02 (0.04)	0.09** (0.04)	0.03 (0.04)	0.09** (0.04)
Constant	0.34*** (0.05)	0.38*** (0.04)	0.33*** (0.03)	0.39*** (0.03)
Observations	954	966	954	966
R-squared	0.00	0.01	0.00	0.01
Composite	Yes	Yes	Yes	Yes
Reversed	Yes	Yes	Yes	Yes
Round	No	No	Yes	Yes
Estimator	OLS	OLS	FE	FE
Sample	Outcome Lotteries	Prob Lotteries	Outcome Lotteries	Prob Lotteries
Number of iid			120	120

Note: Clustered (id) std err in parenthesis. \*p<0.1 \*\*p<0.05 \*\*\*p<0.01

## 5. Discussion and Conclusions

This article presents experimental evidence on dark pattern manipulation and the effectiveness of a transparency-based remedy. Participants reveal their trade-off between money and data protection and then are offered a service that does not meet those stated preferences. Falling for the offer implies showing inconsistent choices. With respect to a placebo condition, subjects increase their likelihood of inconsistency from one-sixth to one-fourth of a standard deviation. This holds for the average consumer, induced to be reasonably attentive via an incentive compatible motivated delay. Vulnerable consumers, put under pressure by a time pressure, are more inconsistent and are prone to fall for dark patterns, although to a lower extent. Transparency-based remedies are ineffective for countering dark patterns and manipulative personalization practices.

In evaluating the size effect of the dark patterns, one should consider two elements. Even small rents extracted through dark patterns can represent significant welfare losses in aggregate, because malignant users' interfaces are common in platform with larger scale of operations. Second, for the consumer, dark patterns look like a social dilemma, where the individual cost to raise a complaint is too high given that benefits are shared across the market.

In trying to assess dark patterns, we had to face a methodological challenge. The EU and US law affirm the following principles based on common sense. (1) A practice cannot be banned on the sole argument of being effective, because this would stifle marketing innovation; (2) A practice should be fair to the consumer, who pays enough attention and does not blindly trust the seller's words. However, they take different approaches. US law deems a practice unfair if it is effective and leads to a financial loss that outweighs any benefits. In contrast, EU law deems a practice unfair if it is effective and can be characterized as deceptive or coercive. In this article, we propose evaluating commercial practices from the perspective of consumer preferences to find a common ground. This leads us to introduce the transaction test: determining whether a practice influences or is likely to influence consumers to make transaction decisions that contradict their individual preferences. The transaction test allows for a controlled assessment of the practice and transforms the manipulation conundrum into a relative straightforward empirical problem. This is a major contribution of this article. Our second experiment shows that the transaction test is a relatively general and robust approach.

We close with two considerations, one related with policy and one with theory.

The law evaluates the fairness of a practice from the perspective of an average consumer, assuming no party has a clear asymmetric advantage. In presence of digital asymmetry, the mean consumer looks like a convex combination between the average consumer and the vulnerable consumer. In our setting, more than one third of consumer makes an inconsistent choice, even when they are paid to motivate their decisions and we induce value for consistency. This weights in favor of a new benchmark.

In this article, we distance ourselves from most of the existing literature on the taxonomy of dark patterns (Luguri & Strahilevitz, 2021; Mathur et al., 2019; UK Competition and Market Authority, 2022) which heavily rely on System I-System II (Kahneman, 2011). This explanation posits that participants tend to make impulsive decisions most of the time, even though they are



equipped with an alternative deliberative system which processes all the information and is activated only under stronger incentives to focus. Dual process model is an appealing explanation: simply stated dark patterns exploit impulsivity nudging towards a more intuitively appealing option.

Data from our experiment do not provide support for this mechanism. Under dual process, three stylized facts should hold. First, the impact of dark patterns should be larger in vulnerable consumers than in average consumers. Second, the best protective measure should indeed be a cool-down period. Third, participants classified as more likely to choose inconsistently should be more severely impacted by dark patterns. None of the three are supported in our data. Instead, the impact of dark patterns for average consumers is one and a half to two times higher than for vulnerable consumers, the cool-down intervention was ineffective, and participants classified as less present biased are more prone to dark pattern driven inconsistency. Future research should address manipulative practice in the context of alternative behavioral explanations.

## References

- Akerlof, G. A., & Shiller, R. J. (2015). *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton University Press.
- Alós-Ferrer, C., & Garagnani, M. (2020). The cognitive foundations of cooperation. *Journal of Economic Behavior and Organization*, 175, 71–85. <https://doi.org/10.1016/j.jebo.2020.04.019>
- Angner, E. (2012). *A Course in Behavioral Economics*. Palgrave Macmillan UK.
- Angrist, J., & Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- BEUC. (2022). “DARK PATTERNS ” AND THE EU CONSUMER LAW ACQUIS.
- Blake, T., Moshary, S., Sweeney, K., & Tadelis, S. (2021). Price Salience and Product Choice. *Marketing Science*, 40(4), 619–636. <https://doi.org/10.1287/mksc.2020.1261>
- Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies*, 2016(4), 237–254. <https://doi.org/10.1515/popets-2016-0038>
- Brignull, H. (2010). *Deceptive Design*. <https://www.deceptive.design/>.
- Calo, R. (2014). Digital market manipulation. *George Washington Law Review*, 82(4), 995–1051. <https://doi.org/10.2139/ssrn.2309703>
- Cara, C. (2019). Dark Patterns in the Media: a Systematic Review. *Network Intelligence Studies*, VII(14), 105–113.
- Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., & Camerer, C. (2021). *On the Relation between Willingness to Accept and Willingness to Pay*. [www.jnchapman.com](http://www.jnchapman.com)[www.columbia.edu/~md3405/](http://www.columbia.edu/~md3405/)[www.snowberg.hss.caltech.edu/~camerer/](http://www.snowberg.hss.caltech.edu/~camerer/)
- Cialdini, R. B. (2009). *Influence: Science and Practice* (5th ed.). Pearson Education.
- Della Vigna, S., & Gentzkow, M. (2010). Persuasion: Empirical Evidence. *Annual Review of Economics*, 2(1), 643–669. <https://doi.org/10.1146/annurev.economics.102308.124309>
- Dertwinkel-Kalt, M., Köster, M., & Sutter, M. (2020). To buy or not to buy? Price salience in an online shopping field experiment. *European Economic Review*, 130, 103593. <https://doi.org/10.1016/j.euroecorev.2020.103593>
- Diaz, L., Houser, D., Ifcher, J., & Zarghamee, H. (2021). Estimating Social Preferences Using Stated Satisfaction: Novel Support for Inequity Aversion. *Ssrn*, 14347. <https://doi.org/10.2139/ssrn.3846691>
- Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge University Press.
- Echenique, F., Lee, S., & Shum, M. (2013). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119(6), 1201–1223. <https://doi.org/10.1086/674077>
- Esposito, G., Hernández, P., Van Bavel, R., & Vila, J. (2017). *Nudging to prevent the purchase of*

*incompatible digital products online: An experimental study.*  
<https://doi.org/10.1371/journal.pone.0173333>

- Guidance on the interpretation and application of Directive 2005/29/EC of the European Parliament and of the Council concerning unfair business-to-consumer commercial practices in the internal market, Official Journal of the European Union. C 526 (2021).
- European Commission, D.-G. for J. and C., Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F., Liva, G., Lechardoy, L., & de las Heras Ballell, T. (2022). *Behavioural study on unfair commercial practices in the digital environment : dark patterns and manipulative personalisation : final report*. <https://doi.org/doi/10.2838/859030>
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences\*. *The Quarterly Journal of Economics*, 133(4), 1645–1692. <https://doi.org/10.1093/qje/qjy013>
- FTC Policy Statement on Unfairness, 104 FIC 949 (1984) (testimony of Federal Trade Commission.).
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150(3), 700–709.
- Grassl, P., Schraffenberger, H., Zuiderveen Borgesius, F. J., & Buijzen, M. (2021). Dark and Bright Patterns in Cookie Consent Request. *Journal of Digital Social Research*, 3(1), 1–35.
- Gu, Y., & Wenzel, T. (2015). Putting on a tight leash and levelling playing field: An experiment in strategic obfuscation and consumer protection. *International Journal of Industrial Organization*, 42, 120–128. <https://doi.org/10.1016/j.ijindorg.2015.07.008>
- Gu, Y., & Wenzel, T. (2020). Curbing obfuscation: Empower consumers or regulate firms? *International Journal of Industrial Organization*, 70, 102582. <https://doi.org/10.1016/j.ijindorg.2020.102582>
- Hartzog, W. (2018). *Privacy's Blueprint: The Battle to Control the Design of New Technologies*. Harvard University Press.
- Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). lab.js: A free, open, online study builder. *Behavior Research Methods*, 54(2), 556–573. <https://doi.org/10.3758/s13428-019-01283-5>
- Huck, S., & Wallace, B. (2015). The impact of price frames on consumer decision making A price frame refers to the way a price is presented. In *Mimeo*.
- Kahneman, D. (2011). Thinking fast, thinking slow. In *Interpretation, Tavistock, London*.
- Kalayci, K., & Potters, J. (2011). Buyer confusion and market prices. *International Journal of Industrial Organization*, 29(1), 14–22. <https://doi.org/10.1016/j.ijindorg.2010.06.004>
- Kalaycı, K. (2016). Confusopoly: competition and obfuscation in markets. *Experimental Economics*, 19(2), 299–316. <https://doi.org/10.1007/s10683-015-9438-z>

- Kaptein, M., De Ruyter, B., Markopoulos, P., & Aarts, E. (2012). Adaptive Persuasive Systems: A Study of Tailored Persuasive Text Messages to Reduce Snacking. *ACM Trans. Interact. Intell. Syst.*, 2(2). <https://doi.org/10.1145/2209310.2209313>
- Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1), 43–109. <https://doi.org/10.1093/jla/laaa006>
- 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. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW). <https://doi.org/10.1145/3359183>
- Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, 114(48), 12714–12719. <https://doi.org/10.1073/pnas.1710966114>
- Netherlands Authority for Consumers & Markets. (2020). *Boundaries of online persuasion*.
- Norwegian Consumer Council. (2018). *Deceived by design*. <https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf>
- Rasch, A., Thöne, M., & Wenzel, T. (2020). Drip pricing and its regulation: Experimental evidence. *Journal of Economic Behavior and Organization*, 176, 353–370. <https://doi.org/10.1016/j.jebo.2020.04.007>
- Richards, N., & Hartzog, W. (2019). The Pathologies of Digital Consent. *Washington University Law Review*, 96, 1461. [https://openscholarship.wustl.edu/law\\_lawreviewhttps://openscholarship.wustl.edu/law\\_lawreview/vol96/iss6/11](https://openscholarship.wustl.edu/law_lawreviewhttps://openscholarship.wustl.edu/law_lawreview/vol96/iss6/11)
- SERNAC. (2021). *INFORME DE RESULTADOS DE LEVANTAMIENTO DE DARK PATTERN* (Vol. 2021).
- SERNAC. (2022). *CONSENTIMIENTO EN EL USO DE COOKIES: EVIDENCIA EXPERIMENTAL SOBRE EL IMPACTO DE LA PRIVACIDAD POR DEFECTO Y LOS PATRONES OSCUROS EN LAS DECISIONES DE LOS CONSUMIDORES*.
- Spiller, S. A., Fitzimons, G. J., Lynch JR., J. G., & McClelland, G. H. (2013). Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression. *Journal of Marketing Research*, L(April), 277–288.
- Thaler, R. H. (2018). Nudge, not sludge. *Science*, 361(6401), 431. <https://doi.org/10.1126/SCIENCE.AAU9241>
- Thaler, R. H., & Sunstein, C. (2008). *Nudge*. Yale University Press.
- Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, 93(2), 175–179. <https://doi.org/10.1257/000282803321947001>
- Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2012). Choice Architecture. In E. Shafir (Ed.), *The Behavioral Foundation of Public Policy* (pp. 428–439). Princeton University Press.
- DIRECTIVE 2005/29/EC OF THE EUROPEAN PARLIAMENT AND OF THE*

COUNCIL “*Unfair Commercial Practice Directive*,” L149/22 (2005) (testimony of THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE UNION.).

UK Competition and Market Authority. (2022). *Evidence review of Online Choice Architecture and consumer and competition harm*. <https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers/evidence-review-of-online-choice-architecture-and-consumer-and-competition-harm>

Varian, H. R. (2006). Revealed Preferences. In M. Szenberg, L. Ramrattan, & Aron A. Gattesman (Eds.), *Samuelsonian Economics and the Twenty- First Century* (pp. 99–115). Oxford University Press.

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*, *19*(5), 373–390. <https://doi.org/10.1002/per.542>