CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

Discussion Paper No. 2025-04

Richard Mills, Stuart Mills and Cass R. Sunstein

December 2025

ManipulationDetect: An AI Auditing Tool for Online Choice Architecture

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Samantha Stapleford-Allen
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 74 86214
Samantha.Stapleford-Allen@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

ManipulationDetect: An AI Auditing Tool for Online Choice Architecture

Richard Mills[†‡] Stuart Mills [§] Cass R. Sunstein [¶]

December 17, 2025

Abstract

Policymakers and regulators are increasingly interested in behavioural auditing tools to counteract manipulative designs in Online Choice Architecture (OCA). To date, auditing tools have been largely manual, creating a trade-off between time, cost, and scale. This article presents a tool called 'ManipulationDetect', an internet browser plug-in that uses AI to detect, highlight, and record potentially manipulative OCA techniques in real-time. We offer a technical overview of how ManipulationDetect works, present an example audit which demonstrates the tool's advantages, and highlight important practical next steps for further development.

The authors would like to thank Dilip Soman, Anna Hochleitner, and Joris Schröder for their generous and detailed feedback on the manuscript. We also thank regulators Ceren Cibik, Sazeda Yasmin, Stuart Hossack, and Jonathan Porter (Office of Communications - Ofcom), as well as Sam Taylor and Aisling Ni Chonaire (Competition and Markets Authority - CMA) for constructive comments, valuable discussions on regulatory applications, and feedback on drafts of the manuscript. The study was preregistered on OSF (link: https://osf.io/shw65/). All errors are our own.

[†]Correspondence concerning this manuscript should be addressed to Richard Mills

[‡]University of Nottingham, School of Psychology, United Kingdom, Email: richard.mills2@nottingham.ac.uk.

[§]University of Leeds, School of Economics, United Kingdom, Email: s.mills1@leeds.ac.uk.

[¶]Harvard University, Harvard Law School, United States. Email: csunstein@law.harvard.edu.

1 Introduction

As everyday services shift online, regulators and policymakers have become interested in Online Choice Architecture (OCA). OCA captures the various design techniques that online services might use to influence the choices of citizens and consumers (Sugg and Lesic, 2022). Of particular interest is how OCA techniques can harm people (Mills, 2024), including through manipulation (Akerlof and Shiller, 2015; Sunstein, 2025). The behavioural science literature has developed the idea of *sludge*—choice architecture that leverages behavioural science insights to make it harder for people to achieve their objectives (Sunstein, 2021; Thaler, 2018). Likewise, the public administration literature has explored *administrative burden*—excessive requirements that make it harder for people to access public provisions (Herd and Moynihan, 2019). Finally, the user interface (UI) literature has developed various taxonomies around *dark patterns* and *deceptive designs*—broadly, UI designs which lead users to engage in behaviours which primarily benefit the designer, not the user (Brignull, 2010; Gray et al., 2018, 2024; Lewis and Vassileva, 2024; Mathur et al., 2019). While each of these terms offers particular nuances, all fall under the broad description of *manipulation*, defined as "a form of influence, intended to affect thought or action (or both), that does not respect its victim's capacity for reflective and deliberative choice" (Sunstein, 2025, p. 19).

Manipulative OCA can cause people to pay more for products and services, prevent people from receiving public provisions to which they are entitled, and undermine positive competition between vendors (Akerlof and Shiller, 2015; Competition and Markets Authority, 2021; Federal Trade Commission, 2022). Such harms have driven growing calls for means of tackling manipulative OCA. In particular, the behavioural public policy literature has proposed developing auditing methodologies to document these practices and to target regulatory resources at the most problematic OCA techniques (Mills et al., 2023; Sunstein, 2022). This auditing perspective has generally been seized upon by policymakers and regulators, particularly regarding the measurement of sludge (Varazzani et al., 2024). However, important challenges remain.

One is that of *scale*. A regulator might have to audit a large number of online services, exhausting their available resources (Mills, 2024). This is to say nothing of the complexity of some online processes, owing simply to the variety of products and services a vendor or government body might be responsible for. As a result, audits of OCA can be challenging, and manipulative OCA is often investigated only after people have suffered it (Federal Trade Commission, 2022). Another is that of *variety*. There are many different, potentially manipulative, OCA design techniques (e.g., Gray et al. (2018); Mathur et al. (2019)). This compounds the problem of scale, as auditing is often not a matter of looking for one technique many times, but many techniques, many times. Online services may also evolve their strategies over time, risking a cat-and-mouse approach to online protection. A third challenge is that of *autonomy*. It is impractical, and perhaps undesirable, to suppose that policymakers and regulators should have a divining hand in the online safety of individuals. Individuals themselves might also experience OCA techniques differently, depending on their goals and individual psychology (Mills, 2024). Tools that support individuals to identify and judge OCA *for themselves* may achieve many—though not all—of the goals of auditing, while also empowering people (Reijula and Hertwig, 2022).

We present a novel tool that responds to these challenges. ManipulationDetect is a free, internet browser plug-in which uses a Large Language Model (LLM) to scan webpages for OCA techniques in real-time (manipulationdetect.com). The tool then categorises techniques in terms of manipulative severity based on a traffic-light system (green, amber, red). ManipulationDetect thus enables auditing at scale by leveraging AI technologies, responds to variety by undertaking real-time scanning, and supports autonomy by providing helpful prompts to users, without directing their choices or behaviours.

2 What is ManipulationDetect?

ManipulationDetect is a free, AI-powered browser extension that uses LLM technology to audit and highlight OCA techniques. Using prompt engineering based on OCA frameworks found in the literature, we enable the LLM to identify different OCA techniques. Techniques are then communicated to a user through the universally recognisable traffic-light system, a method that has been used in prior research from nutrition labelling (Kunz et al., 2020) to dementia risk communication (Matovic et al., 2024). Anonymised data from the audit of the webpage is also collected by ManipulationDetect and saved in a database, enabling future use by regulators and scholars.

2.1 How Does ManipulationDetect Work?

We begin with a high-level explanation of how ManipulationDetect works, before elaborating on specific details in subsequent sections. ManipulationDetect's analysis is a five-step, screenshot-based approach (see Figure 1a).

The process begins when a *user activates a scan* (Step 1) (see video for active demonstration). This triggers the *capture stage* (Step 2), where the tool collects and stitches together three screenshots of the active page.[1] This composite image is then scaled and resized to improve processing efficiency.

Next is the *prompt assembly stage* (Step 3). The stitched screenshot is merged with a prompt template containing a taxonomy of common OCA techniques. For each technique, this template provides a detailed definition, illustrative examples, a default severity rating, and a crucial 'tip' which gives the LLM useful guidance on how to find each technique on the webpage (see Figure 1b and the SM).

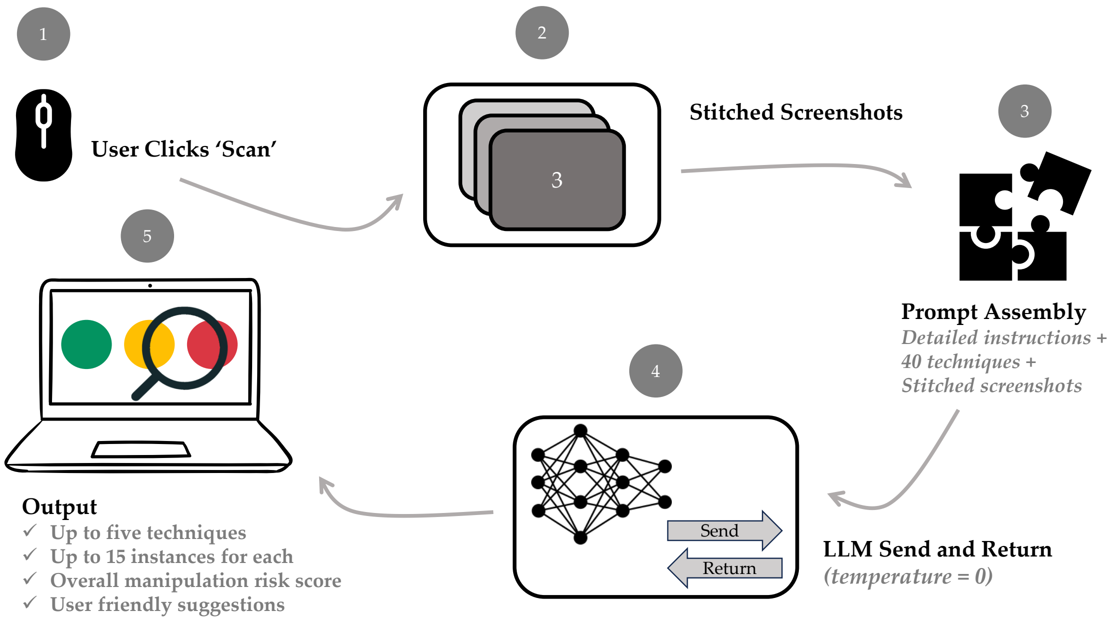
In the *send and return stage* (Step 4), the final payload (stitched screenshots + OCA taxonomy) is sent to Google's Gemini Flash 2.5 LLM, with the system temperature set at 0.[2] The LLM processes the final prompt and returns a list of *candidate* techniques possibly present on the page. For each candidate, the response includes (i) the specific technique name, (ii) instances of this technique through verbatim quotes of the text that triggered the match, (iii) a justification for its reasoning surrounding each technique, and (iv) a severity rating.

Finally, there is the *output stage* (Step 5). The identified techniques are flagged to the user through an on-screen traffic-light indicator. ManipulationDetect's pop-up window also offers a summary report, listing up to five of the techniques identified. This cap is simply to mitigate the risk of a user being overwhelmed, which might inhibit the user experience and thus counteract the tool's ultimate objective. ManipulationDetect also provides the number of instances of each OCA technique, an overall manipulation risk for the webpage, and suggestions for the user to consider. These are all outlined in more detail below.

[1] The tool is flexible in the number of screenshots it can stitch together. Three has been chosen given the trade-off between capturing the "full-set" of information on the page and the processing time of the LLM.

[2] The 'temperature' parameter governs the stochastic nature of LLM outputs. A temperature of 0 forces a largely deterministic approach, selecting the highest-probability words (maximum a posteriori) (Renze, 2024).

a) Flow Diagram of ManipulationDetect



b) OCA Techniques, with default severity rankings.

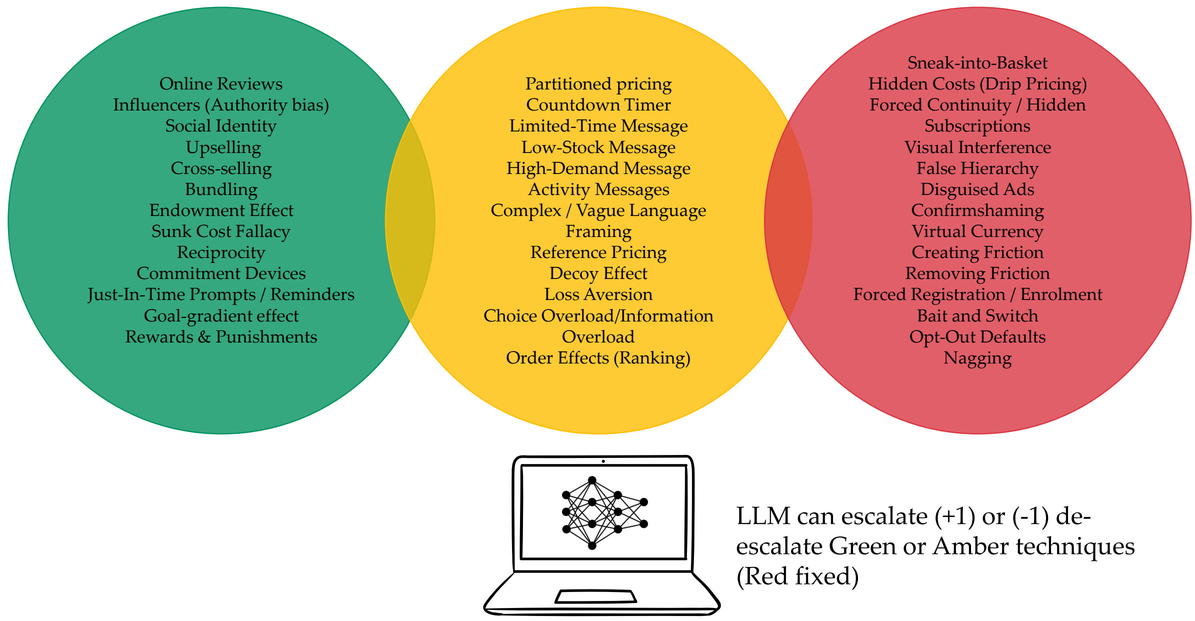


Figure 1: Under the hood of ManipulationDetect

Note. (a) the main mechanics of the tool, and (b) the OCA techniques currently being identified (left – green, middle – amber, right – red). The full working definitions, examples, severity ratings and suggestions are provided in the Supplementary Material (see SM Sections A, and B).

2.2 Severity Ranking Framework

The digital landscape is constantly evolving. As such, ManipulationDetect has purposely been designed to be a dynamic tool. New techniques can be continuously integrated as they are identified by academic researchers, regulators, and the public. As it stands, 40 unique OCA techniques have been integrated, drawing from three foundational sources: Mathur et al. (2019), Li et al. (2024), Competition and Markets Authority (2021) (see Figure 1b).

The most recent of these, Li et al. (2024), draws on the OECD (2022) report on dark commercial patterns. This work views OCA techniques from the perspective of six "vantage points" of harm (labelled H1 to H6), which can be broadly summarised into three groups: *harm to user autonomy* (H1), where choices are forced or obfuscated; personal user detriment, which includes *financial loss* (H2), *privacy breaches* (H3), and *psychological strain or time loss* (H4); and structural user detriment, which involves *distorted market competition* (H5) and the *erosion of consumer trust* (H6).

To translate this *in-depth* classification system into a simple, actionable metric, we use the well-known 'traffic light system' of green, amber and red. When classifying red, ManipulationDetect prioritises the harms identified by Li et al. (2024) as the most 'severe': financial loss (H2) and privacy harms (H3).

To translate the OECD framework to the traffic light system, we employ a simple rule-of-thumb: (1) *Red*: techniques that aim to *directly* and *intentionally* cause financial loss (H2) or privacy harm (H3); (2) *Amber*: techniques that may *indirectly* lead to financial loss (H2) or privacy harm (H3); (3) *Green*: techniques that fall into the *other harm categories* (H1, H4, H5, H6) but are not directly linked to financial or privacy loss. Some techniques with nominal financial loss or privacy harm are also marked as green.

Figure 1b shows the classifications of various prominent OCA techniques following this approach. These are not necessarily fixed, with techniques potentially leading to different harms depending on different contexts. To this end, the LLM is instructed to escalate or de-escalate green and amber techniques based on context. However, red techniques remain fixed given the outsized risks such techniques may pose if erroneously de-escalated. We acknowledge that these classifications may also reflect researchers' subjective judgements, and return to this limitation later in this article.

2.3 User Interface

ManipulationDetect is designed to inform and empower users. Users are provided with a brief description of what each colour ranking of the traffic-light indicator means. The tool itself is designed to be easy to use, with three immediate options: (1) *Scan This Page*; (2) *Last Scan*; and (3) *Example*. These options allow a user to initiate a scan, review their previous scan, and review an example of a given technique, respectively (see Figure 2a, Homepage).

(a) Homepage (b) Results

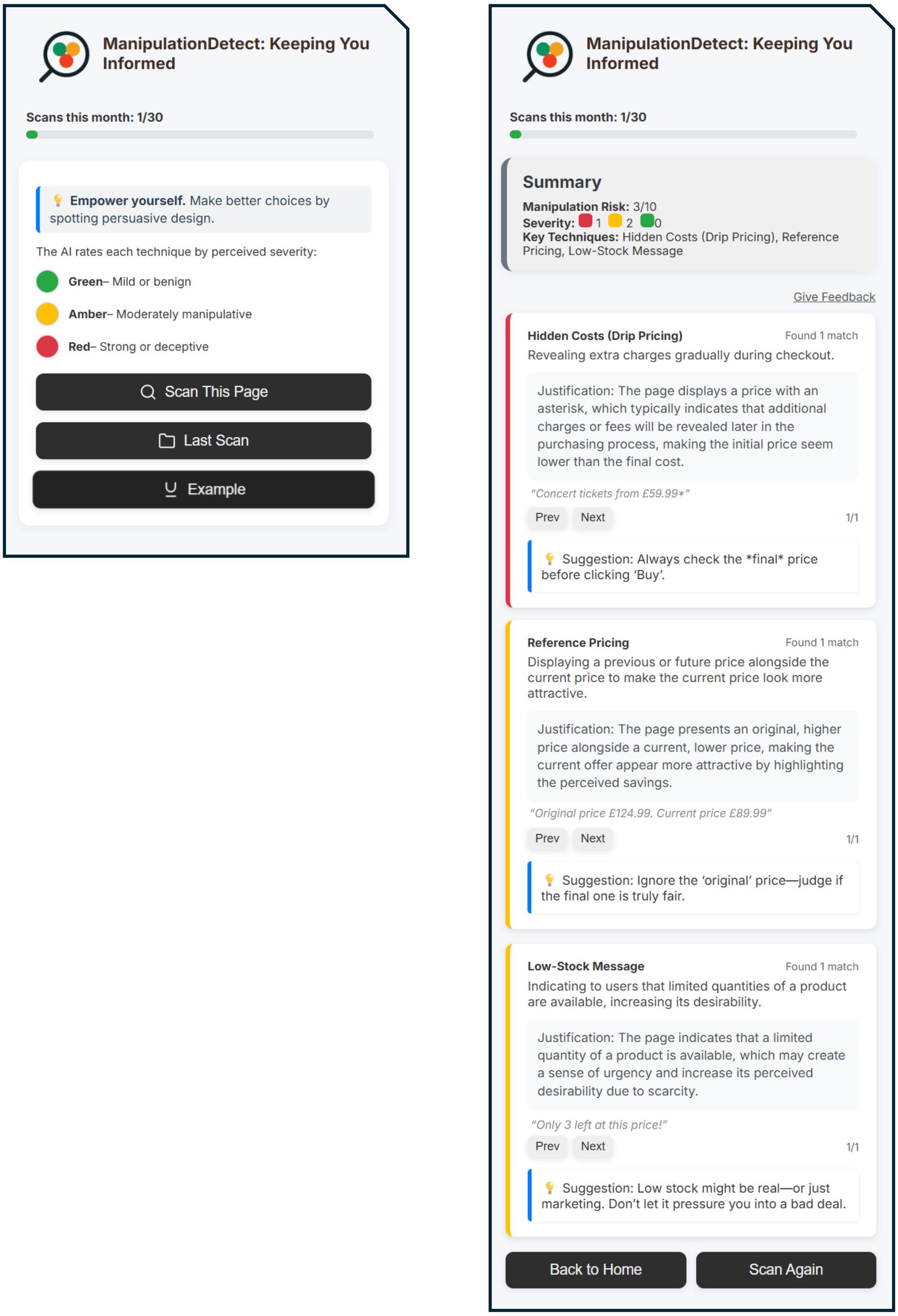


Figure 2: ManipulationDetect User Interface

After a scan, ManipulationDetect provides a variety of information to a user (see Figure 2b, Results). The *Summary box* displays a headline manipulation risk for the website. A user can hover over this box to see how this rating is calculated (also see below). The number of OCA techniques identified is also displayed and ordered from most to least severe (red through to green), followed by a high-level description of these techniques. Below the Summary box is an *Output box* that provides (1) the name of the technique; (2) its definition as given to the LLM; (3) the specific example found on the webpage; (4) a justification from the LLM for highlighting it; and (5) a suggestion to the user as to how they might counteract it.

2.4 Manipulation Risk

The manipulation risk (MR) is calculated through a two-step process that accounts for the severity of a given technique and the frequency of that technique on the webpage. We classify this metric as a 'risk' as not all techniques will necessarily lead to manipulation, but the presence of a technique nevertheless creates the risk that some people will be manipulated. Firstly, ManipulationDetect calculates the cumulative severity of techniques found on a webpage:

$$\mathrm{IS} = \sum_{i} (f_i \times s_i) \tag{1}$$

where f_i is the frequency of technique i on the webpage, and s_i is technique i's corresponding severity ($s_i \in \{1, 2, 3\}$). Following Equation 1, every finding, regardless of severity, contributes to the Instance Score (IS), but more severe findings contribute more. This approach reflects what the Federal Trade Commission (2022, p. 6) have called the "net impression conveyed," and what the Competition and Markets Authority (2021, p. 4) have called the "combined effect" of OCA, recognising that while individual techniques may have important influences, a user's experience of OCA is also a product of the totality of techniques encountered. From the perspective of risk, the more techniques a user encounters, the more opportunities they have to be manipulated. To this end, the IS score is then transformed into a more intuitive 0-10 manipulation risk following Equation 2:

$$\mathrm{MR} = \lfloor 10 \times \frac{\mathrm{IS}}{\mathrm{IS} + k} \rceil \tag{2}$$

where k is a constant that 'tunes' how quickly the MR increases. A simple heuristic is to set k to a value that represents the mid-point for a 10-point MR scale (5 out of 10). For instance, a website with five 'red' techniques (e.g., $k = 3 \times 5$). The effect of this tuning parameter can be observed through some example figures. Where the IS $= 10$, and $k = 15$, the MR $= 4$. Where IS $= 50$, the MR $= \lfloor 7.692 \rceil$ or, after rounding, 8. This aligns with the intuition that a webpage with weighted instances of 50 is more likely to be manipulative compared to one with weighted instances of only 10, because there are many more techniques, and thus many more opportunities for manipulation to arise.

There are two advantages to constructing the MR this way. Firstly, the rating is always monotonic. While there might be disagreements about how a 'manipulation risk' should be conceptualised, and the parameters which may go into it (e.g., the severity ranks and value of k), the manipulation ranking of webpages implied by the MR will be preserved regardless of what parameters are chosen (provided webpages are compared with the same set of parameters). Secondly, Equation (2) ensures that while more techniques increase the overall MR, this occurs at an ever-diminishing rate. For instance, an IS $= 10$ corresponds to a rating of 4, but one of 50 (a quintupling) corresponds to a rating of only 8, and one of 100 (a further doubling) only 9. This reflects the idea of risk, with more techniques increasing the likelihood of manipulation, without the MR ever suggesting manipulation is a certainty.

2.5 What Data Are Collected?

To ensure that ManipulationDetect can be used for OCA research, the tool collects data from each scan. The collection process prioritises (and promises) user anonymity, with each user being assigned a random, non-personally identifiable user ID upon installation. For every scan initiated, ManipulationDetect logs the URL of the webpage, the raw JSON response from the LLM model, and the final, verified list of detected OCA techniques.

Beyond the core scan results, ManipulationDetect gathers anonymised data on how users interact with the findings. This includes which techniques users choose to investigate further, and their use of navigation features. These behavioural data provide crucial insights into usability, and which techniques users find most salient. Furthermore, users can voluntarily submit feedback through a dedicated form, providing information on the tool's perceived usefulness and its impact on their online behaviour.

These data create numerous opportunities for OCA research. Logs of websites allow practitioners to rapidly audit online services, and over time, this database of results will enable longitudinal analysis of OCA techniques and how they are being deployed. The collection of LLM outputs, too, enables third-party verification and (in principle) replication of results. The collection of behavioural data offers valuable opportunities for researchers and practitioners to understand if and how individual techniques are influencing behaviour. These data will also provide insights for developing the parameters of ManipulationDetect over time. Longitudinal monitoring of a technique's prevalence, as well as the MR, provide powerful tools for regulators to monitor compliance with new rules, principles, and regulatory standards.

3 Illustrative Audit with ManipulationDetect

To demonstrate the capabilities of the tool, we present findings from a pre-registered (see Mills et al. (2025)) preliminary audit of 60 of the most visited commercial websites in the UK. The audit simulated a typical shopping experience, which is a common approach in previous auditing studies (Behavioural Insights Team, 2022; Hodson et al., 2025; Mills et al., 2023). The protocol involved the auditor (i) scanning a website's landing page (Step A), and then simulating a typical buying experience from product/service 'category' page (e.g., a list of multiple laptops), to product/service 'product' page (e.g., information about a specific laptop), to product/service 'basket/cart' page (steps B-D). Steps B-D were repeated for a total of three different product and services randomly selected. Moreover, to evaluate the tool's test-retest reliability, each page was scanned sequentially, establishing a planned upper bound of 1,200 total scans.[3]

In total, 60 commercial websites in the UK were scanned (between 16-20th October 2025), with a total of 1,052 scans.[4] The order of the websites audited was randomised, and websites were categorised

[3]The version of the tool used in the audit differed slightly from the version that is publicly available. The main difference is that the audit version included drop-down boxes to select the Scan #, Page Type, and Repetition #, and a screenshot of the stitched page was locally saved to the device for logging purposes. This version of the tool will be made available in the OSF documentation. A screenshot of this tool is provided in the SM Figure C1.

[4]As documented in the OSF, some webpages do not have basket/cart pages for certain product/services, particularly those that require an application (e.g. financial loans, or car insurance).

into Standard Industrial Classification (SIC) sections, which designate the sector of the economy the company belongs to (Office for National Statistics, 2022) (see SM Table C.2 for summary statistics). The mean scan duration was 35.39 seconds (95% CI [34.77, 36.02]), with total costs amounting to £17.66 (~£0.017 per scan).

3.1 Landscape of OCA in the UK

Figure 3 shows the overall prevalence of OCA techniques across the 60 websites, grouped by their severity rating (green, amber or red). Since each page was scanned twice, a technique is counted only if both scans detected it. Bars report the percentage (%) of websites on which a technique appeared in both sequential scans.[5]

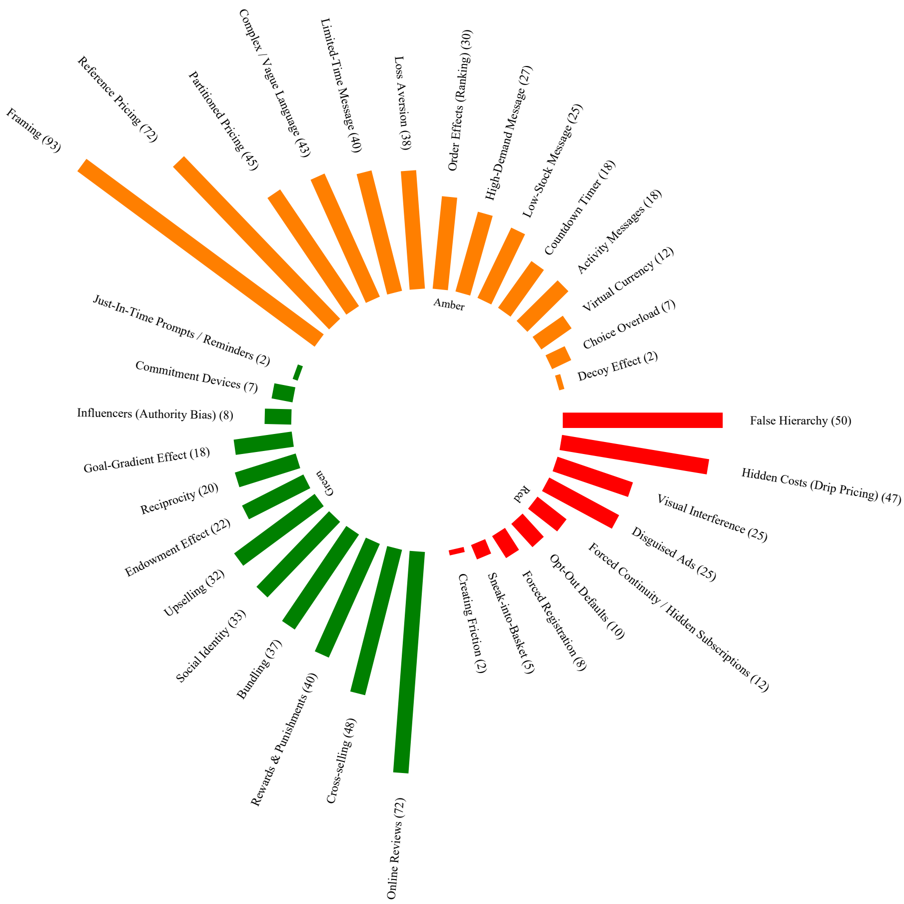


Figure 3: Prevalence of 40 OCA Techniques

Note. Bar length = % of websites where technique was found in both scans (the intersection of the two scans). N=60 websites (N=526 unique observations).

[5] This method is the most conservative as the technique needs to have been detected in both Scan 1 and 2. In a Venn Diagram, this method would be the intersection of Scan 1 and 2. Moreover, with the intersection method, there were only three instances where a green or amber technique was escalated or de-escalated—showing that broadly, the LLM has a strong preference for the pre-defined severity rating given in the engineered prompt.

Two red techniques are strikingly common, with False hierarchy being detected on 50% of sites and Hidden costs (drip pricing) on 47%. Other red techniques were detected less frequently (Visual Interference (25%), Disguised Ads (25%), and Forced Continuity / Hidden Subscriptions (12%), though still at non-trivial rates. Amber techniques were pervasive, with Framing being detected on 93% of sites, followed by Reference Pricing (72%), Partitioned Pricing (45%), Complex/Vague Language (43%), and Limited Time Messages (40%). Green techniques remain widespread as well, with Online Reviews (72%), Cross-selling (48%), Rewards & Punishments (40%), Bundling (37%), and Social Identity (33%) being detected at notable frequencies.

Splitting these results by page-type (see SM, Table C3), shows that certain techniques were significantly more likely to appear on particular pages. For instance, Opt-out Defaults (p<0.01), Cross-selling (p<0.01) and Partitioned Pricing (p<0.01) were significantly more likely to be present on latter stages of the consumer journey (i.e., the product/service or cart/basket pages), whereas False Hierarchy (p<0.01), and Online Reviews (p<0.01) were most prevalent in the middle stages (highest proportions were found in the product/service 'category' pages, p<0.01). Lastly, techniques such as Framing (p<0.01) had their highest proportions in the earlier stages, such as the landing page (p<0.01).

Lastly, across SIC groupings, techniques used vary significantly by Sector type (see SM, Table C4). For instance, Section K (Financial and insurance activities) have the highest prevalence of Complex/-Vague Language (p<0.01), and Section G (Wholesale and retail trade) have the highest prevalence of Cross-selling, Reference Pricing and Limited Time Messages (all p<0.01).

Figure 4 presents average manipulation risk by the 60 webpages. As before, a technique is only counted if detected in both scans, with the score derived by first averaging the risk scores from both scans for each page, and then averaging these page scores across the entire website. Commercial website information has been redacted. The results reveal substantial variation in manipulation risk across pages, with a mean score of 5.11 (SD = 1.02) and a median of 4.98, indicating that approximately half of all pages score below 5 out of 10. Nine webpages (15%) score above 6, whereas the most extreme pages sit near the upper tail (95th percentile), at values of greater than 7.2/10. Such variation is not unsurprising given the variety of different industries captured in our sample.

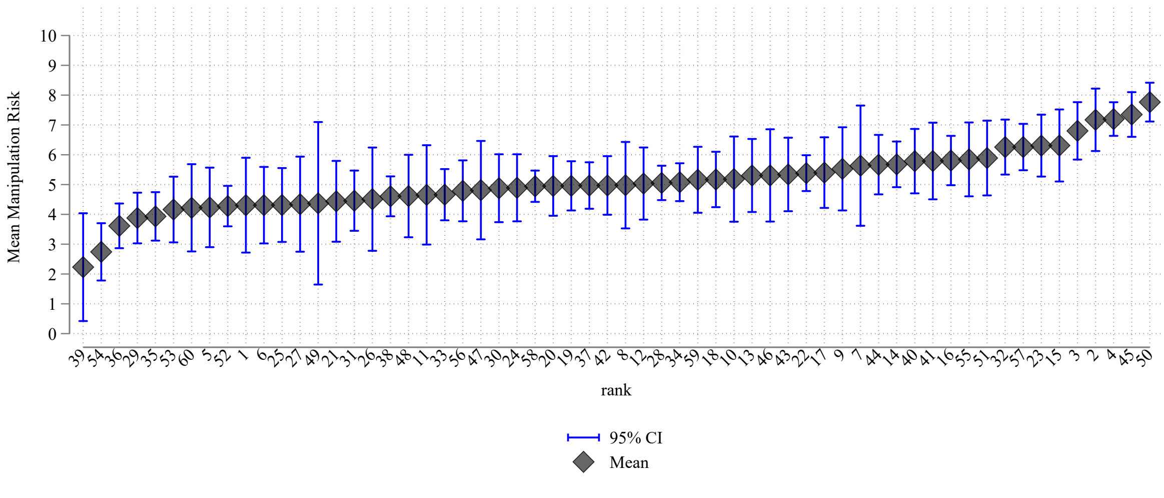


Figure 4: Mean manipulation risks for websites (average of averages between two scans)

Note. N = 60 websites. Dot is mean risk (averaged across pages). Whiskers are 95% CI of the mean.

Regression analyses (SM Table C5) indicate that category pages had a significantly higher MR than landing pages ($b = 0.87$, $SE = 0.30$, $p < 0.01$), while product ($b = 0.151$–0.127, $p > 0.1$) and basket ($b = 0.149$–0.111, $p > 0.1$) pages did not differ reliably. Differences across industry sectors were minimal, with only Administrative Support Service Activities (Section N) showing lower risk than Wholesale and Retail Trade (Section G) (b −0.67, $p < 0.05$). Visual 'density' on the page (proxied by image size sent to the LLM) was positively associated with MR ($b = 0.39$–0.43, $p < 0.05$).

3.2 Reliability of the Tool:

A key test of ManipulationDetect is its test-retest reliability: does it produce a consistent score when auditing the same page multiple times? We first assessed the reliability of the MR score. A two-way random-effects intraclass correlation (ICC [2,1]) on the 526 page-pairs revealed good reliability, with an ICC of 0.733 (95% CI [0.691, 0.770]). The average-measure ICC (ICC [2, k]), which reflects the consistency when averaging scans, was excellent at 0.846 (95% CI [0.817, 0.870]) (Koo and Li, 2016).

This statistical agreement is visualised in the Bland-Altman plot in Figure 5 (Panel A). The plot confirms there is no systematic bias between scans: the mean difference (solid line) is 0.0. The random measurement error, represented by the standard deviation (SD) of the differences, was 1.2, resulting in 95% limits of agreement (dashed lines) at approximately ± 2.4.

a) Raw Scores (M=0.0, SD=1.2)

b) Intersection Scores (M=0.0, SD=1.1)

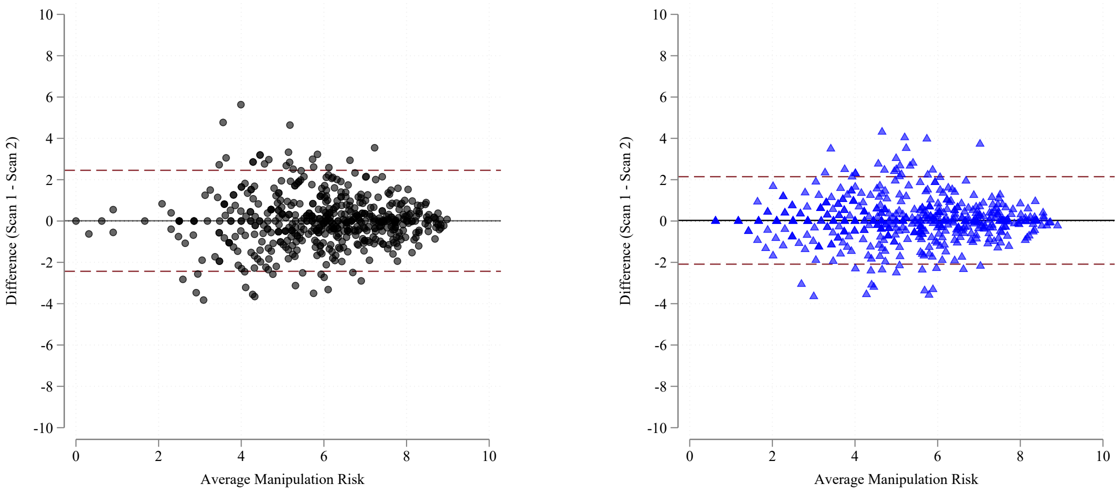


Figure 5: Bland-Altman Plot of Manipulation Risk Scores

While the final scores are reliable, we also consider the measurement "noise" as demonstrated by the variance shown in Panel A. A Jaccard similarity analysis, which measures the overlap of the detected technique lists (see SM, Table C.5), revealed that repeated scans of the same page had a median Jaccard similarity of 0.50. This indicates that while the final manipulation scores were similar, the scans often arrived at that score by detecting different combinations of OCA techniques. The instability was not random: some techniques (e.g., Reference Pricing, J=0.86: Low-Stock Message, J=0.74) were highly stable, while others (e.g., Creating Friction, J=0.25; Decoy Effect, J=0.08) were not particularly stable.

Given these results, we calculated a conservative "Intersection-Only Score" using only those techniques present in *both* scans in the analysis described above. This analysis effectively models the tool's reliability *if* the tool's detection in the techniques found were perfect.

With this method, the individual-measure ICC increases from 0.733 to 0.836 (95% CI [0.808, 0.860]). The average-measure ICC likewise improved to 0.911 (95% CI [0.894, 0.925]), moving firmly into the "excellent" reliability category (Koo and Li, 2016). This improvement is visualised in Figure 5 (Panel B), with standard deviation of the differences being reduced from 1.2 to 1.1, resulting in visibly tighter 95% limits of agreement.

3.3 Tuning Parameter 'k'

Another crucial aspect of the tool is the k parameter, a flexible tuning constant in the manipulation risk formula that can be adjusted by the 'auditor'. The k parameter functions as the "half-point": it is the raw score required to achieve a manipulation risk of 5 out of 10. While our main analysis used a default $k = 15$, this can be adjusted by the auditor.

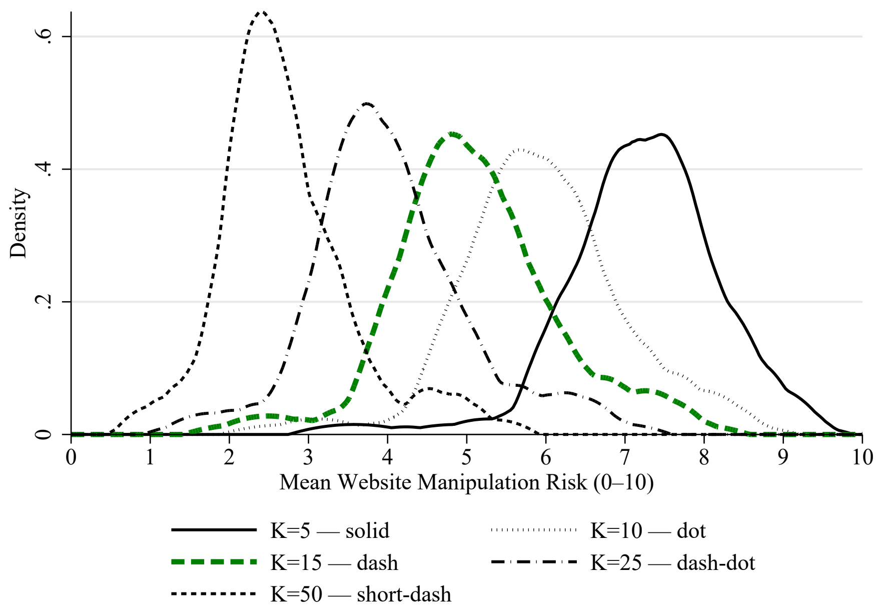


Figure 6: Overlaid Densities of Manipulation Risk by Parameter 'k'

To demonstrate this flexibility, we re-computed the mean webpage risk scores (using the intersection method) across a range of k values ($k \in \{5, 10, 15, 25, 50\}$). As shown in Figure 6, the choice of k directly influences the shape and central tendency of the score distribution, while preserving the rank-order of the websites. Larger k values (e.g., 50) compress the scores toward zero, while smaller k values (e.g., 5) result in higher mean scores.

Our default of $k = 15$ provides a balanced distribution. It is also one of only two distributions (along with $k = 10$) for which the null hypothesis of normality could not be rejected by a Shapiro-Wilk test (W=0.966, p=0.092). While this flexibility allows auditors to select a k that aligns with their own theoretical assumptions, our findings indicate k=15 as a useful starting point.

4 Limitations

It is essential to interpret the findings of this preliminary audit within the context of its design. The audit serves primarily as a proof-of-concept to demonstrate the capabilities of ManipulationDetect, rather than as a comprehensive, systematic study of manipulative OCA on UK websites.

We recognise that, firstly, there is a focus on reliability within the audit. Our analysis evaluates the tool's test-retest reliability (i.e., its consistency), demonstrating good-to-excellent ICC scores. However, this audit did not establish the tool's *validity*. That is, whether the tool is identifying the same techniques as a human auditor would. It is intuitive that a valid tool would align with human judgement around manipulative OCA. Establishing such validity is an important and interesting next step for future research.

Secondly, this audit likely systematically underestimates the prevalence of techniques. The prevalence figures reported here should be considered a conservative lower bound on the true number of manipulative techniques present. This underestimation is a direct result of three key methodological choices: (i) the protocol was limited to scanning the first three viewports of each page, meaning any techniques appearing after these were missed; (ii) the tool was constrained to report a maximum of five unique techniques and 15 total instances per scan. We frequently observed scans hitting this 15-instance cap, strongly indicating that more techniques were present on the page but not captured in the output; (iii) our primary analysis relied on a conservative "intersection method", counting only those techniques detected in both sequential scans. While this increases confidence in the reported findings, it systematically excludes any technique that was (potentially correctly) identified in only one of the two scans.

In summary, our example audit demonstrates the practical advantages of ManipulationDetect and provides encouraging evidence of the tool's reliability. Critiques of the audit design are largely secondary to the study's primary purpose: demonstrating the capabilities of ManipulationDetect.

5 Directions for Future Development

The development of ManipulationDetect presents several methodological challenges but also valuable future directions for research. Firstly, determining the parameters for ManipulationDetect remains an important area of development. Broadly, there are three parameters to consider. Firstly, there is the taxonomy of OCA techniques that are included in the prompt engineering of the tool. Secondly, there is the severity ranking of each of these techniques. Thirdly, there is the value of the constant k in the MR calculation.

At present, these parameters have been determined by researcher judgement, based on a broad engagement with the literature. To this extent, they reflect researcher discretion and could be challenged and improved upon by other scholars and practitioners. We broadly invite wider collaboration in this regard. While the setting of parameters is a challenge, broad collaboration between scholars and practitioners represents a compelling opportunity to further test and refine ManipulationDetect. In particular, we call for greater collaboration to determine a standard taxonomy of OCA techniques. Taxonomies of choice architecture have been proposed in behavioural science (e.g., Münscher et al. (2016)), but regulators have also developed their own internal taxonomies—for instance, the Competition and Markets Authority has proposed a taxonomy of 21 OCA techniques (Sugg and Lesic, 2022).

Within the UI literature, Gray et al. (2018) offer a taxonomy of 5 broad techniques, Mathur et al. (2019) 15, and Li et al. (2024) 32 (though, grouped into six broader categories). There is thus ample space for discretion and disagreement within the literature. A standard taxonomy of OCA techniques would support the development of ManipulationDetect, in terms of prompt engineering, and more broadly promote the widespread development of behavioural auditing approaches by standardising methodologies and fostering a shared language (Gray et al., 2024).

We see great potential in an iterative approach to determining the severity rankings of different OCA techniques. The prevalence of a technique, determined through longitudinal data collection, coupled with the behavioural impact of a technique, determined through behavioural data collection, will—over time—enable much greater refinement of severity rankings, and thus of the tool overall. This effort could be boosted by user engagement and feedback, and by further findings within the behavioural auditing and OCA research community. Similar efforts can already be seen in the approach to categorising dark patterns outlined by the Organization for Economic Cooperation and Development (OECD, 2022), championed by Li et al. (2024).

Setting the parameter k is perhaps most difficult, as it is not clear that either broad collaborative research or iterative data collection can offer compelling insights into what this value ought to be. Nonetheless, it may also be a mistake to believe that there exists a single appropriate k value. Different industries may legitimately engage in different OCA techniques, warranting more nuance in how some services should be evaluated (Mills et al., 2023). The refinement of k, while less straightforward than the refinement of other parameters, nevertheless must entail broad engagement with the wider researcher and practitioner communities. To this end, we invite these groups to engage with ManipulationDetect, and to investigate how the tool aligns with ongoing auditing efforts in different sectors and industries. Our example audit demonstrates how practical results can be leveraged to gain insights into appropriate values of k.

Another challenge, and a surmountable one, is the small ManipulationDetect user-base at present. The tool is currently in the early-adoption phase, and given this, it has already seen promising user adoption. Nevertheless, many of the advantages of ManipulationDetect can be realised only through wider adoption, to build up a critical mass of data for empirical studies of OCA, and for further refinement of the tool itself. As the tool sees wider adoption, the opportunities for future research represent one of the most exciting promises of ManipulationDetect. We believe the tool can address a range of theoretically and practically important questions in the behavioural science and OCA literatures. For instance, ManipulationDetect can catalogue different techniques across different industries, and so can help scholars in, say, the retail investing literature understand how techniques co-occur within, say, the gambling literature. The prospect of collecting longitudinal data may also help scholars track which industries originate techniques, and how techniques 'migrate' across industries over time.

Another critical question that ManipulationDetect might address concerns salience and behavioural spillovers: does making users aware of OCA techniques actually change their behaviour? A long-standing open question remains as to whether transparency regarding 'nudges' (or OCA in our case) nullifies their effectiveness (see meta-analysis by Bruns et al. (2025)). This question closely relates to the matter of autonomy. Again, longitudinal data gathered through ManipulationDetect offers unique opportunities to investigate this question. Furthermore, online experiments, and A/B testing of different versions of the tool, could help design effective interventions to protect individuals from deceptive

OCA techniques. ManipulationDetect would enable these interventions to be linked to real behavioural outcomes, offering further advantages for scholars and practitioners.

6 Conclusion

This article has introduced ManipulationDetect, a free, AI-powered browser plug-in for detecting OCA techniques across the internet. ManipulationDetect is motivated by the growing demand for behavioural auditing tools to help people avoid manipulative OCA techniques. Interest in behavioural auditing is growing, but methodologies remain constrained by a lack of scale, the problem of variety, and the demand for individual autonomy. ManipulationDetect responds to each of these problems and represents an important practical development for scholars and practitioners within the OCA space.

We have provided technical details of how ManipulationDetect works. This includes a procedural breakdown of the tool when a user scans a webpage; a methodological breakdown of how key metrics are calculated; and an explanation of the data collected by the tool, as well as the parameters used within it. We have demonstrated how a practitioner might use the tool through an example audit of 60 commercial websites.

ManipulationDetect offers several promising avenues for future research, but it also faces important challenges. We have outlined these challenges and strategies to overcome them. Questions around the parameters of the tool necessitate broad collaboration between OCA researchers and practitioners. Developing a standard taxonomy of OCA techniques will be essential for standardising a tool like ManipulationDetect. Wider adoption offers opportunities to refine the tool iteratively with data, while unlocking numerous opportunities for research into OCA and auditing approaches, particularly in terms of longitudinal insights.

References

Akerlof, G. A. and Shiller, R. J. (2015). *Phishing for phools: The economics of manipulation and deception.* Princeton University Press.

Behavioural Insights Team (2022). Behavioural risk audit of gambling operator platforms. Technical report, Behavioural Insights Team.

Brignull, H. (2010). Dark patterns: inside the interfaces designed to trick you.

Bruns, H., Fillon, A., Maniadis, Z., and Paunov, Y. (2025). Comparing transparent and covert nudges: A meta-analysis calling for more diversity in nudge transparency research. *Journal of Behavioral and Experimental Economics*, 116:102350.

Competition and Markets Authority (2021). Algorithms: How they can reduce competition and harm consumers. Research paper, Competition and Markets Authority.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Federal Trade Commission (2022). Bringing dark patterns to light. Staff report, Federal Trade Commission.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Gray, C. M., Santos, C. T., Bielova, N., and Mildner, T. (2024). An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Herd, P. and Moynihan, D. P. (2019). *Administrative burden: Policymaking by other means.* Russell Sage Foundation.

Hodson, N., Onyeaso, O. O., Mills, S., Sunstein, C. R., and de Bruin, Wändi Bruine (2025). Evaluating adherence to patient registration paperwork guidelines: a mystery shopper study in english primary care. *BMJ Open*, 15(11):e100719.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98(6):1325–1348.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Kunz, S., Haasova, S., Rieß, J., and Florack, A. (2020). Beyond healthiness: the impact of traffic light labels on taste expectations and purchase intentions. *Foods*, 9(2):134.

Lewis, F. and Vassileva, J. (2024). Integrating dark pattern taxonomies. *arXiv preprint arXiv:2402.16760.*

Li, M., Wang, X., Nie, L., Li, C., Liu, Y., Zhao, Y., Xue, L., and Said, K. S. (2024). A comprehensive study on dark patterns. *arXiv preprint arXiv:2412.09147*.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., and 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):1–32.

Matovic, D., Ahern, M., Lei, X., and Wuthrich, V. M. (2024). The influence of traffic lights presentation of dementia risk screening information on older adults' motivations for risk reduction in primary care settings. *Alzheimer Disease & Associated Disorders*, 38(1):70–76.

Mills, R., Mills, S., and Sunstein, C. R. (2025). A preliminary audit with ManipulationDetect: An AI auditing tool for online choice architecture. OSF Preprints.

Mills, S. (2024). Deceptive choice architecture and behavioral audits: A principles-based approach. *Regulation & Governance*, 18(4):1426–1441.

Mills, S., Whittle, R., Ahmed, R., Walsh, T., and Wessel, M. (2023). Dark patterns and sludge audits: an integrated approach. *Behavioural Public Policy*, pages 1–27.

Münscher, R., Vetter, M., and Scheuerle, T. (2016). A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*, 29(5):511–524.

OECD (2022). Dark commercial patterns. Technical Report 336, OECD.

Office for National Statistics (2022). UK SIC 2007.

Reed II, A., Forehand, M. R., Puntoni, S., and Warlop, L. (2012). Identity-based consumer behavior. *International Journal of Research in Marketing*, 29(4):310–321.

Reijula, S. and Hertwig, R. (2022). Self-nudging and the citizen choice architect. *Behavioural Public Policy*, 6(1):119–149.

Renze, M. (2024). The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356.

Sugg, O. and Lesic, V. (2022). Online choice architecture – how do we end up making decisions we don't want to make?

Sunstein, C. R. (2021). *Sludge: Bureaucratic burdens and why we should eliminate them.* MIT Press.

Sunstein, C. R. (2022). Sludge audits. *Behavioural Public Policy*, 6(4):654–673.

Sunstein, C. R. (2025). *Manipulation: What it Is, why It's Bad, what to Do about it.* Cambridge University Press.

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

Varazzani, C., Pathak, P., Khan, H., Trudinger, D., Koromilas, E., Galassi, A., and Chan, S. (2024). Fixing frictions: Sludge audits around the world. Oecd public governance policy papers, OECD.

Online Supplementary Material for

ManipulationDetect: An AI Auditing Tool for Online Choice Architecture

Richard Mills[†‡], Stuart Mills[§], Cass R. Sunstein[¶]

December 17, 2025

Contents

[†]Correspondence concerning this manuscript should be addressed to Richard Mills
[‡]University of Nottingham, School of Psychology, United Kingdom, Email: richard.mills2@nottingham.ac.uk.
[§]University of Leeds, School of Economics, United Kingdom, Email: s.mills1@leeds.ac.uk.
[¶]Harvard University, Harvard Law School, United States. Email: csunstein@law.harvard.edu.

The following information will be made available on OSF (link: https://osf.io/shw65/) along with the preregistration document already uploaded before data collection:

1. The code for the auditing tool (for an individual to run in their Chrome browser as an extension), as well as
2. Cleaned audit data and analysis code, with company information redacted.

A OCA Technique Sources and OECD Harms

Table A1: Technique List and OECD Harms Framework to Traffic Light Colour Operationalisation

			OECD. 2022. Dark commercial patterns						
			H1	H2	H3	H4	H5	H6	
Consolidated Techniques	**Source of Definition**	**Severity**	**Autonomy**	**Financial**	**Privacy**	**Psych. & Time**	**Comp.**	**Trust & Eng.**	**Traffic Light Colour**
Sneaking & Hidden Charges									
Sneak-into-Basket	Mathur et al. (2019)	H1, H2, H5	Yes	Yes	No	No	Yes	No	Red
Hidden Costs (Drip Pricing)	Adapted from Mathur et al. (2019)	H1, H2, H5, H6	Yes	Yes	No	No	Yes	Yes	Red
Forced Continuity / Hidden Subscriptions	Adapted from Mathur et al. (2019)	H1, H2, H5	Yes	Yes	No	No	Yes	No	Red
Partitioned pricing	Adapted from Competition and Markets Authority (2021)	H2, H5, H6	No	Yes	No	No	Yes	Yes	Amber
Urgency & Scarcity									
Countdown Timer	Mathur et al. (2019)	H2, H6	No	Yes	No	No	No	Yes	Amber
Limited-Time Message	Mathur et al. (2019)	H2, H6	No	Yes	No	No	No	Yes	Amber
Low-Stock Message	Mathur et al. (2019)	H2, H6	No	Yes	No	No	No	Yes	Amber
High-Demand Message	Mathur et al. (2019)	H2, H6	No	Yes	No	No	No	Yes	Amber
Social Influence									
Activity Messages	Mathur et al. (2019)	H1, H2, H5, H6	Yes	Yes	No	No	Yes	Yes	Amber
Online Reviews	Adapted from Competition and Markets Authority (2021)	H2, H6	No	Yes	No	No	No	Yes	Green
Influencers (Authority bias)	Adapted from Competition and Markets Authority (2021)	H2, H6	No	Yes	No	No	No	Yes	Green
Social Identity	Adapted from Reed II et al. (2012)	H4, H6	No	No	No	Yes	No	Yes	Green
Misdirection & Visual Manipulation									
Visual Interference	Mathur et al. (2019)	H1, H2, H3	Yes	Yes	Yes	No	No	No	Red
False Hierarchy	Li et al. (2024)	H2	No	Yes	No	No	No	No	Red

Continued on next page

Table A1 – continued from previous page

Consolidated Techniques	Source	Severity	H1 Autonomy	H2 Financial	H3 Privacy	H4 Psych.	H5 Comp.	H6 Trust	Colour
Disguised Ads	Adapted from Li et al. (2024)	H2, H4, H6	No	Yes	No	Yes	No	Yes	Red
Complex / Vague Language	Competition and Markets Authority (2021)	H2, H4, H6	No	Yes	No	Yes	No	Yes	Amber
Framing & Language									
Framing	Competition and Markets Authority (2021)	H2, H3, H4	No	Yes	Yes	Yes	No	No	Amber
Confirmshaming	Mathur et al. (2019)	H2	No	Yes	No	No	No	No	Red
Upselling	Adapted from Mathur et al. (2019)	H2	No	Yes	No	No	No	No	Green
Cross-selling	Adapted from Mathur et al. (2019)	H2	No	Yes	No	No	No	No	Green
Pricing & Value Perception									
Reference Pricing	Competition and Markets Authority (2021)	H2, H5, H6	No	Yes	No	No	Yes	Yes	Amber
Decoy Effect	Adapted from Competition and Markets Authority (2021)	H2, H4, H6	No	Yes	No	Yes	No	Yes	Amber
Bundling	Adapted from Competition and Markets Authority (2021)	H1	Yes	No	No	No	No	No	Green
Loss Aversion	Adapted from Competition and Markets Authority (2021)	H2, H4	No	Yes	No	Yes	No	No	Amber
Endowment Effect	Adapted from Kahneman et al. (1990)	H4	No	No	No	Yes	No	No	Green
Sunk Cost Fallacy	Adapted from Competition and Markets Authority (2021)	H4	No	No	No	Yes	No	No	Green
Virtual Currency	Adapted from Competition and Markets Authority (2021)	H1, H2, H5	Yes	Yes	No	No	Yes	No	Red
Frictions									
Creating Friction	Adapted from Sunstein (2021)	H1-H5	Yes	Yes	Yes	Yes	Yes	No	Red
Removing Friction	Adapted from Sunstein (2021)	H2, H3, H4	No	Yes	Yes	Yes	No	No	Red
Forced Actions & Lack of Choice									
Forced Registration / Enrolment	Mathur et al. (2019)	H1, H5	Yes	No	No	No	Yes	No	Red

Continued on next page

Table A1 – continued from previous page

Consolidated Techniques	Source	Severity	H1 Autonomy	H2 Financial	H3 Privacy	H4 Psych.	H5 Comp.	H6 Trust	Colour
Bait and Switch	Competition and Markets Authority (2021)	H4, H6	No	No	No	Yes	No	Yes	Red
Opt-Out Defaults	Adapted from Li et al. (2024)	H2, H3	No	Yes	Yes	No	No	No	Red
Other Techniques									
Choice / Info Overload	Adapted from Competition and Markets Authority (2021)	H4	No	No	No	Yes	No	No	Amber
Nagging	Adapted from Li et al. (2024)	H3, H4, H5	No	No	Yes	Yes	Yes	No	Red
Reciprocity	Adapted from Falk and Fischbacher (2006)	H4	No	No	No	Yes	No	No	Green
Commitment Devices	Adapted from Competition and Markets Authority (2021)	H4	No	No	No	Yes	No	No	Green
Just-In-Time Prompts	Competition and Markets Authority (2021)	H4	No	No	No	Yes	No	No	Green
Goal-Gradient Effect	Adapted from Competition and Markets Authority (2021)	H4	No	No	No	Yes	No	No	Green
Rewards & Punishments	Adapted from Competition and Markets Authority (2021)	H4	No	No	No	Yes	No	No	Green
Order Effects (Ranking)	Competition and Markets Authority (2021)	H2, H4, H6	No	Yes	No	Yes	No	Yes	Amber

Notes: To identify behavioural techniques, ManipulationDetect uses a taxonomy that includes 40 OCA techniques, each defined by a name, a conceptual definition, a real-world example, an initial severity rating (green/amber/red), a tip, and a suggested user response. These initial classifications were generated and iteratively refined based on a broad-engagement with the literature and regulatory guidance. While not yet validated formally, they serve as a transparent, open starting point for discussion and improvement. Moreover, the severity ratings are there to serve as initial guides for the LLM (however, they are flexible).

B Prompt Engineering

B.1 OCA Techniques

The list of techniques below is provided to the LLM, each with a name, definition, examples (or examples), tip, and suggestion. This set framework has specifically been designed such that new techniques can be added. An important feature is also the “tip” section, which provides the model with some guidance on how to find various techniques.

```

export const techniques = [

{

name: "Sneak-into-Basket",
definition: "Adding additional products to users’ shopping carts without their consent.",
example: "Travel insurance auto-added.",
severity: "red",
tip: "This is hard to detect before the final checkout. Look for text implying an item will be
    auto-added, or for pre-ticked checkboxes where the label offers an extra product or
    service like 'gift wrapping' or 'travel insurance'.",
suggestion: "Check your cart - and remove anything you didn’t ask for."

},
{

name: "Hidden Costs (Drip Pricing)",
definition: "Revealing extra charges gradually during checkout.",
example: "“Tickets £25!” but final price higher at checkout, or “Each month, prices rise by £X
    (often noted with an asterisk (∗ or †)“",
severity: "red",
tip: "Look for a prominent price that seems incomplete. Scan for asterisks (∗ or †) next to
    prices or text like 'plus fees', 'excl. taxes', or 'service charge not included', which
    indicate more costs will be revealed on a later screen.",
suggestion: "Always check the ∗final∗ price before clicking ‘Buy’."

},
{

name: "Forced Continuity / Hidden Subscriptions",
definition: "Enrolling users in a recurring payment plan, such as an automatic renewal, after
    an initial purchase or free trial. This is especially manipulative when the renewal terms
    are not clearly and prominently displayed.",
example: "“Free trial auto-renews at £9.99/month,” or a product purchase that states “(
    automatic renewal)” in smaller print.",
severity: "red",
tip: "Scan the image for text with keywords like 'auto-renews', 'recurring', 'subscription', '
    monthly plan', or '/mo' and '/yr'. Pay close attention when these appear near offers for a
     'free trial' or a low introductory price.",
suggestion: "Cancel ∗before∗ the trial ends - set a calendar reminder now."

},
```

```
{

name: "Partitioned Pricing",
definition: "Separating the total cost into a base price and additional or implied fees, all shown together. Often used to make a high total seem smaller.",
example: "“Flights from £136*” - but bags (£25.99), seat selection (£4.99), and boarding (£5.99) add another £37.",
severity: "amber",
tip: "Look for multiple prices being displayed together that make up a total cost. This is often a main 'base price' accompanied by separate line items like 'booking fee', 'service charge', 'taxes', or 'delivery'. The key is that the separate parts of the price are all visible at once.",
suggestion: "Add it all up! These “small extras” can quietly inflate the final cost."

},
{

name: "Countdown Timer",
definition: "Indicating to users that a deal or discount will expire using a counting-down timer.",
example: "“Flash sale ends in 2 hours - shop now!”",
severity: "amber",
tip: "Scan for text that combines numbers with time-based keywords like 'ends in', 'offer expires', 'hours', 'minutes', or 'seconds'. Look for formats like HH:MM:SS or text that explicitly mentions a timer.",
suggestion: "Don’t rush just because the clock’s ticking - take a moment to compare prices first."

},
{

name: "Limited-Time Message",
definition: "Indicating to users that a deal or sale will expire soon without specifying a deadline.",
example: "“Sale ends soon!”",
severity: "amber",
tip: "Look for phrases creating urgency that don't include a specific number or timeframe. Scan for keywords like 'ends soon', 'last chance', 'limited time only', or 'don't miss out'. This is different from a Countdown Timer, which is specific.",
suggestion: "“Limited-time” doesn’t always mean limited. Only buy it if you’d want it anyway."

},
{

name: "Low-Stock Message",
definition: "Indicating to users that limited quantities of a product are available, increasing its desirability.",
example: "“5 left in stock!”, “Only 1 left available at this price”",
severity: "amber",
tip: "Scan for text that combines a low number (e.g., under 10) with keywords like 'left in stock', 'only X available', or 'few remaining'. This focuses on the scarcity of the product inventory itself.",
```

```
suggestion: "Low stock might be real - or just marketing. Don’t let it pressure you into a bad deal."

},
{

name: "High-Demand Message",
definition: "Indicating to users that a product is in high demand and likely to sell out soon, increasing its desirability.",
example: "“14 people have added this to their bag in the last 24 hours!”",
severity: "amber",
tip: "Look for social proof that implies scarcity due to popularity. Scan for phrases like 'in X people's carts', 'X people bought this recently', 'selling fast', or badges like 'Popular Pick'.",
suggestion: "Popularity does not always equate to quality - check the reviews, not just the hype."

},
{

name: "Activity Messages",
definition: "Informing the user about other users' activity on the website (e.g., purchases, views, visits).",
example: "”Abigail from Michigan just bought a new stereo system”, ”35 people added this item to cart”, ”90 people have viewed this product”",
severity: "amber",
tip: "Look for specific, real-time social proof. Scan for patterns like 'Someone in [Location] just bought...', 'Purchased X minutes ago', or 'X people are viewing this right now'. This shows current user activity.",
suggestion: "Just because others are buying doesn’t mean it’s right for you. Decide based on your needs - not theirs."

},
{

name: "Online Reviews",
definition: "Displaying reviews from other consumers to influence a user's decision-making process through social proof.",
example: "85% of visitors rated this 5 stars” encourages conformity.",
severity: "green",
tip: "Scan for star symbols (e.g., ⋆⋆⋆⋆⋆), rating formats like '4.5/5', or keywords like 'reviews', 'ratings', or 'customer score'. Also check the `badges` field for review-related information.",
suggestion: "Others’ choices aren’t always ∗your∗ best choice. It is sometimes important to decide independently."

},
{

name: "Influencers (Authority Bias)",
definition: "Featuring influencers or public figures promoting products on-site, often without clear disclosure that the endorsement is paid or sponsored.",
```

```
example: "“Top Picks from @StyledBySophie” - no #ad or disclosure.",
severity: "green",
tip: "Scan for social media handles (e.g., text starting with '@'), celebrity names, or phrases like '[Name]'s Picks', 'As seen on', or 'In collaboration with'. Look for signs of a personal endorsement being used to promote a product.",
suggestion: "If it looks like a personal opinion, check whether it’s really an ad. Look for small-print labels like 'sponsored' or 'paid content'."

},
{

name: "Social Identity",
definition: "Appealing to a user's values or aspirations to make them feel that a purchase will reinforce a desirable identity.",
example: "“Join other eco-conscious drivers and make the switch.”, “Smart shoppers choose Brand X.”",
severity: "green",
tip: "Look for aspirational language that groups users into a desirable category. Scan for phrases like 'For the serious...', 'Smart shoppers choose...', 'Join other eco-conscious ...', or text that appeals to a user's values, status, or lifestyle.",
suggestion: "You define your values - don’t let brands do it for you."

},
{

name: "Visual Interference",
definition: "Using style and visual presentation to steer users to or away from certain choices.",
example: "Two buttons, one brightly-coloured “YES (I do want to hear about exclusive offers and discounts” button, while the other button is greyed out (“NO, I'd  rather not hear about exclusive offers and discounts”).",
severity: "red",
tip: "Visually analyse choices with a **company-preferred action** (e.g., 'Accept All', 'Yes to Marketing'). Look for the preferred button being brightly coloured or larger, while the **alternative, dispreferred choice** (e.g., 'no thanks', 'continue as guest', 'Reject All') is **visually suppressed** as plain text or a low-contrast button. If the prominent button is for a standard, non-persuasive action (e.g., 'Add to cart', 'View Cart', 'Check-out', 'Sign-up' etc), it is **not** this technique.",
suggestion: "Don’t be tricked by colour or bold text - read all options carefully, even the less prominent ones."

},
{

name: "False Hierarchy",
definition: "Giving one option visual precedence to convince the user it is the best or only choice.",
example: "“Expensive option labelled 'Most Popular' despite no evidence.“, “Recommended Plan“ or “Best Value” highlighted without explanation.“ , “Free/basic option buried under a dropdown or shown in small print.“",
severity: "red",
tip: "Compare the `styles` objects across multiple, similar cards (like pricing plans). Look
```

```
for one card that has a visually distinct style, such as a different `backgroundColor`, a thicker `border`, a prominent `boxShadow`, or a badge that says 'Recommended' or 'Most Popular'.",
suggestion: "Don’t assume “Recommended” means best for you - always compare features and terms."
},
{
name: "Disguised Ads",
definition: "Ads styled to look like articles, navigation links, or organic site content to mislead clicks.",
example: "“Sponsored content” under a news-style headline, or ad blocks labelled ‘You might also like’.",
severity: "red",
tip: "Look for content that mimics the site's own editorial style but is an advertisement. Scan for subtle keywords like 'Sponsored', 'Promoted', 'Ad', or 'From Our Partners' that lead to product pages.",
suggestion: "If it looks like a recommendation, hover or check for a ‘Sponsored’ label - it’s probably an ad."
},
{
name: "Complex / Vague Language",
definition: "Using obscure words or confusing sentence structures to make information difficult to understand.",
example: "“We’d love to send you emails... but if you do not wish to receive these updates, please tick this box.” , long unreadable privacy notices.",
severity: "amber",
tip: "Scan for confusing phrasing, especially in checkbox labels or fine print. Look for double negatives (e.g., 'Do not uncheck this box to opt out'), technical jargon, or unnecessarily long sentences designed to make comprehension difficult.",
suggestion: "Don’t skim - some boxes are worded in reverse. Read carefully before clicking."
},
{
name: "Framing",
definition: "Describing or presenting information in a way that influences how it is perceived.",
example: "“Don’t miss these savings!” vs. “Sign up to save.”",
severity: "amber",
tip: "Look for emotionally loaded language that presents a choice in a biased way. Scan for persuasive phrases that emphasise a gain or loss, such as 'Don't miss out on savings!' or 'Unlock your full potential', rather than neutral, factual descriptions.",
suggestion: "Reframe it yourself - would you still act if it was worded differently?"
},
{
name: "Confirmshaming",
```

```
definition: "Using language to evoke shame and steer users away from a certain choice.",
example: "“No thanks, I don't want to save money.”",
severity: "red",
tip: "Focus specifically on the text for opt-out links or buttons. Look for dismissive phrases that are reworded to make the user feel foolish, such as 'No thanks, I like paying full price' or 'I don't want to save money'.",
suggestion: "Don’t fall for emotional manipulation. It's just guilt-trip wording."

},
{

name: "Upselling",
definition: "Tactics that steer users into purchasing a more expensive version of a product.",
example: "“Upgrade to Premium for just £5 more!” ,“Want more features? Choose Pro instead of Basic.”",
severity: "green",
tip: "This is a relational nudge. Visually compare similar product sections and look for language encouraging a move to a higher tier. Scan for keywords like 'Upgrade to Premium', 'Most Popular Plan', or phrases like 'For just £X more...'.",
suggestion: "You may not need the upgrade - check what’s actually included before paying more."

},
{

name: "Cross-selling",
definition: "Tactics that steer users into purchasing additional, related products.",
example: "“Customers also bought...” , “Add travel insurance to your booking for £8.” , “Other items perfect for you“",
severity: "green",
tip: "Look for suggestions to add different, complementary items to a purchase. Scan for headings like 'Customers also bought', 'Frequently bought together', or specific prompts like 'Add travel insurance for £X'.",
suggestion: "Only add extras you really need - skip the ones that don’t add value for you."

},
{

name: "Reference Pricing",
definition: "Displaying a previous or future price alongside the current price to make the current price look more attractive.",
example: "“Was £120, now just £79.”",
severity: "amber",
tip: "Look for a strikethrough price next to a current price. Also scan for keywords like 'was ', 'now', 'RRP', 'save', or the '%' symbol next to a price to indicate a discount.",
suggestion: "Ignore the ‘original’ price-judge if the final one is truly fair."

},
{

name: "Decoy Effect",
definition: "Adding a third, less appealing option to influence the perception of the original two choices.",
```

```
example: "“Phone A - £25, 32 GB | 8 MP | 1-yr/Phone B - £50, 64 GB | 12 MP | 2-yr(Target)/Phone C - £105, 256 GB | 48 MP | 3-yr(Extreme decoy - makes B look like the sensible compromise)“, “Phone A - £25, 32 GB | 8 MP | 1-yr/Phone B - £50, 64 GB | 8 MP | 6-mo(Inferior decoy - strictly worse on storage, camera & warranty)/Phone C - £60, 128 GB | 12 MP | 2-yr(Target)“",
severity: "amber",
tip: "This is a relational nudge. Visually compare three or more similar options (like pricing plans) and look for an 'asymmetrically dominated' choice-one that is clearly worse value than another (e.g., lower performance for the same price). If decoy is found, only display the 'quote' for the extreme or inferior decoy.",
suggestion: "Ignore the odd one out-pick the option that genuinely fits your needs and budget."

},
{

name: "Bundling",
definition: "Grouping two or more products or services into a single “package” at a special price.",
example: "“Lipstick: £15” vs. “Makeup Set: £25 (includes lipstick + mascara + blush)”",
severity: "green",
tip: "Look for offers that combine multiple items into one purchase. Scan for keywords like 'bundle', 'package', 'set', 'kit', or text that joins items with '+' or '&' (e.g., 'Phone + Case').",
suggestion: "Don’t assume the bundle is better value-check if you actually need everything in it."

},
{

name: "Loss Aversion",
definition: "Using the fact that people fear losses more than they value equivalent gains to steer decisions.",
example: "“Avoid the £5 cancellation fee - switch now!“, “Buy insurance so that you do not lose £400“",
severity: "amber",
tip: "Scan for language that frames a choice in terms of avoiding a negative outcome. Look for keywords like 'avoid', 'don't lose', 'prevent', or text that mentions a 'fee', 'penalty' or 'extra charge' for inaction.",
suggestion: "Check if you’re driven by fear of loss more than real benefit."

},
{

name: "Endowment Effect",
definition: "Using the fact that people value something more highly once they feel a sense of ownership.",
example: "“Enjoy your 7-day free trial”, “Your saved items are still waiting”, “Continue with your personalised plan”",
severity: "green",
tip: "Look for possessive language that creates a premature sense of ownership. Scan for words like 'your' or 'my' (e.g., 'Your saved items'), or phrases like 'Enjoy your free trial' or 'Claim your gift'.",
```

```
suggestion: "Just because it feels like it’s yours doesn’t mean it is-free trials can be a tactic to get you attached."

},
{

name: "Sunk Cost Fallacy",
definition: "Exploiting the tendency to continue with an endeavour because of previously invested resources (time, money).",
example: "“You’ve already used half your membership - don’t waste it!”, “You’re almost finished - complete your course!” , “You’ve already spent £60 this month - get the most from it!”" ,
severity: "green",
tip: "Scan for language that references a user's past investment (time, money, or effort) to encourage them to continue. Look for phrases like 'You've already spent...', 'Don't waste your progress', or 'You're almost finished'.",
suggestion: "Don’t throw good money (or time) after bad-re-evaluate if it’s still worthwhile."

},
{

name: "Virtual Currency",
definition: "Replacing real money with virtual points or tokens to obscure true costs and encourage spending.",
example: "“Top up 500 credits for £4.99”, “Only 120 coins to unlock premium access!”",
severity: "amber",
tip: "Scan for non-standard currency names being used as a price. Look for keywords like 'coins', 'gems', 'credits', 'points', or 'tokens' instead of real currency symbols (£, $, €).",
suggestion: "Always convert virtual currency back into real money in your head before spending."

},
{

name: "Creating Friction",
definition: "Deliberately making a process more difficult (sludge) to discourage an action not in the company's interest.",
example: "“Having to untick over 20 'legitimate interest' cookies with no way to just reject all“, “To close your account, please call our customer service line.“, “Claim your £50 cashback by printing this form and mailing it with the original receipt.“",
severity: "red",
tip: "This is difficult to detect from one page. Scan the text for instructions that describe a cumbersome process for an undesirable action (like cancelling or returning). Look for phrases like 'To cancel, you must call...', 'visit a store', or 'fill out this form'.",
suggestion: "If you're sure, push through-this friction is designed to make you give up."

},
{

name: "Removing Friction",
definition: "Making a harmful or costly choice deceptively easy, exploiting the path of least resistance.",
```

```
example: "“Buy now with 1-Click“ (bypassing a final cart review), “Your premium trial starts automatically after creating an account.“",
severity: "red",
tip: "Look for buttons with text that implies an immediate purchase and bypasses review steps. Scan for keywords like 'Buy now with 1-Click', 'Instant Checkout', or 'Quick Buy'.",
suggestion: "Ask yourself if you're buying just because it's easy, or if you'd still want it after a moment's thought."

},
{

name: "Forced Registration",
definition: "Forcing users to create an account or share information to complete basic tasks.",
example: "“Continue to checkout” is disabled until you create an account - no guest checkout option.“, “Streaming services, rental agencies, credit providers that force you to register before showing available options“",
severity: "red",
tip: "Look for signs that a guest checkout is unavailable. Scan for text like 'You must create an account to purchase' or a 'Continue as Guest' option is missing.",
suggestion: "Ask yourself whether it’s worth giving away your data-not just your cash."

},
{

name: "Bait and Switch",
definition: "Advertising a desirable product to lure a user in, then switching it for a more expensive or inferior alternative.",
example: "“A retailer advertises a laptop “from £499“ - but that model is out of stock or unavailable, and only more expensive models can be purchased“, “Clicking “No, thanks“ to dismiss a pop-up still results in the user being signed up or redirected (e.g., “X” closing the box actually means “accept”).“",
severity: "red",
tip: "Look for text indicating an advertised offer is now unavailable, especially when presented alongside more expensive options. Critically, distinguish this from a simple ' out of stock' message with helpful substitutes. Bait and Switch implies a deceptive pattern, not just a logistical issue. Scan for phrases like 'Offer expired', 'Promotion ended', or a suspiciously unavailable low price.",
suggestion: "If the deal changes once you click-stop. Check if it’s still what you were promised."

},
{

name: "Opt-Out Defaults",
definition: "Using a pre-ticked checkbox or other pre-selected default choice to automatically include a user in a service or data collection process, requiring them to manually unselect it (opt out).",
example: "“Your plan includes auto-renewal (uncheck to disable),” or ”a pre-checked box stating [✓] Yes, sign me up for special offers”",
severity: "red",
tip: "Look for a pre-ticked checkbox or a pre-selected radio button. This is a strong indicator ONLY IF its label text suggests the user is being signed up for something extra (e.g.,
```

```
newsletters, marketing, insurance), or opening another link ('Open places to stay').",
suggestion: "Always look for pre-checked boxes and uncheck them if you're unsure."

},
{

name: "Choice Overload",
definition: "Presenting too many options to overwhelm the user, making a considered decision difficult.",
example: "A comparison site lists 150+ broadband plans with no clear summary or filter.",
severity: "amber",
tip: "Analyse the total number of similar, competing choices presented at once. If there is a very large number (e.g., more than 36) of product or plan options visible on the screen without clear filtering, it may indicate this pattern.",
suggestion: "Too many choices? Step back, filter by what matters most, and don’t let overload lead to a bad decision."

},
{

name: "Nagging",
definition: "Repeated prompts or pop-ups that interrupt the user’s task flow to push unrelated actions.",
example: "“Enjoying the app? Leave a review!”-triggered after every login, even when dismissed.",
severity: "red",
tip: "The best clue is an element that looks like a pop-up or modal window and uses language suggesting a repeated request, like 'Are you sure?' or 'Enjoying the app?'",
suggestion: "If it's pushing too hard, it's probably not worth your time."

},
{

name: "Reciprocity",
definition: "Offering a free gift or resource to create a feeling of indebtedness, encouraging the user to 'repay' the favour.",
example: "“Here’s a free guide-consider joining today.”, “Enjoy 10% off your first order-just tell us your email.”",
severity: "green",
tip: "Look for offers of something 'free' (e.g., 'free guide', 'free e-book', 'free gift', or a discount) that are conditional on the user giving something in return, such as their email address ('in exchange for your email').",
suggestion: "It’s okay to accept a gift, but only give back if it feels right for you."

},
{

name: "Commitment Devices",
definition: "Nudging users to commit to a future action that may be hard to reverse or costly to break.",
example: "“Auto-renew and lock in today’s rate.” (but the price may increase later).",
severity: "green",
```

```
tip: "Look for options that lock a user into a future agreement. Scan for keywords like 'auto-renew', 'lock in this rate', 'pre-order', or 'subscribe and save' that imply a long-term or recurring arrangement.",
suggestion: "If commitment comes with fine print or effort to cancel, set a reminder-or think twice."
},
{
name: "Just-In-Time Prompts / Reminders",
definition: "Prompts or reminders aimed at grabbing attention to trigger specific, often urgent, behaviours.",
example: "“Hurry! Only 2 hours left to complete your order.”, “Still interested? Check out now before they’re gone.”",
severity: "green",
tip: "Scan for text that reminds the user about an incomplete action. Look for phrases like 'Still interested?', 'Complete your order', 'Forgot something?', or 'Your cart is about to expire'.",
suggestion: "Pause before clicking. These alerts feel urgent by design-your decision doesn’t have to be."
},
{
name: "Goal-Gradient Effect",
definition: "Exploiting the tendency to increase effort as a goal gets closer, often using progress bars.",
example: "“You’re 80% of the way to earning a bonus-just one more purchase!”",
severity: "green",
tip: "Look for visual or textual indicators of progress. Scan for phrases like 'You're almost there', 'Just one step left', text containing percentages (e.g., '80% complete'), or mentions of a 'progress bar'.",
suggestion: "Ask if finishing truly benefits you, or if you’re just compelled by being ‘almost done’."
},
{
name: "Rewards & Punishments",
definition: "Using positive or negative incentives to influence behaviour.",
example: "“Early-bird discount” vs. “£10 late renewal fee.”, “If you cancel, you will lose your Gold member status.”",
severity: "green",
tip: "Scan for language offering a clear incentive (reward) or disincentive (punishment). Look for rewards like 'early-bird discount' or 'bonus points', and punishments like 'late fee', 'penalty', or 'lose your status'.",
suggestion: "Focus on the final outcome. Whether it's a discount or a penalty, is the service a good value for you at that final price?"
},
{
```

```
name: "Order Effects (Ranking)",
definition: "Displaying options in a particular order to influence choice.",
example: "On a flight site, the top result is listed as “Our Top Pick” but is actually a
    sponsored placement.",
severity: "amber",
tip: "This is a relational nudge. Look for badges or labels on the first few items in a list
    that are not on others. Scan for keywords like 'Our Top Pick', 'Recommended', 'Best Seller
    ', or 'Sponsored' applied to items at the top of a list.",
suggestion: "The default order is rarely the best for you. Actively re-sort the list by 'Price'
     or 'Rating'."

}
];
```

B.2 LLM Instructions

```

<ROLE>
You are an expert in behavioural science and identifying Online Choice Architecture (OCA) techniques.
Your sole purpose is to analyse the provided webpage SCREENSHOT to find evidence of the techniques listed in the reference section.
</ROLE>

<PRIMARY_GOAL>
Identify OCA techniques used to pressure users into a purchase, subscription, or data submission.
</PRIMARY_GOAL>

<WORKFLOW>
1. Analyse the provided webpage SCREENSHOT. Identify visual and textual evidence of the techniques listed in the reference section.
2. For each technique you find, you MUST transcribe the exact text from the image that demonstrates the pattern.
3. Group and Consolidate: For each technique you identify, find and list **up to a maximum of 15 distinct instances**. If more than 15 exist, provide the first 15 you find. This is a strict limit.
4. Construct a JSON object for each technique, using the format in <OUTPUT_FORMAT>.
5. Sort and Limit: Sort by severity (red > amber > green) and return a maximum of five (5) techniques.
6. If the webpage does not appear to be for buying, subscribing, or collecting personal data, return an empty array: [].
</WORKFLOW>

<OUTPUT_FORMAT>
Return ONLY a valid JSON array of objects:

[
  {
    "name": "Reference Pricing",
    "severity": "amber",
    "justification": "The page displays several instances of a crossed-out higher price ('Was £120') next to the current price ('now just £79')",
    "instances": [
      { "quote": "Was £120, now just £79"},
      { "quote": "Was £200, now just £149" }
    ]
  }
]
</OUTPUT_FORMAT>

<TECHNIQUES_REFERENCE>
${techniqueText}
</TECHNIQUES_REFERENCE>

<EVALUATION_CRITERIA>
```

```
- Only report independent techniques.
- Validate carefully against the definitions.
</EVALUATION_CRITERIA>

<OTHER IMPORTANT INFORMATION>
- Use non-definitive language: these are potential patterns only. Use phrases like “this may
    indicate...”, “could reflect...”, or “appears to be...” in the justification.
- Use the rating given in the reference unless the context on the page makes it clearly more or
     less severe (you may Escalate an AMBER to RED, or De-Escalate an AMBER to GREEN). The RED
     rating is absolute. If a technique is marked RED in the reference list, you MUST ALWAYS
    output it as RED. This rule cannot be overridden.
- Use British English Spelling.
- Human-Readable Justification: The 'justification' text must be a user-friendly summary. **Do
    not** list every single quote in this field; that raw data belongs in the 'instances'
    array.
- For the 'instances' array, provide a quote for each **distinct and separate occurrence** of
    the technique on the page. A single pricing table or product card that uses a technique
    counts as **one instance**, even if it contains multiple pieces of related text. Do not
    list multiple quotes from the same single element; choose the most representative text.
- Do not reuse quotes across different techniques.
</OTHER IMPORTANT INFORMATION>
```

C Preliminary Audit

C.1 ManipulationDetect Audit Version

(a) Homepage I (Non-active Session)

(b) Homepage II (Active Session)

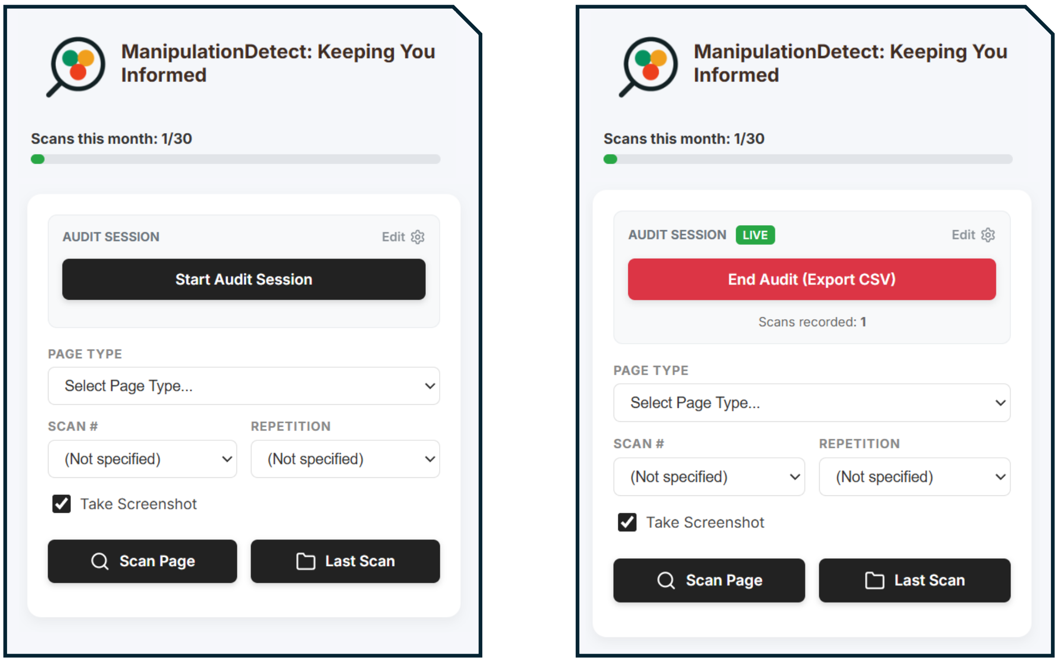


Figure C1: ManipulationDetect Audit Version Interface.

Note. Panel (a) displays the initial session controls used to tag metadata (Page Type, Scan #, Repetition #) and the 'Start Audit Session' button. Panel (b) shows the interface during an active session, displaying the 'LIVE' session indicator, the 'Scans recorded' counter, and the 'End Audit' button which triggers the CSV export. The 'Edit' button (visible in both panels) allows the auditor to add or remove custom metadata tags (i.e., removing Landing Page).

C.2 Summary Statistics

Table C2: Summary Statistics of Page Type and SIC (2007) Classifications

	Summary
N (%)	**1,052**
Page Type	
Landing Page	122 (11.6%)
Category Page	358 (34.0%)
Product Page	358 (34.0%)
Basket	214 (20.3%)
SIC (2007) Section	
Section G (Wholesale and retail trade; repair of motor vehicles and motorcycles)	450 (42.8%)
Section H (Transportation and storage)	110 (10.5%)
Section J (Information and communication)	188 (17.9%)
Section K (Financial and insurance activities)	158 (15.0%)
Section N (Administrative and support service activities)	146 (13.9%)

C.3 Landscape of OCA in the UK

Table C3: Technique Prevalence by Page Type and Fisher's Exact Test Results.

Technique Name	A. Landing Page	B. Category Page	C. Product Page	D. Basket /Cart	Total	Fisher's Exact p
Activity Messages	0.016	0.061	0.039	0.028	0.042	0.434
Bundling	0.082	0.073	0.050	0.019	0.055	0.157
Choice Overload	0.000	0.011	0.011	0.000	0.008	0.806
Commitment Devices	0.016	0.017	0.022	0.009	0.017	0.961
Complex / Vague Language	0.082	0.067	0.117	0.140	0.101	0.178
Countdown Timer	0.066	0.034	0.039	0.065	0.046	0.488
Creating Friction	0.000	0.000	0.000	0.009	0.002	0.319
Cross-selling	0.033	0.028	0.212	0.168	0.120	0.000***
Decoy Effect	0.000	0.000	0.000	0.009	0.002	0.319
Disguised Ads	0.049	0.084	0.039	0.028	0.053	0.172
Endowment Effect	0.049	0.028	0.034	0.056	0.038	0.579
False Hierarchy	0.000	0.162	0.134	0.075	0.116	0.001***
Forced Continuity / Hidden Subscriptions	0.016	0.028	0.039	0.047	0.034	0.741
Forced Registration	0.000	0.000	0.006	0.047	0.011	0.004***
Framing	0.754	0.475	0.447	0.327	0.468	0.000***
Goal-Gradient Effect	0.000	0.000	0.006	0.103	0.023	0.000***
Hidden Costs (Drip Pricing)	0.115	0.084	0.117	0.178	0.118	0.137
High-Demand Message	0.049	0.101	0.117	0.178	0.116	0.077
Influencers (Authority Bias)	0.049	0.006	0.011	0.000	0.011	0.049***
Just-In-Time Prompts	0.000	0.000	0.006	0.000	0.002	1.000
Limited-Time Message	0.213	0.089	0.106	0.150	0.122	0.059
Loss Aversion	0.115	0.045	0.061	0.131	0.076	0.031***
Low-Stock Message	0.016	0.095	0.028	0.131	0.070	0.001***
Online Reviews	0.164	0.341	0.318	0.065	0.257	0.000***
Opt-Out Defaults	0.000	0.000	0.011	0.047	0.013	0.012**
Order Effects (Ranking)	0.000	0.101	0.022	0.028	0.048	0.001***
Partitioned Pricing	0.033	0.039	0.156	0.271	0.125	0.000***
Reciprocity	0.098	0.034	0.039	0.037	0.044	0.228
Reference Pricing	0.295	0.436	0.358	0.364	0.378	0.199
Rewards & Punishments	0.115	0.084	0.061	0.103	0.084	0.439
Sneak-into-Basket	0.000	0.000	0.017	0.000	0.006	0.137
Social Identity	0.131	0.061	0.061	0.009	0.059	0.010**
Upselling	0.049	0.056	0.056	0.056	0.055	1.000
Virtual Currency	0.000	0.000	0.028	0.047	0.019	0.011**
Visual Interference	0.049	0.045	0.067	0.028	0.049	0.513
Total	**0.076**	**0.077**	**0.081**	**0.084**	**0.080**	

Notes: ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

Table C4: Technique Prevalence by SIC (2007) and Fisher's Exact Test Results.

Technique Name	**Sec G** Wholesale & Retail	**Sec H** Transport & Storage	**Sec J** Info & Comm.	**Sec K** Finance & Insur.	**Sec N** Admin & Supp.	**Total**	**Fisher's Exact p**
Activity Messages	0.058	0.018	0.043	0.000	0.055	0.042	0.152
Bundling	0.062	0.091	0.064	0.025	0.027	0.055	0.399
Choice Overload	0.000	0.018	0.000	0.013	0.027	0.008	0.045**
Commitment Devices	0.004	0.000	0.074	0.013	0.000	0.017	0.001***
Complex / Vague Lang.	0.044	0.109	0.064	0.354	0.041	0.101	0.000***
Countdown Timer	0.044	0.018	0.053	0.051	0.055	0.046	0.872
Creating Friction	0.000	0.018	0.000	0.000	0.000	0.002	0.105
Cross-selling	0.213	0.036	0.074	0.025	0.055	0.120	0.000***
Decoy Effect	0.000	0.018	0.000	0.000	0.000	0.002	0.105
Disguised Ads	0.076	0.000	0.032	0.013	0.096	0.053	0.017**
Endowment Effect	0.031	0.036	0.053	0.051	0.027	0.038	0.809
False Hierarchy	0.044	0.218	0.138	0.152	0.192	0.116	0.000***
Forced Continuity / Hidden Subs.	0.004	0.000	0.149	0.038	0.000	0.034	0.000***
Forced Registration	0.009	0.000	0.011	0.000	0.041	0.011	0.169
Framing	0.400	0.527	0.489	0.747	0.301	0.468	0.000***
Goal-Gradient Effect	0.000	0.055	0.032	0.000	0.082	0.023	0.000***
Hidden Costs (Drip)	0.062	0.164	0.181	0.190	0.096	0.118	0.002***
High-Demand Msg	0.200	0.000	0.128	0.013	0.041	0.116	0.000***
Influencers	0.018	0.000	0.011	0.013	0.000	0.011	0.964
Just-In-Time Prompts	0.000	0.000	0.000	0.013	0.000	0.002	0.394
Limited-Time Msg	0.231	0.091	0.011	0.013	0.068	0.122	0.000***
Loss Aversion	0.036	0.182	0.043	0.165	0.068	0.076	0.000***
Low-Stock Message	0.076	0.127	0.032	0.000	0.137	0.070	0.001***
Online Reviews	0.320	0.145	0.128	0.165	0.411	0.257	0.000***
Opt-Out Defaults	0.018	0.018	0.011	0.000	0.014	0.013	0.886
Order Effects (Ranking)	0.040	0.127	0.021	0.025	0.068	0.048	0.045
Partitioned Pricing	0.071	0.309	0.149	0.101	0.151	0.125	0.000***
Reciprocity	0.067	0.000	0.074	0.013	0.000	0.044	0.008***
Reference Pricing	0.604	0.200	0.319	0.063	0.233	0.378	0.000***
Rewards & Punish.	0.089	0.036	0.043	0.177	0.055	0.084	0.018
Sneak-into-Basket	0.013	0.000	0.000	0.000	0.000	0.006	0.819
Social Identity	0.022	0.018	0.053	0.228	0.027	0.059	0.000***
Upselling	0.004	0.164	0.106	0.051	0.068	0.055	0.000***
Virtual Currency	0.027	0.055	0.011	0.000	0.000	0.019	0.117
Visual Interference	0.018	0.055	0.138	0.076	0.000	0.049	0.000***
Total	**0.083**	**0.082**	**0.078**	**0.080**	**0.070**	**0.080**	

Notes: ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

C.4 Manipulation Risk

Table C5: OLS and Mixed-Effects Regressions on Manipulation Risk

	(1) OLS		(2) Mixed (random intercept by domain)	
Variable	***b***	**(SE)**	***b***	**(SE)**
Page Type				
(ref = Landing Page)				
Category Page	0.868***	(0.302)	0.873***	(0.292)
Product Page	0.151	(0.337)	0.127	(0.318)
Basket	0.149	(0.478)	0.111	(0.438)
SIC Section				
(ref = Section G)				
Section H (Transport)	0.034	(0.262)	-0.046	(0.274)
Section J (Info/Comm)	-0.270	(0.386)	-0.323	(0.409)
Section K (Finance)	-0.081	(0.320)	-0.125	(0.323)
Section N (Admin)	-0.663**	(0.315)	-0.673**	(0.318)
Controls				
Image size (std)	0.430***	(0.204)	0.393***	(0.178)
Constant	4.866***	(0.326)	4.909***	(0.332)
Random Effects				
Var(intercept)			0.588	(0.196)
Var(residual)			2.717	(0.265)
Model Statistics				
N	526		526	
Clusters	60 domains		60 domains	
Model Fit	$R^2 = 0.117$ $F(8, 59) = 4.530$ Prob $> F = 0.000$		Log likelihood $= -1040.997$ $\chi^2(8) = 37.440$ Prob $> \chi^2 = 0.000$	

Notes: Model (1) estimates a standard Ordinary Least Squares (OLS) regression with standard errors clustered at the domain level. Model (2) re-estimates the relationship using a mixed-effects specification with a random intercept for each domain, accounting for unobserved between-site heterogeneity. The dependent variable is the average conservative intersection risk score. Robust standard errors in parentheses.
***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

C.5 Reliability of the Tool

Table C6: Jaccard Index by Technique (Reliability of the Tool)

Technique Name	**Both**	**Either**	**Jaccard Index**
Removing Friction	0	15	0.000
Bait and Switch	0	3	0.000
Sunk Cost Fallacy	0	1	0.000
Just-In-Time Prompts / Reminders	1	15	0.067
Decoy Effect	1	13	0.077
Choice Overload	4	20	0.200
Creating Friction	1	4	0.250
Commitment Devices	9	33	0.273
Rewards & Punishments	44	158	0.278
Forced Registration	6	21	0.286
Order Effects (Ranking)	25	84	0.298
Sneak-into-Basket	3	10	0.300
Endowment Effect	20	66	0.303
Loss Aversion	40	125	0.320
Reciprocity	23	69	0.333
Influencers (Authority Bias)	6	18	0.333
Opt-Out Defaults	7	21	0.333
Virtual Currency	10	29	0.345
Visual Interference	26	73	0.356
Social Identity	31	87	0.356
Bundling	29	80	0.363
Activity Messages	22	56	0.393
Partitioned Pricing	66	159	0.415
Forced Continuity / Hidden Subscriptions	18	41	0.439
Complex / Vague Language	53	119	0.445
Upselling	29	65	0.446
Cross-selling	63	136	0.463
Hidden Costs (Drip Pricing)	62	131	0.473
Goal-Gradient Effect	12	25	0.480
Limited-Time Message	64	126	0.508
Disguised Ads	28	54	0.519
False Hierarchy	61	115	0.530
Countdown Timer	24	44	0.545
Framing	246	406	0.606
High-Demand Message	61	99	0.616

Continued on next page...

Table C6: (Continued) Jaccard Index by Technique

Technique Name	**Both**	**Either**	**Jaccard Index**
Online Reviews	135	189	0.714
Low-Stock Message	37	50	0.740
Reference Pricing	199	231	0.861

Notes: The Jaccard Index is used here to measure the inter-scan reliability of the tool. It quantifies the consistency of agreement between two independent scans (Scan 1 and Scan 2) for each technique on the same set of webpages. **Both** represents the intersection ($|S_1 \cap S_2|$): where both Scan 1 and Scan 2 identified the technique. **Either** represents the union ($|S_1 \cup S_2|$): where at least one of the two scans identified the technique. The Jaccard Index is the ratio of agreement, calculated with the formula: $J(A, B) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$ (in words, the number of techniques that are in both Scan 1 and Scan 2 divided by the number of techniques that are in either Scan 1 or Scan 2).