Autoplay on Trial: An Online Experiment on the Suspect Behind Excessive Screen Time

Reha Tuncer*

November 11, 2023

Abstract

Interface design can be strategically employed to 'nudge' consumers to take certain actions and is suspected of promoting addictive online behavior. A prevalent interface element across numerous popular social media and streaming platforms is the default autoplay feature. In this study, we present an incentivized online experiment to investigate whether the autoplay feature alone can cause an unwarranted increase in desired viewing times. While controlling for content, we demonstrate that the autoplay feature, in isolation, does not override participants' previously indicated preferences for media consumption. We discuss the our findings in terms of what actually renders online platforms addictive. Our results have direct policy implications for the proposed blanket ban on autoplay by the US Congress (SMART Act).

1 Introduction

The rise of social media and streaming platforms have generated discussion around potential societal benefits and concerns about harms such as addiction and self-control issues. A number of recent economic studies have found convincing evidence that people spend significantly more time than they would like to on their smartphones in general (Hoong, 2021) and on social media (Allcott, Gentzkow, & Song, 2022), and social media use was associated with decreased subjective well being (Allcott, Braghieri, Eichmeyer, & Gentzkow, 2020). The reason behind such findings need not be elusive. Online content serving platforms, like Facebook or Netflix, monetize consumer attention and have the

*I am grateful to Ernesto Reuben and Kerstin Bongard-Blanchy for their guidance in the design and conception of this experiment, to Sophie Doublet for her help in interface design, to Suvadeep Mukherjee for his guidance in building web experiments. I am also thankful for the invaluable comments from the Behavioral Economics Summer School at the University of Copenhagen, the doctoral workshop at UC Louvain, the Experimetrics Workshop Soleto at Sapienza University Rome, Samuel Greiff, Manuel Munoz Herrera, and from countless colleagues. ChatGPT was used for reformulating passages in this text as well analyzing the data.

ability to control the setting in which consumption decisions occur (e.g., format of information, presentation of choices). Such firms are potentially incentivized to increase the 'addictiveness' of their platforms by altering the content and design provided to users, as occupying more of consumers' free time can lead to higher revenues (Ichihashi & Kim, 2023; Kleinberg, Mullainathan, & Raghavan, 2022). Designing for addiction, in turn, would lead to increased content consumption but decreased utility from consumption, as implied by the extensive literature on present bias (see Ericson and Laibson (2019) for a review). Wary policymakers across both sides of the Atlantic have already made efforts to restrain the negative effects such incentives could have on consumer welfare (see SMART Act by US Congress and Digital Services Act by EU). Both acts are attempts to regulate the increasingly relevant digital environments which contain effective ways to solicit consumers' attention and influence them to make transactional decisions that may go against their best interests (Lupiáñez-Villanueva et al., 2022). Although both regulators and scientists alike successfully identify the problem, it remains to be seen what specific aspects of such online platforms are problematic, as a blanket regulation over the entire landscape of digital services could offset the many benefits users derive from them. Our study constitutes the first such attempt to our knowledge in disentangling the many interrelated features online platforms have and pinpoints to what specifically requires the attention of the regulators. Relevant to our experimental setting is the proposed ban on defaults regarding media player controls, where the platform serves content "without an express, separate prompt by the user (such as pushing a button or clicking an icon)".

Interface designers and computer scientists have already documented the many ways in which interface design can be instrumentalized to 'nudge' users in practice (see Caraban, Karapanos, Gonçalves, and Campos (2019) for a review). Nudge theory suggests knowledge about systematic biases in decision-making can be leveraged to support people in making better decisions for themselves (Thaler & Sunstein, 2009). Much of the existing literature is dedicated to improving decision-making in the context of benevolent government-sponsored nudges (see Mertens, Herberz, Hahnel, and Brosch (2022) for a review). More recently, however, discussions also extended to private sector interventions. This branch of the literature addresses the implications of changing behavior for 'better' decisions from the firm's point of view (Beggs, 2016; Mills, 2023; Smith, Goldstein, & Johnson, 2013). When firms nudge, there should be a distinction between whether the firm or the recipient of the nudge benefits from it. In other words, nudges can become problematic if the objective of the intervention is to maximize the benefit received by the firm instead of the decision-maker. Media player defaults, such as autoplay, are potentially one such example that could be driving the concerning results attested by literature. If these defaults are conceived to increase welfare of consumers instead, they should be preferred to active decisions or other alternatives. Carroll, Choi, Laibson, Madrian, and Metrick (2009) make a similar comparison between defaults and active decisions in the context of saving decisions, with a focus on modelling decision cost attached to making explicit choices.

The understudied 'autoplay' feature, is an interface nudge prevalent across social media and streaming platforms that automatically transitions users to the next video content without providing an explicit opportunity to discontinue consumption. Unless opted-out by users, platforms including Facebook, Netflix, YouTube and TikTok provide autoplay as the *default* media player configuration as of mid-2023. Defaults are effective because they (i) reduce effort, (ii) imply endorsement, and (iii) work hand in hand with cognitive biases such as loss aversion (Smith et al., 2013). In case of autoplay, the default makes watching successive videos more salient as it removes the opportunity from users to decide whether or not to continue watching. Studies have shown that media player design promotes "mindless" content consumption and reduces user agency (Lukoff et al., 2021). Removing the opportunity to decide whether or not to continue watching leads to increased watching times in children (Hiniker, Heung, Hong, & Kientz, 2018), known to exhibit less self-control than adults. If indeed autoplay promotes digital addiction as suggested by the proposed legislation and the literature, we would expect it to alter the decisions of viewers with self control issues even when proposed content is kept same. We therefore hypothesize that holding content constant, autoplay should result in increased viewing times compared to the control group making active decisions. Further, the concerns of addiction could be verified by contrasting actual watching choices to participants' previous commitment choices regarding how much they initially wanted to watch videos across conditions. In sum, we conceptualize autoplay as facilitating continuous watching, and compare it in an incentivized experiment to a 'neutral' setting where users must actively select videos to watch them.

To determine whether autoplay alters viewers' past commitment choices, we designed an online experiment where participants spent 20 minutes on a higher-paying real effort task (retyping CAPTCHAs)[1] or a lower-paying leisure task (watching videos). In the treatment condition, videos played automatically, while in the control condition, participants had to click to watch each video. To investigate the effects of autoplay on viewers with heterogeneous self-control issues, we included an additional task to elicit self control issues in our sample. In this task, participants visited the experiment website one day before the 20 minute session and practiced both the real effort task and the leisure task for 2 minutes in a so-called "practice session". Participants were already randomly assigned to the treatment during the practice session and were able to familiarize themselves with the tasks and the media player treatment. After the practice session, participants decided how long they would like to spend on the next day for each task with the help of a slider indicating their payoffs. This decision was reminded to the participants on the second day, but it was not binding for most.[2] In a separate data collection with a smaller sample, we also measured the demand for a commitment device using a multiple price list to see if participants would pay to turn off the autoplay feature. Our two task experimental setup was inspired by Kool and Botvinick (2014), while

[1] Charness, Gneezy, and Henderson (2018) categorize CAPTCHAs as a decoding task similar to Augenblick, Niederle, and Sprenger (2015) with the following useful properties: Low difficulty of implementation, small differences regarding the skill required to accomplish the task, and limited learning opportunities to improve at accomplishing the task.

[2] 5% of participants who are excluded from the main analysis were bound by the choice they made on the first day to add weight to the decision.

the real effort task is akin to Augenblick et al. (2015), and the leisure task resembles Bonein and Denant-Boèmont (2015). The contribution of our study is on its novel experimental design which allows testing both established and recent theories on societally relevant topics such as digital addiction. The entire task interface is constructed from scratch with the help of professional designers, and it aims to mimic the user experience on a social media platform. The design brings together extensive possibilities of data collection on behavior with the ease of scaling, while holding internal validity at a high standard. In sum, our design welcomes a plethora of new research questions that can be studied through economic experiments.

2 Experimental Design

To explore the relationship between default video player settings and self control, we created a decision making task that is boring, offers limited learning opportunities, and does not allow performance based compensation. We also offer participants a second, less paying leisure task to which they can switch at any time. The experiment is conducted over 2 days, with a cooling off period of 24 hours (up to 48 hours) between the two dates. Participants are asked to hypothetically allocate time on the first day, and then realize their allocation on the next day. Hypothetical time allocation made on the first day is contrasted with the realized time allocated on the second day to identify potential self control issues. Participants who completed all required elements of the experiment were paid within seven days after the second day of the experiment.[3] There was no completion bonus for completing the experiment and payment was dependent on the seconds participants actively spent on each task. If a participant failed to meet the quality criteria for the typing task they were not paid for the seconds spent on the task.

We present the design in four subsections. First, we describe the online decision environment and its built-in validity checks. Second, we describe tasks to be completed during the timed sessions. In the third subsection we present a timeline of the experiment. The fourth subsection addresses design details including recruitment, selection, attrition.

2.1 Online Interface

The experiment was conducted on a specialized website designed to guide participant navigation. Upon entry, a description of the experiment was immediately presented to participants. Subsequent access to the informed consent form and providing their Prolific identity was contingent on passing certain criteria, including checks on internet speed, device operating system, and browser type. Once verified, participants received further instructions, underwent attention checks, engaged in two time-limited tasks, and then completed a questionnaire. A key feature of the website was its linear navigation: participants could only move forward. To further maintain this flow, right-click functions

[3] Prolific policy gives researchers 21 days to transfer funds to participants. The typical delay between the study completion and payment is however only a couple of days as indicated on the Prolific website.

and keyboard shortcuts were also disabled. Crucially, any attempt to navigate backward would reset the participant's progress to the landing page. This design feature was communicated to participants at the beginning of the experiment.

Time-limited tasks were presented within a single web page under two distinct tabs. Participants could switch between tasks at any time using these tabs while saving their progress from the previous task. During the Practice Session, when instructions did not allow switching between tasks, tabs were grayed out and unclickable. Above the tabs, a progress bar and timer helped participants monitor time, indicating how long they have already spent in a session. When participants began a session, the default task was always the Typing Task. When a participant wanted to switch to the Watching Task, they had to click the according tab. Critical interactions, i.e., (i) tab switches, (ii) key presses during the Typing Task, (iii) video play or pause clicks, (iv) mouse cursor movements out of the experimental website (e.g., switching to another browser tab), were precisely logged in real time.

2.1.1 Internal Validity

A major concern when running online experiments is the lack of control over the environment. Online experiments can be influenced by participants multitasking or engaging in other activities, technical glitches, manipulative behaviors to increase earnings, and varied physical and temporal settings. To enhance the internal validity of our experiment, we have put in place several stringent measures. First, using browser-specific cookies, we selectively admit participants at the beginning of the experiment. This ensures that only individuals using stable internet connections (a minimum threshold of 30 Mbps on the first day and of 25 Mbps on the second day was ascertained through a speed-test on the first page) are included. Moreover, to standardize the user experience and minimize browser-specific variations, we confined our study to participants using Windows, Linux, or Mac operating systems exclusively with Chromium-based browsers. We purposefully excluded browsers such as Safari and Mozilla Firefox due to their restrictive autoplay and cookie configurations, and disallowed portable devices such as tablets and mobile phones. Second, after the completion of the experiment, participants were asked for any potential disruptions in their connectivity and their genuine engagement in the video watching task. Third, adopting a method previously implemented by Purohit, Bergram, Barclay, Bezençon, and Holzer (2023), we used a JavaScript package to detect any obstructions in the interface's visibility.[4] Using this tool, we detected actions such as minimizing the browser, toggling between web pages, or launching other applications and logged it. These logs provided a granular, second-by-second account of participants' interactions, enabling us to filter out data that may have been compromised by distractions or divided attention.

[4]For more information about the package, click here.

2.2 Tasks

2.2.1 Typing Task

Typing Task requires retyping successive CAPTCHA's. CAPTCHA's are computer generated images that contain random uppercase and lowercase letters as well as randomly placed white spaces.[5] Each CAPTCHA has a total length of 35 characters, including the white spaces. The images are blurred and distorted by lines and dots added on top to increase comprehension difficulty.[6] These CAPTCHA images are always presented in the interface with a text box and a submit button. When the submit button is clicked, the next CAPTCHA image comes and the text input for the previous one is stored. Typing Task is identical across all conditions.

There are easy-to-attend quality requirements to make sure participants are active: (i) an overall typing accuracy of at least 70 percent in submitted CAPTCHA's,[7] (ii) a minimum of one CAPTCHA submission per minute spent on task. Remuneration condition for the Typing Task is to meet these quality requirements. Typing Task is remunerated at 0.38 pennies per second participants work on it. A hidden timer records each second participants keep the Typing Task in the visible part of their screen. Participants in both conditions spent on average 41 seconds per CAPTCHA.

2.2.2 Watching Task

In this task participants watch successive short videos in a customized media player. Since we are interested in isolating the media player's default settings, the video content and the order in which videos appear to participants are identical across treatments. The media player controls are completely disabled: Participants are not allowed to skip videos (user scrolling is disabled) or use the media player's slider to advance or go back in a particular video. Watching Task is remunerated at 0.33 pennies per second participants work on it. In the Autoplay treatment, videos play automatically when the Watching tab is visible and pause with a click anywhere on the video. While the Watching task is visible, the media player keeps scrolling on to next videos and keeps playing them unless it is interrupted. In the control treatment, participants need to have the Watching tab visible and click on each consecutive video to play them. At the end of each video, the scrolling animation brings the next video but the new video only starts playing when the participant clicks on it.

Videos are curated from YouTube and TikTok, using tags such as "funny animal videos/short" and "cute animal videos/shorts". Manual selection of videos isolates the difference in the media player design, and help avoid algorithmic biases resulting from

[5] White spaces were added to improve readability after pretests. Number of white spaces were limited to maximum 5 instances per CAPTCHA in order to control the typing time per image.

[6] The randomization procedure to generate CAPTCHA's is available here.

[7] We use Python's `difflib` string-matching algorithm to measure accuracy. It consists of finding the longest common substring and then finding recursively the number of matching characters in the non-matching regions on both sides of the longest common substring. Overall accuracy of 70 percent implies an average of 10.5 ($0.3 * 35$) mistakes across all submitted CAPTCHA's.

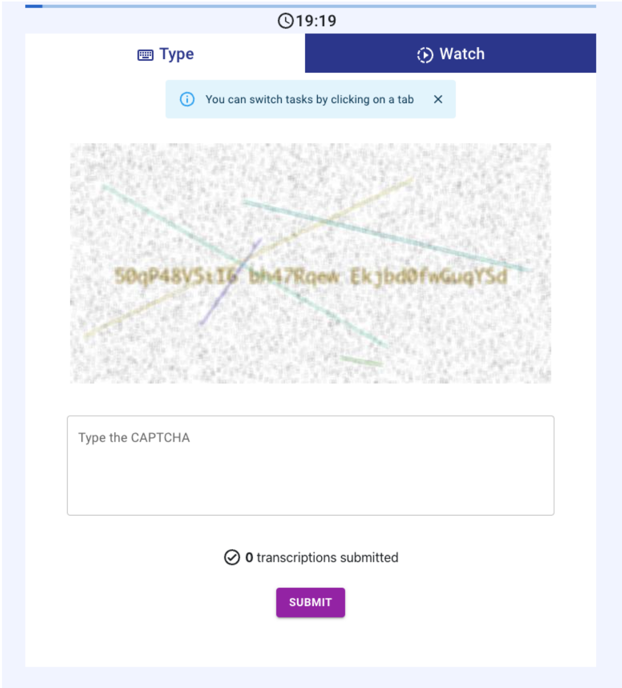


Figure 1: Typing Task tab

the recommender system's personalized suggestions. The reason for choosing animal videos was their popularity during pilot studies.[8] The selected videos were manually reviewed to ensure that they did not contain any harmful or violent actions towards the animals involved. The videos are kept short to maximize the frequency of experiencing the Autoplay treatment during the watching session.

2.3 Experimental Procedure

2.3.1 Practice Session

There are two sessions during which participants interact with the tasks described above. First, the Practice Session in day one: It comes right after the written descriptions of each task and serves to familiarize participants with the tasks and the interface. Participants are not remunerated for the time they spend here because the purpose of the practice session is to understand the difficulty and the pacing of the tasks, as well as navigation within the interface. During the practice session, each task is available only for 60 seconds and participants are not allowed to switch between tasks. On the first tab participants are required to retype CAPTCHAs. The 60 second period is enough to submit one CAPTCHA and see the second CAPTCHA image. After 60 seconds, a pop-up appears on the screen blocking any other interaction and takes participants

[8] During the pilots, the experiment was followed by a separate questionnaire asking participants to rate the videos they have seen and suggest what types of videos they would have liked to see more of.

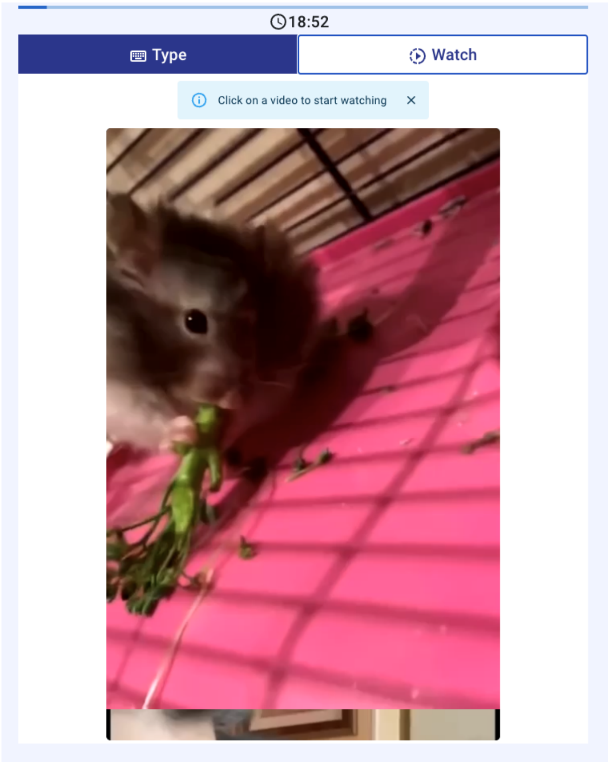


Figure 2: Watching Task tab

to the second page. Participants are required to close this pop-up to advance to the second task. The rationale behind the pop-up is to assure participants spend equal time on both tasks. Whether or not this pop-up is closed is verified to check participant activity, and serves as an additional attention check.[9] Once closed, the Watching Task begins for 60 seconds, enabling participants to watch 4 to 5 short videos depending on whether they were assigned to the Autoplay or the control treatment. Notice how in the Autoplay treatment videos played continuously and in the control treatment videos required manual start.

2.3.2 Multiple Price List

After the Practice Session on day one, participants in the Multiple Price List (MPL) condition made 9 binary decisions. They chose between their preferred media player setting (Autoplay off or on) and a bonus payment. The MPL condition was introduced after the initial data collection phase, and its bonuses were determined by the per-second earning differences across tasks, multiplied by the observed time differences between the Autoplay and control groups. Specifically, the average difference between the Autoplay and control treatments, in terms of time spent across tasks, as can be seen in Table 3, was approximately 60 seconds. This resulted in an average earning difference of $60 * (0.38 - 0.33) = 3$ pennies. To capture this in the MPL choices offered, we set the symmetric bonuses accordingly: £$[0.05, 0.1, 0.25, 0.5]$.

Once these decisions were completed, participants received information on the media player setting—specifically for day two, if Autoplay would be on or off—and the associated bonus. This information was displayed on the next page. Armed with this knowledge, they then decided on their Time Choice. Participants who were not in the MPL condition proceeded directly to the Time Choice page after their Practice Session. More details on the Time Choice can be found in the subsequent section.

2.3.3 Time Choice

Following the Practice Session, participants transitioned to the Time Choice section. This section required them to specify their preferred time allocation between the two tasks for the main 20-minute session scheduled for the following day. Participants knew the main session would be 20 minutes long and they were free in choosing corner solutions, even if that meant focusing solely on one task. Time Choice served as a benchmark for participants, and, accordingly, participants received a textual reminder of it on day two. From a research perspective, the Time Choice section was crucial, offering insights into participants' informed preferences and gauging their commitment to these choices after the 24-hour cooling period.

Participants used a slider tool with 20-second steps to indicate their time allocation preferences. Initially, the slider appeared grayed out. Requiring a click to activate, this design aimed to ensure participants considered their choices in an active manner. The click-to-activate design had two benefits: It guaranteed participants interacted with

[9] All participants closed this pop-up and switched to the Watching Task.

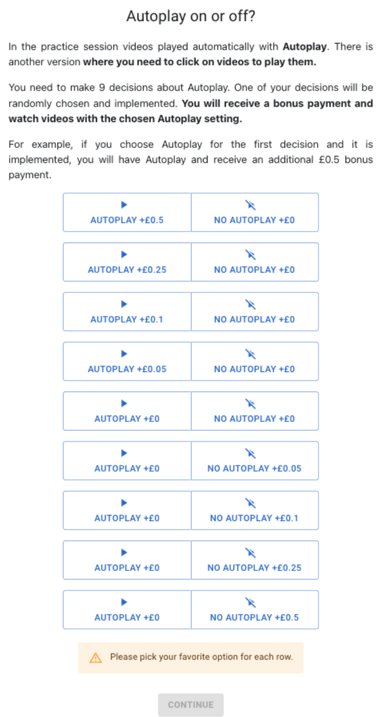


Figure 3: Multiple Price List page

the slider and helped participants understand the task-related payoff structure. To emphasize the importance of this decision, the choices of 5 percent of participants were binding in the main session. Binding choices led to payments based on the initial Time Choice, but only if participants met the quality standards described under the Typing Task. Data from these participants were excluded from the final dataset accordingly. Participants were informed whether their allocation was binding with their Time Choice reminder on day two, right before the start of the main 20-minute session.

Table 1: Summary of experimental design and timeline

	Practice Session	Time Choice	Main Session	Multiple Price List	Decision Horizon
Day 1 (May 10)	x	x			24-hour
Day 2 (May 11)			x		immediate
Day 1 (May 18)	x	x		x	24-hour
Day 2 (May 19)			x		immediate

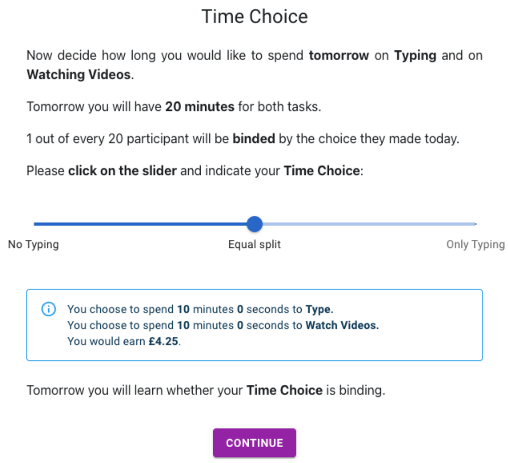


Figure 4: Time Choice page

2.4 Behavioral Hypotheses

We preregistered our main hypotheses before beginning data collection.[10] Our preregistration included hypotheses 1 and 4. We have added hypotheses 2 and 3 regarding our expectations about the Time Choice section for the purpose of this report.

- **Hypothesis 1**: The Autoplay treatment will lead to an increase in the time spent watching videos when compared to the control condition where participants actively need to click on each video to watch. This will be measured by:

  1.1 The total time participants dedicate to the Watching task.

  1.2 The number of videos watched per condition.

- **Hypothesis 2**: No significant difference is expected in the time allocation decisions made by participants between the two tasks, as detailed in the Time Choice section. We test for this difference by comparing the initial Time Choice allocations across conditions.

- **Hypothesis 3**: Participants in the Autoplay condition will deviate more towards watching videos from their initial Time Choice, where they made their preferred time allocation between the two tasks at least 24 hours prior to the main session. We measure this difference by contrasting the initial Time Choice allocations to actualized allocations during the main session across conditions.

- **Hypothesis 4**: There will be a positive demand for a commitment device. Specifically, when given a choice, participants will choose to have the autoplay feature

[10] To access the preregistered hypotheses, click here.

turned off. This demand will be measured through the participant choices in the MPL condition, specifically looking at the decision where participants switch from one column to the other.

3 Results

The experiment was run entirely online, using widely available open-source tools.[11] The sampling was made from the Prolific participant pool, an online recruitment platform commonly used for academic research, targeting adult British residents. We further imposed the following criteria: fluency in English, balanced gender representation, having a stable internet connection (> 30 Mbps on the first and > 25 on the second day of the experiment). We did not allow mobile devices and restricted access to the experiment to Chromium-based browsers to enhance the internal validity of the experiment as previously discussed in detail here. Only participants that succeeded both attention checks, reported not having engaged in other activities, and did not have connection issues during the experiment were included in the final dataset.

For the first stage of data collection, that is to say excluding MPL treatment, we pre-registered a two-sample parallel design using our prototype data and tested for equality of means across treatments. We found that the required sample size to test Hypothesis 1 with the usual parameter values ($\alpha = .05$, $\beta = .2$) would be 114 per treatment, with an effect size of 65 seconds calculated by the difference in the average time spent between conditions ($\mu_{Autoplay} - \mu_{Control}$) and population standard deviation of 165 seconds. Since attrition in longitudinal experiments are commonplace, we invited 301 participants on the first day. On day two, we invited the same 301 participants and received 276 complete responses. No participants failed the attention checks more than twice, and therefore were not removed from the dataset as per Prolific policy. 17 participants who self-reported having connection issues and 24 additional participants that stated having engaged in other activities were dropped. 11 participants who had their Time Choice enforced randomly were also removed and the final dataset on which we base the analysis consisted of 224 individuals. Table 2 shows the demographic characteristics of the observed individuals and Table 3 shows sample statistics for variables of interest.

For the MPL treatment, applying the same selection criteria as before, we invited 60 participants on day one. On day two, we received 56 complete answers. Once again, no participant failed the attention checks. No participant reported having connection issues. This improvement is likely because we had put a higher network speed requirements ($>$ 35 Mbps on the first and > 30 on the second day). 5 participants reported having engaged in other activities and were dropped. As we were interested in the demand for a commitment device, participants did not have their Time Choice enforced in this treatment. We therefore ended up with 51 individuals in the final dataset.

[11] We used the MERN stack to develop and host the website in an entirely free way.

	Autoplay N = 104	**Control** N = 120	**Overall** N = 224
Gender			
Female	48 (46.15%)	54 (45.00%)	102 (45.54%)
Male	56 (53.85%)	66 (55.00%)	122 (54.46%)
Age			
Mean (SD)	40.56 (11.66)	40.24 (10.98)	40.39 (11.27)
Median [Min, Max]	37.00 [19.00, 70.00]	39.00 [20.00, 69.00]	38.50 [19.00, 70.00]
Employment			
Unemployed looking for work	4 (3.85%)	8 (6.67%)	12 (5.36%)
Homemaker	4 (3.85%)	4 (3.33%)	8 (3.57%)
Self employed	8 (7.69%)	14 (11.67%)	22 (9.82%)
Retired	3 (2.88%)	3 (2.50%)	6 (2.68%)
Employed full time	62 (59.62%)	67 (55.83%)	129 (57.59%)
Employed part time	20 (19.23%)	14 (11.67%)	34 (15.18%)
Unable to work	1 (0.96%)	2 (1.67%)	3 (1.34%)
Student	2 (1.92%)	7 (5.83%)	9 (4.02%)
Unemployed not looking for work	0 (0.00%)	1 (0.83%)	1 (0.45%)
Income			
0-10	14 (13.46%)	23 (19.17%)	37 (16.52%)
12-15	6 (5.77%)	4 (3.33%)	10 (4.46%)
15-20	8 (7.69%)	10 (8.33%)	18 (8.04%)
20-30	22 (21.15%)	26 (21.67%)	48 (21.43%)
10-12	8 (7.69%)	7 (5.83%)	15 (6.70%)
50-70	10 (9.62%)	6 (5.00%)	16 (7.14%)
30-50	31 (29.81%)	37 (30.83%)	68 (30.36%)
70+	5 (4.81%)	7 (5.83%)	12 (5.36%)
Marital Status			
Married or domestic partnership	56 (53.85%)	60 (50.00%)	116 (51.79%)
Divorced	3 (2.88%)	6 (5.00%)	9 (4.02%)
Widowed	1 (0.96%)	1 (0.83%)	2 (0.89%)
Separated	2 (1.92%)	1 (0.83%)	3 (1.34%)
Single	42 (40.38%)	52 (43.33%)	94 (41.96%)

Table 2: Descriptive Statistics for the full sample

	Autoplay N = 104	Control N = 120	Overall N = 224
Allocated Type time			
Mean (SD)	573.85 (388.30)	520.17 (400.76)	545.09 (395.05)
Median [Min, Max]	600.00 [0.00, 1200.00]	600.00 [0.00, 1200.00]	600.00 [0.00, 1200.00]
Type time			
Mean (SD)	718.24 (386.57)	657.84 (413.38)	685.88 (401.40)
Median [Min, Max]	765.21 [3.04, 1200.00]	608.71 [2.57, 1200.00]	707.28 [2.57, 1200.00]
Watch time			
Mean (SD)	466.67 (383.58)	524.78 (408.00)	497.80 (397.03)
Median [Min, Max]	423.13 [0.00, 1196.96]	562.03 [0.00, 1197.43]	473.18 [0.00, 1197.43]
Content rating			
Mean (SD)	6.92 (2.29)	6.44 (2.53)	6.67 (2.43)
Median [Min, Max]	7.00 [2.00, 10.00]	7.00 [0.00, 10.00]	7.00 [0.00, 10.00]
Nb. of task switches			
Mean (SD)	4.00 (4.50)	3.69 (5.03)	3.83 (4.78)
Median [Min, Max]	2.00 [0.00, 16.00]	2.00 [0.00, 28.00]	2.00 [0.00, 28.00]
Nb. of retyped images			
Mean (SD)	18.01 (11.97)	16.26 (11.59)	17.07 (11.78)
Median [Min, Max]	17.00 [0.00, 53.00]	14.50 [0.00, 45.00]	15.50 [0.00, 53.00]
Nb. of videos watched			
Mean (SD)	35.75 (27.21)	34.40 (26.35)	35.03 (26.70)
Median [Min, Max]	36.00 [0.00, 80.00]	40.50 [0.00, 77.00]	37.50 [0.00, 80.00]

Table 3: Sample statistics for variables of interest

3.1 Testing Behavioral Hypotheses

3.1.1 Effects of Autoplay on watching videos (H1)

Our primary hypothesis was investigating whether the Autoplay feature influences the time participants spend on watching videos. We scrutinized this using two distinct measures. The first one considered the time span participants allocated to the Watching Task, captured through second-by-second logs during the 20-minute main session. Unfortunately, a technical glitch with the countdown timer inadvertently caused most participants to spend an additional 30 to 60 seconds on the session. A deeper dive into this issue unveiled that the timer *slowed down* in the absence of user interactions, such as clicks, cursor movements or typing. Nevertheless, post-experiment feedback reassured us that participants remained unaware of this anomaly. To understand the overall implications of this glitch on our experiment, we analyzed the distribution of the total time spent across the two treatments, visualized in Figure 5. Herein, the x-axis delineates the total time allocated to both tasks, while the y-axis depicts its frequency. The distribution of ending times appears roughly symmetrical for both treatments on the histogram, and a Mann-Whitney U test confirms that the glitch affected both conditions uniformly ($U = 6331, p = 0.85$).

To remedy for the differences in time spent across tasks resulting from the glitch, we normalize observed task data to 1200 seconds. Figure 6a shows the distribution

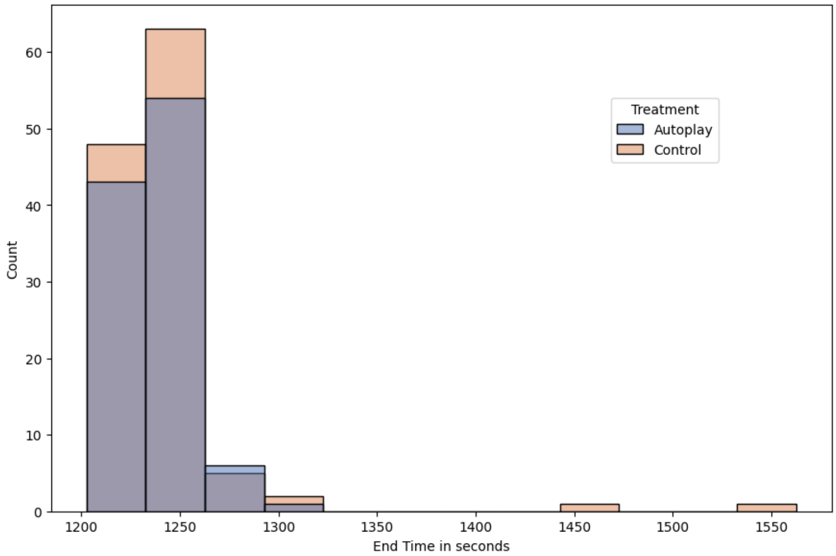
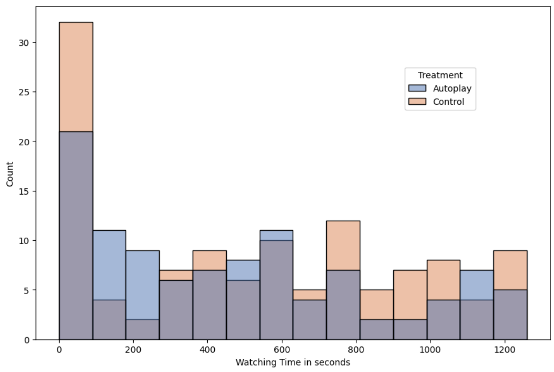
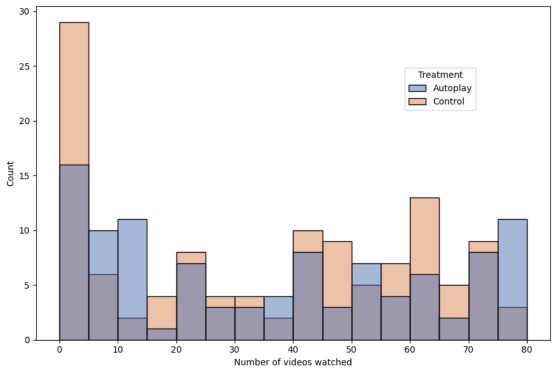


Figure 5: Distribution of main session ending time across treatments

of the normalized time spent on watching videos across treatments. We test and reject Hypothesis 1.1 using the Mann-Whitney U test ($U = 5825, p = 0.39$), meaning Autoplay treatment did not result in an increased time spent watching videos. We then proceed with Hypothesis 1.2, focusing on the number of videos each participant watched across treatments. Figure 6b shows the distribution of the number of videos watched per condition. Using a Mann-Whitney U test again, we reject that Autoplay treatment increased the number of videos watched ($U = 6494, p = 0.599$).



(a) Watching time in seconds



(b) Number of videos watched

Figure 6: Distributions of variables relating to watching time across treatments

3.1.2 Effects of Autoplay on Time Choice (H2)

We next examine Hypothesis 2, which posits there are no significant differences in the initial Time Choice made by participants across two treatments. For clarity, Time Choice refers to how much time participants preferred to spend on the Typing Task on day one. As depicted in Figure 7, the distribution of Time Choices across treatments appears roughly symmetrical for both treatments. To validate this observation, we employed the Mann-Whitney U test, where the null hypothesis suggests there is no statistically significant difference in Time Choice distributions across treatments. The outcome leads us to confirm our hypothesis ($U = 6713.5, p = 0.323$), suggesting that participants' initial Time Choice were statistically indistinguishable across treatments. Participants seem to have been uniformly affected by the two treatments, which is important ascertain before we delve further into the analysis of Time Choice.

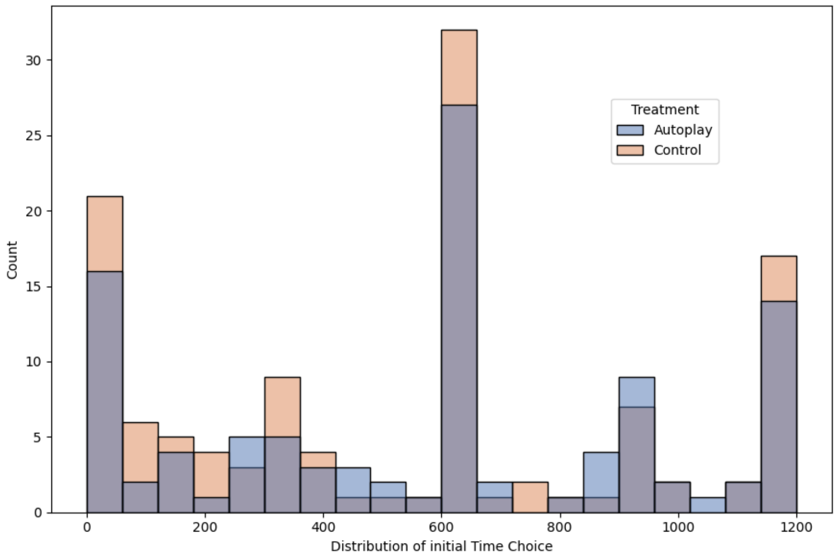


Figure 7: Distribution of Time Choice across treatments

3.1.3 Effects of Autoplay on deviations from Time Choice (H3)

We then move on to the Hypothesis 3, which suggests participants in the Autoplay condition would deviate more towards watching videos from their initial Time Choice. To assess this hypothesis, we transformed the Time Choice variable by normalizing to percentages and calculated the deviation by subtracting the percentage of actual time spent on the Typing Task from it. Figure 8 shows the distribution of deviations from Time Choice across treatments which appears left-skewed, suggesting increases in actual typing instead. Specifically, negative values in the figure represent increases in typing, whereas positive values are increases in watching. Applying the Mann-Whitney U test, our data ($U = 6134, p = 0.827$) did not support a significant deviation for the Autoplay condition towards watching. Further descriptive statistics reveal a mean deviation of -11.73, a median of -8.46, and a mode of 0, emphasizing the observed left skew. To strengthen our findings, we used the Wilcoxon Signed-Rank test to compare the distribution of deviations to a 0 median distribution, and rejected the null ($W = 5152$, $p < 0.001$). In summary, contrary to our hypothesis, participants generally spent more time on the Typing Task than they initially intended regardless of treatment.

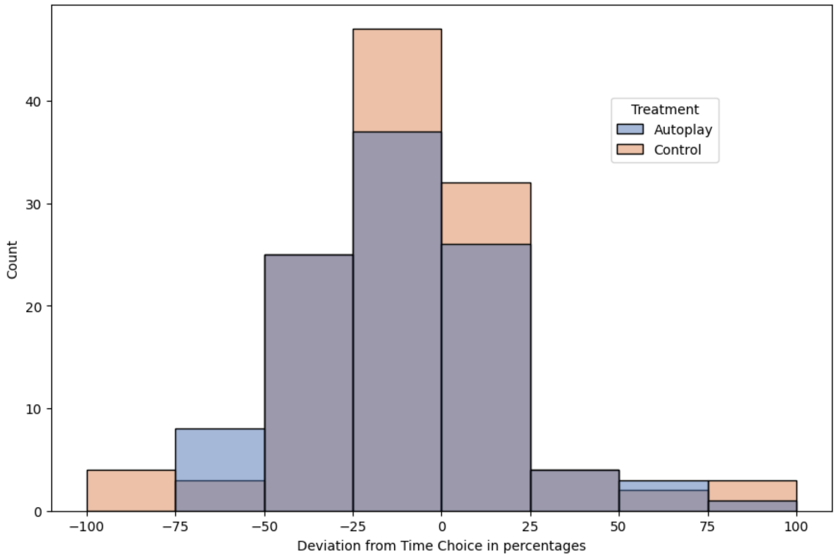


Figure 8: Distribution of the deviation from initial Time Choice in percentages across treatments

3.1.4 Demand for the commitment device (H4)

Hypothesis 4 suggests participants would choose to turn Autoplay off when given the possibility, meaning there is positive demand for the commitment device. We assess this demand by comparing the share of participants choosing to turn Autoplay on or off along with the associated bonus payments for each of the nine decisions in the price list. Figure 9 shows the percentages of participants preferring Autoplay in blue and those preferring to turn Autoplay off in orange. Distributions clearly show that when given a choice, participants have a preference for Autoplay. Specifically, more than 75 percent of participants chose Autoplay without bonus, 12 percent chose Autoplay if it meant giving up £0.05, and 7 percent chose Autoplay if it meant giving up £0.1.

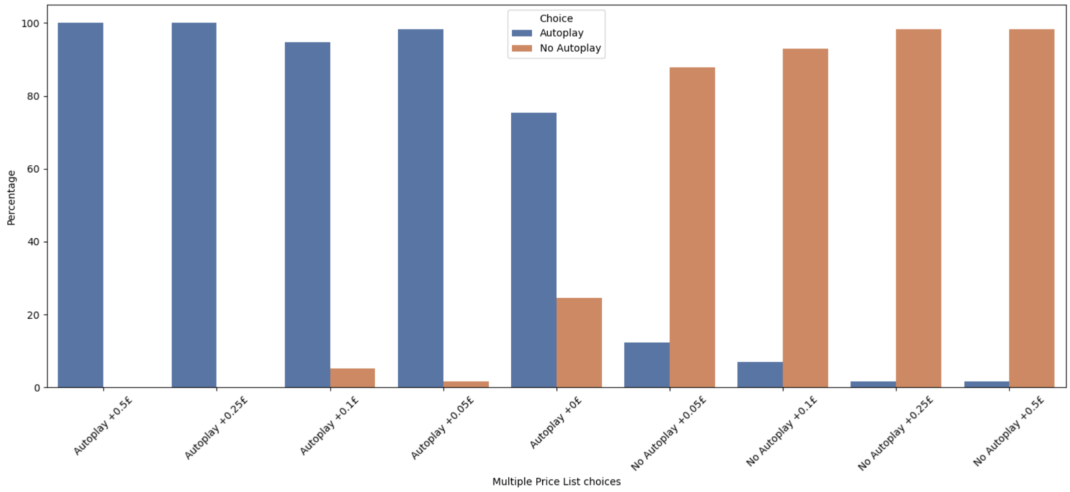


Figure 9: Distribution of Autoplay preferences in percentages

4 Discussion

There are two ways to interpret our findings. We begin our discussion first with a simplistic understanding of the results and build up for a more nuanced explanation. The straightforward conclusion based on the lack of support for Hypothesis 1 is that Autoplay does not have a distinct impact on participants' actualized decisions when contrasted to active decisions. Combined with the lack of support for Hypothesis 3 on deviations from how long participants wanted to watch videos initially and the actual preference for Autoplay in Hypothesis 4 further supports the idea that Autoplay is a 'harmless' default when it comes to media control settings and does not provoke increased consumption. Our findings allow for this conclusion, and show instead that users actually prefer autoplay to active decisions when there is money at stake. This would contradict the presumed welfare benefits of the proposed ban in the SMART Act by the US Congress. However, this 'simple' interpretation of results does not consider the complexity of the issue at hand and leaves out a crucial design choice: The content of videos. To reassure, how the content was perceived by participants in our experiment was controlled by a survey question which asked them to rate the content of the videos. The median response was a 7 out of 10 rating across conditions for the content (see Table 3 for sample statistics).[12] Yet, the selection (manual) and the presentation (randomized order) of the content are clearly different from the algorithmically curated personalized experience participants are accustomed to have on daily basis. To have a clear understanding of

[12]Using Shapiro-Wilk tests we cannot reject the null that distributions are significantly different from a normal distribution for both autoplay ($W = 0.932$, $p < 0.001$) and control ($W = 0.933$, $p < 0.001$) treatments. A Mann-Whitney U test between treatments cannot reject the null that distributions are significantly different from one another ($U = 6894$, $p = 0.173$), although there is some indication (see Figure 10 for the density distribution) that participants in autoplay condition gave higher ratings to the content.

the implications of the experiment, it is preferable to look at the larger context in which content and design together shape the digital environment where decisions are made.

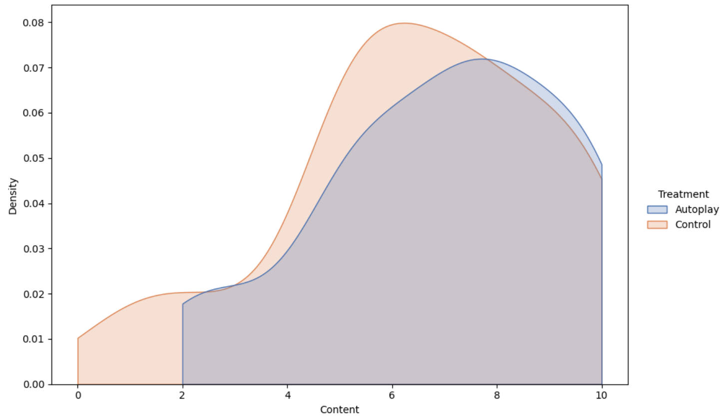


Figure 10: Density distribution of content ratings across treatments

In order to take into account this interplay between the content and the design, we rely on the insights from a recent model on media consumption (Kleinberg et al., 2022) where firms optimize over a large set of potential content and interface designs parametrized by their underlying features. Consumers optimize utility for their two selves: The impulsive and myopic one (self 1) and the forward-looking and thoughtful one (self 2). Content is consumed so long as self 1 derives some positive utility, and only after self 1 receives negative utility can system 2 make the decision to stop consuming content in case system 2 was already receiving negative utility. In this model, autoplay, a feature that facilitates seamless content consumption, has heterogeneous effects over different types of content. The intuition is as follows: If the existing content already is such that it has a low probability that utility for impulsive self 1 is larger than zero, autoplay has little impact on the user behavior. In other words, on a content landscape with low addictiveness, autoplay would have minimal impact on engagement. Applying this to our setting implies that the content provided for the experiment, although itself curated from social media, is much less interesting for the impulsive self 1 of participants.

Further, when a user is actively engaged and desires to consume more content, the autoplay feature aligns with the user's intent. This supports the preference for autoplay found in the MPL condition regarding Hypothesis 4, implying that participants were aware of their actual content consumption and desired to consume more content. This is entirely in accordance with the inference on the selection of videos not being tempting enough. In other words, because videos were not tempting, participants were in control of their consumption and could end sessions whenever their system 2 received negative utility from videos. In this setting, autoplay is actually helpful to participants in achieving their desired session length as it reduces decisions costs attached to it, fully aligned with previous the modelization of active decisions compared to defaults by Carroll et al. (2009). Putting the dual insights of the model together builds a coherent picture in terms of the policy implications of the present study, and we present convincing evidence that autoplay itself does not promote addiction in the absence of engaging content. Our findings go contrary to the proposed legislation which implicitly assumes all autoplayed content is equally addictive and welfare decreasing.

5 Limitations and Conclusion

EXPERIMENTAL DEMAND EFFECTS AND SOCIAL IMAGE CONCERNS. THE EXPERIMENT IS SET AS A WORKING ENVIRONMENT. PARTICIPANTS PERCEIVED THE DECISION AS ONE TO WORK AS MUCH AS POSSIBLE.

References

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–76.

Allcott, H., Gentzkow, M., & Song, L. (2022). Digital addiction. *American Economic Review*, *112*(7), 2424–2463.

Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over time: Dynamic inconsistency in real effort tasks. *The Quarterly Journal of Economics*, *130*(3), 1067–1115.

Beggs, J. N. (2016). Private-sector nudging: the good, the bad, and the uncertain. In *Nudge theory in action: Behavioral design in policy and markets* (pp. 125–158). Springer.

Bonein, A., & Denant-Boèmont, L. (2015). Self-control, commitment and peer pressure: a laboratory experiment. *Experimental Economics*, *18*, 543–568.

Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–15).

Carroll, G. D., Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2009). Optimal defaults and active decisions. *The quarterly journal of economics*, *124*(4), 1639–1674.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, *149*, 74–87.

Ericson, K. M., & Laibson, D. (2019). *Intertemporal choice, vol. 2, chap. 1.* Elsevier.

Hiniker, A., Heung, S. S., Hong, S., & Kientz, J. A. (2018). Coco's videos: an empirical investigation of video-player design features and children's media use. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).

Hoong, R. (2021). Self control and smartphone use: An experimental study of soft commitment devices. *European Economic Review*, *140*, 103924.

Ichihashi, S., & Kim, B.-C. (2023). Addictive platforms. *Management Science*, *69*(2), 1127–1145.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2022). The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *arXiv preprint arXiv:2202.11776*.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of experimental psychology. General*, *143*(1), 131.

Lukoff, K., Lyngs, U., Zade, H., Liao, J. V., Choi, J., Fan, K., ... Hiniker, A. (2021). How the design of youtube influences user sense of agency. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–17).

Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F., Liva, G., Lechardoy, L., & de las Heras Ballell, T. R. (2022). *Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation.* Publications Office of the European Union.

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, *119*(1), e2107346118.

Mills, S. (2023). Nudge/sludge symmetry: on the relationship between nudge and sludge and the resulting ontological, normative and transparency implications. *Behavioural Public Policy*, *7*(2), 309–332.

Purohit, A. K., Bergram, K., Barclay, L., Bezençon, V., & Holzer, A. (2023). Starving the newsfeed for social media detox: Effects of strict and self-regulated facebook newsfeed diets. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–16).

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, *32*(2), 159–172.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness.* Penguin.