Active Online Choices: Designing to Empower Users

Technical Report, August 2021

THE BEHAVIOURAL INSIGHTS TEAM

doteveryone

Centre for Data Ethics and Innovation

Contents

Introduction

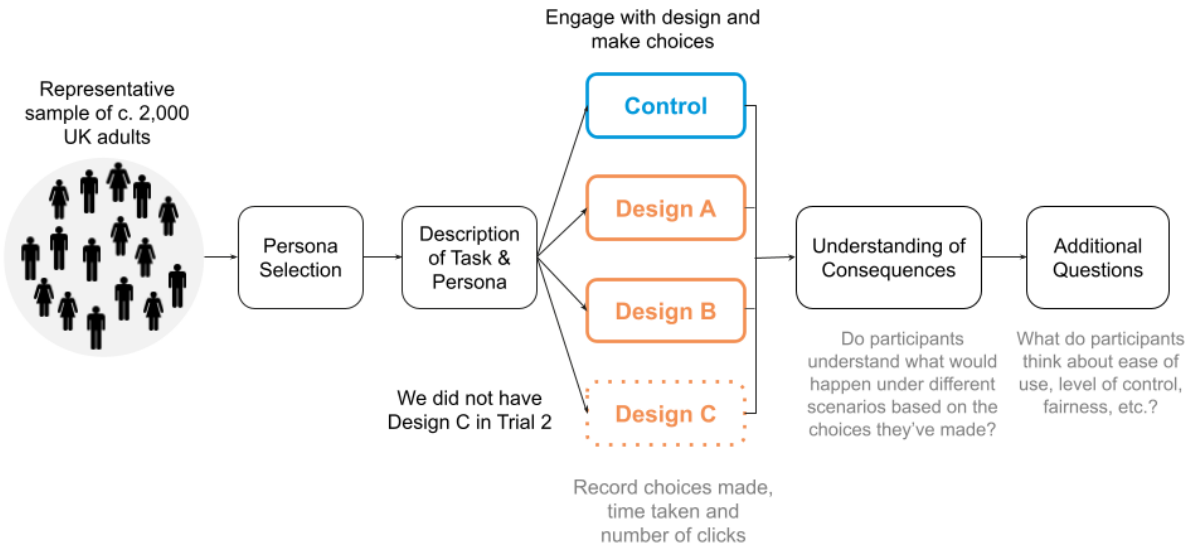
This Technical Report supplements the main research report titled Active Online Choices: Designing to Empower Users. We encourage you to read the main report either before or alongside this document.

All policy implications and recommendations are discussed in the main report. The purpose of this Technical Report is to share the full details of the methodology used to run three Randomised Controlled Trials (RCTs) and the findings from them. This maximises research transparency and provides readers with technical details to help develop and run similar experiments in their field.

1. Overview of the participant journey

All three experiments were conducted entirely online using BIT's Predictiv platform. Participants in the study were members of an online research panel who could decide to participate in return for a small financial reward. All three experiments had the same structure (see Figure 1).

Figure 1: Overview of the structure of the experiments



Once a participant entered the experiment, they were then taken through several stages:

- **Eligibility check**: In the smartphone trial, participants were asked if they were active Android users; those who did not own an Android phone or those that had not used it in the month prior to the experiment were screened out. Additionally, only participants using a mobile device were eligible. In the web browser and social media trials, eligible participants had to be using a computer or a tablet, not a mobile phone. Participants in the social media trial were additionally asked which social media platforms they actively used and those who did not select "Facebook" were screened out.
- **Attention check**: The following question was used as an attention check: "People are very busy these days. We are testing whether people read questions. To show that you've read this much, answer both 'Extremely interested' and 'Very interested'." Participants who failed the attention check were terminated.

- **Persona selection**: Participants were presented with a general description of three personas: one comfortable with sharing data (Persona 1, 'least concerned'), one uncomfortable with sharing data (Persona 3, 'most concerned'), and one in between (Persona 2, 'partly concerned'). They were asked to select the persona they most closely identified with.

- **Description of task and persona**: Participants were shown a description of the main task. They were told that they would need to adjust settings so that they match the persona's preferences in an interactive mock-up. They were told that for each correct setting, they would receive an additional small financial reward. They were then presented with a detailed description of the persona they had chosen.

- **Engagement with the designs**: Participants were then randomly assigned to see one of the experimental designs. For more information on the designs, please see the main report. Participants could interact with the designs for as long as they wanted. The persona description was available for display throughout the task so there was no need for memorisation.

- **Understanding of consequences**: This consisted of a series of multiple-choice questions based around mini-scenarios, intended to test participants' understanding of the consequences of the choices that they had just made.

- **Follow-up questions**: Participants were asked additional questions about the interface that they had just seen, such as an assessment of their experience and whether they would want to see similar interfaces in the future. They could also optionally provide free-text comments.

- **Closing page:** Participants were thanked for their participation and given both their base pay and their performance-linked reward.

2. Outcome measures

Table 1 below lists the outcome measures used in this trial. In addition to those variables, we collected the following covariates:

- Gender
- Age
- Household income
- Location
- Education level
- Ethnicity

Table 1: Primary and secondary outcome variables

PRIMARY			
Measure	**Smartphone**	**Web browser**	**Social media**
Task Accuracy	Sum of correct choices (based on persona choice) for the following settings: ● Notification frequency ● Personalised advertising ● Private browsing (by default or not) ● Location tracking by an app	Sum of correct choices (based on persona choice) for the following settings: ● Blocking cross-site trackers ● Regular clearing of cookies ● Flagging untrustworthy news sources	Sum of correct choices (based on persona choice) for the following settings: ● Trustworthy information sources only ● Personalised Feed order ● Personalised ads ● Whether posts are shared publicly
SECONDARY			
Understanding of Consequences Sum of correct answers to the following questions Correct answers were based on the settings participants' selected even if these choices were not in line with the persona's preferences.	Please answer the following questions based on the settings you have just selected. 1. Alex has the Instagram app installed on their phone. Based on your choices, when will the phone give Instagram access to location data? *[All the time (including when not using the app) / Whenever using the app / Each time Alex gives the app permission to access their location data]* 2. Based on your choices, what types of advertisement might be shown on this device when browsing the internet? *[Advertisements that are based on your device information / Advertisements that are NOT based on your device information*	Please answer the following questions based on the settings you have just selected. 1. Based on your choices, would a website that Alex visited on their browser collect data on… ● Alex's activity on the website itself, e.g. which products they looked at *[Yes / Yes, but it would be deleted after Alex closes their browser / No]* ● Alex's activity on other websites, e.g. what other websites they visited and how long they spent on them *[Yes / Yes, but it would be deleted after Alex closes their browser / No]*	Please answer the following questions based on the settings you have just selected. 1. Based on your choices, will The Network collect data about Alex's behaviour on shopping websites? *[Yes / No]* 2. Based on your choices, how will the Network order Alex's Feed? *[Most recent posts will be shown first / Feed order will be personalised based on Alex's browsing data]* 3. Imagine Alex has posted a new photo. Based on your choices, who will be able to see it?

	/ No advertisements will be shown on the device] 3. If Alex were to now open a new web browser window on this phone (e.g. in Chrome), would data on browsing history be collected by the browser? *[Yes / No]* 4. Alex has the Instagram app installed on their phone. How often will they now *receive notifications from Instagram?* *[Alex will receive notifications immediately / Alex will receive notifications once an hour / Alex won't receive any notifications from Instagram]*	• Alex's activity on his computer outside of their browser *[Yes / Yes, but it would be deleted after Alex closes their browser / No]* 2. If Alex were to now open their browser on this computer, would… • they remain logged in (where they've logged in previously)? • items they've previously added still be in their basket? • their cookie preferences on individual websites be saved? *[Yes / No]* 3. Will Alex see content from potentially untrustworthy sources? *[Yes, but such content will be flagged / Yes and such content won't be flagged / No, such content won't be shown]*	*[Only Alex / Alex's friends / Any user of The Network]* 4. Based on your choices, what kinds of ads will Alex see? *[Ads based on [his] basic profile information as well as [his] online behaviour / Ads based only on his basic profile information / Ads that are not personalised in any way]* 5. If Alex's friend shares an article from an untrustworthy news source (according to NewsGuard), what will Alex see, based on your choices? *[The article will appear in Alex's Feed / The article will not appear in Alex's Feed / The article will appear in Alex's Feed but it will be flagged as untrustworthy]*
Feelings of control	1. How much control do you feel you had over the privacy and notification settings when making your choices? *[No control / Little control / Some control / A lot of control / Complete control]*	1. How much control do you feel you had over the privacy settings when making your choices? *[No control / Little control / Some control / A lot of control / Complete control]* 2. How much control do you feel you had over the quality of the news that	1. How much control do you feel you had over the privacy and content settings when making your choices? *[No control / Little control / Some control / A lot of control / Complete control]*

		you see when making your choices? *[No control / Little control / Some control / A lot of control / Complete control]*	
EXPLORATORY			
Experiences	The next set of questions is about your experiences doing the task. We want to know how **<u>you</u>** feel, not what you think Alex would feel. 1. To what extent was it easy or hard to make choices on behalf of Alex? *[Very easy / Quite Easy / Neutral / Quite Difficult / Very Difficult]* 2. How much do you trust that the choices were presented to you with your best interests in mind? *[Not at all / A little / Somewhat / Very much / Completely]* 3. To what extent do you feel like the choices were presented in a fair way, allowing you to form your own opinions without being influenced? *[Not at all / Slightly / Moderately / Considerably / Completely]* 4. To what extent did you think that the available settings were explained in 'lay terms' (as in, easy for people to understand)? *[Not at all / Slightly / Moderately / Considerably / Completely]* 5. Did you have as much control over the settings as you would have liked when making choices for Alex? *[No - I want much more control / No - I want a little more control / Yes - it was enough / No - I want a little less control / No, I want much less control]* 6. Please have a look at an image of the interface again and answer the questions below. Would you like to see an interface like this one in future?	

	[No, definitely not / No, probably not / Unsure / Yes, probably / Yes, definitely] 7. Given the opportunity, would you choose to use the options in this interface rather than the [current settings on Android / your browser / social media websites]? *[No, definitely not / No, probably not / Unsure / Yes, probably / Yes, definitely]*
Optional comments	Free-text boxes: 1. Was anything about the settings interface particularly confusing? (optional) 2. Do you have any suggestions for improving the interface? (optional) [not asked in the control arm]
Concerns about technology	How concerned, if at all, would you say you are about each of the following? ● Companies selling on data about me ● How addictive technology can be ● Fake news or disinformation online ● Decisions being made about individuals by artificial intelligence *(1) Not at all concerned / (2) / (3) / (4) / (5) Very concerned*
Digital comfort	To what extent are you comfortable using a computer, a tablet or a smartphone to access government or commercial services online? *[Not at all comfortable / A little comfortable / Somewhat comfortable / Very comfortable / Completely comfortable]*

3. Analytical strategy

Primary outcome

We used a linear regression (Ordinary Least Squares) model to test the effect of our treatments on main task accuracy. We fitted two models: one treatment assignment and persona terms only and one with an additional set of covariates. We then chose the better-fitting model (using the AIC and adjusted R^2) for reporting.

Model 1a: $$Accuracy_i = \alpha + \beta_1 Treatment_i + \beta_2 Persona_i + \varepsilon_i$$

Model 1b: $$Accuracy_i = \alpha + \beta_1 Treatment_i + \beta_2 Persona_i + A_i \Gamma + \varepsilon_i$$

- $Accuracy_i$ is treated as a continuous variable representing the number of settings adjusted correctly.
- $Treatment_i$ is a dummy-coded variable set to
    - 1 if the *i*th participant saw design A
    - 2 if the *i*th participant saw design B
    - 3 if the *i*th participant saw design C (we did not have design C in the web browser trial)
    - and 0 otherwise (control)
- $Persona_i$ is a dummy-coded variable equal to 1 if participant *i* selects Persona 1 ('least concerned'), equal to 2 if they select Persona 2 ('partly concerned'), and equal to 3 if they select Persona 3 ('most concerned').
- A_i is a vector of controls which indicate the gender, age bracket, income bracket, location, and education of participant *i*. These variables were treated as dummy variables.
- β_1 is a vector of regression coefficients associated with the treatment-assignment dummy variables.
- β_2 is a vector of regression coefficients associated with the persona dummy variables.
- α is the regression constant.
- ε_i is the error term.

We estimated standard errors using a heteroskedasticity-robust method. We adjusted *p*-values for comparisons using the Hochberg-Benjamini method.

Secondary outcomes

We used linear (OLS) regressions to test the effect of our treatments on the two secondary outcomes (understanding of consequences and feelings of control). For consistency across analyses, we used the equivalent model specification (i.e. either adjusted or unadjusted) as selected in the analysis of the primary outcome.

$$Y_i = \alpha + \beta_1 Treatment_i + \beta_2 Persona_i + A_i \Gamma + \varepsilon_i$$

- Y_i is either:
  - a continuous variable representing the number of settings correct answers to the 'understanding of consequences' questions
  - a continuous variable representing the answer to the 'feelings of control' question
- $Treatment_i$ is a dummy-coded variable set to the same values as in the primary analysis
- A_i is a vector of controls which indicate the gender, age bracket, income bracket, location, and education of participant *i*. These variables were treated as dummy variables.
- α is the regression constant.
- ε_i is the error term.

We estimated standard errors using a heteroskedasticity-robust method. We adjusted *p*-values for comparisons using the Hochberg-Benjamini method.

Robustness checks

To ensure that our results do not materially change when using different model specifications, we ran several robustness checks.

For the primary outcome of task accuracy, we used the following two checks:

1. Equivalent specification to the main analysis but using a quasi-binomial model. This was because the primary outcome consisted of a small number of binary components (setting choice either correct or incorrect) and a linear model may not have provided a good fit, for instance if the distribution of residuals was very skewed. In such cases, a quasi-binomial model is expected to fit the data better and potentially have higher statistical power than a linear model.

2. A linear model where raw task accuracy is replaced with 'accuracy increase'. This was defined as the difference between the actual achieved score and either the default score (in the control design) or the average score achieved by a 'minimum effort' approach (i.e. random button clicking). This was to make sure that our results were not driven by how easy or hard it was to achieve a particular score based on the

selected persona. Although we designed the task so that the default/'minimum effort' scores were as similar as possible for all persona-design combinations, as Table 2 below shows, there was still some variation we weren't able to eliminate.

Table 2: Average scores achieved in each design by a "minimum effort"/"button mashing" approach

Smartphone			
	Persona 1	Persona 2	Persona 3
Control	50%	50%	50%
1A Slider	50%	58%	50%
1B Private mode	62%	50%	37%
1C Trusted third party	42%	50%	58%
Web browser			
Control	33%	33%	33%
2A Graduated control options	50%	50%	50%
2B Four-box grid	50%	50%	50%
Social media			
Control	50%	50%	50%
3A Filtering slider	58%	33%	33%
3B Private mode	50%	50%	50%
3C Responsive toggles	50%	50%	50%

For the secondary outcome of understanding of consequences, we again replaced the linear model with a quasi-binomial one, with the same reasoning as in the primary analysis robustness check.

Finally, for feelings of control, we ran a series of three Mann-Whitney *U* tests to test the null hypothesis that data in the treatment designs come from the same distribution as the control design data. In our main analysis, we treated the rating-scale 'feelings of control' question as if it was a continuous variable. However, it was actually an ordinal variable with potentially unequal distances between answer options. As such, a linear model may result in misleading inference. The Mann-Whitney *U* test, in contrast, is a non-parametric test that only treats data points based on their *rank*, without making assumptions about the distances between answer options.

4. Implementation

The three trials were implemented as planned, except for some minor changes made to the pre-specified trial protocol following the pilot results. Most of these changes related to minor changes to the wording of the questions and the response options to improve clarity and remove potential ambiguity.

In addition, in Trial 1 we needed to make a decision about what to count as “correct” settings in the task. Specifically, there were two settings (location tracking by apps and app notifications) that had a single toggle in some trial designs but four single-app toggles in other designs. This means that satisfying a persona’s preferences – for instance, not to receive notifications – required less interaction with the UI in some designs than in others.

There are two scoring approaches we could take in designs with app-level settings:

- Strict scoring: Require the setting to be changed for all relevant apps
- Lenient scoring: Require at least one app’s settings to be changed

Neither of these is clearly better, given the task instructions and the set-up. However, since our goal was to contrast performance across the different designs, we decided that the lenient scoring provided a fairer comparison. This choice was also supported by the fact that the text describing the personas’ preferences may have been partly ambiguous: Persona 3 (‘most concerned’) had a preference not to receive notifications from “email and messaging apps” but it is debatable whether Instagram counts as a messaging app. By treating unsubscribing from at least one email or messaging app (Instagram, WhatsApp, Gmail) as correct, we ensured that we did not unfairly disadvantage some participants.

5. Smartphone trial results

5.1 Sample characteristics and balance checks

Table 3 below shows a breakdown of the sample characteristics compared with the distribution in the general UK population. Quotas were applied for gender, age, household income, and location; no quotas were used for education and ethnicity and desired quotas were met. Note that the table only includes those who were eligible to participate.

Table 3: Sample characteristics: Smartphone trial

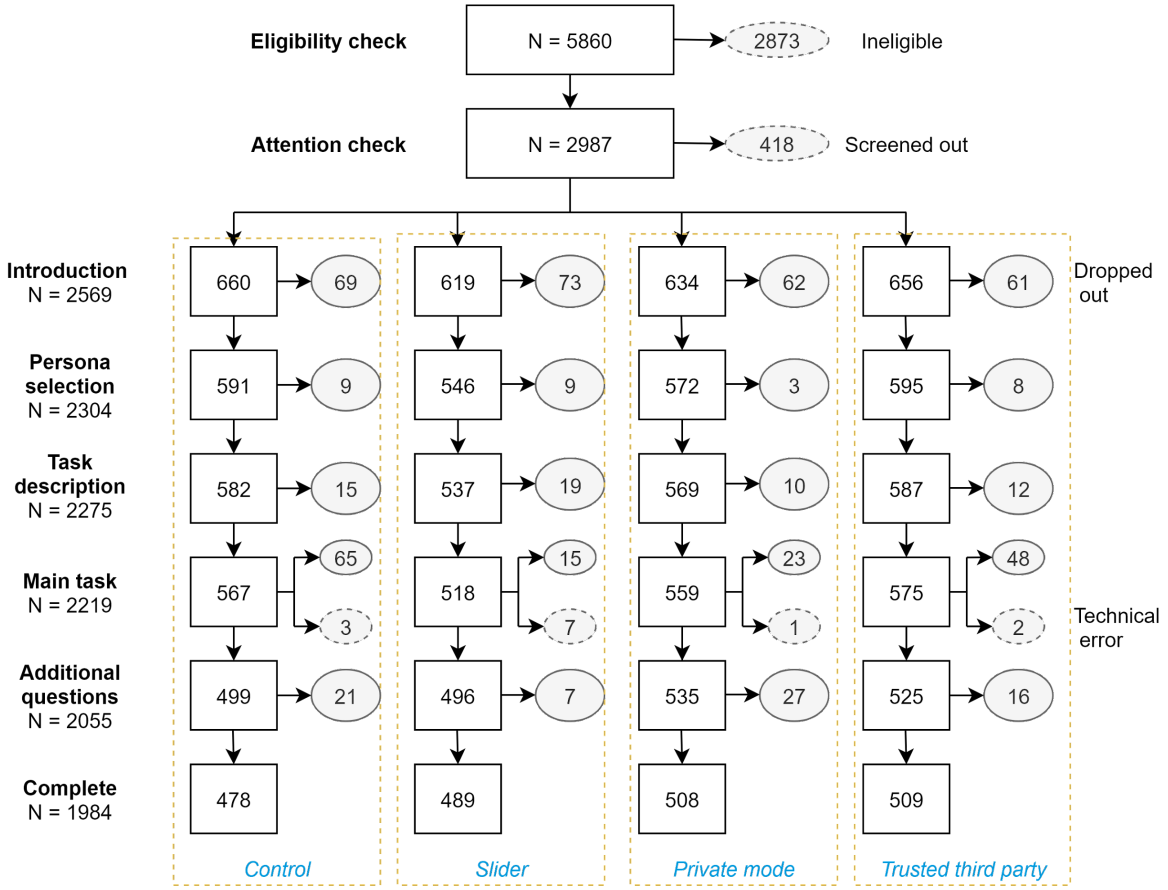
	Final sample	Target
Total n	1,984	2,000
Gender	**%**	**%**
Male (n=1,013)	51	50
Female (n=968)	49	50
Other (n=3)	<1	-
Age		
18-24 (n=246)	12	12
25-54 (n=1,124)	57	52
55+ (n=614)	31	36
Household Income		
Less than £30,000 (n=1,053)	53	50
More than £30,000 (n=931)	47	50
Location		
London (n=254)	13	13
North (n=487)	25	23
South & East (n=570)	29	32
Midlands (n=362)	18	16
Wales, Scotland & N.Ireland (n=311)	16	16
Education Level		

No degree (n=1,449)	73	74
Degree (n=517)	26	26
Prefer not to say (n=18)	1	-
Ethnicity		
White (n=1,726)	87	86
Black (n=66)	3	3
Asian (n=116)	6	7
Other (n=76)	4	3

Differential attrition and balance checks

We collected data from 5,860 individuals, of whom 49% were ineligible and a further 7% failed our attention check. Of the remaining 2,569 participants, 585 (23%) either dropped out at some point during the experiment (572) or encountered a technical error (13) where participants' settings were not being saved. The CONSORT diagram below provides a detailed breakdown.

Figure 2: CONSORT diagram of the numbers of participants present at the various stages of the online experiment.



We further checked for differential attrition. We ran two models, one looking at attrition for those who dropped out at any point and one looking at attrition amongst those who had seen the designs. In both cases, we found that participants were less likely to drop out in the treatment designs than in the control – by 2 to 5 percentage points – which was likely due to the fact that the control design featured a more complicated user interface and the task took more time and effort (as measured by number of clicks) to complete. The differences in drop-out rates between the treatment designs were not statistically significant.

This means that there is a small chance that the samples of participants who completed the trial were systematically different in different designs, as there was somewhat more self-selection in the control design than in the other designs. However, the participants who dropped out were likely the less committed ones. In consequence, the leftover participants in the control design were, on average, likely to be somewhat more committed and attentive than those in the other designs. As such, our estimate of the performance of the control design is more likely to be an overestimate than an underestimate, so it is unlikely that this differential attrition is driving the treatment effects we observed.

5.2 Descriptive statistics

Table 4 compares summary statistics between designs, including outcome variables, time taken and clicks taken to complete the task. As expected, we found that on average, participants took more time and effort to complete the task in the control design than in the treatment designs. Participants in the treatment designs performed better on all of the outcome variables (task accuracy, feelings of control and understanding of consequences), with one exception: the feelings of control in the trusted third party design was the same as in the control design.

Table 4: Descriptive statistics for time taken, number of clicks, and the primary and secondary outcomes. Note that all outcomes have been converted to percentages.

Design	Time spent on task (Median + Q1 and Q3*)	N clicks (Median + Q1 and Q3*)	Task accuracy (Mean + SD)	Feelings of control † (Mean + SD)	Understanding of consequences (Mean + SD)
Control design (n = 478)	109 s (61 - 175 s)	23 (9 - 39)	0.66 (0.21)	2.24 (0.96)	0.53 (0.29)
Slider (n = 489)	66 s (45 - 97 s)	7 (5 - 9)	0.87 (0.21)	2.44 (0.92)	0.66 (0.24)
Private mode (n = 508)	49 s (29 -75 s)	4 (2 - 7)	0.79 (0.21)	2.40 (0.96)	0.66 (0.25)
Trusted third	65 s (45 - 100 s)	7 (4 - 11)	0.80 (0.24)	2.24 (0.96)	0.65 (0.26)

party (n = 509)					

Note: accuracy and understanding of consequences are expressed as proportions (1 = full score)

*Q1 (first quartile) is the 25th percentile; Q3 (third quartile) is the 75th percentile

†0 = “no control”, 4 = “complete control”

Table 5 shows persona selection among eligible participants, alongside mean scores for the three outcome variables. Persona 2 (‘partly concerned’) was the most popular choice, selected by almost two-thirds of participants. Those who chose Persona 3 (‘most concerned’) tended to perform worse and in terms of accuracy and understanding. Those who chose Persona 2 or 3 gave a lower rating for their feeling of concern. We return to these observations in the ‘Subgroup analysis by persona’ section.

Table 5: Proportion of participants choosing each persona.

	Proportion of participants	Task accuracy (Mean + SD)
Persona 1	25%	82% (0.26)
Persona 2	62%	78% (0.20)
Persona 3	13%	73% (0.28)

5.3 Task accuracy

Participants achieved significantly higher accuracy on the main task (changing settings in line with persona preferences) in the treatment designs compared to the control. The slider design showed the highest increase, 21pp compared to the control design. The private mode design and the trusted third party design performed similarly, increasing accuracy by around 14pp. Comparing the unadjusted (1a) and adjusted (1b) models, we see that model 1b has slightly higher R^2 and higher AIC, meaning the models provide a very similar fit to the data. Our graphs and all later analyses are based on adjusted model specifications.

Table 6 also shows that mean performance varied by persona, with participants selecting Persona 2 (‘party concerned’) or Persona 3 (‘most concerned’) performing worse than those who selected Persona 1 (‘least concerned’). We return to these findings in the ‘Subgroup analysis by persona’ section.

Table 6: Primary analysis – main analysis (linear models with Huber-White standard errors, adjusted for 5 comparisons)

Coefficient	Model 1a	Model 1b
1A: Slider	0.209** (0.014)	0.210** (0.014)
1B: Private mode	0.135** (0.013)	0.134** (0.013)

1C: Trusted third party	0.141** (0.015)	0.140** (0.014)
Persona 2	-0.040** (0.012)	-0.040** (0.013)
Persona 3	-0.091** (0.020)	-0.089** (0.020)
Other covariates	No	Yes
Adjusted R^2	0.117	0.121
AIC	-438.8	-437.5
Observations	1,984	1,984

Table shows the effects of the treatment designs (compared to the control design) on a linear 0-1 scale (1 = all four settings correct). In parentheses are the standard errors of these effects.

Figure 3 below visualises the results from the primary regression model (model 1b). We extended the pre-specified analysis by comparing all designs with each other, using the Tukey test for post-hoc contrasts to adjust for the additional comparisons. The slider design performed significantly better than all other designs. The private mode design and the trusted third party design did not significantly differ in their average accuracy.

Figure 3: Primary analysis – effect of treatment design assignment on task accuracy

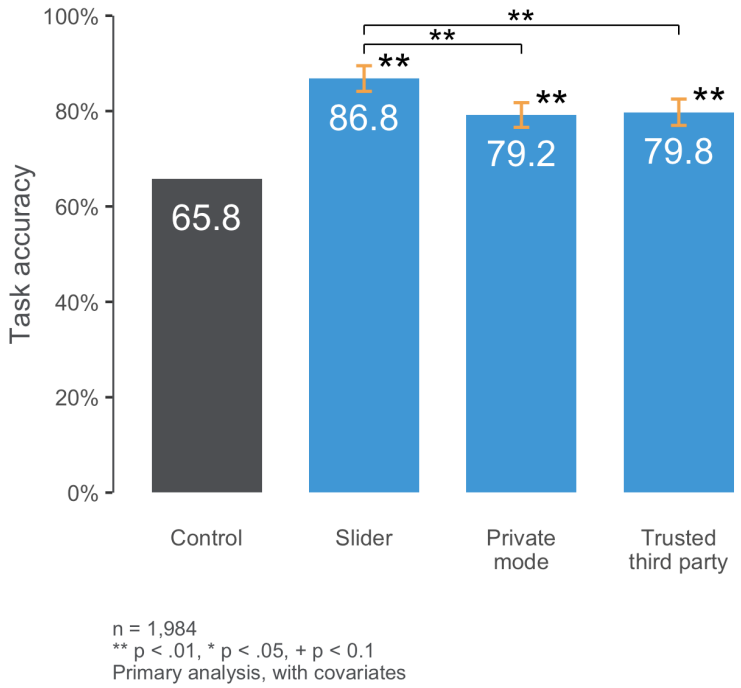


Figure 4 shows the performance of each design in more detail. All designs led to a large decrease of 50% scores (i.e. 2 out of 4 correct) compared to the control design (yellow bar in Figure 4). However, while over 60% of participants in the slider design achieved a full score, only 38% and 43% did so in the private mode and trusted third party designs, respectively.

Moreover, the trusted third party design had a relatively high proportion (4%) of zero scores, a point that we return to in the exploratory analysis.

Figure 4: Distribution of accuracy scores across treatments

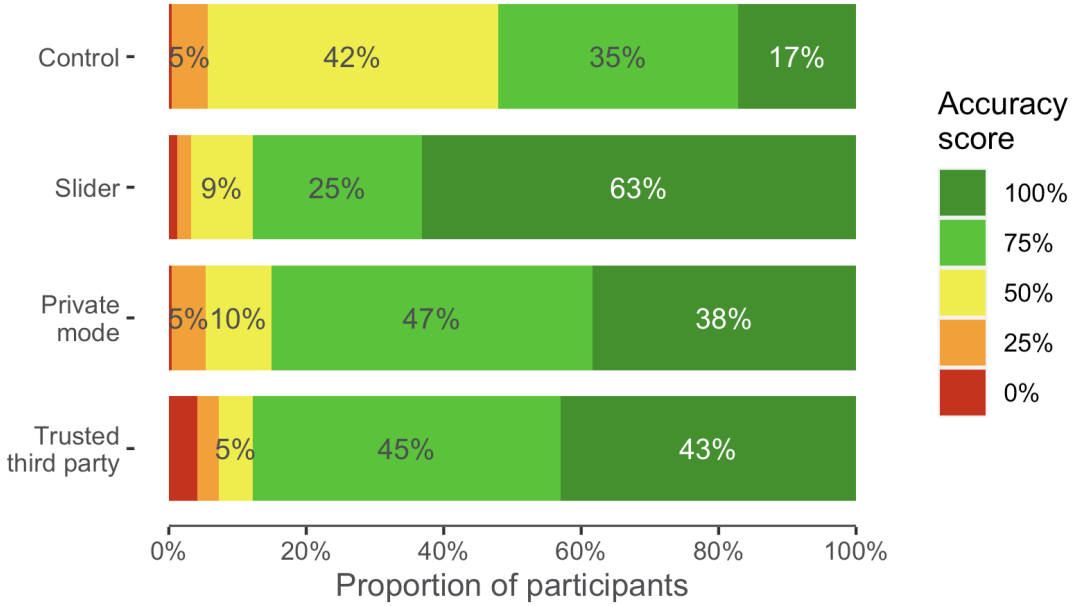


Table 7 shows how accuracy varied with different demographic characteristics. Overall, we found little variation. The only significant differences were by age (18-24-year-olds performing better than 25-54-year-olds and 55+-year-olds, and income (above-median earners performing 3pp better than below-median earners).

Table 7: Associations between task accuracy (%) and the covariates. p-values are the results of univariate linear regressions with Huber-White standard errors.

	Accuracy (%)	**p-value** compared to reference group
Total sample	**78%**	
Gender		
Female (n=968)	78%	Reference group
Male (n=1013)	79%	p > .10
Other (n=3)	67%	*Sample too small*
Age		
18-24 (n=246)	81%	Reference group
25-54 (n=1124)	77%	p < .05*
55+ (n=614)	78%	p < 0.10^{+}

Household Income		
Less than £30,000 (n=1053)	77%	*Reference group*
More than £30,000 (n=931)	80%	p < 0.01**
Location		
London (n=254)	77%	Reference group
North (n=487)	78%	p > .10
South & East (n=570)	79%	p > .10
Midlands (n=362)	77%	p > .10
Wales, Scotland & N.Ireland (n=311)	77%	p > .10
Education Level		
No degree (n=1449)	78%	Reference group
Some degree (n=517)	79%	p > .10
Prefer not to say (n=18)	73%	*Sample too small*
Ethnicity		
White (n=1726)	78%	Reference group
Black (n=66)	75%	p > .10
Asian (n=116)	76%	p > .10
Other (n=76)	80%	p > .10

Primary analysis robustness check

We ran two robustness checks (see Table 8). The first check used a quasi-binomial model. It showed results consistent with the main analysis: all designs significantly increased accuracy compared to the control, with the slider design resulting in the greatest increase.

The second robustness check used as the outcome variable the *accuracy increase* from a default/minimum-effort baseline, instead of the raw accuracy score. In this case, the slider and trusted third party designs performed similarly well whereas the private mode design performed slightly worse than these two designs but still better than control (see Figure 5). The reason is that a “button mashing” approach would have resulted in different average scores across the three designs: 55.1% for the slider design, 51.5% for the private mode

design, and 49.0% for the trusted third party design[1], resulting in increases of roughly 86.8 – 55.1 = 31.7, 79.2 – 51.5 = 27.7, and 79.8 – 49.0 = 30.8 percentage points, respectively.[2]

This implies that the patterns seen in our main analysis, with the slider design outperforming the other two designs (which performed comparably) is not robust to this change in the specification of the outcome variable. Although neither model is clearly right or wrong, we used the main analysis for reporting. This is because the calculation of the "accuracy increase" score depends on certain assumptions that probably weren't met – for instance, that inattentive participants would have clicked on buttons at random, whereas, as shown in Table 10, very few people clicked on the mental health charity logo in the trusted third party design.

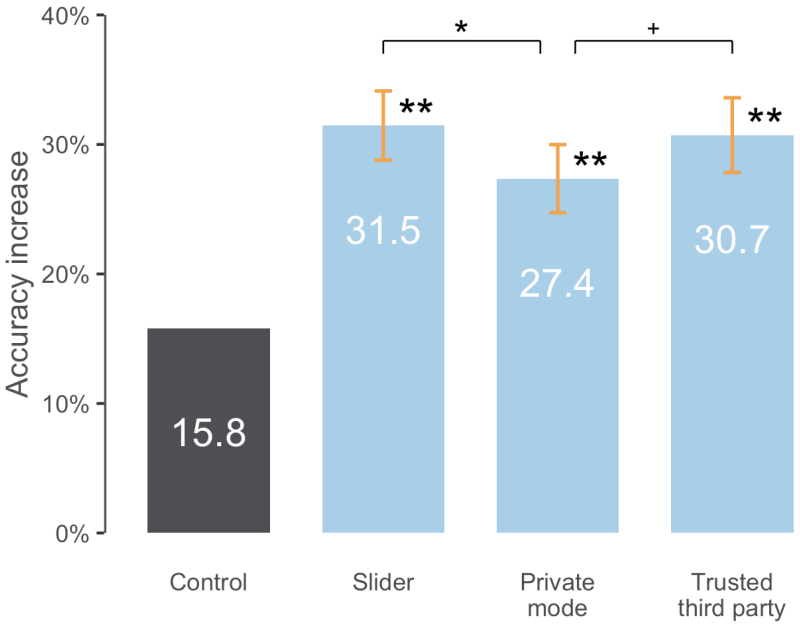
Table 8: Primary analysis robustness checks. Adjusted for 5 comparisons.

Coefficient	(1) Quasibinomial model, exponentiated coefficients (-1 SD, + 1 SD)	(2) Linear model of accuracy increase (SD)
1A: Slider	3.44** (-0.30, +0.33)	0.157** (0.014)
1B: Private mode	1.99** (-0.16, +0.17)	0.112** (0.014)
1C: Trusted third party	2.06** (-0.16, +0.18)	0.149** (0.014)
Persona 2	0.77** (-0.07, +0.06)	-0.048** (0.012)
Persona 3	0.59** (-0.06, +0.06)	-0.066** (0.017)
Other covariates	Yes	Yes
Adjusted R^2	-	0.077
Observations	1,984	1,984

Table shows the regression coefficients from the two models. In parentheses are the standard errors of these coefficients.

[1] These exact figures depend on how many participants selected each persona so we couldn't have calculated them prior to data collection.
[2] Exact numbers differ from those in Figure 5 due to rounding errors and covariate adjustment.

Figure 5: Primary analysis robustness check (check #2).



n = 1,984
** p < .01, * p < .05, + p < 0.1
Primary analysis robustness check, with covariates

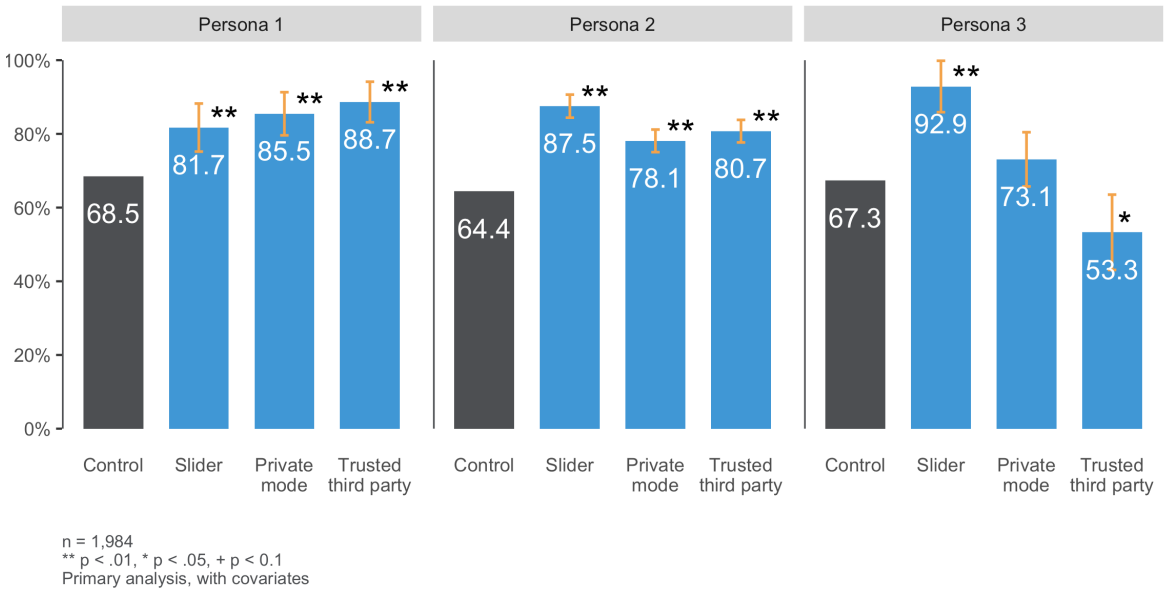
Subgroup analysis by persona

In an exploratory follow-up analysis, we repeated the main regression modelling separately for participants choosing each of the personas. Figure 6 visually presents the results. We make several observations:

- The slider design showed relatively stable average performance across all personas.
- The private mode design performed no better than control for Persona 3 (‘most concerned’) because most participants failed to select bundled notifications. That’s likely the consequence of the notification option being only shown on the customisation screen. Participants with Persona 3 who selected ‘Private’ and then submitted would’ve got this setting wrong. It is possible that because the design was so simple, many users did not feel compelled to customise and just accepted one of the pre-selected bundles.
- The trusted third party design did very well for Persona 1 (‘least concerned’) and poorly for Persona 3 (‘most concerned’). This was affected by the fact that the majority of participants chose the set of recommendations from a well-recognised technology company across all personas (see Table 10). The preferences of Persona 1 (‘least concerned’) were highly aligned with recommendations from the technology company, while the preferences of Persona 3 (‘most concerned’) were the least aligned. This means that participants who chose Persona 1 (‘least concerned’) and the bundle from the technology company needed very few manual adjustments to get the highest score, while those with Persona 3 (‘most concerned’) needed a lot more

adjustments[3]. In line with this explanation, those who initially chose the technology company achieved an average accuracy of 47.7% whereas those who chose a consumer organisation or Mind achieved on average 75.0% and 100%, respectively. This result should be treated with caution because of the small number of people who chose the mental health charity and the consumer organisation for Persona 3 (3 and 24 people respectively).

Figure 6: Subgroup analysis – task accuracy by persona.



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 9: OLS models of task accuracy as a function of treatment, split by persona

	(1) Persona 1	(2) Persona 2	(3) Persona 3
1A: Slider	0.13**	0.23**	0.26**
	(0.03)	(0.02)	(0.04)
1B: Private mode	0.17**	0.14**	0.06
	(0.03)	(0.02)	(0.04)
1C: Trusted third party	0.20**	0.016**	-0.14*
	(0.03)	(0.02)	(0.05)
Raw control mean	0.68	0.64	0.67
Covariates	Yes	Yes	Yes
Observations	500	1,220	264

[3] Participants could have also selected another organisation's recommendations after clicking on one of them (this wasn't tracked). However, based on how the final scores depended on the initial organisation selection, we suspect that not many participants did that.

Adjusted R^2	0.08	0.18	0.24

Table 10: Number of users choosing each of the trusted parties by persona

Third-party choice[4]	A large technology company	A mental health charity	A consumer organisation	Total users
Persona 1	92 (70%)	3 (2%)	36 (28%)	**131**
Persona 2	180 (58%)	14 (5%)	114 (37%)	**308**
Persona 3	43 (61%)	3 (4%)	24 (34%)	**70**
Total users	**315**	**20**	**174**	**509**

Task accuracy - scores for individual settings and for each persona

Table 11 further breaks down this performance by individual settings. We observe that:

- The control design had comparatively low scores for notifications and private-browsing settings. This is because the default for these settings was always different from the persona's preferences, resulting in 2 out of the 4 settings always requiring manual customisation in this design. We had made this choice in order to achieve similar task difficulty for all personas, across all designs.[5]
- The slider design performed very well for all settings, though not as well on notifications (which was still better than control). This is likely due to the fact that the persona's preferences were expressed in terms of *states* (how many notifications they wanted to receive) but the interface's choices are expressed as a *change* (e.g. "reduce the frequency of notifications").
- In the private mode design, participants performed worse for the notifications (71%) and personalised ads (72%) compared to the private browsing (87%) and location tracking (86%).

[4] Our mock up designs had names and logos of real organisations however these were used for illustrative purposes and do not constitute any organisation endorsing any design.
[5] The treatment designs did not feature default settings per se (since they contained a series of forced choices) but, as shown in Table 2 in Section 3, the scores achieved by random button pressing were, on average, close to 50%.

- The low score for notifications is driven by Persona 3 ('most concerned') with a score of only 13%. It is possible that this setting was confusing given this persona's preferences description.[6,7]
- The low score for personalised ads is driven by Persona 2 ('partly concerned') with a score of 61%. This persona was the only one that needed to customise a setting from the high-level 'Regular' or 'Private' choice, meaning making the correct choice here required more effort (in terms of clicking through to the customisation display)

Table 11: Mean accuracy of each setting, split by design and persona

Design	Notifications			Personalised ads			Private browsing			Location tracking		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Control design	49%			73%			58%			83%		
	49%	52%	39%	29%	87%	90%	96%	35%	97%	100%	84%	44%
1A: Slider	73%			85%			94%			94%		
	75%	69%	88%	63%	89%	100%	95%	95%	89%	95%	95%	89%
1B: Private mode	71%			72%			87%			86%		
	87%	77%	13%	85%	61%	94%	84%	87%	93%	85%	86%	94%
1C: Trusted third party	89%			67%			87%			76%		
	95%	94%	50%	82%	60%	71%	97%	93%	43%	81%	75%	73%

Note: P1 = Persona 1 = 'least concerned', P2 = Persona 2 = 'partly concerned', P3 = Persona 3 = 'most concerned'

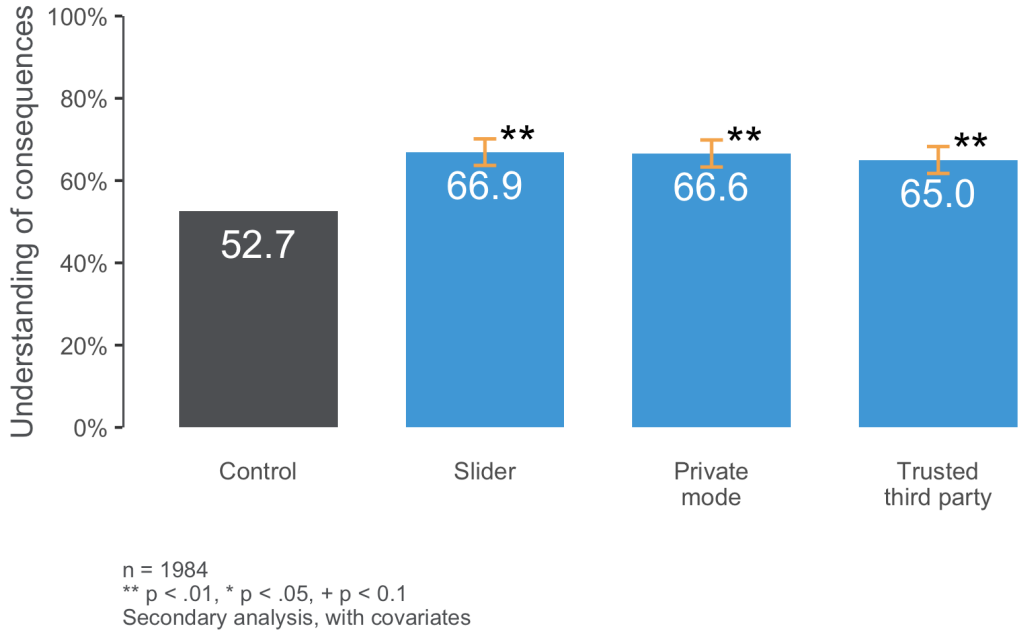
5.4 Understanding of consequences

Participants had significantly better understanding of the consequences of their choices (measured by four comprehension questions) in the treatment designs compared to the control design. Descriptively, the best performing design was the slider, followed by the private mode and the trusted third party. However, our follow-up tests found that differences between the treatment designs were not statistically significant (see Figure 7 and Table 12). The robustness check (using a quasi-binomial model) showed a consistent pattern of results with the main analysis and is thus not presented here.

[6] The persona text said "Alex feels like he receives too many notifications and finds them distracting (especially email and messaging apps)." The correct choice was to bundle notifications. However, the explanatory text for bundling said "All notifications arrive together once an hour, except for calls and messages which are notified immediately."

[7] Our sensitivity analysis suggests that even if the majority of participants had made the correct choice here, it would have only improved the performance of the private mode design by about 2pp and wouldn't have substantively changed our overall findings.

Figure 7: Secondary analysis: Understanding of consequences. Adjusted for 6 comparisons.



n = 1984
** p < .01, * p < .05, + p < 0.1
Secondary analysis, with covariates

Table 12: Secondary analysis: Understanding of consequences. Adjusted for 6 comparisons.

Coefficient	Main analysis - linear model (SD)
1A: Slider	0.143** (0.016)
1B: Private mode	0.140** (0.017)
1C: Trusted third party	0.123** (0.017)
Persona covariates	Yes
Other covariates	Yes
Adjusted R^2	0.094
Observations	1,984

Table 13: Understanding of consequences, unadjusted mean scores for each question, by design.

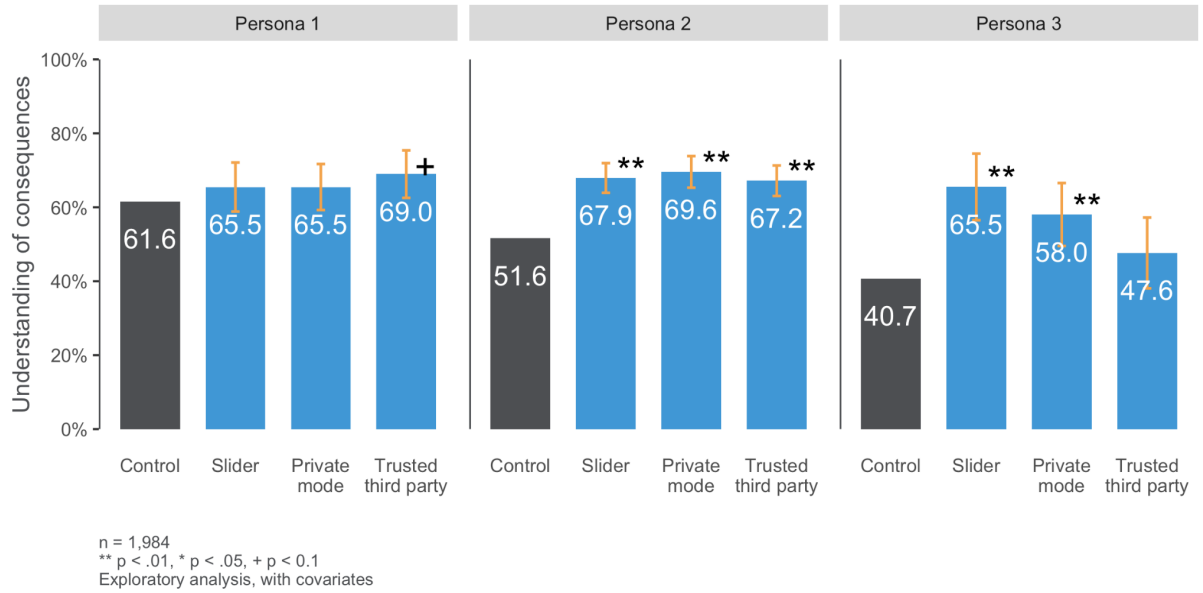
Design	Notifications	Personalised ads	Private browsing	Location tracking	Total
Control	55%	40%	59%	57%	**53%**
1A: Slider	75%	55%	85%	50%	**66%**
1B: Private mode	68%	64%	84%	51%	**67%**
1C: Trusted third party	72%	62%	78%	48%	**65%**

Subgroup analysis by persona

Figure 8 below breaks down the 'understanding of consequences' results by persona and trial design:

- In the control design, understanding of consequences was notably poorer for personas 2 ('partly concerned') and 3 ('most concerned') than for Persona 1 ('least concerned').
- The trusted third party design was the only treatment design that led to a marginal (significant at the 10% level) improvement of understanding for Persona 1 but also the only design that was not significantly better than control for Persona 3.

Figure 8: Subgroup analysis – Understanding of consequences by persona



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 14: OLS models of Understanding of consequences as a function of treatment, split by persona.

	(1) Persona 1	(2) Persona 2	(3) Persona 3
1A: Slider	0.04	0.16**	0.25**
	(0.03)	(0.02)	(0.05)
1B: Private mode	0.04	0.17**	0.17**
	(0.03)	(0.02)	(0.04)
1C: Trusted third party	0.07+	0.16**	0.07
	(0.03)	(0.02)	(0.05)
Raw control mean	0.62	0.52	0.41
Covariates	Yes	Yes	Yes
Observations	500	1,220	264
Adjusted R^2	0.01	0.12	0.13

Table shows the effect sizes of the treatment designs (compared to the control design) on the understanding score, measured on a linear 0-1 scale (1 = all four understanding questions answered correctly). In parentheses are the standard errors.

5.5 Feelings of control

Participants reported significantly higher feelings of control (measured by a single sentiment question) in two of the treatment designs, slider and private mode, compared to the control. The trusted third party design did not perform better than the control on this outcome measure. Descriptively, the best performing design was the slider design, followed by the private mode design. In pairwise post-hoc comparisons (using the Tukey adjustment), the slider arm outperformed the control arm and the trusted third party arm but not the private mode arm. The robustness check (using non-parametric test) showed a consistent pattern of results with the main analysis and is thus not presented here.

Figure 9: Effect of treatment on feelings of control

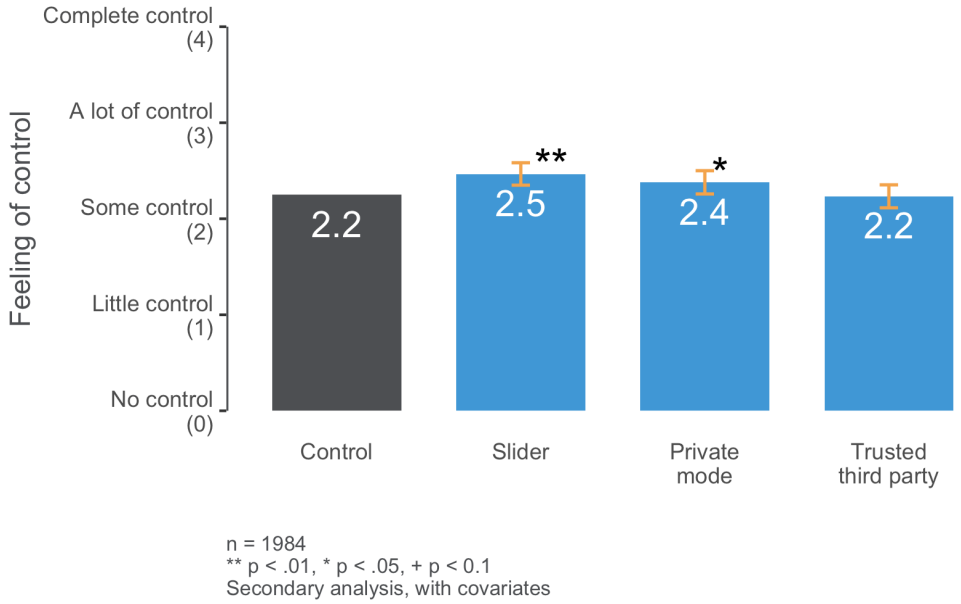


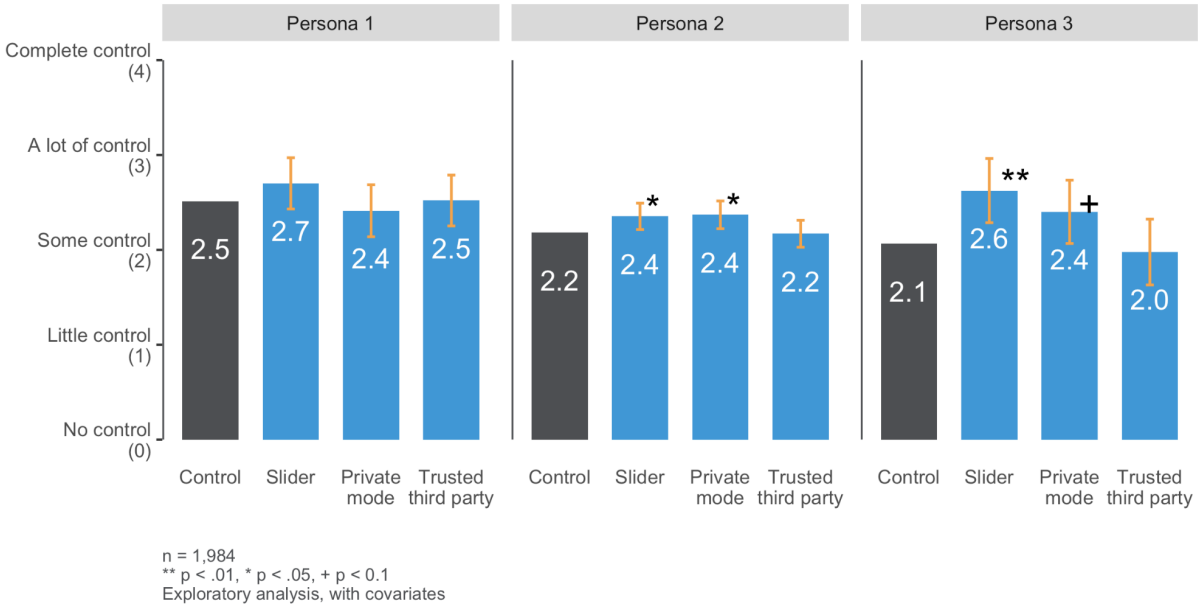
Table 15: Secondary analysis, feelings of control. Adjusted for 6 comparisons.

Coefficient	Main analysis - linear model (SD)
1A: Slider	0.22** (0.06)
1B: Private mode	0.13* (0.06)
1C: Trusted third party	-0.02 (0.06)
Persona covariates	Yes
Other covariates	Yes
Adjusted R^2	0.020
Observations	1,984

Subgroup analysis by persona

Figure 10 below presents the 'feelings of control' results separately for each persona. The slider and private mode designs outperformed the control and trusted third party designs for personas 2 and 3. There were no significant differences for Persona 1.

Figure 10: Subgroup Analysis - Feelings of control by persona



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 16: OLS models of feelings of control as a function of treatment, split by persona

	(1) Persona 1	(2) Persona 2	(3) Persona 3
1A: Slider	0.18	0.17*	0.56**
	(0.14)	(0.07)	(0.17)
1B: Private mode	-0.10	0.19*	0.34+
	(0.14)	(0.07)	(0.17)
1C: Trusted third party	0.08	-0.01	-0.09
	(0.14)	(0.07)	(0.18)
Raw control mean	2.51	2.18	2.06
Covariates	Yes	Yes	Yes
Observations	500	1,220	264

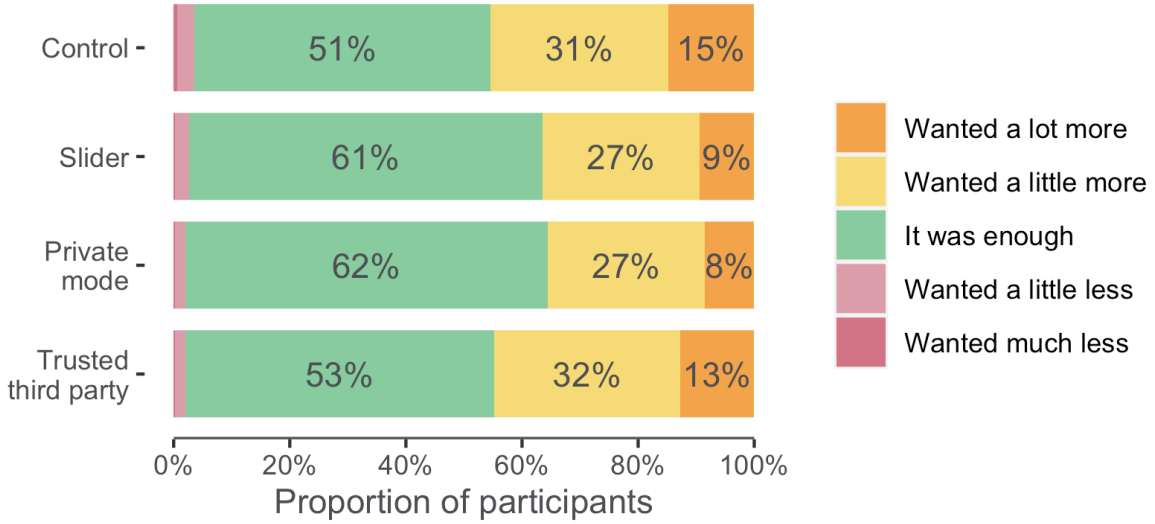
Adjusted R^2	0.002	0.01	0.04

Additional (comparative) feelings of control question

We additionally asked participants whether they had *as much control as they wanted* ("Did you have as much control over the settings as you would have liked when making choices for Alex?"). Figure 11 presents the descriptive results.

Across all designs, the majority of participants reported having "enough control". However, c. 45% of participants in the control and trusted third party design wanted "a little more control" or "a lot more control" compared to c. 36% in the slider and private mode designs.

Figure 11: Comparative feelings of control by treatment



5.6 Interaction variables

This analysis was only conducted for Trial 1. It was purely exploratory and we did not run it for Trials 2 and 3 because of budget and time limitations.

We measured how much time it took our participants to finish the task and how many clicks on buttons within the interface they needed to make. It took participants significantly less time to complete the task in all of the tested designs compared to the control design. Additionally, participants interacting with the private mode design needed significantly less time than participants using the slider and trusted third party designs.

Figure 12: Time spent on task (in seconds). The thick horizontal lines show the medians of each group, the boxes show the 25th and 75th percentiles, the 'whiskers' (vertical lines) show data lying within 1.5 times the height of the boxes (from the boxes' edges) and the dots show data points outside this range. Significance tests are based on a median regression.

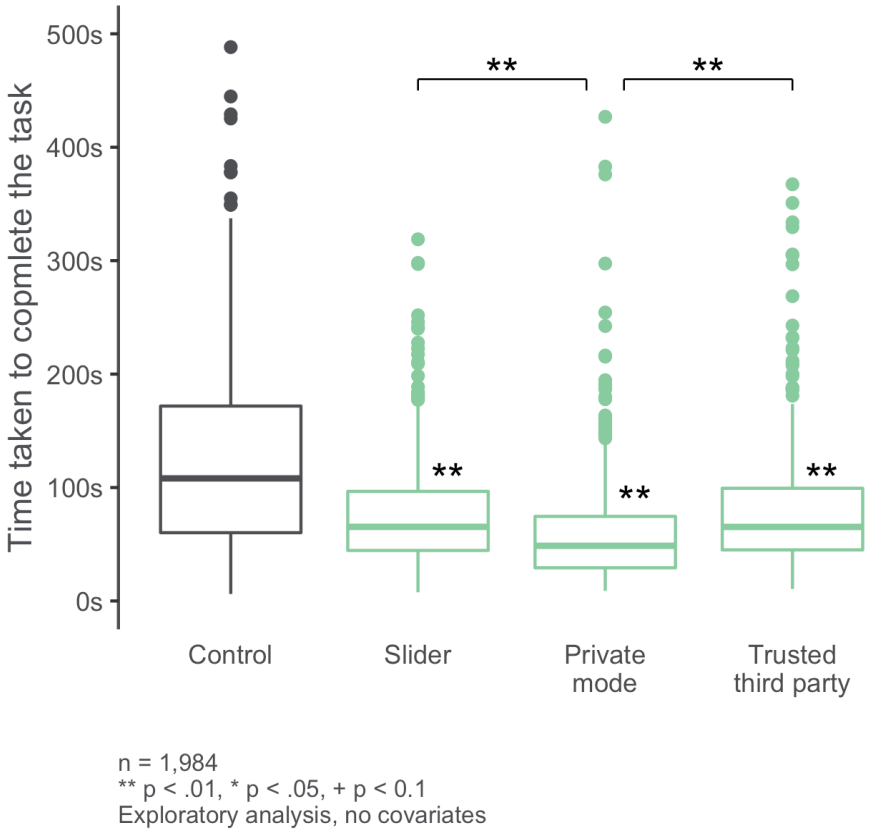
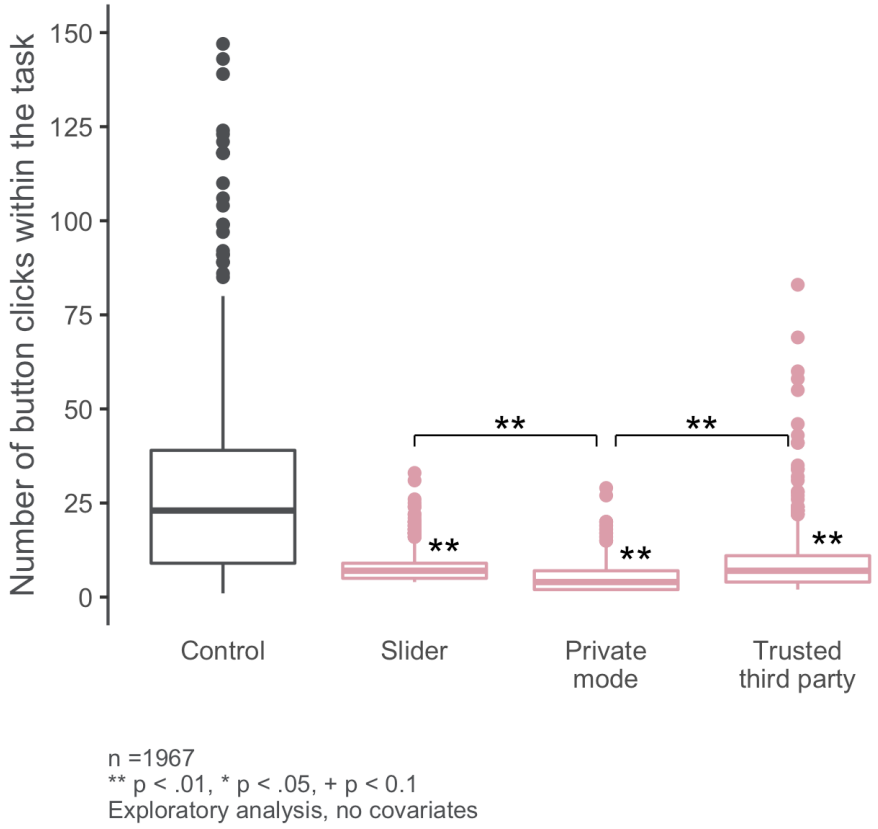


Figure 13 shows the equivalent results for the number of clicks in the task. We can see that there is a large difference between the control design and the other designs: the median number of clicks in the control design was 23 (see Table 4) with some participants requiring more than 100 clicks; in the treatment designs, the medians were never greater than 7, and few participants needed more than 50 clicks. The private mode design required significantly fewer clicks than the other designs (in line with the findings on time taken). The trusted third party design had somewhat more outliers (who clicked on buttons more than 50 times) than the slider and private mode designs. However, its results weren't statistically significant from those of the slider design.

Figure 13: Number of button clicks within the task. The thick lines show the medians of each group, the boxes show the 25th and 75th percentiles. Significance tests are based on a median regression.



Taken together, these findings suggest that our tested designs not only achieved higher accuracy and understanding but also did so more efficiently for users. The private mode design, which had the simplest interface, required the least user interaction.

5.7 Additional results

Other questions about the interfaces

We also asked participants 6 additional rating-scale questions and 2 (optional) free-text questions. The results from the rating-scale questions are summarised in Table 17 below. To construct it, we converted the answers from the 5-point answer scales to scores from 1 to 5 and calculated the mean score.

The slider and private mode designs were joint best-performers on all questions, except for the “Would you like to see an interface like this one in future?” question where the control design was also a joint best-performer.

Table 17: Summary of results for all the additional rating-scale questions.

Average answer to the following question*	**Control** (n=478)	**Slider** (n=489)	**Private mode** (n=508)	**Trusted third party** (n=509)
Was it easy or hard to make choices on behalf of Alex?	3.42	3.88	3.88	3.60
Do you trust the choices were presented with your best interests in mind?	3.02	3.22	3.15	3.10
Were the choices presented in a fair way, without trying to influence you?	3.17	3.47	3.40	3.27
Were the choices explained in easy-to-understand terms?	3.01	3.54	3.46	3.23
Would you like to see an interface like this one in future?	3.78	3.86	3.80	3.58
Would you choose to use this interface rather than the current Android settings screens?	-	3.60	3.60	3.37

*Green indicates the jointly best performing arms for each outcome (at the 5% significance level).

Concern about technology & digital comfort

At the end of the experiment, we asked participants a set of four questions about their concern about technology and one question about their comfort using digital technologies.

Participants answered the *concern about technology* questions on a 5-point scale from “1 (not at all concerned)” to “5 (very concerned)”. As expected, those who had chosen Persona 2 (‘partly concerned’) tended to be more concerned than those who had chosen Persona 1 (‘least concerned’), and those who had chosen Persona 3 (‘most concerned’) tended to be the most concerned (see Table 18). Based on overall scores, participants tended to be most concerned about fake news or disinformation, and companies selling their data.

Table 18: Mean answers by persona to the four sub-questions of the questions “How concerned, if at all, would you say you are about each of the following?”

sub-question	**Persona 1** (n=500)	**Personal 2** (n=1220)	**Persona 3** (n=264)	**Overall** (n=1984)
How addictive technology can be	2.16	2.48	2.73	**2.44**
Decisions being made about individuals by artificial intelligence	2.25	2.73	3.13	**2.66**
Companies selling on data about me	2.23	3.03	3.46	**2.89**
Fake news or disinformation online	2.66	3.04	3.12	**2.96**

For the *digital comfort* question, we converted the 5-point verbal answer scale (from “Not at all comfortable” to “Completely comfortable”) to a numerical 1-5 scale. Results by persona are shown in Table 19. Those who chose Persona 1 tended to be the most comfortable, and those who chose Persona 3 tended to be the least comfortable.

Table 19: Mean answers by persona to the question “To what extent are you comfortable using a computer, a tablet or a smartphone to access government or commercial services online?”

	Persona 1 (n=500)	**Personal 2** (n=1220)	**Persona 3** (n=264)	**Overall** (n=1984)
Digital comfort	3.79	3.64	3.29	**3.63**

We next looked at the relationship between concern about technology, digital comfort, and task accuracy. We calculated a *total concern score* by adding up answers to the four sub-questions and rescaling to 0-1. Table 20 shows the results of exploratory regression analysis of the relationship between this total concern score and accuracy in the main task. Rather surprisingly, even after controlling for trial design assignment, persona, design-persona interaction, and the other collected covariates, there was a significant negative relationship between concern and accuracy of -6.5pp. Given that for the majority of participants’ concern lay between 0.40 and 1.00 (see Figure 14), this translates into roughly 3.9pp lower performance for participants with a concern score of 1.00 compared to (otherwise similar) participants with a concern score of 0.40.

Similarly, we found that those who reported being more comfortable using digital technologies scored better in the main task. Those who reported being “completely comfortable” are estimated to have scored 3.2×4=12.8pp better than those who were “not at all comfortable” (see Figure 15 for the distribution of answers).

Note, however, that this correlation may be the result of different causal processes:

1. Concern about technology and digital comfort may be causally influencing participants’ performance. This seems plausible for digital comfort improving task performance, though it’s unclear why concern about technology would lead to worse performance.

2. Participants’ performance influences their answers to these questions. Since the questions were asked after the task, it is possible that participants' answers were affected by their perceived performance. For instance, participants who did well and were aware of this would have subsequently rated their digital comfort high. We couldn’t test this hypothesis within this trial; however, in Trial 3 (social media) we asked these questions *before* the task in one half of the sample, allowing us to inspect how much the task affects answers. Indeed, we found that the association of task performance with digital comfort – but not concern about technology –

decreased when asked prior to the task (see Concern about technology & digital comfort in section 7 for details).

3. The constructs (concern and comfort) and task performance may have an unmeasured common cause. For instance, both high concern about technology and low task performance may be the results of low levels of knowledge about digital technologies.

Table 20: Regression table for the analysis of the relationships between task accuracy and (1) the total concern about technology score or (2) the digital comfort score (exploratory analysis). No corrections for multiple comparisons applied.

Term	(1) **Concern model** Coefficient (Huber-White SE)	(2) **Comfort model** Coefficient (Huber-White SE)
Intercept	71.6 (6.5)	60.7 (5.9)
Concern about technology (difference between full and zero score)	-6.5* (2.8)	-
Digital comfort (difference of 1 point on the 5-point scale)	-	3.2** (0.5)
Treatment: Slider	14.0** (3.2)	13.8** (3.2)
Treatment: Private mode	17.1** (2.9)	17.5** (2.9)
Treatment: Trusted third party	16.7** (3.1)	16.9** (3.1)
Persona 2	-3.8+ (2.2)	-3.6+ (2.2)
Persona 3	0.1 (3.2)	0.8 (3.2)
Treatment: Slider * Persona 2	10.1** (3.6)	10.0** (3.6)
Treatment: Private mode * Persona 2	-2.6 (3.3)	-3.1 (3.3)
Treatment: Trusted third party * Persona 2	-6.9+ (3.6)	-7.6* (3.5)
Treatment: Slider * Persona 3	11.1* (4.7)	10.4* (4.6)
Treatment: Private mode * Persona 3	-10.9* (4.5)	-11.2* (4.4)
Treatment: Trusted third party * Persona 3	-30.2** (5.9)	-30.8** (5.9)

Gender: Male	0.2 (1.0)	0.2 (1.0)
Gender: Other	-13.2 (13.0)	-15.1 (12.1)
Age category: 25-54	-5.1** (1.6)	-5.5** (1.6)
Age category: 55 and over	-4.6** (1.6)	-5.5** (1.6)
Income: £30,000 and over	3.0** (1.0)	2.6* (1.0)
Location: Midlands	-0.8 (1.9)	-0.9 (1.8)
Location: North of England	0.1 (1.7)	-0.3 (1.7)
Location: South and East England	1.4 (1.6)	1.0 (1.6)
Location: Wales, Scotland and Northern Ireland	-0.7 (1.9)	-1.3 (1.9)
Education: No degree	2.8 (5.7)	2.0 (5.3)
Education: Degree	2.8 (5.7)	1.8 (5.4)

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$

Figure 14: Distribution of the total digital concern score.

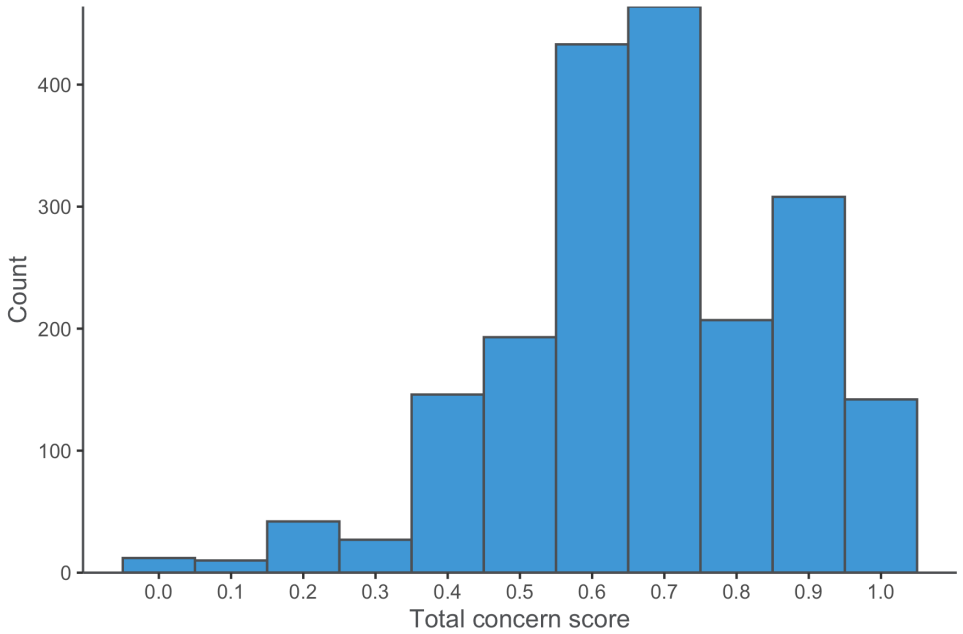
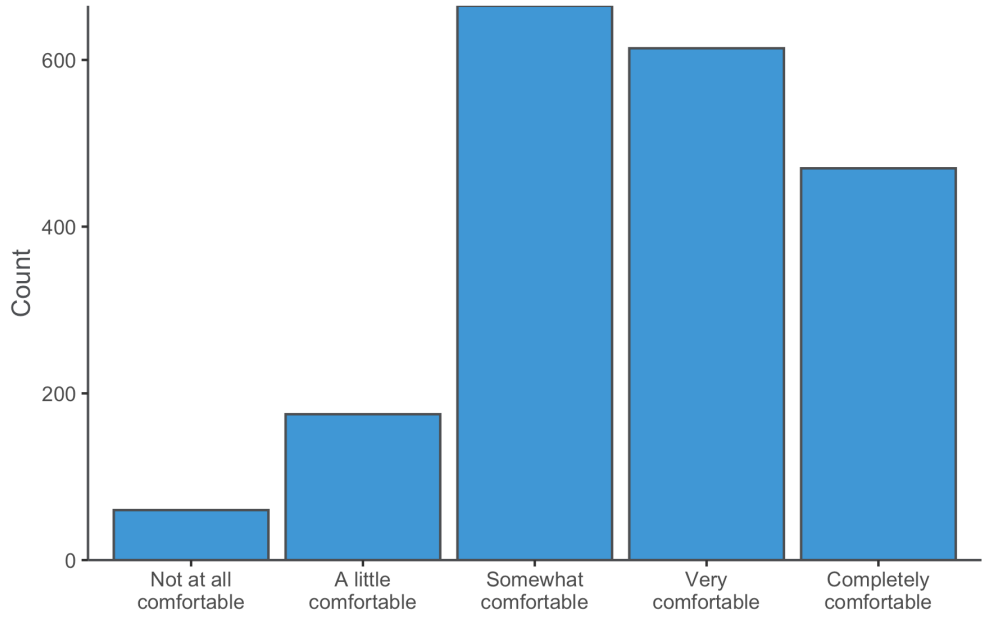


Figure 15: Distribution of comfort with digital technologies.



6. Web browser trial results

6.1 Sample characteristics and balance checks

All participants completed the experiment on a computer or a tablet. Table 21 below shows a breakdown of the sample characteristics compared with the distribution in the general UK population. Quotas were applied for gender, age, household income, and location; no quotas were used for education and ethnicity.

Table 21: Sample characteristics: Web browser trial

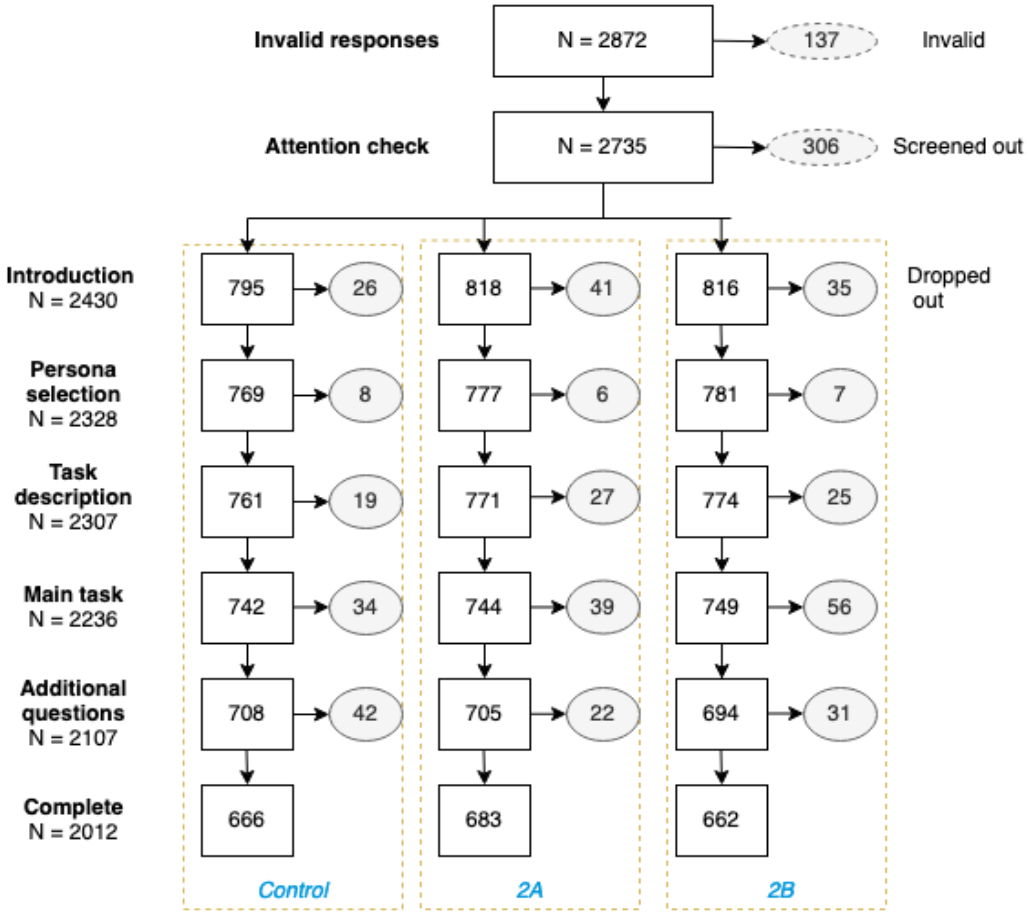
	Final sample	Target
Total n	2,012	2,000
Gender	**%**	**%**
Male (n=995)	49	50
Female (n=1012)	50	50
Other (n=5)	<1	-
Age		
18-24 (n=234)	12	12
25-54 (n=1057)	52	52
55+ (n=721)	36	36
Household Income		
Less than £30,000 (n=1048)	52	50
More than £30,000 (n=964)	48	50
Location		
London (n= 264)	13	13
North (n=457)	23	23
South & East (n=666)	33	32
Midlands (n=348)	17	16
Wales, Scotland & N.Ireland (n=277)	14	16
Education Level		

No degree (n=1325)	66	74
Degree (n=687)	33	26
Prefer not to say (n=18)	1	-
Ethnicity		
White (n=1766)	88	86
Black (n=68)	3	3
Asian (n=113)	6	7
Other (n=65)	3	3

Differential attrition and balance checks

We collected data from 2,872 individuals, of which 137 were invalid responses (e.g. duplicates) and a further 306 failed our attention check. Of the remaining 2,429 participants, 418 dropped out at some point during the experiment. The CONSORT diagram provides a detailed breakdown.

Figure 16: CONSORT diagram of the numbers of participants present at the various stages of the online experiment.



We further checked for differential attrition. We ran two models, one looking at attrition for those who dropped out at any point and one looking at attrition amongst those who had seen the designs. The difference in drop-out rates between were not statistically significant.

6.2 Descriptive statistics

Table 22 shows the summary statistics, including outcome variables, time taken and clicks needed to complete the task. As expected, we find that participants took more time and effort to complete the task in the control design than in the treatment designs. We also find that participants in the treatment arms performed better on the outcome variables.

Table 22: Descriptive statistics for time taken, number of clicks, and the primary and secondary outcomes. Note that all outcomes have been converted to percentages.

Design	Time spent on task (Median + Q1 and Q3*)	N clicks (Median + Q1 and Q3*)	Task accuracy (Mean + SD)	Feelings of control † (Mean + SD)	Understanding of consequences (Mean + SD)
Control arm (n = 666)	66 s (34 - 105 s)	10 (5 - 18)	0.37 (0.29)	2.08 (0.83)	0.42 (0.25)
2A: Graduated settings (n = 683)	43 s (28 - 69 s)	5 (4 - 6)	0.71 (0.32)	2.36 (0.84)	0.58 (0.27)
2B: Four-box grid (n = 662)	54 s (33 - 80 s)	5 (3 - 7)	0.72 (0.30)	2.14 (0.78)	0.58 (0.27)

Note: accuracy and understanding of consequences are expressed as proportions (1 = full score)

*Q1 (first quartile) is the 25th percentile; Q3 (third quartile) is the 75th percentile

†0 = “no control”, 4 = “complete control”

Table 23 shows persona selection among the eligible sample and the mean task performance. Overall, the split between personas was more equal than in Trial 1, although Persona 2 (‘partly concerned’) was still the most popular choice, selected by 48% of participants (62% in Trial 1).

Table 23: Proportion of participants choosing each persona.

	Proportion of participants	Task accuracy (Mean + SD)
Persona 1	30%	53% (35.2)

Persona 2	48%	56% (32.9)
Persona 3	22%	78% (30.2)

6.3 Task accuracy

Participants achieved significantly higher accuracy on the main task in the treatment arms compared to the control. The two treatment arms performed similarly, increasing accuracy by around 35pp. Comparing the unadjusted (1a) and adjusted (1b) models, we see that model 1b has slightly higher R^2 and lower AIC, meaning model 1b provides a slightly better fit. Our graphs and all later analyses are based on adjusted model specifications.

Table 24 also shows that mean performance varied by persona, with participants selecting Persona 3 ('very concerned') performing better than those who selected Persona 1 ('least concerned') or Persona 2 ('partly concerned').

Table 24: Primary analysis - main analysis (linear models with Huber-White standard errors, adjusted for 3 comparisons)

Coefficient	Model 1a	Model 1b
2A (Graduated control options)	0.354** (0.016)	0.353** (0.016)
2B (Four-box grid)	0.353** (0.015)	0.353** (0.016)
Persona 2	0.027 (0.017)	0.026 (0.016)
Persona 3	0.252** (0.019)	0.253** (0.019)
Other covariates	No	Yes
Adjusted R^2	0.309	0.319
AIC	675.6	657.5
Observations	2,011	2,011

Figure 17 visualises the results from the primary regression model (model 1b). We extended the pre-specified analysis by comparing all arms with each other, using the Hochberg-Benjamini test for post-hoc contrasts to adjust for the additional comparisons. The treatment arms performed better than the control, but did not significantly differ in their average accuracy when compared to each other.

Figure 17: Primary analysis – effect of treatment design assignment on task accuracy

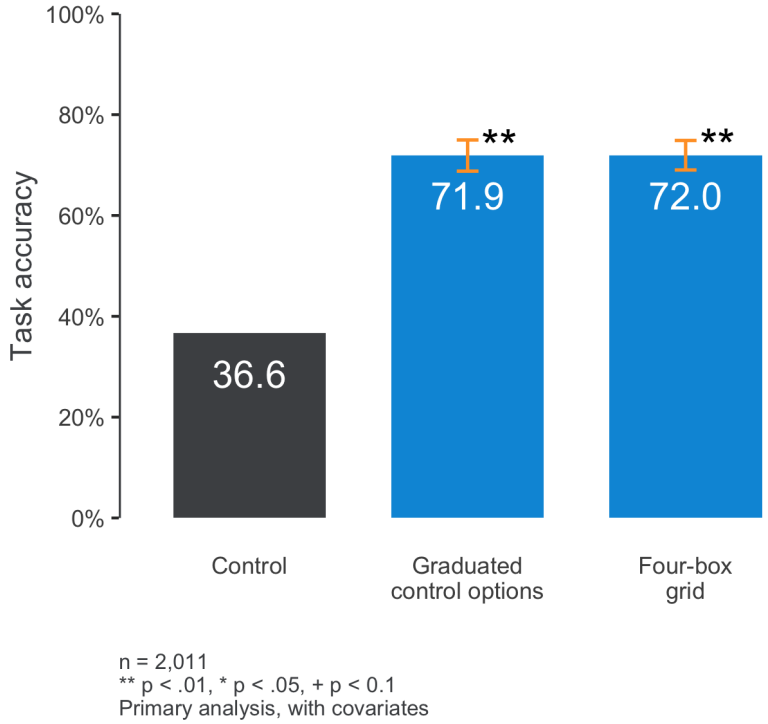


Figure 18 shows the performance of each arm in more detail. We can see that all arms led to a large decrease of 33% scores (i.e. 1 out of 3 correct) compared to the control arm, and also led to an increase in 67% and 100% scores in the two treatment arms.

Figure 18: Distribution of accuracy scores across treatments

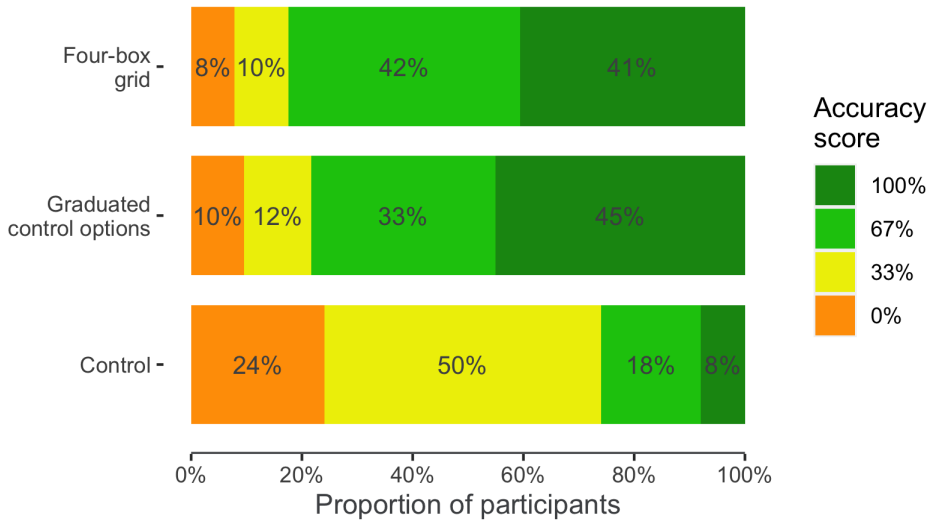


Table 25 shows how accuracy varied with different demographic characteristics. Overall, we found little variation. This analysis was purely exploratory as our research design was not set up to answer whether and why main task accuracy varies across different demographic groups.

Table 25: Associations between task accuracy (%) and the covariates. p-values are the results of univariate linear regressions with Huber-White standard errors.

	Accuracy (%)	**p-value** compared to reference group
Total sample	60%	
Gender		
Male (n=995)	59%	Reference group
Female (n=1012)	61%	p > .10
Other (n=5)	17%	*Sample too small*
Age		
18-24 (n=234)	57%	Reference group
25-54 (n=1057)	61%	$p < 0.10^{+}$
55+ (n=721)	60%	p > .10
Household Income		
Less than £30,000 (n=1048)	60%	Reference group
More than £30,000 (n=964)	60%	p > .10
Location		
London (n= 264)	53%	Reference group
North (n=457)	62%	p < 0.01**
South & East (n=666)	61%	p < 0.01**
Midlands (n=348)	59%	p < .05*
Wales, Scotland & N.Ireland (n=277)	62%	p < 0.01**
Education Level		
No degree (n=1325)	59%	Reference group
Degree (n=687)	62%	p < .05*
Prefer not to say (n=18)	41%	p < .05*

Ethnicity		
White (n=1766)	61%	Reference group
Black (n=68)	53%	$p < .10^{+}$
Asian (n=113)	51%	$p < 0.01^{**}$
Other (n=65)	57%	$p > .10$

Primary analysis robustness check

We ran two robustness checks (see Table 26). The first one replaced the linear-model specification with a quasibinomial model, the second one used a linear model but replaced raw accuracy with *accuracy increase* from a default/minimum-effort score. The findings from these models are consistent with the main analysis. As expected, the model of *accuracy increase* shows a more modest difference between the control and intervention arms, due to the fact that the control arm had a lower default/minimum-effort score.

Table 26: Primary analysis robustness checks. Adjusted for 3 comparisons.

Coefficient	**(1) Quasibinomial model, exponentiated coefficients (-1 SD, + 1 SD)**	**(2) Linear model of accuracy increase (SD)**
2A: Graduated control options	4.87** (-0.37, +0.40)	0.183** (0.016)
2B: Four-box grid	4.91** (-0.37, +0.41)	0.183** (0.016)
Persona 2	0.12 (-0.08, +0.09)	0.026 (0.015)
Persona 3	1.33** (-0.35, +0.39)	0.253** (0.017)
Other covariates	Yes	Yes
Adjusted R^2	-	0.181
Observations	2,011	2,011

Table shows the regression coefficients from the two models. In parentheses are the standard errors of these coefficients.

Subgroup analysis by persona

In an exploratory follow-up analysis, we repeated the main regression modelling separately for participants choosing each of the personas. Figure 19 visually presents the results. We make several observations:

- The results from our primary analysis hold, even if we look at one persona at a time. Both treatment arms performed significantly better than the control but neither treatment arm did significantly better or worse than the other.

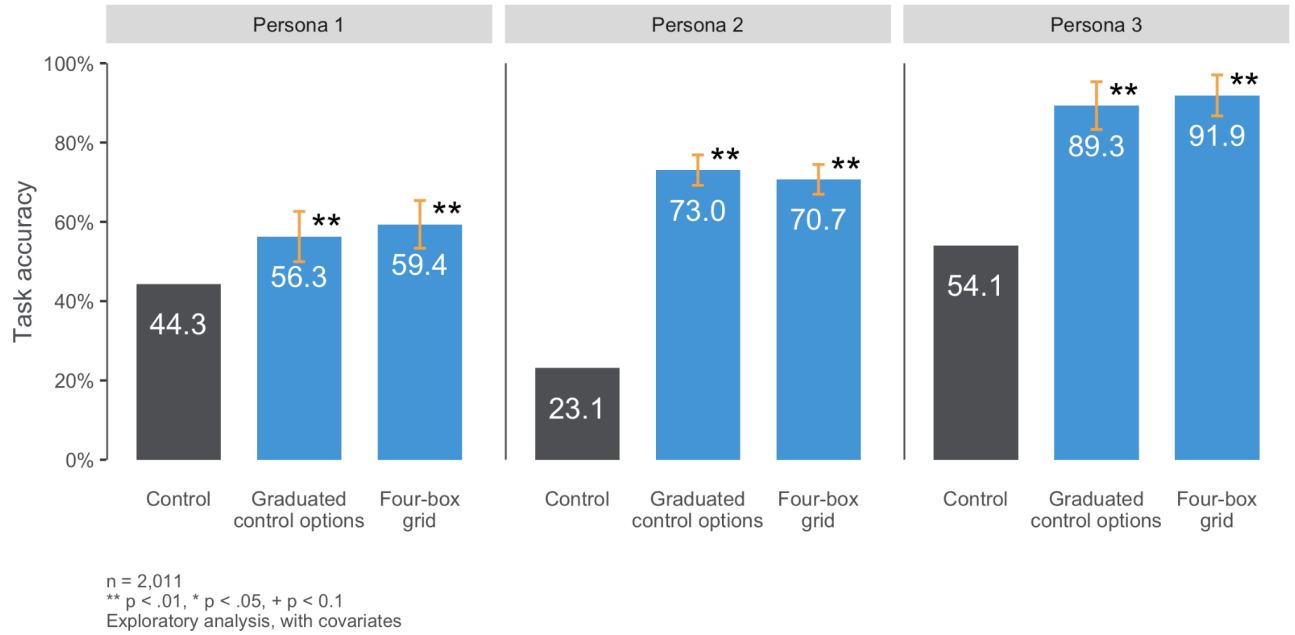
- Baseline performance (i.e. accuracy in the control) differs across personas, as does the observed treatment effect. We observe the lowest baseline (23%) but highest increase (almost 50pp) for Persona 2 (‘partly concerned’), while Persona 1 (‘least concerned’) has the lowest accuracy increase in the treatment arms at merely 12-13pp.

- Our research design was not set up to provide an explanation to this but the following explanation is plausible: The task required a more nuanced understanding of the difference between clearing cookies and blocking third-party trackers to get the majority of choices right. This was hard to do in the control where the set-up was very confusing but much easier to do when the language and the options were simplified as in our treatment arms.

- A potential explanation for the weaker treatment effect for Persona 1(‘least concerned’) is that those who are unconcerned about privacy also know less about privacy settings and thus a stronger or different intervention might be needed to achieve better accuracy, e.g. education aimed at understanding privacy controls. Another factor might be that, in real life, those who are less concerned about privacy and prefer convenience would usually be able to rely on the defaults to cater to their preferences. They might therefore be less used to looking into settings (which was necessary in the treatment arms) as much as people whose preferences differ from the default.

- The robustness check, where we use accuracy increase above the lowest-effort score instead of raw accuracy, supports our results, except for Persona 1 (‘least concerned’), where the treatment arms do not outperform the control anymore. This is still largely in line with our previous results that Persona 1 does not seem to benefit as much (or at all) from the treatment as the two other personas.

Figure 19: Subgroup analysis – task accuracy by persona.



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 27: OLS models of task accuracy as a function of treatment, split by persona

	(1) Persona 1	(2) Persona 2	(3) Persona 3
2A: Graduated control options	0.120**	0.499**	0.353**
	(0.032)	(0.018)	(0.031)
2B: Four-box grid	0.150**	0.476**	0.378**
	(0.031)	(0.019)	(0.026)
Raw control mean	0.443	0.231	0.541
Covariates	Yes	Yes	Yes
Observations	611	957	444
Adjusted R^2	0.047	0.498	0.119

Task accuracy - scores for individual settings and for each persona

Table 28 further breaks down this performance by individual settings. We observe that:

- The control arm had comparatively low scores across all three settings, reflecting our design choice whereby the default for these settings was always different from the persona's preferences, resulting in 2 out of the 3 settings always being wrong by

default in this arm. By comparison, in the treatment arms participants could get 50% accuracy by randomly making a selection.

- We find lower scores for Persona 2 ('partly concerned') for the 'Clearing Cookies' setting across all of the arms, which could be because a more nuanced understanding of the difference between this setting and the 'Third-party trackers' setting was needed to get this setting correct for Persona 2 (who liked to have trackers disabled but did not like to clear cookies) whereas Personas 1 and 3 prefered both settings on or off. This is perhaps an indication that most people do not have a good understanding of how these two differ.

Table 28: Mean accuracy of each setting, split by design and persona

Design	Third-party trackers			Clearing Cookies			Content Filtering		
	P1	P2	P3	P1	P2	P3	P1	P2	P3
Control arm	47%			31%			32%		
	100%	22%	32%	16%	6%	99%	19%	41%	31%
2A: Graduated control options	77%			58%			79%		
	59%	83%	85%	62%	42%	91%	48%	94%	88%
2B: Four-box grid	79%			62%			75%		
	67%	81%	91%	67%	42%	93%	43%	89%	90%

Note: P1 = Persona 1 = 'least concerned', P2 = Persona 2 = 'partly concerned', P3 = Persona 3 = 'most concerned'

6.4 Understanding of consequences

Participants had significantly better understanding of the consequences of their choices (measured by three comprehension questions) in the treatment arms compared to the control arm. However, we find no difference between treatment arms (see Figure 20 and Table 9). The robustness check (using a quasi-binomial model) showed a consistent pattern of results with the main analysis and is thus not presented here.

Figure 20: Secondary analysis: Understanding of consequences. Adjusted for 3 comparisons.

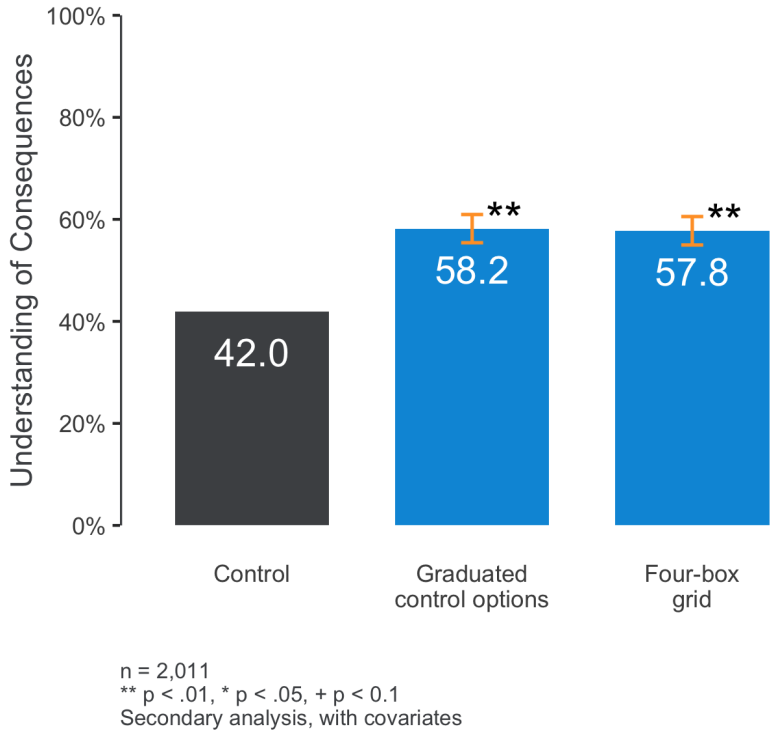


Table 29: Secondary analysis: Understanding of consequences. Adjusted for 3 comparisons.

Coefficient	Main analysis - linear model (SD)
2A: Graduated control options	0.162** (0.014)
2B: Four-box grid	0.158** (0.014)
Persona covariates	Yes
Other covariates	Yes
Adjusted R^2	0.097
Observations	2,011

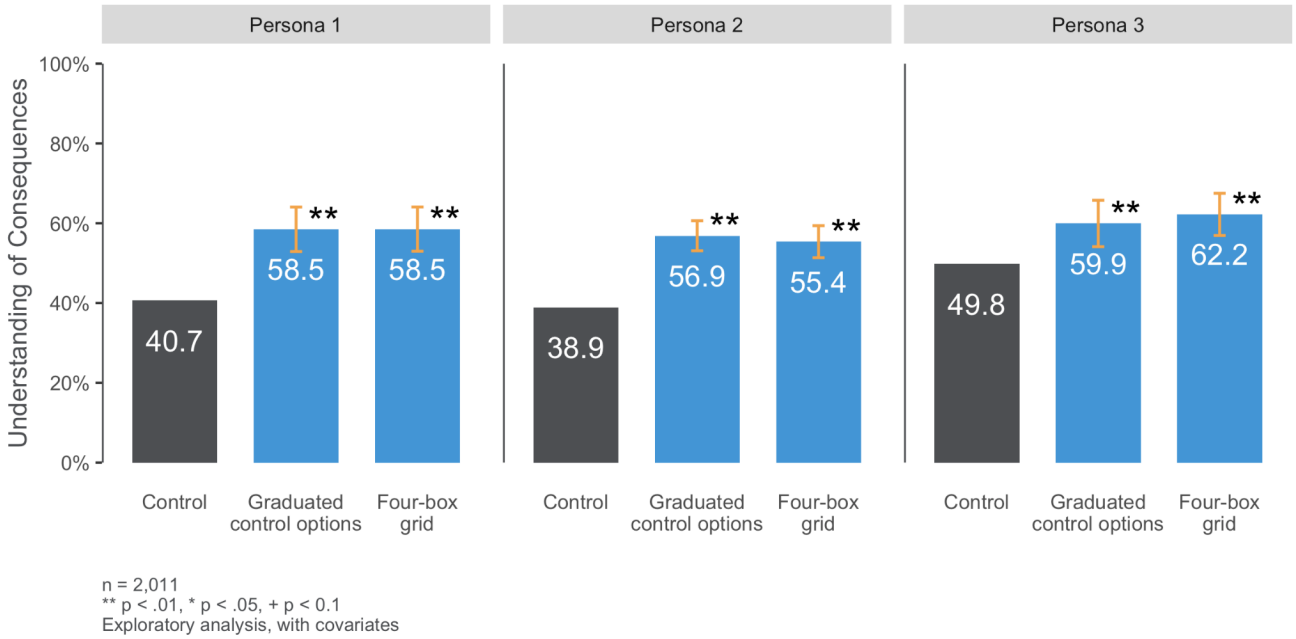
Table 30: Understanding of consequences, unadjusted mean scores for each question, by arm.[8]

Design	Third-party trackers	Clearing Cookies	Content Filtering	Total
Control	45%	39%	41%	**42%**
2A: Graduated settings	55%	60%	60%	**58%**
2B: Four-box grid	55%	65%	54%	**58%**

Subgroup analysis by persona

Figure 21 below breaks down the 'understanding of consequences' results by persona and trial design. The treatment arms both performed similarly on this outcome, with limited variation across personas. Overall, this supports our main result from the secondary analysis with the pooled data.

Figure 21: Subgroup analysis – Understanding of consequences by persona



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

[8] Table 30 and Figure 20 slightly differ due to covariate adjustment in the figure.

Table 31: OLS models of understanding of consequences as a function of treatment, split by persona

	(1) Persona 1	(2) Persona 2	(3) Persona 3
2A: Graduated control options	0.178 (0.028)	0.180** (0.019)	0.102** (0.030)
2B: Four-box grid	0.179 (0.028)	0.165** (0.020)	0.124** (0.027)
Raw control mean	0.407	0.389	0.498
Covariates	Yes	Yes	Yes
Observations	611	957	444
Adjusted R^2	0.073	0.108	0.078

Table shows the effect sizes of the treatment designs (compared to the control design) on the understanding score, measured on a linear 0-1 scale (1 = all four understanding questions answered correctly). In parentheses are the standard errors.

6.5 Feelings of control

Participants reported significantly higher feelings of control (measured by a single sentiment question) in the graduated control options arm compared to the control. The Four-box grid arm did not perform better than the control.

Figure 22: Effect of treatment on feelings of control

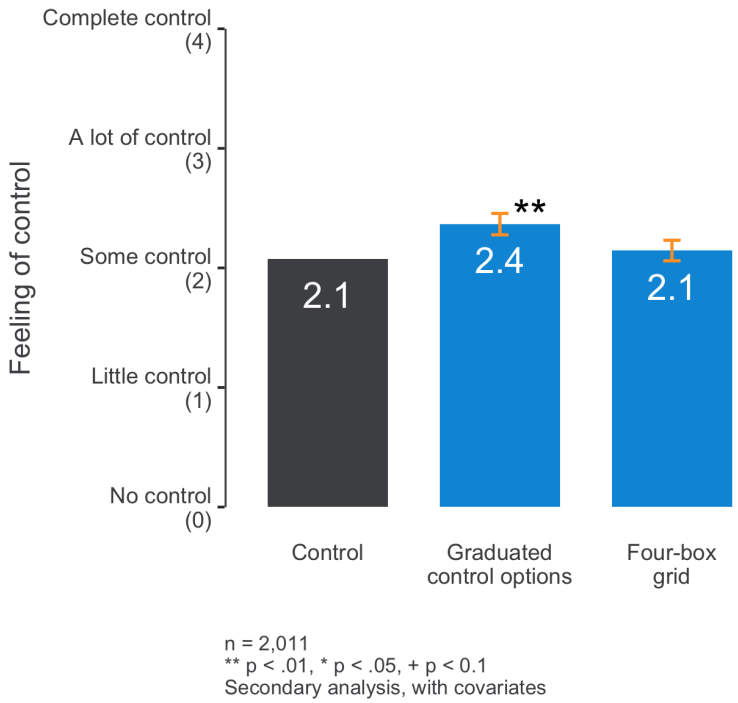


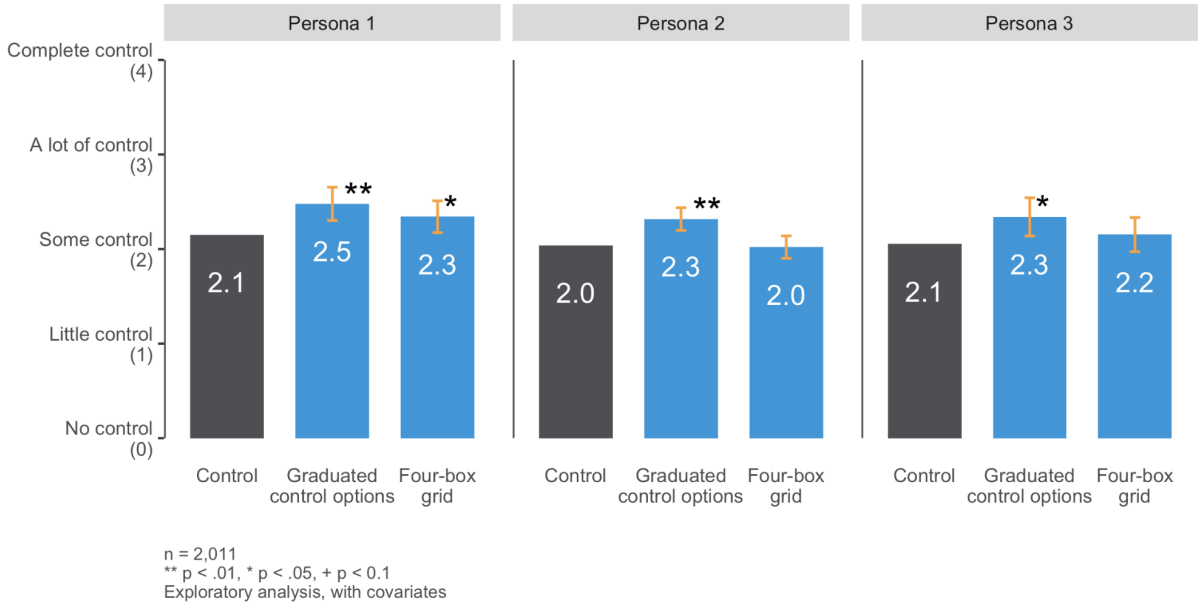
Table 32: Secondary analysis, feelings of control. Adjusted for 3 comparisons.

Coefficient	Main analysis - linear model (SD)	Robustness check - series of two Wilcoxon rank-sum tests, *W* statistic
2A: Graduated control options	0.290** (0.046)	184,068**
2B: Four-box grid	0.070 (0.044)	212,614
Persona covariates	Yes	No
Other covariates	Yes	No
Adjusted R^2	0.039	-
Observations	2,011	2,011

Subgroup analysis by persona

Figure 23 below presents the 'feelings of control' results separately for each persona. Overall, the subgroup analysis supports our main result from the secondary analysis with the pooled data. The graduated control options treatment outperformed the control, but the four-box grid did not, except for Persona 1 ('least concerned'). This is in line with our qualitative findings from the feedback participants left at the end of the survey which suggested that they found it counterintuitive to use the four-box grid to make their choice.

Figure 23: Subgroup Analysis - Feelings of control by persona



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 33: OLS models of feelings of control as a function of treatment, split by persona

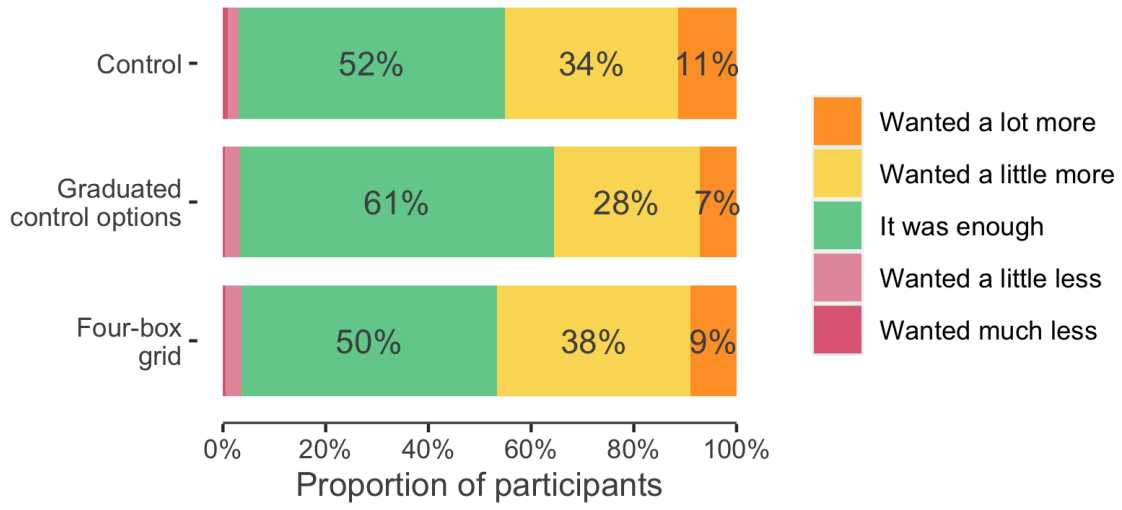
	(1) Persona 1	(2) Persona 2	(3) Persona 3
2A: Graduated control options	0.328**	0.277**	0.285*
	(0.090)	(0.061)	(0.010)
2B: Four-box grid	0.192*	-0.019	0.099
	(0.086)	(0.060)	(0.098)
Raw control mean	2.15	2.04	2.05
Covariates	Yes	Yes	Yes
Observations	611	957	444
Adjusted R^2	0.038	0.030	0.009

Additional (comparative) feelings of control question

We additionally asked participants whether they had *as much control as they wanted* ("Did you have as much control over the settings as you would have liked when making choices for Alex?"). Figure 24 presents the descriptive results.

The results for the control and the Four-box grid look similar with around half of participants saying they had enough control and 45%-47% wanting more control. Participants in the graduated control options arm were more likely to say they had enough control (61%) less likely to say that they wanted more control (35%)

Figure 24: Comparative feelings of control by treatment



6.6 Additional results

Other questions about the interfaces

We also asked participants 6 additional rating-scale questions and 2 (optional) free-text questions. The results from the rating-scale questions are summarised in Table 34 below.

Table 34: Summary of results for all the additional rating-scale questions.

Average answer to the following question*	**Control** (n=666)	**2A: Graduated Control Settings** (n=683)	**2B: Four-box Grid** (n=663)
Was it easy or hard to make choices on behalf of Alex?	3.23	3.63	3.34
Do you trust the choices were presented with your best interests in mind?	3.02	3.21	3.02
Were the choices presented in a fair way, without trying to influence you?	3.21	3.48	3.27
Were the choices explained in easy-to-understand terms?	3.07	3.44	3.13
Would you like to see an interface like this one in future?	3.75	3.96	3.37
Would you choose to use this interface rather than your current browser settings screens?	-	3.73	3.31

*Green indicates that the arm performed significantly better than the control. Red indicates that it performed significantly worse than the control. Dark green indicates that the treatment arm significantly outperformed other treatment arms and is the best performer. We use a 5% significance level throughout.

Concern about technology & digital comfort

We asked participants the same set of questions about *concern about technology* and *digital comfort* as in the smartphone trial. As before, we found that participants' self-reported level of concern mapped onto their chosen persona's concern level (Table 35) and that those who chose the more concerned personas tended to report lower digital comfort (Table 36). Those who reported higher concern were less comfortable using technology (Pearsons' correlation $r = -.15$, $p < .001$).

Table 35: Mean answers by persona to the four sub-questions of the questions "How concerned, if at all, would you say you are about each of the following?"

sub-question	**Persona 1** (n=611)	**Personal 2** (n=956)	**Persona 3** (n=444)	**Overall** (n=2011)
How addictive technology can be	2.08	2.10	2.36	**2.15**

Decisions being made about individuals by artificial intelligence	2.45	2.83	3.14	**2.78**
Companies selling on data about me	2.54	3.05	3.35	**2.96**
Fake news or disinformation online	2.69	2.89	2.95	**2.84**

Table 36: Mean answers by persona to the question "To what extent are you comfortable using a computer, a tablet or a smartphone to access government or commercial services online?"

	Persona 1 (n=611)	**Personal 2** (n=956)	**Persona 3** (n=444)	**Overall** (n=2011)
Digital comfort	3.68	3.51	3.19	**3.49**

Table 37 below shows the results of our analysis of the relationship between concern about technology, digital comfort, and task accuracy. As in the smartphone trial, after controlling for trial design assignment, persona, design-persona interaction, and the other collected covariates, there was a large negative relationship between concern and accuracy and a strong positive relationship between accuracy and digital comfort.

The same caveats about the interpretation of these results apply. We therefore refer the reader to Concern about technology & digital comfort in Section 5 and the corresponding part of Section 7.

Table 37: Regression table for the analysis of the relationships between task accuracy and (1) the total concern about technology score or (2) the digital comfort score (exploratory analysis). No corrections for multiple comparisons applied.

Term	**(1) Concern model** Coefficient (Huber-White SE)	**(2) Comfort model** Coefficient (Huber-White SE)
Intercept	25.4 (6.7)	4.8 (6.8)
Concern about technology (difference between full and zero score)	-10.1* (3.2)	-
Digital comfort (difference of 1 point on the 5-point scale)	-	4.3** (0.6)
Treatment: Graduated control options	12.2** (3.2)	11.2** (3.2)
Treatment: Four-box grid	15.1** (3.0)	14.3** (3.0)
Persona 2	-20.6** (2.2)	-21.0** (2.1)

Persona 3	11.4** (2.6)	12.3** (2.5)
Treatment: Graduated control options * Persona 2	37.7** (3.7)	38.4** (3.7)
Treatment: Four-box grid * Persona 2	32.9** (3.6)	33.6** (3.5)
Treatment: Graduated control options * Persona 3	22.8** (4.4)	23.0** (4.4)
Treatment: Four-box grid * Persona 3	22.0** (4.0)	21.6** (4.0)
Gender: Male	-1.3 (1.2)	-1.6 (1.2)
Gender: Other	-22.1** (4.1)	-22.9** (4.5)
Age category: 25-54	2.6 (2.1)	1.0 (2.1)
Age category: 55 and over	0.5 (2.1)	-2.1 (2.1)
Income: £30,000 and over	1.3 (1.3)	0.9 (1.2)
Location: Midlands	6.6** (2.4)	6.8** (2.3)
Location: North of England	8.9** (2.3)	8.6** (2.3)
Location: South and East England	7.6** (2.2)	7.6** (2.2)
Location: Wales, Scotland and Northern Ireland	7.5** (2.6)	7.9** (2.6)
Education: No degree	16.5** (5.9)	17.5** (6.0)
Education: Degree	18.8** (6.0)	19.5** (6.1)

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$

7. Social media trial results

7.1 Sample characteristics and balance checks

All participants completed the experiment on a computer or a tablet. Table 38 below shows a breakdown of the sample characteristics compared with the distribution in the general UK population. Quotas were applied for gender, age, household income, and location; no quotas were used for education and ethnicity.

Table 38: Sample characteristics: Social media trial

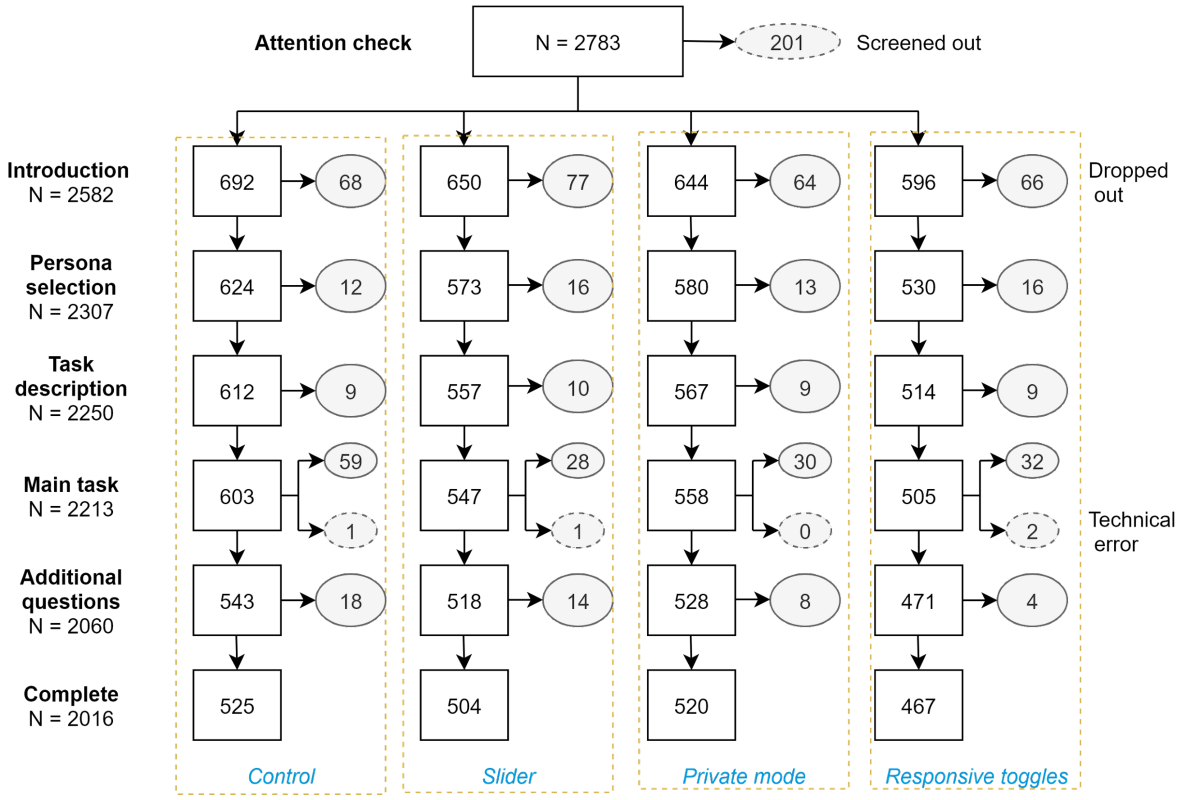
	Final sample	Target
Total n	2,016	2,000
Gender	**%**	**%**
Male (n=943)	47	50
Female (n=1071)	53	50
Other (n=2)	<1	-
Age		
18-24 (n=231)	11	12
25-54 (n=1113)	55	52
55+ (n=6672)	33	36
Household Income		
Less than £30,000 (n=1110)	55	50
More than £30,000 (n=906)	45	50
Location		
London (n=269)	13	13
North (n=487)	24	23
South & East (n=627)	31	32
Midlands (n=326)	16	16
Wales, Scotland & N. Ireland (n=307)	15	16
Education Level		

No degree (n=1272)	63	74
Degree (n=721)	36	26
Prefer not to say (n=23)	1	-
Ethnicity		
White (n=1804)	89	86
Black (n=47)	2	3
Asian (n=99)	5	7
Other (n=66)	3	3

Differential attrition and balance checks

We collected data from 2,783 individuals, of whom 201 (7%) failed our attention check. Of the remaining 2,583 participants, 566 (22%) either dropped out at some point during the experiment (562) or encountered a technical error (4) where participants' settings weren't being saved. Overall, 2,016 participants fully completed the experiment. The CONSORT diagram in Figure 25 provides a detailed breakdown.

Figure 25: CONSORT diagram of the numbers of participants present at the various stages of the online experiment.



We further checked for differential attrition. We ran two models, one looking at attrition for those who dropped out at any point and one looking at attrition amongst those who had seen the designs. Using the first model, we found that participants in the private mode arm were more likely to finish the experiment (less likely to drop out) than participants in the control arm. Using the second model, all treatment arms showed significantly lower attrition than the control arm.

This is in line with the expectations. The control arm featured a more complicated environment in which participants spent more time and used more mouse clicks and this additional required effort presumably made more participants quit the task. The fact that attrition was (descriptively) lowest in the private mode arm is also in line with this explanation, as participants in this arm spent the least time and used the fewest clicks to finish the task.

In line with the conclusion on different attrition rates in Trial 1 in section 5.1, we believe it is unlikely that it is driving the treatment effects we observed.

7.2 Descriptive Statistics

Table 39 compares summary statistics between arms, including outcome variables, time taken and clicks needed to complete the task. As expected, we found that participants took more time and used more clicks to complete the task in the control group than in the treatment arms. Unlike in Trials 1 and 2, treatment arms did not outperform the control on the three main outcome variables.

Table 39: Descriptive statistics for time taken, number of clicks, and the primary and secondary outcomes. Note that all outcomes have been converted to percentages.

Design	Time spent on task (Median + Q1 and Q3*)	N clicks (Median + Q1 and Q3*)	Task accuracy (Mean + SD)	Feelings of control † (Mean + SD)	Understanding of consequences (Mean + SD)
Control arm (n = 525)	132s (72 - 224 s)	13 (6 - 21)	0.72 (0.21)	2.12 (0.86)	0.55 (0.23)
Filtering slider (n =503)	82s (55 - 122 s)	5 (4 - 6)	0.71 (0.24)	2.40 (0.88)	0.51 (0.24)
Private mode (n =521)	78.8s (51 - 128s)	4 (3 - 6)	0.74 (0.24)	2.29 (0.84)	0.55 (0.23)

Responsive toggles (n = 467)	84.5s (55 - 130 s)	4 (3 - 6)	0.73 (0.24)	2.36 (0.86)	0.55 (0.24)

Table 40 shows how many participants in the final sample selected each persona. Persona 2 ('partly concerned') was the most popular choice (53%) which is in line with Trials 1 and 2 (62% and 48% respectively). Table 40 also shows that there seems to be a trend in task performance, with scores increasing as concern about privacy does.

Table 40: Proportion of participants choosing each persona.

	Proportion of participants	Task accuracy (Mean + SD)
Persona 1 (n = 335)	**16%**	65% (0.23)
Persona 2 (n = 1,063)	**53%**	71% (0.22)
Persona 3 (n = 618)	**31%**	78% (0.26)

7.3 Task accuracy

We did not find any significant differences in task accuracy between the four trial arms. Participants in all arms made, on average, just over 70% correct choices. Interestingly, participants who chose Personas 2 or 3 tended to perform better than those who chose Persona 1 ('least concerned').

Table 41 shows the results from our two pre-specified models. Comparing the unadjusted (1a) and adjusted (1b) models, we see that model 1b has higher R^2 and lower AIC, implying better fit. Our graphs and all later analyses are based on adjusted model specifications.

Table 41: Primary analysis - main analysis (linear models with Huber-White standard errors, adjusted for 5 comparisons)

Coefficient	Model 1a	Model 1b
3A: Filtering slider	-0.008 (0.014)	-0.012 (0.014)
3B: Private mode	0.019 (0.014)	0.019 (0.014)
3C: Responsive toggles	0.015 (0.014)	0.012 (0.014)

Persona 2	0.061** (0.016)	0.064** (0.016)
Persona 3	0.130** (0.017)	0.140** (0.018)
Other covariates	No	Yes
Adjusted R^2	0.035	0.048
AIC	-190.76	-204.21
Observations	2,016	2,016

Figure 26 below visualises the results from the primary regression model (model 1b).

Figure 26: Primary analysis – effect of treatment design assignment on task accuracy

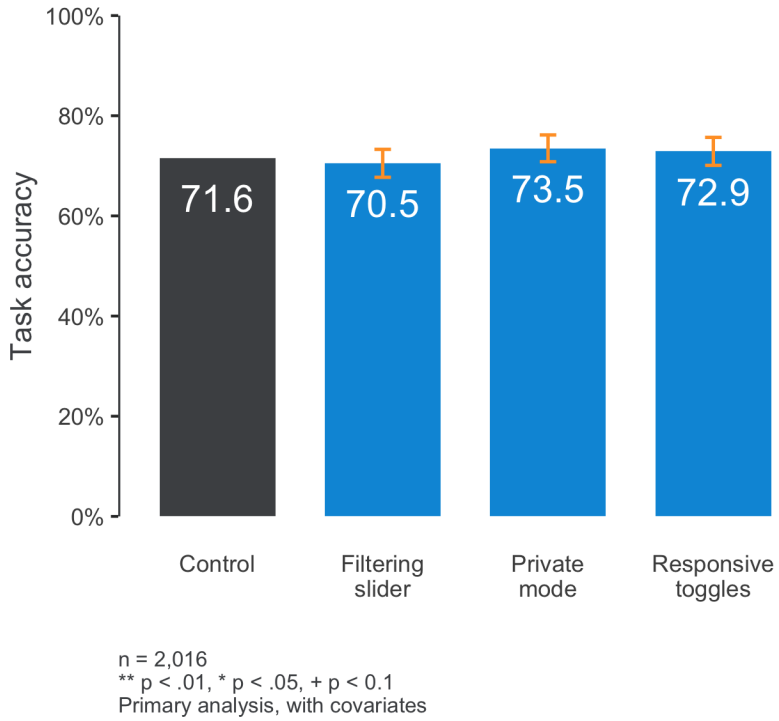


Figure 27 shows the performance of each design in more detail. In the control arm, the majority of participants achieved an accuracy score of 50%, which was the default score a user got without making any changes to the settings. Over one half of the participants achieved a score of either 75% or 100%, and very few participants scored 0 or 25%.

In all three treatment arms, the proportion of participants achieving above-default scores (75% or 100%) was higher than in the control arm, but so was the proportion of participants achieving low scores (0% or 25%). We discuss this finding in more detail in the main report.

Figure 27: Distribution of accuracy scores across treatments

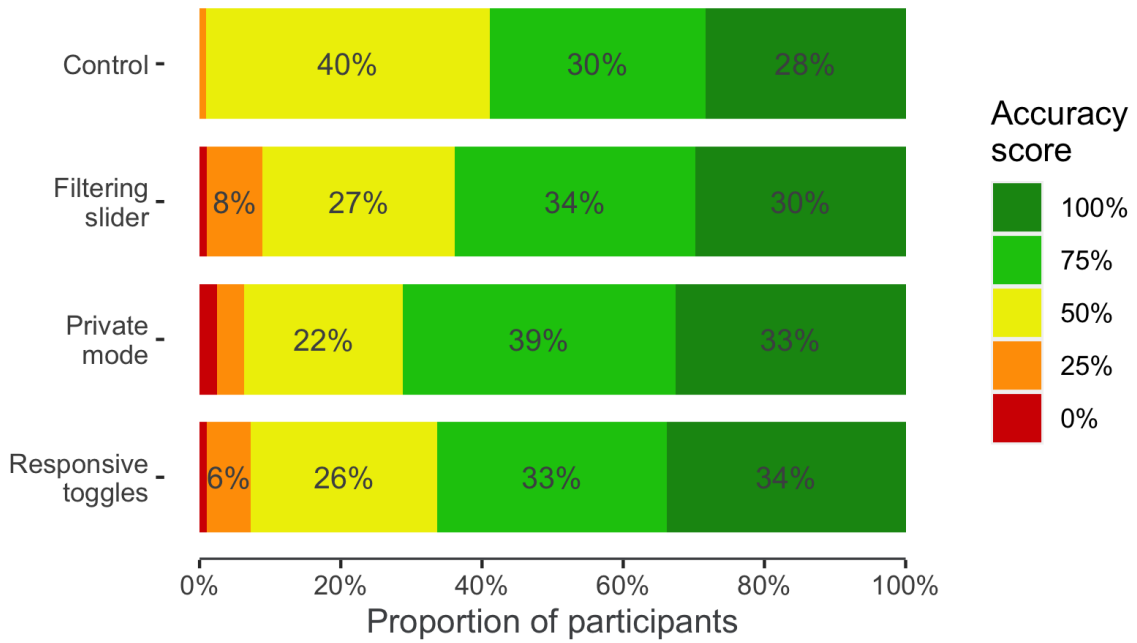


Table 42 shows how accuracy varied with different demographic characteristics. Overall, we found little variation, with the largest differences by participants' geographical location. This analysis was purely exploratory as our research design was not set up to answer whether and why main task accuracy varies across different demographic groups.

Table 42: Associations between task accuracy (%) and the covariates. *p*-values are the results of univariate linear regressions with Huber-White standard errors.

	Accuracy (%)	**p-value** compared to reference group
Total sample	**72.3%**	
Gender		
Female (n=1071)	73.2%	Reference group
Male (n=943)	71.2%	$p < .10^{+}$
Other (n=2)	100%	*Sample too small*
Age		
18-24 (n=231)	73.1%	Reference group
25-54 (n=1113)	72.4%	$p > .10$
55+ (n=672)	71.8%	$p > .10$
Household Income		

Less than £30,000 (n=1110)	72.2%	Reference group
More than £30,000 (n=906)	72.4%	$p > .10$
Location		
London (n=269)	68.3%	Reference group
North (n=487)	72.4%	$p < .05$*
South & East (n=627)	74.8%	$p < .01$**
Midlands (n=326)	71.7%	$p < .10^{+}$
Wales, Scotland & N.Ireland (n=307)	71.2%	$p > .10$
Education Level		
No degree (n=1272)	71.8%	Reference group
Degree (n=721)	73.4%	$p > .10$
Prefer not to say (n=23)	66.3%	$p > .10$
Ethnicity		
White (n=1804)	72.4%	Reference group
Black (n=47)	66.5%	$p < .10^{+}$
Asian (n=99)	70.2%	$p > .10$
Other (n=66)	77.3%	$p < .10^{+}$

Primary analysis robustness check

We ran two robustness checks (see Table 43). The first one replaced the linear-model specification with a quasibinomial model, the second one used a linear model but replaced raw accuracy with *accuracy increase* from a default/minimum-effort score. The results of the quasibinomial model are consistent with the main analysis in finding no significant differences between the trial arms.

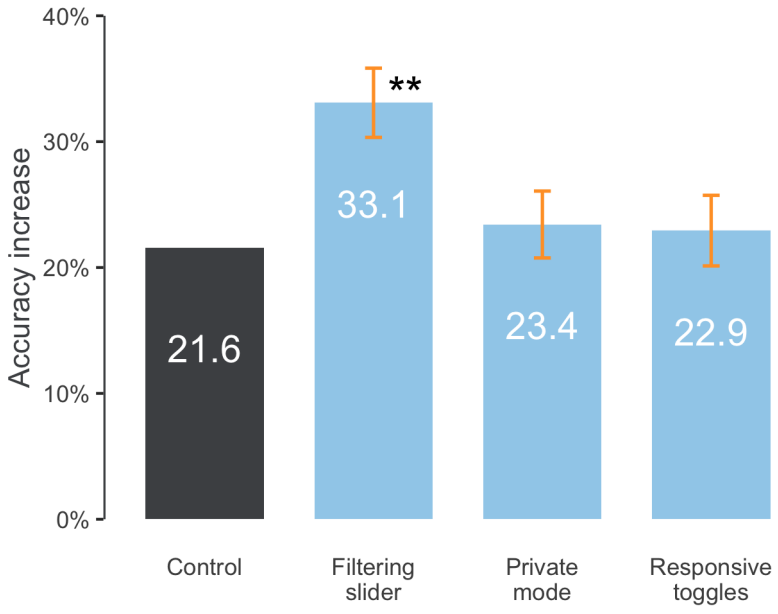
The second robustness check showed that the filtering slider arm significantly outperformed the other three trial arms. Note, however, that these figures were obtained under the assumption that a participant exerting minimum effort would randomly select one of the presented options (namely, one of the three slider positions), with equal probability. This may not be a fair assumption and it may give an unfair advantage to the filtering slider arm. The observed effect is nearly perfectly explainable by this potential unfair advantage. Therefore, we cannot conclude that the filtering slider arm indeed outperformed the other arms.

Table 43: Primary analysis robustness checks. Adjusted for 5 comparisons.

	(1)	(2)

Coefficient	Quasibinomial model, exponentiated coefficients (-1 SD, + 1 SD)	Linear model of accuracy increase (SD)
3A: Filtering slider	0.95 (-0.19, +0.22)	0.115** (0.014)
3B: Private mode	1.10 (-0.07, +0.08)	0.019 (0.14)
3C: Responsive toggles	1.07 (-0.07, +0.08)	0.013 (0.14)
Persona 2	1.35** (-0.10, +0.10)	0.13** (0.16)
Persona 3	2.02** (-0.17, +0.19)	0.20** (0.17)
Other covariates	Yes	Yes
Adjusted R^2	-	0.12
Observations	2016	2016

Figure 28: Primary analysis robustness check (check #2).



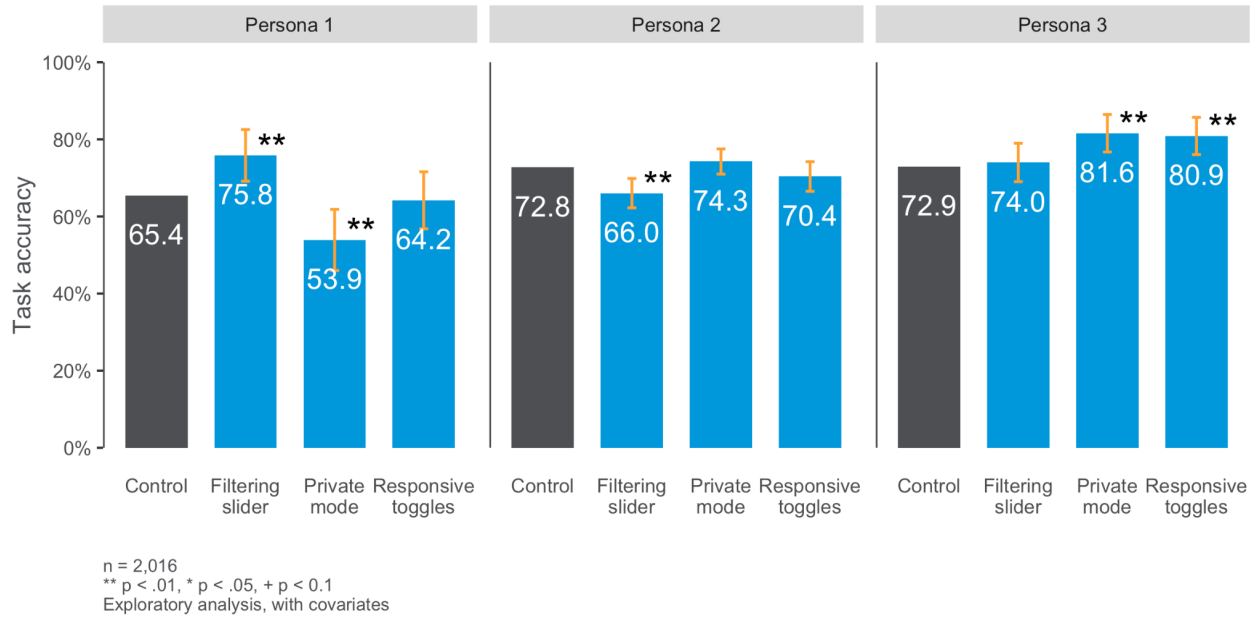
n = 2,016
** p < .01, * p < .05, + p < 0.1
Primary analysis robustness check, with covariates

Subgroup analysis by persona

In an exploratory follow-up analysis, we repeated the main regression modelling separately for participants choosing each of the personas. Figure 29 visually presents the results. We make the following observations:

- We found no consistent pattern. The treatments that performed well for some personas tended to perform poorly for others.
- For instance, unsurprisingly, the private mode arm performed very well for the Persona 3 (‘most concerned’) as it allowed making 3 out of 4 correct choices using a single toggle. However, it did not outperform the control for Perona 2 (‘partly concerned’) and performed significantly worse than control for Persona 1 (‘least concerned’). This could be due to participants wrongly switching on the toggle for this Persona 1.

Figure 29: Subgroup analysis – task accuracy by persona.



Note: Persona 1 = ‘least concerned’, Persona 2 = ‘partly concerned’, Persona 3 = ‘most concerned’

Table 44: OLS models of task accuracy as a function of treatment, split by persona. *p*-values not corrected for multiple comparisons.

	(1) Persona 1	(2) Persona 2	(3) Persona 3
3A: Filtering slider	0.10**	-0.067**	0.011
	(0.034)	(0.019)	(0.025)
3B: Private mode	-0.12**	0.015	0.087**

	(0.040)	(0.017)	(0.025)
3C: Responsive toggles	-0.012 (0.038)	-0.024 (0.020)	0.079** (0.025)
Raw control mean	0.65 (0.20)	0.73 (0.20)	0.73 (0.21)
Covariates	Yes	Yes	Yes
Observations	335	1063	618
Adjusted R^2	0.10	0.027	0.075

Task accuracy - scores for individual settings and for each persona

Table 45 presents mean accuracy scores for each question, split by arm and persona. We observe that:

- Scores tended to be lowest for the post visibility setting and highest when setting preferences for untrustworthy news sources.
- Interestingly, the filtering slider arm did not do particularly well on the untrustworthy sources setting, despite it being the key focus of the arm's design (the other three settings were presented in the form of toggles).

Table 45: Task accuracy, unadjusted mean scores for each question, by arm.

Design	Personalised feed			Personalised ads			Untrustworthy sources			Post visibility		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Control arm	77%			58%			88%			63%		
	93%	93%	40%	96%	50%	53%	34%	98%	99%	39%	49%	99%
3A: Filtering slider	77%			73%			81%			52%		
	89%	79%	68%	91%	67%	75%	68%	79%	90%	58%	41%	68%
3B: Private mode	67%			78%			84%			66%		
	75%	61%	73%	77%	79%	77%	30%	95%	93%	33%	63%	87%
3C: Responsive toggles	77%			78%			80%			57%		
	78%	79%	74%	89%	69%	87%	37%	90%	88%	52%	47%	79%

Note: P1 = Persona 1 = 'least concerned', P2 = Persona 2 = 'partly concerned', P3 = Persona 3 = 'most concerned'

7.4 Understanding of consequences

Participants had low understanding scores compared to Trials 1 and 2, achieving just over 50% of correct responses on average. The following reasons likely contributed to this:

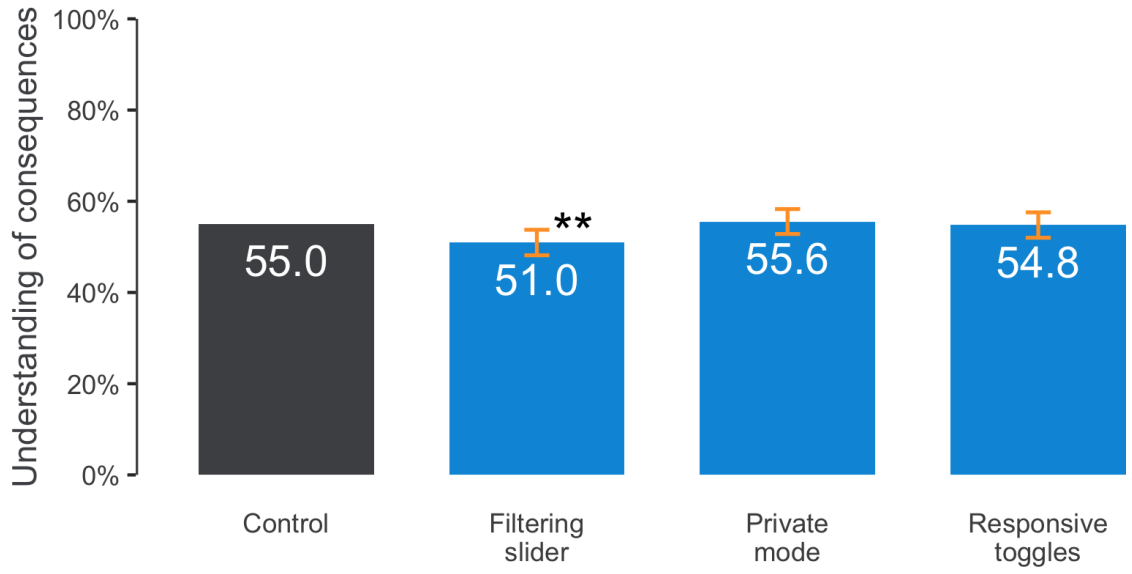
1. Some of the questions were rather difficult (see Table 1). For instance, to answer the question about the browsing data correctly participants needed to combine information about two settings and the information was only shown in a pop-up info box.

2. The questions about personalised ads and trusted news sources had answer options that were incorrect, but plausible. For example, someone with 'personalised ads' turned off would still have ads personalised based on basic profile information, however many might think that their ads would not have been personalised in any way.

There was little variation across trial arms. The filtering slider arm performed significantly worse than the control arm; however, this difference was only 4pp. There was some variability in the understanding scores across individual questions (Table 47). Descriptively, the question about personalised ads had the lowest score. As explained above this was quite a difficult question as there were two types of information about them which could be used in the ads.

The robustness check (using a quasi-binomial model) showed a consistent pattern of results with the main analysis and is thus not presented here.

Figure 30: Secondary analysis: Understanding of consequences. Adjusted for 6 comparisons



n = 2,016
** p < .01, * p < .05, + p < 0.1
Secondary analysis, with covariates

Table 46: Secondary analysis: Understanding of consequences. Adjusted for 6 comparisons.

Coefficient	Main analysis - linear model (SD)
3A: Filtering slider	-0.041** (0.014)
3B: Private mode	0.005 (0.014)
3C: Responsive toggles	-0.003 (0.014)
Persona covariates	Yes
Other covariates	Yes
Adjusted R^2	0.092
Observations	2,016

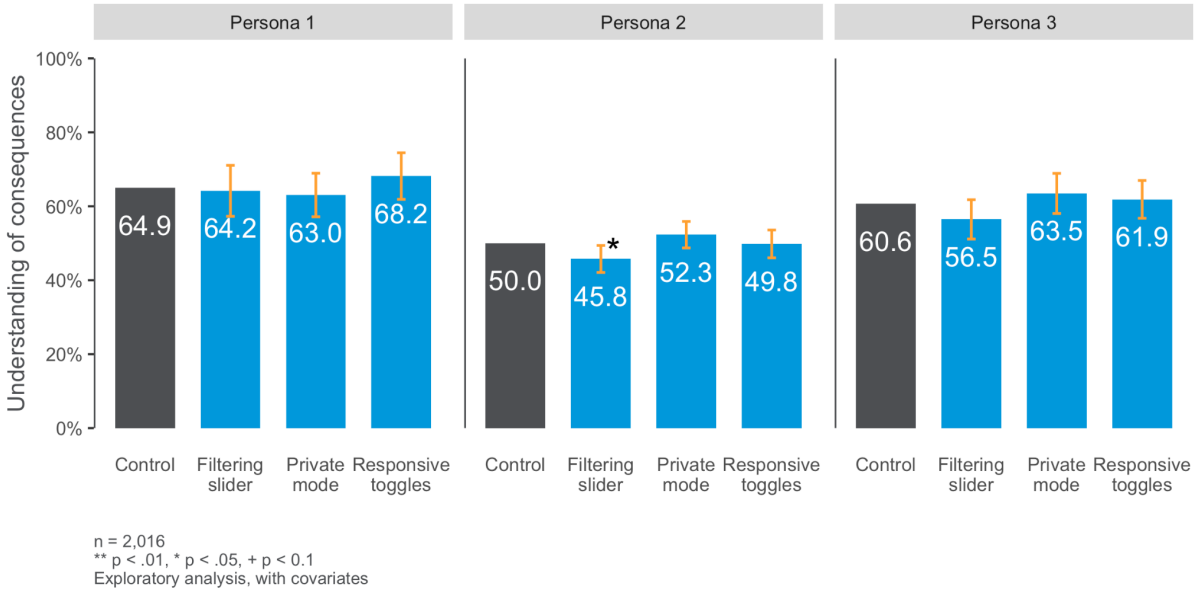
Table 47: Understanding of consequences, unadjusted mean scores for each question, by arm.

Design	Persona-lised Feed	Post visibility	Persona-lised Ads	Trusted News Only	Browsing Data	Total
Control	0.57	0.67	0.51	0.56	0.44	**0.55**
3A: Filtering slider	0.64	0.49	0.34	0.58	0.51	**0.51**
3B: Private mode	0.62	0.66	0.40	0.55	0.55	**0.55**
3C: Responsive toggles	0.67	0.57	0.40	0.59	0.53	**0.55**

Subgroup analysis by persona

Figure 31 below breaks down the ‘understanding of consequences’ results by persona and trial design. There were no significant differences between the designs for each persona, with one exception — the filtering slider arm performed worse for Persona 2 (‘partly concerned’) than control.

Figure 31: Subgroup analysis – Understanding of consequences by persona



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 48: OLS models of understanding of consequences as a function of treatment, split by persona

	(1) Persona 1	(2) Persona 2	(3) Persona 3
3A: Filtering slider	0.009	-0.050**	-0.046
	(0.036)	(0.019)	(0.027)
3B: Private mode	-0.034	0.007	0.017
	(0.045)	(0.018)	(0.028)
3C: Responsive toggles	0.029	-0.016	0.001
	(0.046)	(0.019)	(0.026)
Raw control mean	0.64 (0.21)	0.49 (0.21)	0.60 (0.25)
Covariates	Yes	Yes	Yes
Observations	335	1,063	618
Adjusted R^2	0.011	0.014	0.034

Table shows the effect sizes of the treatment designs (compared to the control design) on the understanding score, measured on a linear 0-1 scale (1 = all four understanding questions answered correctly). In parentheses are the standard errors.

7.5 Feelings of control

Participants reported significantly higher feelings of control (measured by a single sentiment question) in all three treatment designs compared to the control. Those in the control arm gave an average rating of 2.1 out of 4, corresponding to 'some control'. All treatment arms made participants feel significantly more in control, with average scores up to 2.4, roughly halfway between 'some control' and 'a lot of control'. The robustness check (using non-parametric test) showed a consistent pattern of results with the main analysis and is thus not presented here.

Figure 32: Effect of treatment on feelings of control

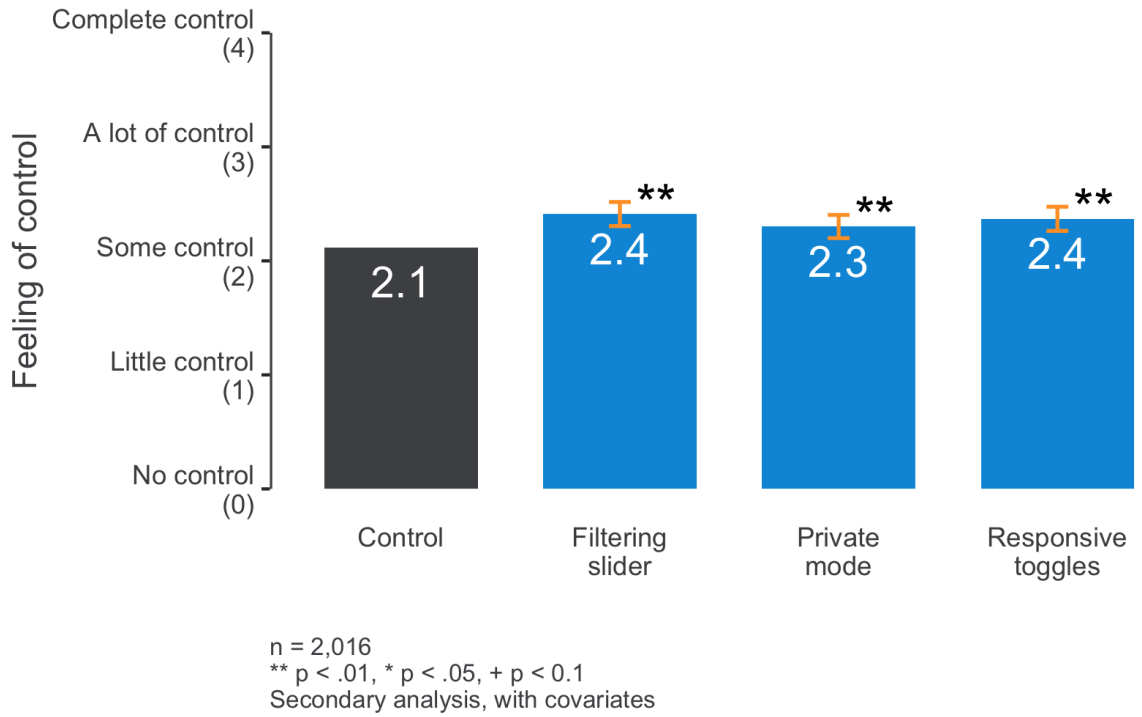


Table 49: Secondary analysis, feelings of control. Adjusted for 6 comparisons.

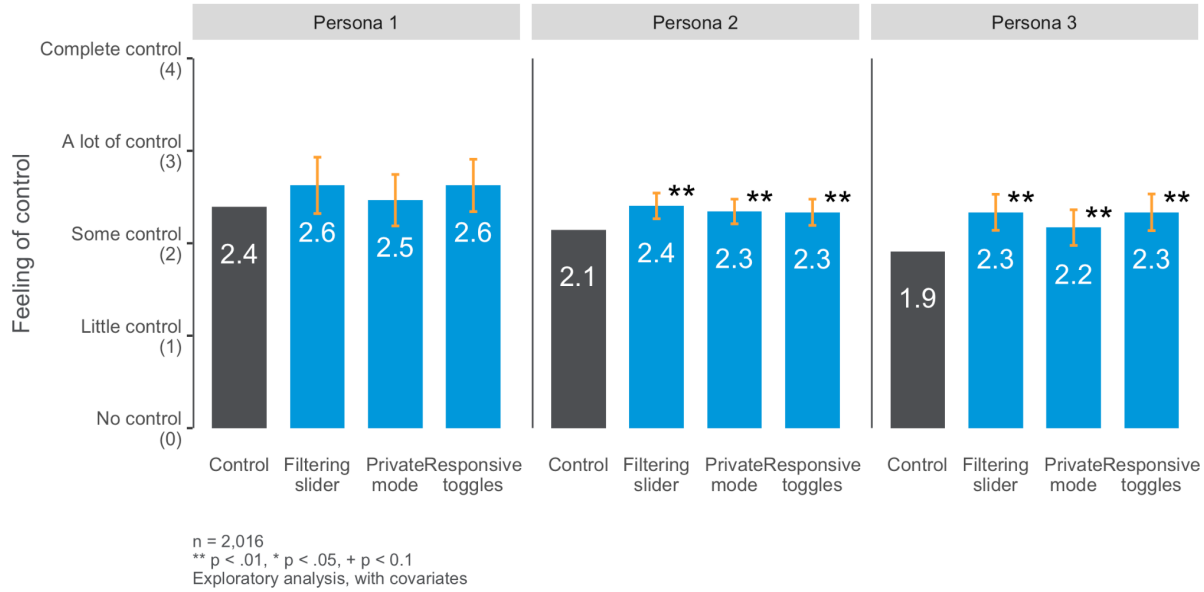
Coefficient	Main analysis - linear model (SD)
3A: Filtering slider	0.29** (0.054)
3B: Private mode	0.19** (0.052)

3C: Responsive toggles	0.25** (0.054)
Persona covariates	Yes
Other covariates	Yes
Adjusted R^2	0.033
Observations	2,016

Subgroup analysis by persona

Figure 33 below presents the 'feelings of control' results separately for each persona. Participants who chose Persona 1 ('least concerned') did not report greater feelings of control in the treatment arms than in the control arm. However, all treatments consistently led to increases in the reported feelings of control for those who chose personas 2 and 3.

Figure 33: Subgroup Analysis - Feelings of control by persona



Note: Persona 1 = 'least concerned', Persona 2 = 'partly concerned', Persona 3 = 'most concerned'

Table 50: OLS models of feelings of control as a function of treatment, split by persona

	(1) Persona 1	(2) Persona 2	(3) Persona 3
3A: Filtering slider	0.23 (0.16)	0.36** (0.07)	0.42** (0.10)
3B: Private mode	0.07	0.20**	0.26**

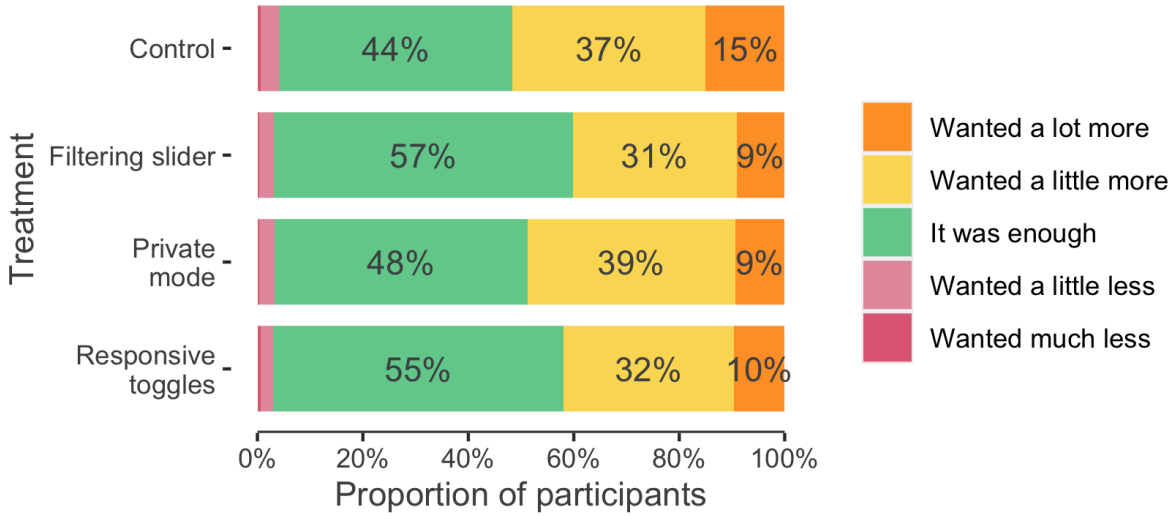
	(0.14)	(0.07)	(0.10)
3C: Responsive toggles	0.23	0.19**	0.42**
	(0.14)	(0.07)	(0.10)
Raw control mean	2.39	2.15	1.91
Covariates	Yes	Yes	Yes
Observations	335	1,063	618
Adjusted R^2	0.010	0.016	0.029

Additional (comparative) feelings of control question

We additionally asked participants whether they had *as much control as they wanted* ("Did you have as much control over the settings as you would have liked when making choices for Alex?"). Figure 34 presents the descriptive results.

The proportion of participants who reported having "enough control" was higher in the Filtering slider arm and the responsive toggles arm than in the private mode arm and the control. Correspondingly, the proportion of respondents who wanted more control was lower in the Filtering slider and the responsive toggles arms.

Figure 34: Comparative feelings of control by treatment



7.6 Additional results

Other questions about the interfaces

We also asked participants 6 additional rating-scale questions and 2 (optional) free-text questions. The results from the rating-scale questions are summarised in Table 51 below. For most of the questions, all treatment arms significantly outperformed the control arm but there were no significant differences between the treatment arms.

Table 51: Summary of findings of our additional outcome variables.

Average answer to the following question*	**Control** (n=525)	**Filtering slider** (n=503)	**Private mode** (n=521)	**Responsive toggles** (n=467)
Was it easy or hard to make choices on behalf of Alex?	2.90	3.37	3.37	3.38
Do you trust the choices were presented with your best interests in mind?	2.91	3.07	3.01	3.02
Were the choices presented in a fair way, without trying to influence you?	3.14	3.34	3.28	3.33
Were the choices explained in easy-to-understand terms?	2.91	3.19	3.19	3.21
Would you like to see an interface like this one in future?	3.81	3.92	3.87	3.89
Would you choose to use this interface rather than the current social media settings screens?	-	3.90	3.78	3.83

*Green indicates the best performing arm(s) for each outcome (at the 5% significance level). If no cell is highlighted, none of the arms outperformed the rest.

Concern about technology & digital comfort

As in the previous two trials, we asked participants a set of questions to capture their *concern about technology* (see Table 52). As expected, those who had chosen Persona 2 ('partly concerned') tended to be more concerned than those who had chosen Persona 1 ('least concerned'). Those who had chosen Persona 3 ('most concerned') tended to be the most concerned. Based on overall scores, participants tended to be most concerned about fake news or disinformation, and companies selling their data.

Since we suspected that participants' answers to these questions may be influenced by their interaction with the task (see Concern about technology & digital comfort in Section 5), in this trial we asked these questions before the task in one half of the sample (randomly selected) and after the task in the other half.

Indeed, we saw a significantly higher reported concern when asked after the task. It appears that the activity served as a reminder to participants about the concerns surrounding technology. By asking them to set preferences regarding data-sharing, news filtering etc, the activity implies that these topics are important and deserve attention, hence the increase in concern.

Table 52: Mean answers by persona to the four sub-questions of the questions "How concerned, if at all, would you say you are about each of the following?"

	Persona 1 (n=335)		**Personal 2** (n=1063)		**Persona 3** (n=618)		**Overall** (n=2016)	
sub-question	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
How addictive technology can be	2.13		2.28		2.30		**2.26**	
	2.14	2.11	2.22	2.35^{+}	2.25	2.35	**3.22**	**3.31^{+}**
Decisions being made about individuals by artificial intelligence	2.47		2.69		2.00		**2.75**	
	2.38	2.56	2.59	2.81**	2.94	34.07	**3.68**	**3.84****
Companies selling on data about me	2.41		2.86		3.36		**2.94**	
	2.33	2.47	2.75	2.98**	3.35	3.39	**3.89**	**4.00****
Fake news or disinformation online	2.53		2.93		3.00		**2.88**	
	2.39	2.65*	2.91	2.96	2.91	3.11	**3.83**	**3.95***

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$

Table 53 shows the results by person for the *digital comfort* question. Those who chose Persona 2 ('partly concerned') tended to be the most comfortable, and those who chose Persona 3 ('most concerned') tended to be the least comfortable. Additionally, there was a significant decrease in digital comfort when the question was asked post-task. Potentially, this comes as a result of the task forcing participants to reconsider their technical capabilities, lowering confidence.

Table 53: Mean answers by persona to the question "To what extent are you comfortable using a computer, a tablet or a smartphone to access government or commercial services online?"

	Persona 1 (n=335)		**Personal 2** (n=1063)		**Persona 3** (n=618)		**Overall** (n=2016)	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Digital comfort	3.65		3.78		3.52		**3.68**	
	3.86	3.47**	3.90	3.64**	3.60	3.41**	**3.80**	**3.54****

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$

Table 54 shows full regressions on both digital concern and comfort. Interestingly, digital comfort is negatively associated with concern, i.e. concern lowers as digital comfort increases. This could perhaps be explained by a potential lack of understanding about technology causing concern, however an argument could have been made apriori that more understanding/comfort with technology could lead to more concern as the issues surrounding it become more salient. Females were generally more concerned and less comfortable, persona's 2 & 3 were unsurprisingly more concerned about technology, while older people were both more concerned, and surprisingly, more comfortable with technology than 18-24 year olds.

Table 54: Regression table for the analysis of the relationships between task accuracy and (1) the total concern about technology score or (2) the digital comfort score (exploratory analysis). No corrections for multiple comparisons applied.

Term	**(1)** **Concern model** Coefficient (Huber-White SE)	**(2)** **Comfort model** Coefficient (Huber-White SE)
Intercept	0.43	0.57
Digital comfort (difference of 1 point on the 5-point scale)	-0.04** (0.005)	-
Concern about technology (difference between full and zero score)	-	-0.04** (0.03)
Treatment: Filtering slider	-0.13** (0.03)	0.11** (0.03)
Treatment: Private mode	-0.11* (0.04)	-0.12* (0.04)
Treatment: Responsive toggles	-0.008 (0.03)	-0.01 (0.04)
Persona 2	0.08* (0.02)	0.08* (0.03)
Persona 3	0.10** (0.03)	0.09** (0.03)
Treatment: Filtering slider * Persona 2	-0.19** (0.04)	- 0.18** (0.04)
Treatment: Private mode * Persona 2	0.13* (0.04)	0.13* (0.04)
Treatment: Responsive toggles * Persona 2	-0.01 (0.04)	-0.007 (0.04)
Treatment: Filtering slider * Persona 3	-0.11* (0.04)	-0.10+ (0.04)
Treatment: Private mode * Persona 3	0.20** (0.05)	0.21** (0.04)

Treatment: Responsive toggles * Persona 3	0.10^{+} (0.04)	0.10^{+} (0.04)
Gender: Male	-0.01 (0.01)	-0.008 (0.01)
Gender: Other	0.29** (0.02)	0.31 (0.02)
Age category: 25-54	-0.02 (0.06)	-0.02 (0.02)
Age category: 55 and over	-0.05* (0.02)	-0.04^{+} (0.01)
Income: £30,000 and over	0.00 (0.01)	0.006 (0.01)
Location: Midlands	-0.03 (0.02)	0.03 (0.02)
Location: North of England	0.04^{+} (0.02)	0.04* (0.02)
Location: South and East England	0.06** (0.02)	0.06** (0.02)
Location: Wales, Scotland and Northern Ireland	0.03 (0.02)	0.03 (0.02)
Education: Degree	0.09^{+} (0.04)	0.10^{+} (0.04)

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$

Finally, we tested whether the strength of the association between these two variables and task performance varied depending on whether we asked the questions before or after the task. Tables 55 and 56 summarise these results: Table 55 contains coefficients from models equivalent to model (1) in Table 54. The regression was run once on the full dataset, once on the subsample who were asked the *concern about technology* questions before the task, and once for the subsample who were asked the questions after the task. Table 56 similarly shows coefficients equivalent to those in model (2) in Table 54 for *digital comfort*.

We can make two observations. Firstly, the association between concern and task performance didn't vary much when the questions were asked before vs after the task. This is interesting, as the correlation stayed the same despite the fact that digital concern increased.

Secondly, the correlation between digital comfort and task performance was about twice as strong when the question was asked after the task vs before. This supports our hypothesis that participants may be using their experience with the task to inform their answer, with those who knew they had performed well indicating higher comfort and those who knew they'd performed poorly stating lower comfort. However, there is still a significant association even for when the question was asked prior to the task, suggesting that this wasn't the only factor behind these associations we have seen across the three trials.

Table 55: Regression coefficients from fully-adjusted linear models of task accuracy, similar to the regression in Table 54 . Only coefficients for concern about technology are shown.

Term	(a) Full dataset (n = 2,016)	(b) Questions before task (n = 1,056)	(c) Questions after task (n = 960)
Concern about technology	-0.046^{+} (0.026)	-0.047 (0.034)	-0.046 (0.041)

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$

Table 56: Regression coefficients from fully-adjusted linear models of task accuracy, similar to the regression in Table 54. Only coefficients for digital comfort are shown.

Term	(a) Full dataset (n = 2,016)	(b) Questions before task (n = 1,056)	(c) Questions after task (n = 960)
Digital comfort	0.036** (0.005)	0.024** (0.007)	0.050** (0.001)

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$