A simple self-reflection intervention boosts the detection of targeted advertising

Philipp Lorenz-Spreen[1*†], Michael Geers[1], Thorsten Pachur[1], Ralph Hertwig[1], Stephan Lewandowsky[2,3], & Stefan M. Herzog[1*]

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[2]School of Psychological Science and Cabot Institute, University of Bristol, Bristol, United Kingdom

[3]School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia

[*]contributed equally; [†]lorenz-spreen@mpib-berlin.mpg.de

Abstract

Online platforms collect and infer detailed information about people and their behaviour, giving advertisers an unprecedented ability to reach specific groups of recipients. This ability to "microtarget" messages contrasts with people's limited knowledge of what data platforms hold and how those data are used. Two online experiments (total $N = 828$) demonstrated that a short, simple intervention prompting participants to reflect on a targeted personality dimension boosted their ability to correctly identify the ads that were targeted at them by up to 26 percentage points. Merely providing a description of the targeted personality dimension did not improve accuracy; accuracy increased when participants completed a short questionnaire assessing the personality dimension—even when no personalized feedback was provided. We argue that such "boosting approaches," which improve peoples' ability to detect advertising strategies, should be part of a policy mix aiming to increase platforms' transparency and give people the competences necessary to reclaim their autonomy online.

Introduction

Advertisers have always sought to maximize the match between their messages and presumed customers. There are few cosmetic ads in motorcycle magazines, and TV commercials rarely advertise toys after children are in bed. However, compared with traditional targeted advertising, online advertising offers advertisers unprecedented ability to reach specific groups of recipients with tailored messages[1;2]. In addition, advertisers receive direct feedback on the reception of their message (e.g., via click-through rates), enabling them to further optimize their message and its targeting via large-scale A/B testing[3;4]. With increasing technological capacity and sophistication, these processes are becoming ever more opaque for the public and for targeted individuals, in particular[5]. This development further amplifies the asymmetry of knowledge between platforms and their users: Platforms collect and infer detailed information about users and their behaviour[1;2]. Users, by contrast, know little about what data the platforms hold and how those data are used to shape their online experience[6;7]. Here we investigate a short, simple intervention that aims to boost people's competence to detect targeted messages and could contribute to counteracting this asymmetry. The intervention raises users' awareness of personality dimensions[8] that might be targeted, and enables them to detect a targeting strategy designed to exploit those dimensions.

Here we define *microtargeting* as the method of addressing users based on "non-observable" features (e.g., partisanship, personality dimensions) rather than easily "observable" demographic features such as age and gender. This type of targeting is by definition difficult to detect unless it is explicitly announced or labelled.

Although the persuasive effect of a single ad on a single individual may be relatively small[9], the potential harms of microtargeting can scale up and propagate to the collective[10]. Political online advertising, for example, generates billions of impressions on social media[11], and it has been shown that even small visual details can affect voting intentions[12]. Facebook's hidden ad-delivery mechanisms can increase biases[13] and polarize political campaigns[14]. More generally, increasingly precise microtargeting can harm the democratic process because manipulative messages directed at a specific, but not publicly known target audience, cannot be scrutinized and rebutted by political opponents in a free marketplace of ideas[10]. Furthermore, microtargeting includes "boosted organic content"[*], that is, seemingly personal content that the platforms deliver to target audiences against payment of a fee. This blending of personal communication and advertising can increase the effect of advertisements[15] and, in political advertising, contributes to a distorted picture of democratic discourse[16].

Tech companies have taken some steps towards transparency in the form of ad libraries[†]. These libraries compile ads run on the platforms along with information on the characteristics of the target audience. Due to their complexity and size, however, these libraries are unlikely to help end users; they mainly serve political analysts, journalists, and researchers. Moreover, the information documented on the target audience does not extend beyond coarse variables such as age group or region of residence. This coarseness prevents quantitative studies[11] and rebuttal messages from political competitors[17]. Ad libraries in their present form therefore cannot counter unduly manipulative microtargeting and its effects on individuals.

In addition to the collective harms for democracy, opaque targeting practices are at odds with attitudes across the political spectrum. In a recent representative survey in Germany, Great Britain, and the United States[18], people were inclined to accept personal-

[*] https://www.facebook.com/business/help/317083072148603

[†] E.g., https://www.facebook.com/ads/library and https://transparencyreport.google.com/political-ads

ization based on information that users typically provide knowingly (e.g., age or gender). However, they rejected the use of other types of data, especially information that cannot be easily observed or was not knowingly provided, such as political and sexual orientation or personality dimensions. These sensitive attributes can be inferred from behavioural data without users' input, knowledge, or explicit consent[1;19] by machine learning methods that are inherently opaque (e.g., Facebook's patent "Determining user personality characteristics from social networking system communications and characteristics" [‡]).

An experiment conducted on Facebook has suggested that inferred personality dimensions can be used to personalize ads: Participants were more likely to buy a product when they were targeted with an advertisement that matched their personality type (extravert vs. introvert)[20]. Other studies found that personality-based targeting increased engagement, but did not consistently change attitudes towards a product[21]. Recent results showed that personality-matching political advertising can be more effective in influencing political attitudes and voting intentions than non-matching advertising[22].

Whatever the persuasive power of current practices, microtargeting lacks transparency and contributes to a growing knowledge gap between platforms and users. While platforms are becoming increasingly more sophisticated in collecting data and customizing user experiences, there is a dearth of effective measures that could help counteract the adverse consequences of these developments and reduce the knowledge gap. Clearly, there is no silver bullet to redress this informational asymmetry, but a wide range of actions can and should be taken to increase people's autonomy online[7;23]. At present, countermeasures include an assortment of regulations. One of the more forceful measures is to shield private data from being collected in the first place, using legislation such as the E.U.'s General Data Protection Regulation (GDPR). Yet platforms often bypass the intent of the

[‡] `https://patents.google.com/patent/US8825764B2/en`

regulation by using so-called "dark patterns," which nudge users to disable the privacy-protecting defaults (e.g., by clicking on the visually more salient button)[24]. In addition, irrespective of the ban on collecting sensitive personal data, Facebook is still able to infer such information from behavioural data, and to segment users accordingly[25].

A different, but complementary, strategy to close the knowledge gap is to enhance users' awareness of microtargeting practices. This approach may be more robust to constantly changing targeting methods than regulation of those can ever be. Awareness about microtargeting could empower users to deliberately ignore advertisements or discount political messages that they identify as having been microtargeted. It has been shown that advertisements are less effective when people find out that unacceptable practices have been used to target them (i.e., using information obtained from outside the platform or inferred without user input)[26]. In contrast, trust and effectiveness may increase when the practices used are deemed acceptable[26]. However, current transparency measures, such as the "Why am I seeing this?" button on Facebook, provide only superficial information (e.g., "the advertiser wants to reach people who may be similar to their customers") and have to be actively requested by users[27]. The GDPR mandates other ways to achieve transparency, such as users' "right of access"[§] to the data that platforms hold on them.[¶] Yet most users lack the technical sophistication, motivation, or time to explore those large, unstructured datasets[28].

Thus, although platforms are required to disclose the data they hold about users, in practice, for most users this requirement fails to open the platforms' "black box". Achieving *effective transparency*—that demonstrably enables users to understand what platforms do with their data and what users' choices imply, and to translate this knowledge

[§] `https://gdpr-info.eu/art-15-gdpr/`

[¶] See, for example, `https://myactivity.google.com/more-activity` or `https://www.facebook.com/your_information`.

into behavior—is an important step towards more acceptable business practices and to regaining autonomy for users (e.g., by prompting people to adjust their privacy settings[29]). However, as reviewed above, most current transparency initiatives seem to be exercises in "nominal transparency" with no real regard for whether or not people actually read and digest the information or whether it has any effect on their behaviour.

Here we investigate a cognitive approach to counteract the information asymmetry, which explicitly aims to help people to cope with the lack of transparency. It is inspired by research showing that people can be psychologically "inoculated" against misinformation. For example, explaining misleading argumentation techniques reduces the influence of subsequently presented misinformation[30;31]. In this study, we test whether it is possible to inoculate people against personality-based microtargeting[20] by alerting them to the personality dimension being targeted and thus increasing their ability to identify whether or not an advertisement is targeting them personally. If the success of the intervention depends primarily on people being aware of the personality dimension being targeted, then it may suffice to provide a description of that personality dimension. However, to the extent that people lack relevant self-knowledge[8] about the targeted personality dimension, or fail to spontaneously connect their self-knowledge with the advertisements shown, the inoculation intervention may need to dig deeper. Against this background, we investigate three interventions that differ in their degree of personalization: (1) merely describing the targeted personality dimension, (2) having participants complete a short personality questionnaire (without providing feedback), and (3) providing participants with feedback on their personality based on their responses to the questionnaire. All three interventions are based on the notion of psychological inoculation, an instance of the class of "boosting" interventions, that are, interventions aimed at improving people's competences to make better decisions in light of their own goals[32;23].

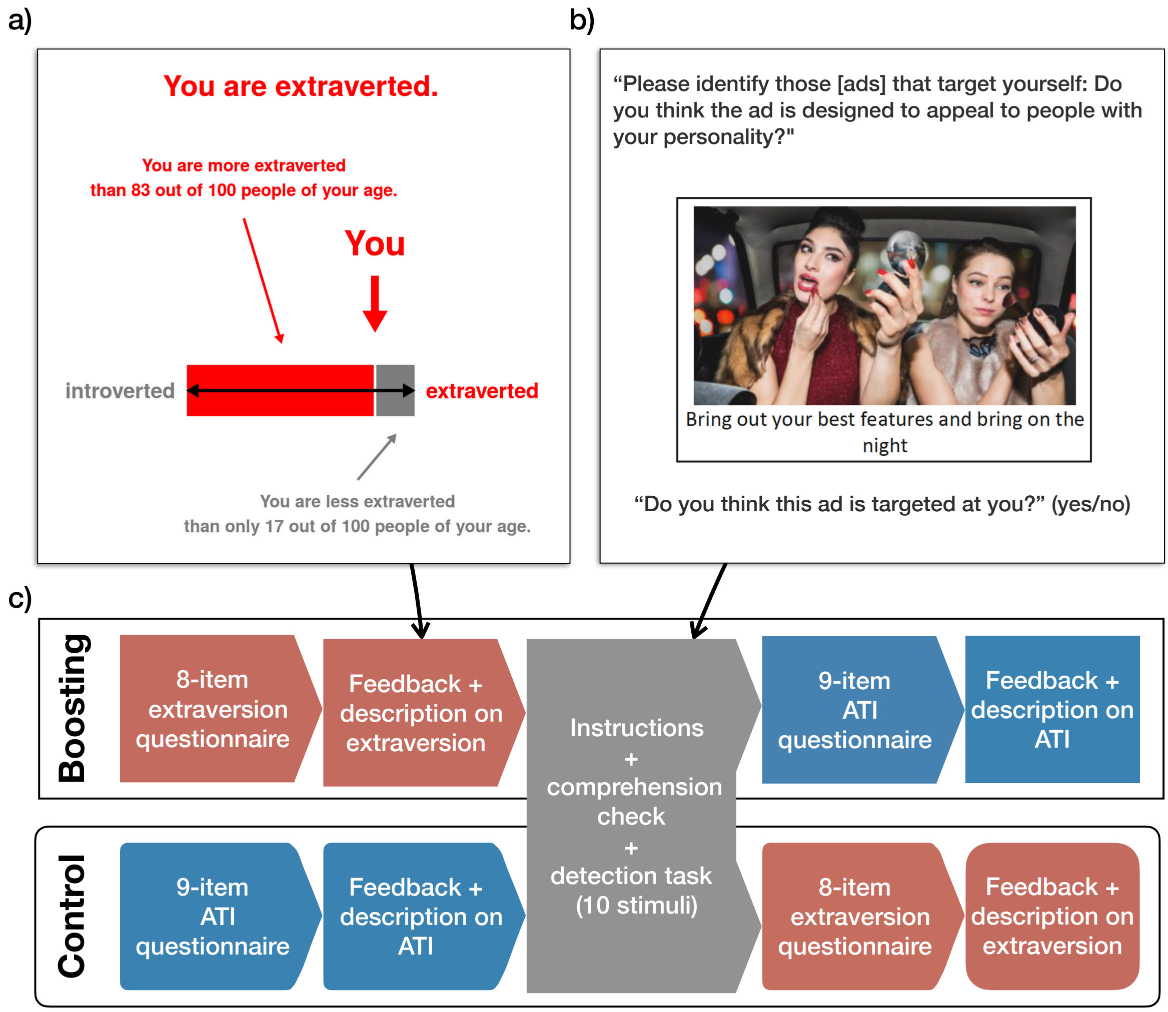


Figure 1: **Elements of the experimental setup used to test the boosting intervention in Experiment 1.** **a** Feedback screen shown to participants after completion of an 8-item personality questionnaire gauging their extraversion level (boosting condition), which includes feedback on their relative rank within an age-matched norm population (from[33]). **b** Instructions of the detection task and example stimulus (for the full set of stimuli, see Fig. S8). **c** Parallel experimental design of the boosting and control conditions—the only difference is that the order of the two personality questionnaires (extraversion and Affinity for Technology Interaction, ATI) and the corresponding feedback were swapped (i.e., before vs. after the detection task).

In two preregistered online studies, we tested the effectiveness of the inoculation approach to boost people's ability to identify ads targeted at their personality in terms of the extraversion–introversion spectrum (N = 828; recruited via Prolific Academic). We used ads developed and validated by Matz and colleagues[20], and therefore recruited from the same population as they did (i.e., female participants from the UK between 18 and 40 years old). In Experiment 1, participants received feedback on their personality (including a general description of the personality dimension), in terms of either their age-matched relative extraversion score (relevant personality feedback, see Fig. 1A and Fig. S3; for full questionnaire, see Fig. S1; items were taken from Srivastava and colleagues[33]) or their affinity for technology interaction (ATI[34]; control feedback, not relevant to the personality dimension in question, see Fig. S4; for questionnaire, see Fig. S2). Participants were then presented with 10 beauty ads (taken from Matz et al.[20]; see Fig. S8); half of which targeted extraverts and the other half introverts. Participants were asked to decide whether each ad was or was not targeted towards their personality (Fig. 1B). A comprehension check ensured that participants understood the instruction (see Fig. S7). However, the specific targeting strategy—that is, that it targeted extraverts vs. introverts—was not revealed to participants. The hypothesis here was:

- **H1**: Participants who reflect on and receive feedback about their relative score on the relevant personality dimension (extraversion; boosting condition) are better able to identify ads that are targeted towards them than are participants who reflect on and receive feedback about their relative score on an unrelated personality dimension (ATI; control condition).

Experiment 2 aimed to disentangle the mechanisms underlying these effects: (1) implicitly hinting at the targeting strategy of the advertiser by describing the relevant personality dimension, (2) encouraging people to reflect on their own position on the relevant personality dimension by having them complete a questionnaire (without providing feedback), and (3) explicitly providing individual feedback on the relevant personality dimension (i.e., degree of extraversion vs. introversion). Experiment 2 was similar to Experiment 1, differing in only two respects. First, half the participants saw only a general description of the relevant personality dimension prior to the detection task (see Fig. S5 and S6 for screenshots). Second, the other half completed the corresponding personality questionnaire (Fig. S1 and S2) after seeing the general description, but did not receive any feedback. Thus, Experiment 2 employed a 2 (control vs. boosting) $\times$ 2 (description only vs. description plus questionnaire) between-subjects design. We tested three mutually exclusive follow-up hypotheses (conditional on hypothesis **H1** being supported):

- **H2a**: The boosting intervention increases accuracy primarily by raising people's awareness of the specific targeting strategy (i.e., differential targeting of extraverts and introverts). This implies that people already have sufficient self-knowledge about their extraversion level and spontaneously apply this knowledge to the task. Thus, fostering self-knowledge is not necessary for boosting accuracy.
- **H2b**: Raising people's awareness of the specific targeting strategy is not sufficient to increase accuracy. People need to actively reflect on their own relevant personality dimensions to recognise that they are being targeted. This also means that simply providing warnings and explanations on platforms will not suffice to enable people to detect microtargeting.
- **H2c**: Neither of the above mechanisms apply; knowledge about one's relative score on the targeted personality dimension (i.e., explicit feedback on one's level of extra- vs. introversion) is required to boost accuracy. This implies that the main reason for people failing to detect microtargeting is a lack of relevant and accurate self-

knowledge about the relevant personality dimension.

Results

Experiment 1. Fig. 2 shows that Experiment 1 supported hypothesis **H1**: Relative to the control condition, participants in the boosting condition on average correctly identified 26 percentage points more ads targeted at them (95% Bayesian credible interval, CI: 18–35)—raising the mean accuracy from 64% (95% CI: 53–73) to 90% (95% CI: 85–94). This difference corresponds to an effect size, expressed in terms of the "common language effect size"[35], of $CL = 0.78$ (95% CI: .70–.84), which here indicates the probability that a randomly selected participant from the boosting condition has higher detection accuracy than a randomly selected participant from the control condition. A value of 0.5 would imply no difference and 1 would imply perfect separation between conditions. Additional analyses, detailed in the Supplementary Information (Supplementary Fig. S9–S11), attest to the robustness of these results. To summarize, the intervention worked (a) for both extraverts and introverts, (b) different levels of education, (c) irrespective of whether participants were clearly or more tentatively classified as extravert or introvert; moreover, the effect (d) was stronger for extraverts than for introverts and (e) also emerged when we measured detection performance independently of any response tendency (lenient vs. strict), in terms of the area under the Receiver Operating Characteristics curve[36] (AUC; based on participants' confidence in their detection decisions). Overall, these results demonstrate that it is possible to improve people's ability to detect targeted advertisements through a short, simple boosting intervention.

Although the results of Experiment 1 were unambiguous, the study left one key question unanswered: What drives the intervention's success? Is it sufficient to hint at the strategy used by the advertiser, thus raising participant awareness (**H2a**)? Or is it neces-

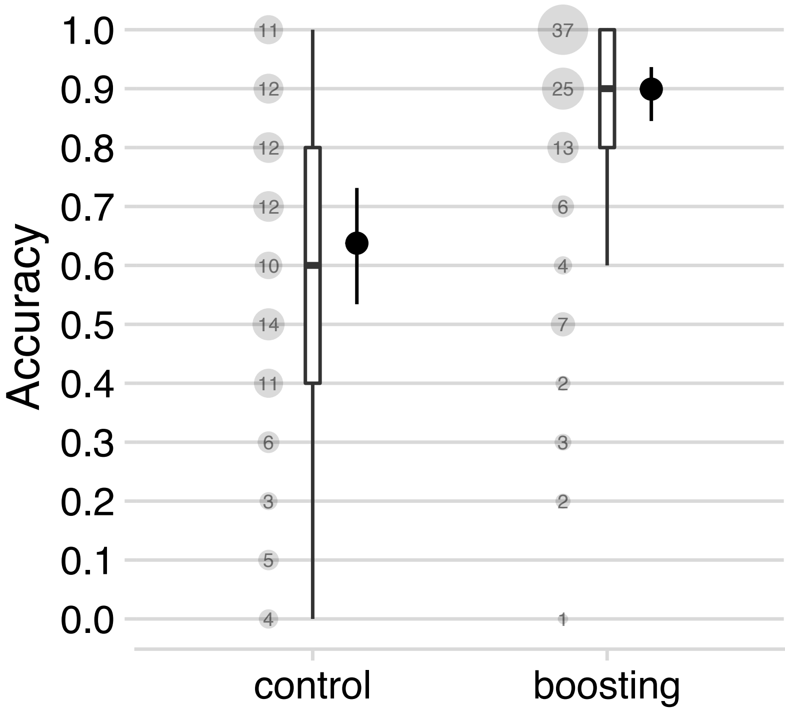


Figure 2: **Effect of boosting and control interventions on the accuracy of detecting targeted advertisements (Experiment 1)**. See Fig. 1 for the experimental setup, where participants in the boosting conditions received feedback about their extraversion prior to the task. Point ranges show the Bayesian point estimate and 95% Bayesian credible interval for the probability of correctly detecting a targeted advertisement (based on a multilevel logistic regression model; see Methods for details). In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed. The area of the dots and their numbers denote the within-condition percentage of participants for each of the 11 possible values for a participant's proportion of correct decisions (given the 10 ads).

sary that participants also reflect on their own relevant personality dimensions (**H2b**)? Or is explicit knowledge of one's relative score on the relevant personality dimension required (**H2c**)? In Experiment 2, we set out to tease apart these three different mechanisms.

Experiment 2. The results of Experiment 2 support hypothesis **H2b** (Fig. 3): reflecting on one's relevant personality dimensions—without receiving any relevant feedback—is necessary, but also sufficient to boost people's ability to identify ads that have been targeted at them. The boosting condition that included the extraversion questionnaire improved participants' performance by, on average, 10 percentage points (95% CI: 2–20) compared to the boosting condition with only the extraversion description, raising mean accuracy from 72% (95% CI: 63–81) to 83% (95% CI: 76–88); this difference corresponds to a common language effect size of $CL = .62$ (95% CI: .52–.71). This positive effect is at odds with hypothesis **H2c**, according to which explicit knowledge of one's level on the relevant personality dimension is necessary for the intervention to work. By contrast, participants who only read the extraversion description performed no better than participants who read the unrelated description of the ATI personality dimension ($CL = .52$, 95%: .43–.62); the latter participants correctly identified 70% of the ads (95% CI: 61–77). This result is at odds with hypothesis **H2a**, according to which hinting at the strategy used by the advertiser is sufficient for the intervention to work. Importantly, the effectiveness of self-reflection was not generic: performance was boosted only when people reflected on the relevant personality dimension. Participants who read the unrelated description of ATI and then completed the ATI questionnaire correctly identified 68% of the targeted ads (95% CI: 57–77)—that is, 15 percentage points (95 CI: 7–24) fewer than the participants who reflected on the relevant personality dimension (i.e., extraversion; $CL = .66$, 95%: 58–74).

Additional analyses, detailed in the Supplementary Information (Supplementary Fig. S12–S14), attest to the robustness of these results. To summarize, the results hold (a) for both extraverts and introverts, (b) different levels of education; moreover, the effect (c) was stronger for extraverts than for introverts, and (d) also emerged when we measured de-

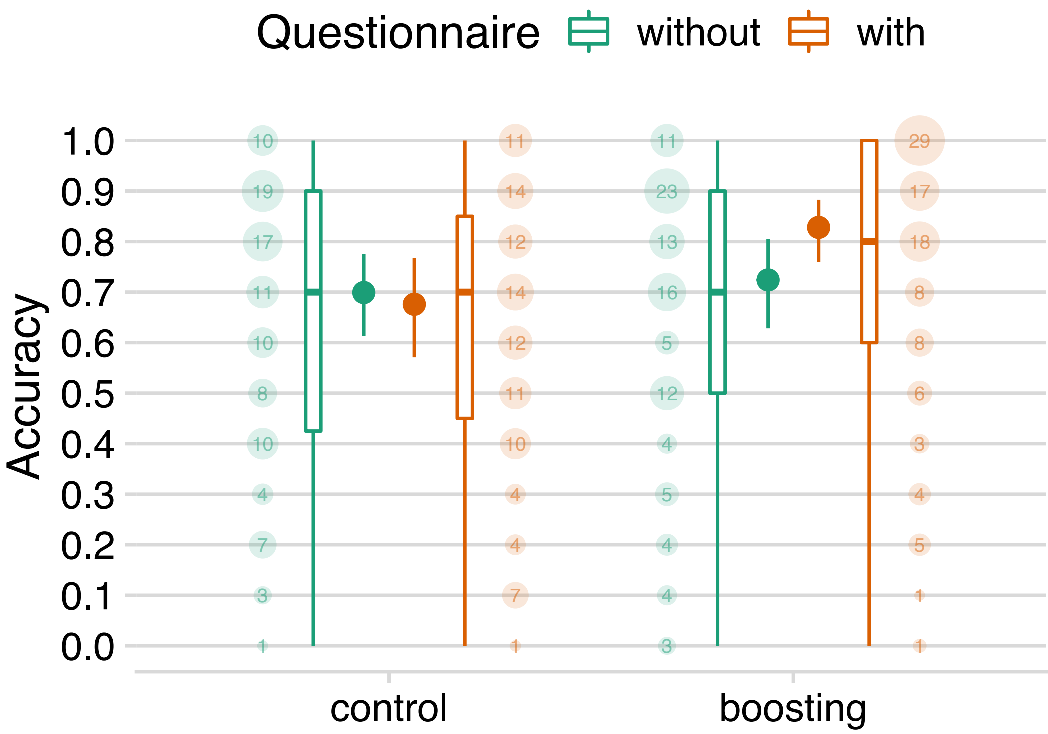


Figure 3: **Effect of boosting and control interventions on the accuracy of detecting targeted advertisements (Experiment 2).** Participants in the boosting conditions either just read a description of the relevant personality dimension prior to the task ("without questionnaire"), or additionally filled out the short questionnaire from Experiment 1, but without feedback ("with questionnaire"). Point ranges show the Bayesian point estimate and 95% Bayesian credible interval for the probability of correctly detecting a targeted advertisement (based on a multilevel logistic regression model; see Methods for details). In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed. The area of the dots and their numbers denote the within-condition percentage of participants for each of the 11 possible values for a participant's proportion of correct decisions (given 10 ads).

tection performance independently of any response tendency (lenient vs. strict), in terms of the AUC[36] (based on participants' confidence in their detection decisions). However, for moderately extraverted participants, we did not observe an effect of filling out the relevant (vs. unrelated) questionnaire (Fig. S12 & S13); for those participants the explicit feedback about their personality seems necessary for improving their detection accuracy (cf. Experiment 1). In summary, Experiment 2 showed that the boosting intervention can improve detection accuracy even without provision of explicit feedback, whereas merely describing the relevant personality dimension was insufficient.

Conclusion

Two experiments demonstrated that prompting people to reflect on a targeted personality dimension—by means of a short and simple intervention—boosts their ability to identify ads that target them on the basis of that personality dimension. Merely providing a description of the targeted personality dimension did not enhance detection accuracy. Completing a short personality questionnaire about the targeted personality dimension was sufficient to increase accuracy—even if people did not receive any feedback. This result resonates with the recent finding that simple interventions, such as exposing misinformation strategies, can help to inoculate people against misinformation strategies[37;38]. Further research needs to clarify the cognitive mechanisms underlying these effects; the extent to which the observed increases in detection ability translate into improved downstream outcomes (e.g., in terms evaluating and responding to ads); and the extent to which the effects generalize to other personality dimensions, domains (e.g., political advertising or misinformation), and populations.

Boosting interventions—which by definition target people's competences—have the advantage that they can often be deployed independently of any platform or technology. That is, they do not need to interface with a platform's information architecture and are therefore not dependent on the platform's cooperation (in terms of access and maintaining interoperability). Compared with, say, an intervention where advertisements are labelled within the platform's interface, an intervention targeting people's competences is therefore more robust with respect to constantly changing technology, advertising strategies, and the tech companies' level of cooperation. Furthermore, as boosting interventions aim to improve people's competences, they have the potential to generalize beyond the immediate context in which they were initially deployed[32;39]. Self-reflection tools aimed at helping people increase their awareness of their vulnerabilities to microtargeting could be deployed on independent websites or apps—or even as "analogue" tools (e.g., a checklist on a printed flyer). Such tools would need to cover a range of the most relevant microtargeting dimensions in order to offer effective protection.

Going one step further: Because the GDPR requires platforms to disclose what data they hold about their users, it is now feasible in many countries to implement tools aiming to raise user awareness of the specific data held on them. The information pages established by some platforms[‖] in principle allow every motivated and technically savvy user to download their personal dataset and explore what is known about them. Digital boosting tools could automatically access this information and provide simple interfaces that encourage active exploration. Such tools could empower less tech-savvy users to find out what platforms know about them—information that might enable the precise targeting of commercial or political advertising. This could be done without processing the user's personal data in any way; as we have shown, the intervention was effective even without personalized feedback—the only necessary condition was active reflection on the part of the user.

The platforms' lack of transparency about their data handling and business practices is an important and much-discussed issue. Our results highlight that researchers, policy

[‖]E.g., `https://myactivity.google.com/more-activity` or `https://www.facebook.com/your_information` makers and other stakeholders also need to consider the issue of "effective transparency," that is, when and how platforms' transparency systems with respect to, say, microtargeting practices actually empower users in practice—and are not simply an exercise in "nominal transparency." Just because something is technically or legally "transparent" does not guarantee that users can or will engage with it—and even if they do, they still may not understand what it means for them.

Our findings showed that merely describing a personality dimension does not suffice to improve people's ability to detect microtargeting. This finding raises the more general question of whether other measures aiming to achieve transparency by merely describing information to users—such as Google's `https://myactivity.google.com` or Facebook's `https://facebook.com/your_information`—may likewise fail to help users understand this data and how it is used.

To conclude, our findings support a vision of *effective transparency*, where platforms—and other players—provide tools that help people to actively explore and reflect on the information held about them, thereby *boosting* their awareness of that knowledge—instead of steering people away from it by burying privacy settings deep within the settings menu or dumping large amounts of unstructured data on users and leaving them to sort it out for themselves. We argue that a boosting approach is a promising component in a broader, evidence-based policy mix aimed at giving people the necessary legal rights and competences to reclaim their individual autonomy in the online world[7;23].

Methods

All data and code are publicly available at `https://osf.io/ne4r9/`.

Experiment 1. The preregistration of the study can be accessed at `http://aspredicted.org/blind.php?x=wu6sk7` and includes, among other things, the research question, hypothesis **H1**, the primary outcome variable, planned sample size, exclusion criteria, and the exact specification of the multilevel logistic regression model detailed below. We report all data exclusions, all manipulations, and all measures used in the study (see Simmons et al.[40]). The experiment was programmed using *formr* (`https://formr.org`)[41].

Participants. We collected responses from 318 participants (boosting condition $N = 158$, control condition $N = 160$, randomly allocated on the fly) via Prolific Academic, an online survey platform whose participants are more diverse and less familiar with experimental procedures than Amazon Mechanical Turk workers[42]. Mirroring the population targeted in[20], we recruited female participants between the ages of 18 and 40 years who were UK residents fluent in English; we did not invite participants who already participated in a pilot study, via the prescreening functionality of Prolific. Participants received £2 for completing the study. Consistent with the preregistered exclusion criteria, we excluded 25 participants for non-completion (13 in the boosting condition, 12 in the control condition), 2 participants for giving different responses to the two age questions (1 in the boosting condition, 1 in the control condition), and 6 participants for failing the comprehension check (4 in the boosting condition, 2 in the control condition). We also excluded 1 participant (from the boosting condition) with a relative extraversion percentile of exactly 50%, as no extraversion personality type can be assigned for participants with this value (see Conditions for further information). The final sample thus comprised 286 participants, $N = 139$ in the control condition and $N = 145$ in the boosting condition. The median age of participants was 30 years (first and third quartile: $Q_1 = 26$ and $Q_3 = 34$ years).

Conditions. In the boosting condition, participants completed an 8-item extraversion questionnaire (see Fig. S1). Based on their responses, they received personalized feedback (see Fig. 1a and Fig. S3) on their extraversion score relative to a large sample of online participants (from Srivastava et al.[33]); this was truthful feedback, calculated for each participant on the fly. In particular, participants were told whether their personality tended more towards extraversion ("You are extraverted") or introversion ("You are introverted"). A participant's percentile was shown both numerically and visually, expressed as how many of 100 random people were more and less extraverted (for participants categorized as extraverts) or introverted (for participants categorized as introverts) than the participant themselves. The feedback was accompanied by a simple definition of extraversion adopted from Wikipedia[**] (see Fig. 1 and S3). We enforced a 1-minute wait on the feedback screen to ensure that participants encoded the feedback. The control condition followed the same procedure, but participants completed an unrelated, 9-item questionnaire tapping their propensity to naturally interact with technical systems (Affinity for Technology Interaction, ATI; for full questionnaire, see Fig. S2). The ATI feedback and the description of the dimension was presented in a format analogous to that used in the boosting condition (see Fig. S5).

Questions and distributional information for the raw scores were adopted from Srivastava and colleagues[33] for extraversion and from Franke et al.[34] for ATI. Srivastava et al.[33] provide the mean and standard deviation (SD) of the raw scores for each age year between 21 and 60; we were thus able to provide age-matched feedback for extra-/introversion (for participants aged 18–20 years, we used the norms for age 21 years). For ATI, we used

[**] `https://web.archive.org/web/20190801042657/https://en.wikipedia.org/wiki/Extraversion_and_introversion`

the mean and SD of the sample "S5-full" reported in Franke et al.[34] (i.e., no age-specific norms were available). To achieve consistency across questionnaires, we presented both questionnaires on a 5-point Likert scale. Because the ATI norm study[34] used a 6-point scale, we rescaled the mean and SD (original norm values $M = 3.61$, $SD = 1.09$, rescaled values $M = 3.09$, $SD = 0.86$). See SI for extraversion and ATI questionnaires and sample feedback (Figs. S1 and S2 for questionnaires; Figs. 1a, S3 and S4 for feedback and definitions).

Ad targeting detection task. We presented the female participants with 10 ads for beauty products (taken from Matz et al.[20]) in random order. Five of the ads were specifically designed to target extraverts; five target introverts (for the full set of stimuli, see Fig. S8). Each ad consisted of a picture and a slogan. "Extraverted" ads emphasized socially stimulating contexts (e.g., "Love the spotlight"), whereas "introverted" ads emphasized socially less stimulating contexts (e.g., "Beauty isn't always about being on show"). The original study[20] validated the stimuli by showing that extraverted ads were rated as more extraverted than introverted ads (and vice versa) and that microtargeting extraverts and introverts on Facebook led to higher sales of actual products in a web shop (relative to mismatched ads).

Right before the beginning of the ad targeting detection task, participants received the following instructions: "In the following you will be shown ads that are all designed for women, but are additionally targeted at different personality types. Please identify those that target yourself: Do you think the ad is designed to appeal to people with your personality? Or do you think it is designed to appeal to people with a different personality?" That is, in this study, microtargeting was defined as addressing participants by tailoring ads to aspects of their personality. This was followed by a comprehension check (see Fig. S7): "Please complete the following sentence. For the following ads, I need to rate whether I think the ad is ...," followed by the options "copied from a previous ad," "targeted towards my personality type," "appealing to me," and "going to be effective when aired." If participants did not select "targeted towards my personality type," the question was repeated (max. two times) with the response options presented in a different order. As per preregistration, we included participants in the analysis only if they passed the comprehension check within the maximum of three attempts. For each ad, participants were then asked whether it was targeted towards their personality type: "Do you think this ad is targeted at you?" ("yes" vs. "no"; see Fig. 1b). Participants also indicated their decision confidence by responding to the question "How confident are you with your choice?" (Likert scale ranging from 1 = "not confident" to 5 = "very confident").

Primary outcome measure. The primary dependent variable was a participant's decision about whether or not a particular ad was targeted towards her personality ("yes" vs. "no"). We classified each participant as either extravert (percentile > 50%) or introvert (percentile < 50%) on the basis of their percentile rank for extraversion. Based on this categorization, each participant's decisions were then scored as either correct or incorrect. Specifically, a decision was scored as correct if an extraverted participant responded that an extraverted ad was targeted at her or an introverted ad was not targeted at her. A decision was scored as incorrect if she responded that an extraverted ad was not targeted at her or that an introverted ad was targeted at her. The opposite coding was used for introverted participants.

Procedure. Participants were randomly assigned to one of two conditions. In the boosting condition, participants first completed the extraversion questionnaire and received feedback on their relative extraversion score (see Fig. 1a), then evaluated the targeting of the ads, and finally completed the ATI questionnaire and were given feedback on their relative ATI score (see Fig. S4). In the control condition, the position of the extraversion and ATI questionnaires (plus their respective feedback) was switched. Participants were asked to indicate their age in both the extraversion and the ATI questionnaire; this measure was used as a response consistency measure (see exclusion criteria). At the end of the study, a question about education was administered.

Statistical analysis. We used a Bayesian multilevel logistic regression model implemented in the R package *brms*[43;44] and its default, vague priors (see code for exact specifications). The preregistered model's syntax is

```
correct ~ 1 + condition + (1 | id) + (1 + condition | stimuli)
```

where `correct` is `1` for correct and `0` for incorrect classification decisions, `condition` is a deviation-coded factor variable for the boosting vs. control condition, `id` is a unique identifier for participants, and `stimuli` is a unique identifier for ads. Note that `(1 + condition | stimuli)` allows the treatment effect to differ in size by ad. Four Markov chain Monte Carlo (MCMC) chains, each with 8,000 samples, were run; the first 8,000 samples were discarded as warm-up. The MCMC diagnostics indicated good convergence (see Supplementary Information).

Posterior distributions were summarized using the median (point estimate) and 95% credible interval (uncertainty interval). Based on the model parameters (see Supplementary Information for a summary table), we derived posterior distributions for several key statistics of interest: (a) the probability of a correct detection decision in both conditions, (b) the percentage point difference, and (c) effect sizes between the two conditions.

We express effect sizes using the "common language effect size" (CL)[35], which indicates the probability that a randomly selected participant from one condition has a higher value than a randomly selected participant from another condition; a value of 0.5 implies no difference and 1 would imply perfect separation between conditions. CL is well suited to compare conditions in a multilevel logistic regression model because—unlike the commonly used measures of effect size based on standardized mean differences—CL is invariant to monotonical transformations. That is, its value does not depend on an arbitrary decision on whether to look at the results in log-odds or probability space. We derive the posterior distribution of a CL-comparison based on the model's posterior distributions for the participant-population mean and standard deviation in each condition (setting the item effects to zero, that is, considering the average item).

Experiment 2. Experiment 2 was identical to Experiment 1, with the exceptions specified here. This study's preregistration can be accessed at `http://aspredicted.org/blind.php?x=39ik6v` and includes, among other things, the research question, hypotheses **H2a–c**, the primary outcome variable, planned sample size, exclusion criteria, and the exact specification of the multilevel logistic regression model detailed below. We report all data exclusions, all manipulations, and all measures in the study (see Simmons et al.[40]).

Participants. 638 participants (boosting condition with questionnaire $N = 173$, boosting condition without questionnaire $N = 130$, control condition with questionnaire $N = 164$, control condition without questionnaire $N = 171$, randomly allocated on the fly), recruited from Prolific Academic, received £2 for completing the study. Experiment 2 involved two additional prescreening criteria on Prolific, namely, that they had not participated in Experiment 1, its pilot, or a pilot study for Experiment 2. Consistent with the preregistered exclusion criteria, we excluded 78 participants for non-completion (16 in the boosting condition with questionnaire, 10 in the boosting condition without questionnaire, 29 in the control with questionnaire, 23 in the control without questionnaire), 5 participants for an extraversion percentile of exactly 0.5 (3 in the boosting condition without questionnaire, 2 in the control with questionnaire), 2 participants for giving different responses for the two age questions (1 in the boosting condition with questionnaire, 1 in the control with questionnaire), and 10 participants for failing the comprehension check (3 in the boosting condition with questionnaire, 3 in the boosting condition without questionnaire, 2 in the control with questionnaire, 2 in the control without questionnaire). Our final sample size was thus 544 participants: boosting condition with questionnaire: $N = 153$ (i.e., 88% retained); boosting condition without questionnaire: $N = 114$ (i.e., 88% retained); control condition with questionnaire: $N = 131$ (i.e., 80% retained); and control condition without questionnaire: $N = 146$ (i.e., 85% retained). The median age of participants was 29 years (first and third quartiles: $Q_1 = 24$ and $Q_3 = 34$ years).

Treatments. We tested two simplifications of the intervention implemented in Experiment 1: providing no feedback on the questionnaire and providing only a relevant definition. Here, before completing the ad targeting detection task, participants were shown a definition of either extraversion (relevant personality dimension, see Fig. S5) or ATI (control personality dimension, see Fig. S6). Within each of these two groups, half of the participants additionally completed the same questionnaire on the respective personality dimension as in Experiment 1, but without any feedback. In contrast to Experiment 1, where the definition of the personality dimension was shown along with the feedback (based on the previously completed questionnaire), all participants in Experiment 2 first saw a definition of the respective personality dimension.

Procedure. Participants were randomly assigned to one of four conditions in a 2 (Intervention relevance: boosting vs. control) $\times$ 2 (Intervention type: Definition only vs. Definition + Questionnaire) between-subjects design. In both boosting conditions, participants first received a description of the relevant personality dimension: extraversion (see Fig. S5). In the questionnaire conditions, participants then additionally completed the relevant extraversion inventory (see Fig. S1). Participants in both boosting conditions were then asked to identify ads targeted towards their personality. After the ad targeting detection task, they were given feedback on their relative extraversion score (as in Experiment 1, see Figs. 1a, S3, and S4); they then completed the ATI questionnaire and were given feedback on their relative ATI score (see Fig. S4). Because all feedback was provided *after* the detection task, it could not have any effect on the detection task; we included the feedback simply to satisfy participants' curiosity. For the two control conditions, the position of the extraversion and ATI descriptions (and, in the case of the condition with questionnaire, the corresponding questionnaire) was switched.

Statistical analysis. The preregistered model's syntax is

```
correct ~ 1 + relevance * questionnaire + (1 | id)
+ (1 + relevance * questionnaire | stimuli)
```

where `correct` is `1` for correct and `0` for incorrect classification decisions, `relevance` is a deviation-coded factor variable for the boosting vs. control conditions (i.e., relevant vs. unrelated personality dimension, respectively), `questionnaire` is a deviation-coded factor variable indicating whether or not participants were administered a questionnaire, `id` is a unique identifier for participants, and `stimuli` is a unique identifier for ads. `relevance * questionnaire` indicates that the model includes the two main effects as well as the interaction `relevance : questionnaire`. Note that

```
(1 + relevance * questionnaire | stimuli)
```

allows the treatment effects (i.e., two main effects and their interaction) to differ in size by ad. Four MCMC chains, each with 8,000 samples, were run; the first 4,000 samples were discarded as warm-up. The MCMC diagnostics indicated good convergence (see Supplementary Information). Based on the model's parameters (see Supplementary Information for a summary table), we derived posterior distributions for several key statistics of interest: (a) the probability of a correct detection decision in each condition, (b) percentage point differences, and (c) effect sizes between conditions. For more information on the analysis approach, see Experiment 1 above.

References

[1] Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* **110**, 5802–5805 (2013).

[2] Hinds, J. & Joinson, A. N. What demographic attributes do our digital footprints reveal? A systematic review. *PLOS ONE* **13**, e0207112 (2018).

[3] Deng, A. *et al.* A/B testing at scale: Accelerating software innovation. In *SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1395–1397 (Association for Computing Machinery, New York, NY, 2017). URL `https://doi.org/10.1145/3077136.3082060`.

[4] Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* **116**, 4156–4165 (2019).

[5] Nissenbaum, H. A contextual approach to privacy online. *Daedalus* **140**, 32–48 (2011).

[6] Auxier, B. *et al.* Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center* (2019, November 15).

[7] Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R. & Hertwig, R. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* **4**, 1102–1109 (2020).

[8] Vazire, S. & Carlson, E. N. Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass* **4**, 605–620 (2010).

[9] Coppock, A., Hill, S. J. & Vavreck, L. The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances* **6** (2020).

[10] Heawood, J. Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity* **23**, 429–434 (2018).

[11] Edelson, L., Sakhuja, S., Dey, R. & McCoy, D. An analysis of United States online political advertising transparency. *arXiv.org* **1902.04385** (2019).

[12] Dan, V. & Arendt, F. Visual cues to the hidden agenda: Investigating the effects of ideology-related visual subtle backdrop cues in political communication. *The International Journal of Press/Politics* 1940161220936593 (2020).

[13] Ali, M. *et al.* Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum.-Comput. Interact.* **3** (2019). URL `https://doi.org/10.1145/3359301`.

[14] Ali, M., Sapiezynski, P., Korolova, A., Mislove, A. & Rieke, A. Ad delivery algorithms: The hidden arbiters of political messaging. *ArXiv* **1912.04255** (2019).

[15] Bakshy, E., Eckles, D., Yan, R. & Rosenn, I. Social influence in social advertising: Evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, 146–161 (Association for Computing Machinery, New York, NY, USA, 2012). URL `https://doi.org/10.1145/2229012.2229027`.

[16] Medina Serrano, J. C., Papakyriakopoulos, O. & Hegelich, S. Exploring political ad libraries for online advertising transparency: Lessons from germany and the 2019 european elections. In *International Conference on Social Media and Society*, SM-Society'20, 111–121 (Association for Computing Machinery, New York, NY, USA, 2020). URL `https://doi.org/10.1145/3400806.3400820`.

[17] Bolden, S. E., McKernan, B. & Stromer-Galley, J. Facebook political advertising transparency report. Tech. Rep., Syracuse University (2020).

[18] Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S. & Herzog, S. M. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from germany, great britain, and the us (2020). URL `psyarxiv.com/3q4mg`.

[19] Youyou, W., Kosinski, M. & Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* **112**, 1036–1040 (2015).

[20] Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* **114**, 12714–12719 (2017).

[21] Winter, S., Maslowska, E. & Vos, A. L. The effects of trait-based personalization in social media advertising. *Computers in Human Behavior* **114**, 106525 (2021). URL `http://www.sciencedirect.com/science/article/pii/S0747563220302776`.

[22] Zarouali, B., Dobber, T., Pauw, G. D. & de Vreese, C. Using a personality-profiling algorithm to investigate political microtargeting: Assessing the persuasion effects of personality-tailored ads on social media. *Communication Research* 0093650220961965 (2020).

[23] Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest* (in press).

[24] Nouwens, M., Liccardi, I., Veale, M., Karger, D. & Kagal, L. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13 (Association for Computing Machinery, New York, NY, 2020). URL `https://doi.org/10.1145/3313831.3376321`.

[25] Cuevas, Á., Cabañas, J. G., Arrate, A. & Cuevas, R. Does Facebook use sensitive data for advertising purposes? Worldwide analysis and GDPR impact. *ArXiv* **1907.10672** (2019).

[26] Kim, T., Barasz, K. & John, L. K. Why am I seeing this ad? The effect of ad transparency on ad effectiveness. *Journal of Consumer Research* **45**, 906–932 (2019).

[27] Andreou, A. *et al.* Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *NDSS 2018 - Network and Distributed System Security Symposium*, 1–15 (San Diego, United States, 2018). URL `https://hal.archives-ouvertes.fr/hal-01955309`.

[28] van Ooijen, I. & Vrabec, H. U. Does the GDPR enhance consumers' control over personal data? An analysis from a behavioural perspective. *Journal of Consumer Policy* **42**, 91–107 (2019).

[29] Parra-Arnau, J., Achara, J. P. & Castelluccia, C. Myadchoices: Bringing transparency and control to online advertising. *ACM Trans. Web* **11** (2017). URL `https://doi.org/10.1145/2996466`.

[30] Cook, J., Lewandowsky, S. & Ecker, U. K. H. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE* **12**, e0175799 (2017).

[31] van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A. & Lewandowsky, S. Inoculating against misinformation. *Science* **358**, 1141–1142 (2017).

[32] Hertwig, R. & Grüne-Yanoff, T. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science* **12**, 973–986 (2017).

[33] Srivastava, S., John, O. P., Gosling, S. D. & Potter, J. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology* **84**, 1041 (2003).

[34] Franke, T., Attig, C. & Wessel, D. A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) scale. *International Journal of Human–Computer Interaction* **35**, 456–467 (2019).

[35] Ruscio, J. A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods* **13**, 19 (2008).

[36] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27** (2006).

[37] Roozenbeek, J. & Van Der Linden, S. The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research* **22**, 570–580 (2019).

[38] Maertens, R., Roozenbeek, J., Basol, M. & van der Linden, S. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied* (2020).

[39] Zimmerman, S., Herzog, S. M., Elsweiler, D., Chamberlain, J. & Kruschwitz, U. Towards a framework for harm prevention in web search. In *BIRDS 2020: Bridging the gap between information science, information retrieval and data science* (2020). URL `https://birds-ws.github.io/birds2020/assets/papers/BIRDS2020_zImmerman.pdf`.

[40] Simmons, J. P., Nelson, L. D. & Simonsohn, U. A 21 word solution. *SSRN* **https://ssrn.com/abstract=2160588** (2012).

[41] Arslan, R. C., Walther, M. P. & Tata, C. S. formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods* **52**, 376–387 (2020).

[42] Peer, E., Brandimarte, L., Samat, S. & Acquisti, A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* **70**, 153–163 (2017).

[43] Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).

[44] Bürkner, P.-C. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* **10**, 395–411 (2018).

Acknowledgments

We acknowledge the Survey Software *formr* (`formr.org`)[41], and thank Ruben Arslan, Anastasia Kozyreva and Briony Swire-Thompson for helpful discussions, Sandra Matz for consulting us with regard to the stimuli[20], and Susannah Goss for editing the manuscript.

We acknowledge financial support from the Volkswagen Foundation (funding initiative "Artificial Intelligence and the Society of the Future")[††]. SL was supported by a Humboldt Award from the Humboldt Foundation in Germany during part of this work. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

PLS and SH contributed equally. All authors conzeptualized the study. PLS, MG, SH, and SL designed the experiments. PLS and MG programmed and conducted the experiments. PLS and SH analyzed and SH visualized the data. PLS, SH, and MG wrote the original draft. All authors reviewed and edited the manuscript.

Competing interest

The authors declare no competing interests.

Ethics declaration

The study was approved by the IRB committee of the Max Planck Institute for Human Development.

Data and code availability

All data and code are publicly available at `https://osf.io/ne4r9/`.

[††]Project description: `http://portal.volkswagenstiftung.de/search/projectDetails.do?ref=95932`

Supplementary material: A simple self-reflection intervention boosts the detection of targeted advertising

Philipp Lorenz-Spreen[1*†], Michael Geers[1], Thorsten Pachur[1], Ralph Hertwig[1], Stephan Lewandowsky[2,3], & Stefan M. Herzog[1*]

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[2]School of Psychological Science and Cabot Institute, University of Bristol, Bristol, United Kingdom

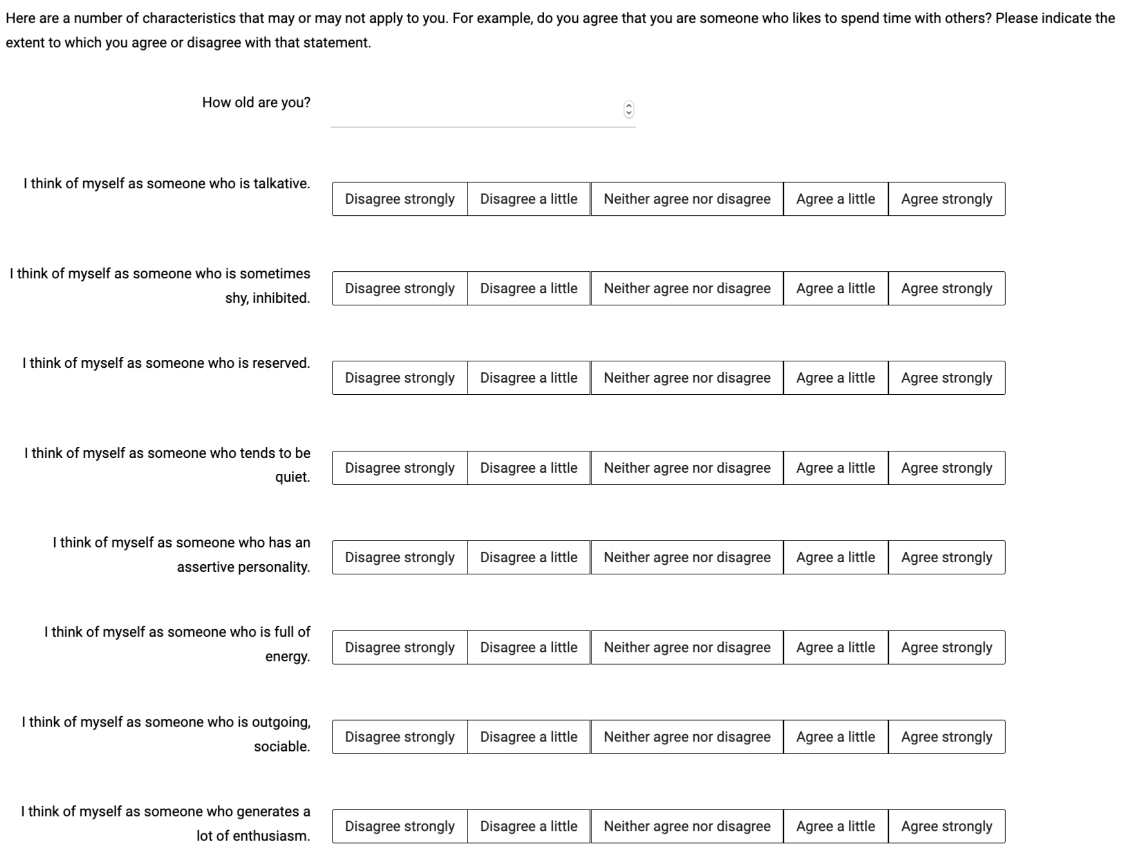
[3]School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia

[*]contributed equally; [†]lorenz-spreen@mpib-berlin.mpg.de

Contents

1 Screenshots of Experiments

1.1 Personality questionnaires



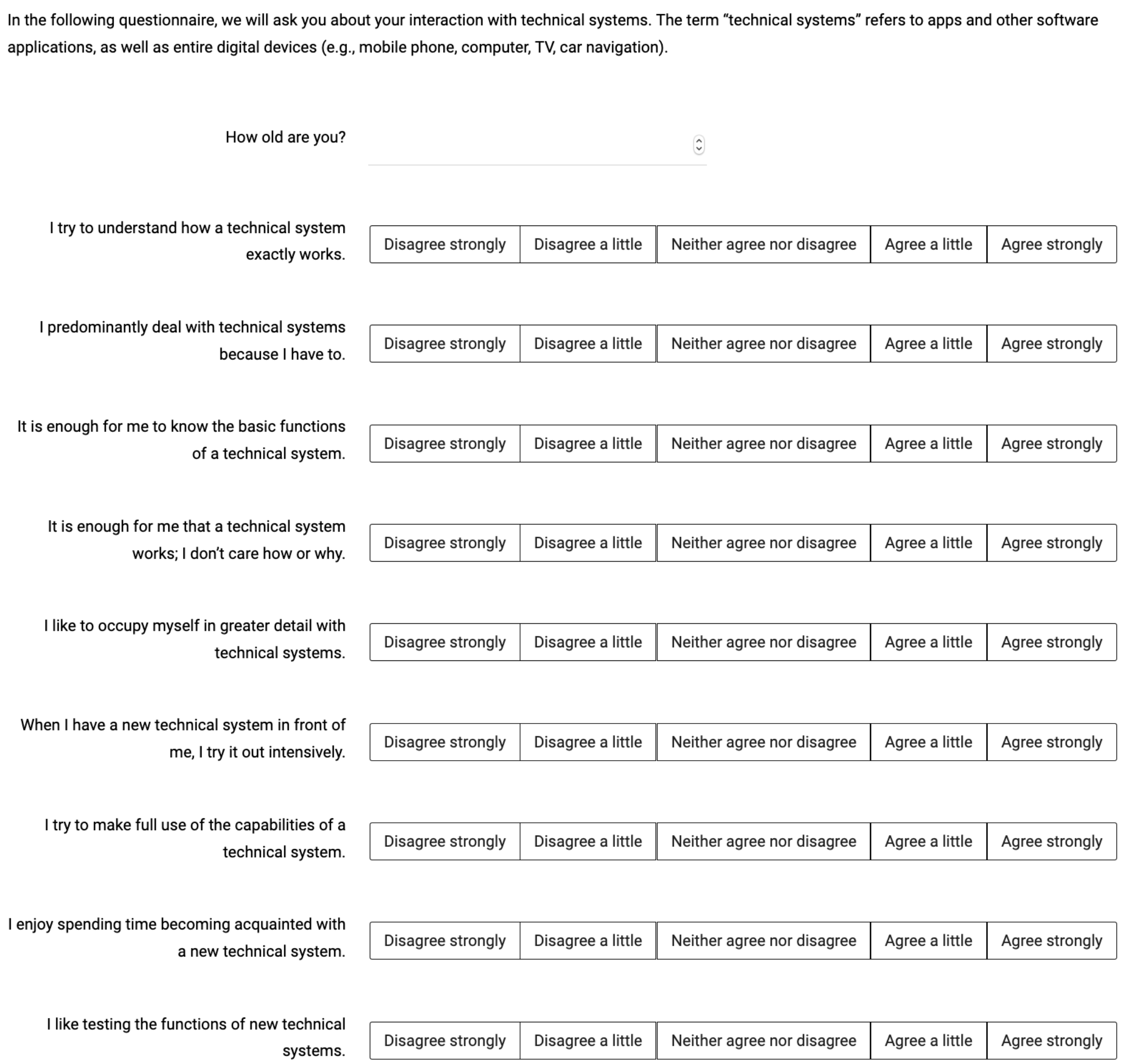
Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please indicate the extent to which you agree or disagree with that statement.

How old are you?

I think of myself as someone who is talkative.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who is sometimes shy, inhibited.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who is reserved.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who tends to be quiet.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who has an assertive personality.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who is full of energy.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who is outgoing, sociable.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I think of myself as someone who generates a lot of enthusiasm.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly

Figure S1: **Extraversion personality questionnaire** used in Experiments 1 and 2. These 8 items are a subset of the 44-items extraversion scale[2]

.



In the following questionnaire, we will ask you about your interaction with technical systems. The term "technical systems" refers to apps and other software applications, as well as entire digital devices (e.g., mobile phone, computer, TV, car navigation).

How old are you?

I try to understand how a technical system exactly works.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I predominantly deal with technical systems because I have to.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
It is enough for me to know the basic functions of a technical system.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
It is enough for me that a technical system works; I don't care how or why.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I like to occupy myself in greater detail with technical systems.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
When I have a new technical system in front of me, I try it out intensively.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I try to make full use of the capabilities of a technical system.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I enjoy spending time becoming acquainted with a new technical system.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I like testing the functions of new technical systems.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly

Figure S2: **Affinity for Technology Interaction (ATI) questionnaire** used in Experiments 1 and 2. Items are taken from Franke et al.[3].

1.2 Personality feedback screens

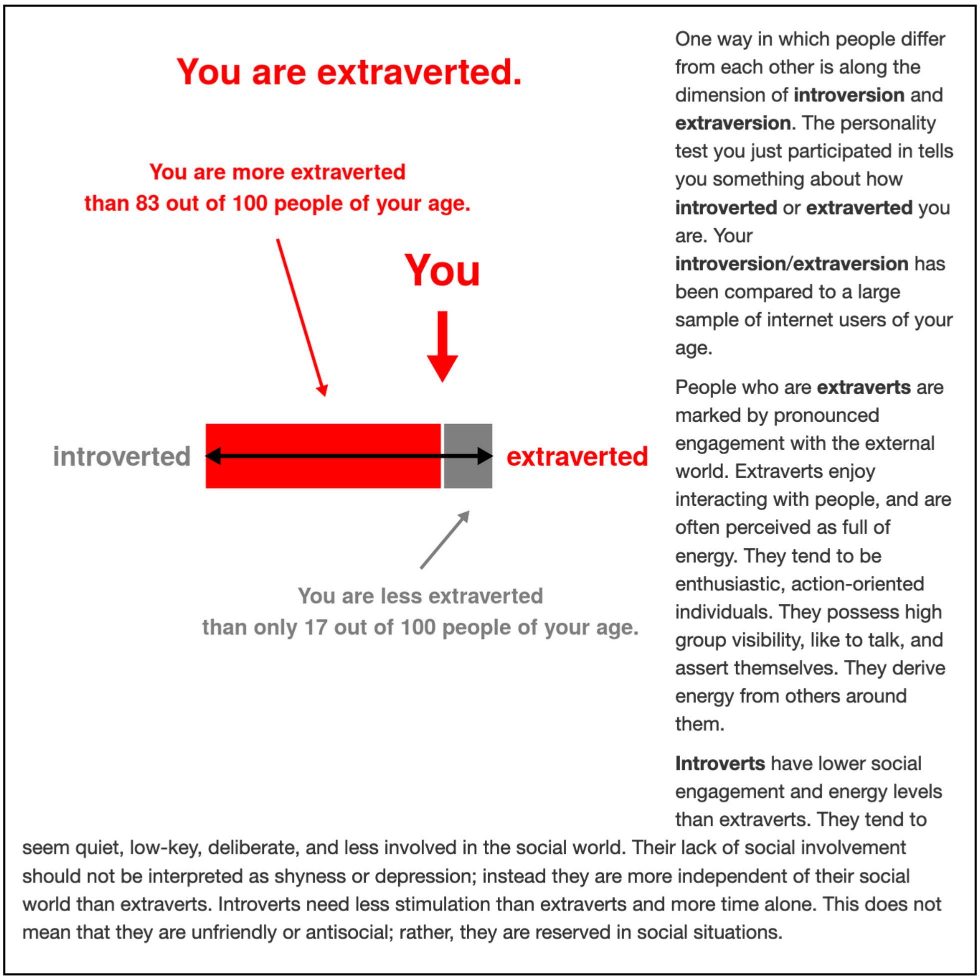


Figure S3: **Personality feedback and description used in the boosting condition in Experiment 1** (i.e., the relevant personality dimension: extraversion). This screenshot is an example for a participant classified as extravert; for participants classified as introverts, the the feedback is reframed in terms of intraversion (i.e., the title reads "You are introverted" and the text below reads "You are more introverted than [XX] out of 100 people of your age" and "You are less introverted than [100 − XX] out of 100 people of your age", where [XX] is the respective percentile). This definition of extraversion is adapted from Wikipedia (`https://web.archive.org/web/20190801042657/https://en.wikipedia.org/wiki/Extraversion_and_introversion`. See Methods in the main text for details on how the percentile was calculated.

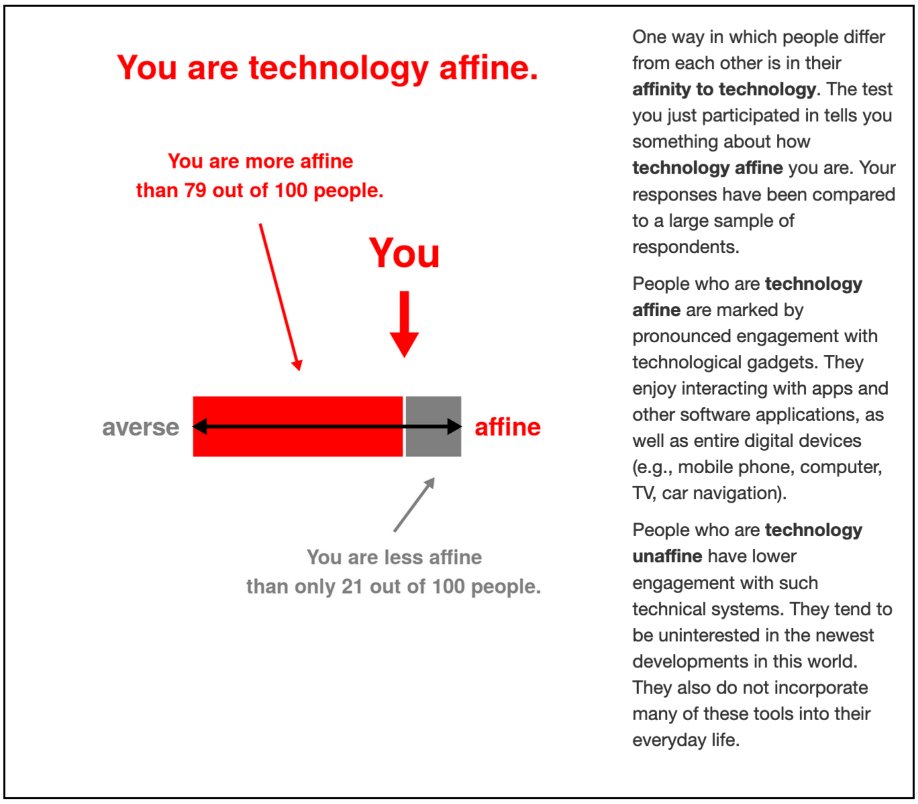


Figure S4: **Personality feedback and description used in the control condition in Experiment 1** (i.e., the irrelevant personality dimension: Affinity for Technology, ATI[3]). This screenshot is an example for a participant classified as technology affine; for participants classified as not technology affine, the the feedback is reframed in terms of technology aversion (i.e., the title reads "You are technology averse" and the text below reads "You are more averse than [XX] out of 100 people" and "You are less averse than [100 – XX] out of 100 people", where [XX] is the respective percentile). See Methods in the main text for details on how the percentile was calculated.

1.3 Descriptions of personality dimensions

Extraversion and introversion

One way in which people differ from each other is along the dimension of **introversion** and **extraversion**.

People who are **extraverts** are marked by pronounced engagement with the external world. Extraverts enjoy interacting with people, and are often perceived as full of energy. They tend to be enthusiastic, action-oriented individuals. They possess high group visibility, like to talk, and assert themselves. They derive energy from others around them.

Introverts have lower social engagement and energy levels than extraverts. They tend to seem quiet, low-key, deliberate, and less involved in the social world. Their lack of social involvement should not be interpreted as shyness or depression; instead they are more independent of their social world than extraverts. Introverts need less stimulation than extraverts and more time alone. This does not mean that they are unfriendly or antisocial; rather, they are reserved in social situations.

Please wait for one minute and read through the above information. Then continue here:

Go on

Figure S5: **Description of the extraversion personality dimension**, used in the boosting condition in Experiment 2. This definition of extraversion is adapted from Wikipedia (`https://web.archive.org/web/20190801042657/https://en.wikipedia.org/wiki/Extraversion_and_introversion`).

Affinity for technology interaction

One way in which people differ from each other is in their **affinity to technology**.

People who are **technology affine** are marked by pronounced engagement with the technological gadgets. They enjoy interacting with apps and other software applications, as well as entire digital devices (e.g., mobile phone, computer, TV, car navigation).

People who are **technology unaffine** have lower engagement with such technical systems. They tend to be uninterested in the newest developments in this world. They also do not incorporate many of these tools into their everyday life.

Please wait for one minute and read through the above information. Then continue here:

Go on

Figure S6: **Description of the Affinity for Technology scale (ATI)** used in the control condition in Experiment 2. This definition is taken from Franke et al.[3].

1.4 Comprehension check

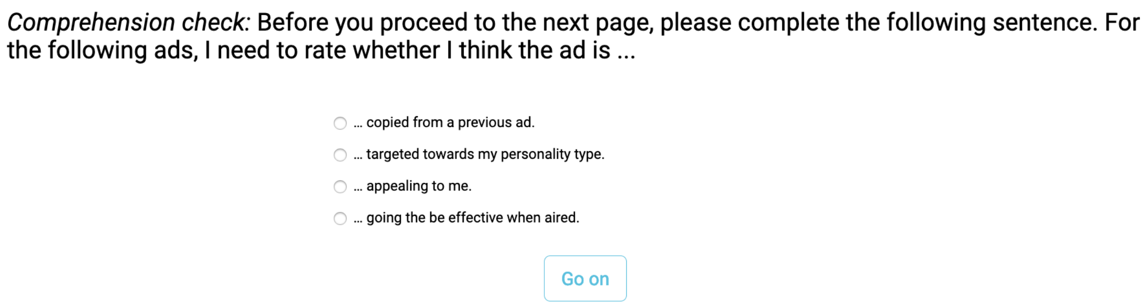


Figure S7: **Comprehension check used in Experiments 1 and 2 prior to starting the detection task.** If a participant did not choose the correct answer (“targeted towards my personality type”), the questions was shown again up to two more times, alongside the note ”The last answer was not correct, please try again:” (i.e., a total maximum of three attempts). The response options were sorted differently after each incorrect response. Only participants who passed the comprehension check within three attempts were included in the analysis (see Methods in the main text and the preregistrations).

1.5 Stimuli: The 10 ads from Matz and colleagues[1]

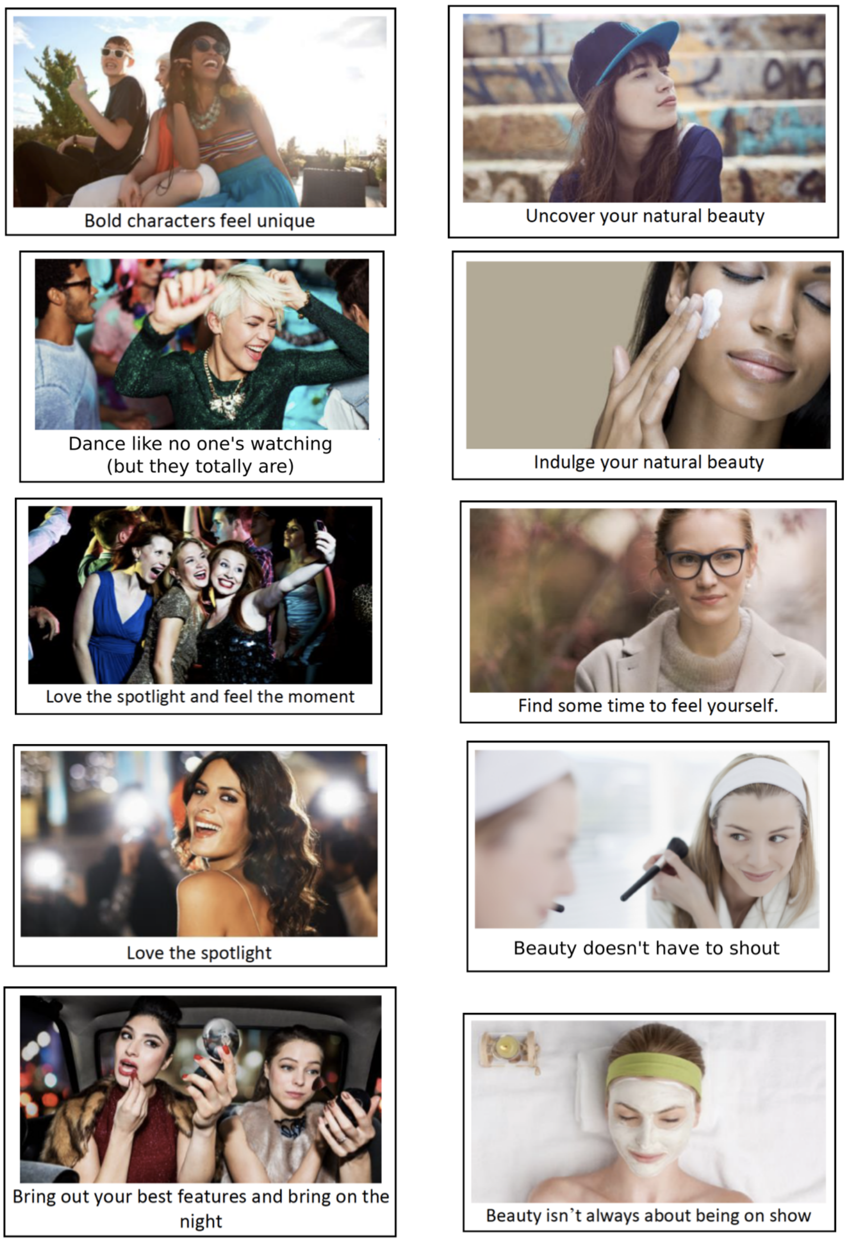


Figure S8: **Stimuli: The 10 ads used in Experiments 1 and 2.** The ads in the left column are tailored to extraverts and the ads in the right column to introverts. Images and text were adopted from Matz et al.[1] and then combined as listed by Matz et al..

2 Additional results

2.1 Experiment 1

2.1.1 Detection performance, boosting intervention, and level of extraversion

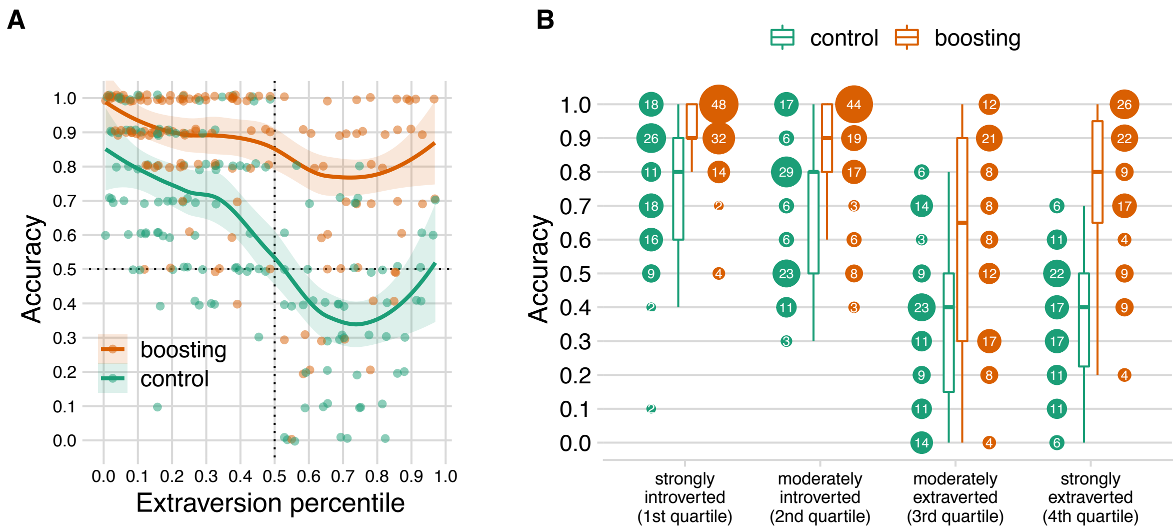


Figure S9: **Detection performance (in terms of proportion of correct decisions), boosting intervention, and level of extraversion (Experiment 1).** **A** Scatterplot of participants' accuracy (i.e., proportion correct decisions; y-axis) and their extraversion percentile (from 0 most introverted to 1 most extraverted; x-axis) for boosting vs. control group (color coded). Dots are slightly jittered vertically to avoid overplotting. Curves and confidence bands show robust LOESS curves (locally estimated scatterplot smoothing using re-descending M estimator with Tukey's biweight function) and their 95% confidence band. **B** Detection accuracy by extraversion quartiles (x-axis) for boosting vs. control group (color coded). In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed. The area of the dots and their numbers denote the within-quartile-and-condition percentage of participants for each of the 11 possible values for a participant's value of proportion of correct decisions (given the 10 ads).

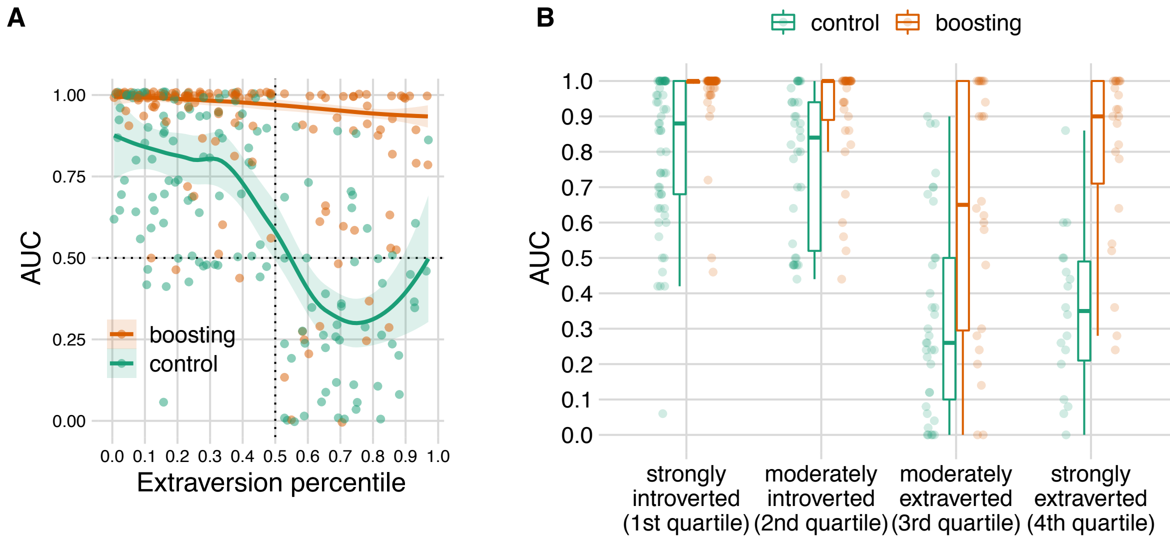


Figure S10: **Detection performance (in terms of the area under the Receiver Operating Characteristics curve, AUC, based on participants' confidence rating), boosting intervention, and level of extraversion (Experiment 1)**. Detection accuracy is quantified using the AUC based on participants' confidence rating, using the trapezoid method (i.e., no kernel- or model-based smoothing)[4]. In particular, this calculation uses a participant's confidence that the ad is targeted towards them (implied by the participant's binary categorization decision and corresponding rating about how confident the respondent is in the correctness of her decision). An AUC value can be interpreted as the probability that a participant's confidence (in the sense described above) is higher for a randomly selected ad that actually targets this participant compared to a randomly selected ad that does not actually target this participant. **A** Scatterplot of participants' detection performance (i.e., AUC; y-axis) and their extraversion percentile (from 0 most introverted to 1 most extraverted; x-axis) for boosting vs. control group (color coded). Dots are slightly jittered vertically to avoid overplotting. Curves and confidence bands show robust LOESS curves (locally estimated scatterplot smoothing using re-descending M estimator with Tukey's biweight function) and their 95% confidence band. **B** Detection performance (i.e., AUC; y-axis) by extraversion quartiles (x-axis) for boosting vs. control group (color coded). In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed.

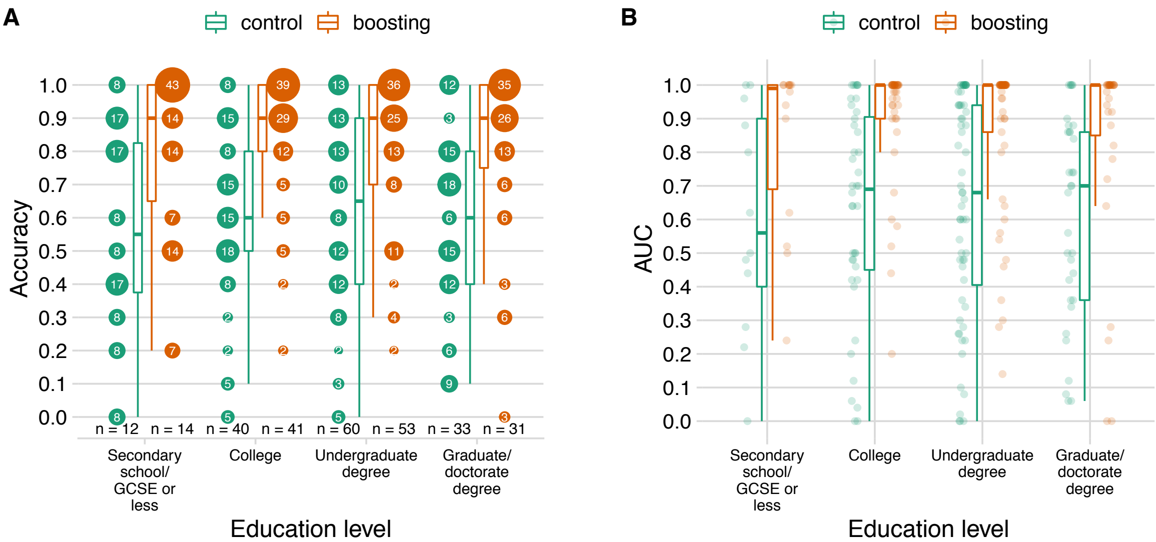


Figure S11: **Detection performance, boosting intervention, and education (Experiment 1)**. **A** Detection accuracy (i.e., proportion correct decisions; y-axis) by education (x-axis) for boosting vs. control group (color coded). The area of the dots and their numbers denote the within-education-and-condition percentage of participants for each of the 11 possible values for a participant's value of proportion of correct decisions (given the 10 ads). **B** Detection performance in terms of AUC (y-axis); see Fig S12 for more details on AUC. Dots represent participants and are slightly jittered to avoid overplotting. In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed. n denotes the number of participants for each combination of education level and condition.

2.1.2 Summary of multilevel logistic regression model

The text below shows the model summary of the *brms* Bayesian multilevel logistic regression model[5;6] reported for Experiment 1. See Methods in the main article for more information on the coding of the variables. `Estimate` shows the median and `l-95%` and `u-95%` show the 95% posterior credibility interval (i.e., the 2.5% and 97.5% percentile, respectively) of the respective marginal posterior distribution. For more details see the R help file `?brms::summary.brmsfit`*

```
 Family: bernoulli
  Links: mu = logit
Formula: dec_correct ~ 1 + condition + (1 | id) + (1 + condition | stimuli)
   Data: tbl_targeting_1 (Number of observations: 2840)
Samples: 4 chains, each with iter = 8000; warmup = 4000; thin = 1;
         total post-warmup samples = 16000

Group-Level Effects:
~id (Number of levels: 284)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.52      0.11     1.32     1.75 1.00     6358    10048

~stimuli (Number of levels: 10)
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sd(Intercept)                 0.48      0.14     0.28     0.92 1.00     5852
sd(condition1)                0.20      0.16     0.01     0.63 1.00     6142
cor(Intercept,condition1)     0.29      0.55    -0.83     0.95 1.00    18907
                          Tail_ESS
sd(Intercept)                 9267
sd(condition1)                7713
cor(Intercept,condition1)    10651

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      1.38      0.20     0.97     1.79 1.00     5350     8842
condition1     1.62      0.24     1.16     2.10 1.00     6967    10456

Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

*E.g., at `https://rdrr.io/cran/brms/man/summary.brmsfit.html`.

2.2 Experiment 2

2.2.1 Detection performance, boosting intervention, and level of extraversion

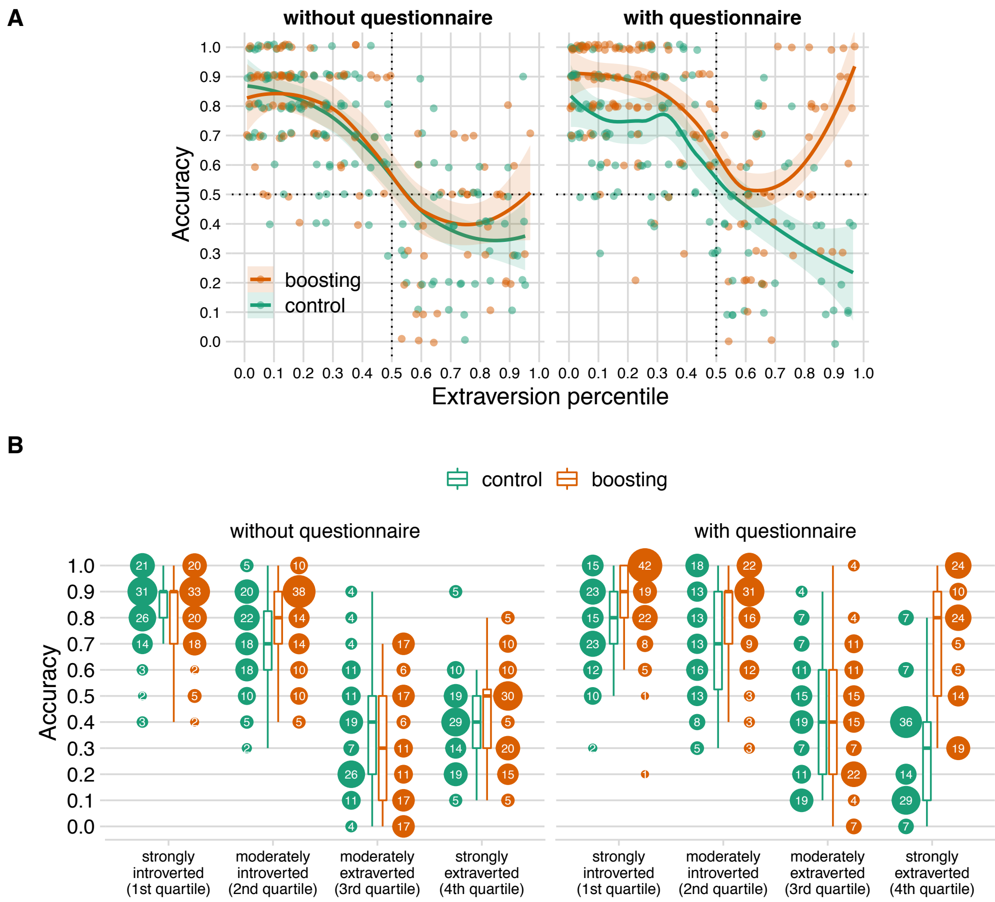


Figure S12: **Detection performance, boosting intervention, and level of extraversion (Experiment 2).** **A** Scatterplot of participants' accuracy (i.e., proportion correct decisions; y-axis) and their extraversion percentile (from 0 most introverted to 1 most extraverted; x-axis) for boosting vs. control group (color coded) and without and with questionnaire (left & right subplot, respectively). Dots are slightly jittered vertically to avoid overplotting. Curves and confidence bands show robust LOESS curves (locally estimated scatterplot smoothing using re-descending M estimator with Tukey's biweight function) and their 95% confidence band. **B** Detection performance by extraversion quartiles (x-axis) for boosting vs. control group (color coded) and without and with questionnaire (left & right subplot, respectively). In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed. The area of the dots and their numbers denote the within-quartile-and-condition percentage of participants for each of the 11 possible values for a participant's value of proportion of correct decisions (given the 10 ads).

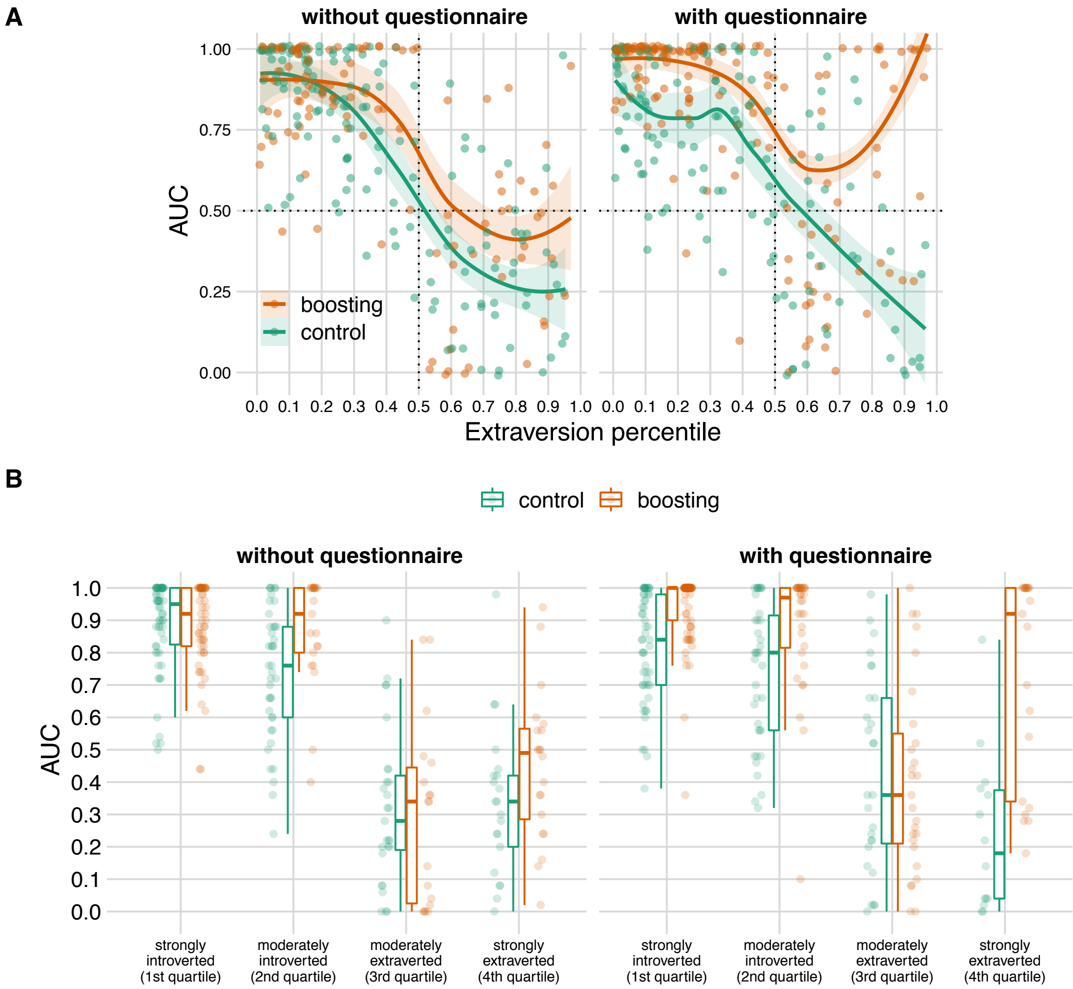


Figure S13: **Detection performance (in terms of the area under the Receiver Operating Characteristics curve, AUC, based on participants' confidence rating), boosting intervention, and level of extraversion (Experiment 2).** Detection accuracy is quantified using the AUC based on participants' confidence rating, using the trapezoid method (i.e., no kernel- or model-based smoothing)[4]. In particular, this calculation uses a participant's confidence that the ad is targeted towards them (implied by the participant's binary categorization decision and corresponding rating about how confident the respondent is in the correctness of her decision). **A** Scatterplot of participants' detection performance (i.e., AUC; y-axis) and their extraversion percentile (from 0 most introverted to 1 most extraverted; x-axis) for boosting vs. control group (color coded) and without and with questionnaire (left & right subplot, respectively). **B** Detection performance (i.e., AUC; y-axis) by extraversion quartiles (x-axis) for boosting vs. control group (color coded) and without and with questionnaire (left & right subplot, respectively). See Fig S10 for more details on AUC and what the two panels show.

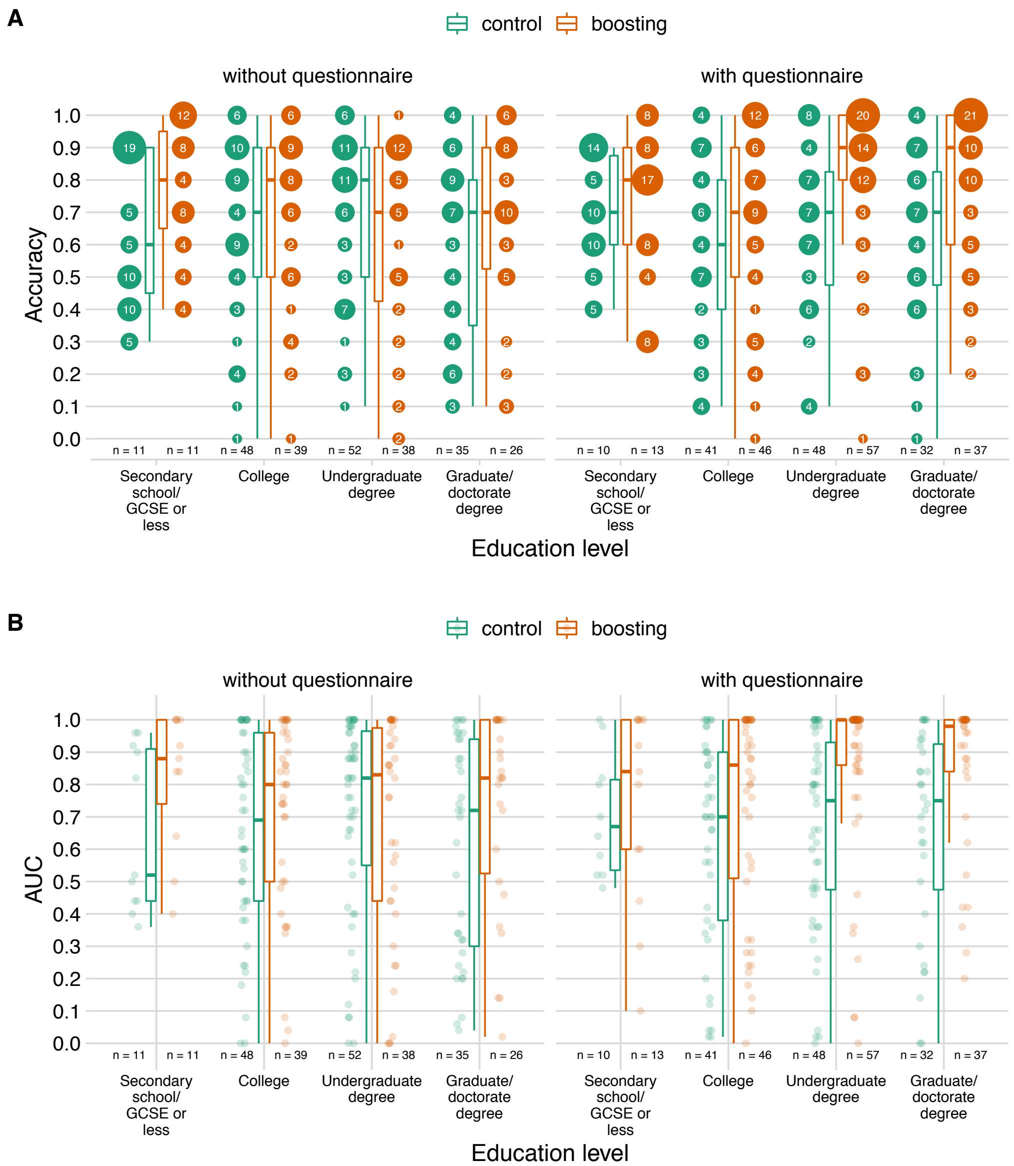


Figure S14: **Detection performance, boosting intervention, and education (Experiment 2).** **A** Detection accuracy (i.e., proportion correct decisions; y-axis) by education (x-axis) for boosting vs. control group (color coded) and without and with questionnaire (left & right subplot, respectively). The area of the dots and their numbers denote the within-education-and-condition percentage of participants for each of the 11 possible values for a participant's value of proportion of correct decisions (given the 10 ads). **B** Detection performance in terms of AUC (y-axis); see Fig S12 for more details on AUC. Dots represent participants and are slightly jittered to avoid overplotting. In the boxplots, the box shows the the first and third quartiles (the 25th and 75th percentiles). The lower and upper whiskers extend from the respective end of the box to the largest value no further than 1.5 × IQR from the box (where IQR is the inter-quartile range, or distance between the first and third quartiles); outliers are not displayed. n denotes the number of participants for each combination of education level and condition.

2.2.2 Summary of multilevel logistic regression model

The text below shows the model summary of the *brms* Bayesian multilevel logistic regression model[5;6] reported for Experiment 2. See Methods in the main article for more information on the coding of the variables. `Estimate` shows the median and `l-95%` and `u-95%` show the 95% posterior credibility interval (i.e., the 2.5% and 97.5% percentile, respectively) of the respective marginal posterior distribution. For more details see the R help file `?brms::summary.brmsfit`[†]

```
 Family: bernoulli
  Links: mu = logit
Formula: dec_correct ~ relevance + questionnaire + (1 | id) + (1 + relevance * questionnaire | stimuli) + relevance:questionnaire
   Data: tbl_targeting_2 (Number of observations: 5440)
Samples: 4 chains, each with iter = 8000; warmup = 4000; thin = 1;
         total post-warmup samples = 16000

Group-Level Effects:
~id (Number of levels: 544)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.40      0.07     1.27     1.55 1.00     6481    10495

~stimuli (Number of levels: 10)
                                              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                                     0.41      0.12     0.25     0.77 1.00     5748     9121
sd(relevance1)                                    0.07      0.07     0.00     0.28 1.00    11664     9450
sd(questionnaire1)                                0.28      0.13     0.04     0.65 1.00     5596     5495
sd(relevance1:questionnaire1)                     0.58      0.26     0.11     1.31 1.00     6164     5441
cor(Intercept,relevance1)                         0.21      0.50    -0.73     0.88 1.00    28097    11260
cor(Intercept,questionnaire1)                     0.26      0.37    -0.50     0.81 1.00    17768    12090
cor(relevance1,questionnaire1)                    0.01      0.51    -0.80     0.80 1.00     7536    11636
cor(Intercept,relevance1:questionnaire1)         -0.30      0.35    -0.82     0.45 1.00    16977    12192
cor(relevance1,relevance1:questionnaire1)        -0.02      0.51    -0.82     0.80 1.00     7980    10940
cor(questionnaire1,relevance1:questionnaire1)     0.01      0.43    -0.74     0.72 1.00    12689    13190

Population-Level Effects:
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                     1.03      0.15     0.71     1.36 1.00     4626     7232
relevance1                    0.48      0.15     0.20     0.77 1.00     7812    10679
questionnaire1                0.25      0.17    -0.10     0.60 1.00     7967    10199
relevance1:questionnaire1     0.72      0.35     0.00     1.44 1.00     8380    10576

Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

[†] E.g., at `https://rdrr.io/cran/brms/man/summary.brmsfit.html`.

References

[1] Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* **114**, 12714–12719 (2017).

[2] Srivastava, S., John, O. P., Gosling, S. D. & Potter, J. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology* **84**, 1041 (2003).

[3] Franke, T., Attig, C. & Wessel, D. A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) scale. *International Journal of Human–Computer Interaction* **35**, 456–467 (2019).

[4] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27** (2006).

[5] Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).

[6] Bürkner, P.-C. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* **10**, 395–411 (2018).