Determinants of willingness to donate data from social media platforms

Zoltán Kmetty*[1,4], Ádám Stefkovics*[1,2], Júlia Számely*[3], Dongning Deng[4], Anikó Kellner[4], Edit Pauló[4], Elisa Omodei[3], Júlia Koltai[5,4]

[1]CSS-RECENS, Center for Social Sciences, Budapest [2]IQSS, Harvard University, Cambridge [3]Department of Network and Data Science, Central European University, Vienna [4]Faculty of Social Sciences, Eötvös Loránd University, Budapest, [5]MTA–TK Lendület "Momentum" Digital Social Science Research Group for Social Stratification, Centre for Social Sciences, Budapest

*These authors contributed equally: Zoltán Kmetty, Ádám Stefkovics, Júlia Számely

Author notes

Conflicts of Interest: The authors declare no conflict of interest.

Data availability: Data is available in the repository of CSS-KDK: https://openarchive.tk.mta.hu/584/ DOI: 10.17203/KDK584

Ethics approval: Ethical review and approval were waived for the Hungarian study because data was anonymized within the fieldwork of the study. The results do not allow identification of the individuals involved in the Hungarian study. The authors and the fieldwork agency managed all information collected in accordance with the General Data Protection Regulation (GDPR). The US study has been approved by the Harvard Institutional Review Board (IRB22-0942).

Corresponding author: Zoltán Kmetty, kmetty.zoltan@tk.hu. Centre for Social Sciences, CSS-RECENS Research Group, 1097 Budapest, Tóth Kálmán u. 4., Hungary

Abstract

Social media data donation through data download packages (DDPs) is a promising new way of collecting individual-level digital trace data with informed consent. Nevertheless, given the novelty of this approach, little is known about whether and how people would share their data with researchers, although this could seriously affect selection bias and thus, the outer validity of the results. To study the determinants of data-sharing and help future data donation studies with detecting the conditions, under which the willingness is the highest, we pre-registered two vignette experiments and embedded them in two online surveys conducted in Hungary and the U.S. In hypothetical requests for donating social media data via DDPs, we manipulated the amount of the monetary incentives (1), the presence or lack of non-monetary incentives (2), the number of requested platforms (3), the estimated upload/download time (4), and the type of requested data (5). The results revealed that data-sharing attitude is strongly subject to the parameters of the actual study, how the request is framed, and some respondent characteristics. Monetary incentives increased willingness to participate in both countries, while other effects were not consistent between the two countries.

Keywords: data donation, DDP, data-sharing, digital trace data, linking survey data

Determinants of willingness to donate data from social media platforms

Digital traces on social media platforms can be promising sources of information for researchers in various fields (Stier et al., 2020). In contrast to self-reports from surveys, which may imply measurement errors due to recall or social desirability bias (Araujo et al., 2017; Ohme et al., 2021; Parry et al., 2021; Scharkow, 2016), digital traces are assumed to provide reliable behavioral data. Surveys, on the other hand, allow researchers to gain access to self-reported beliefs and attitudes, as well as the socio-demographic information of the respondents. Linking individual digital behavioral data with survey responses holds the promise of giving context of behavior on social media platforms, and therefore scientific attention in combining these data sources has substantially grown recently.

However, accessing digital social media data has recently become increasingly difficult. After the Cambridge Analytica scandal in 2018, many platforms, including Facebook decided to strongly restrict the use of Application Programming Interfaces (APIs) (Breuer et al., 2022; Bruns, 2019; Freelon, 2018), which pushed researchers to explore other approaches, such as web-scraping (Christner et al., 2021; Mancosu & Vegetti, 2020), novel partnerships between industry and academia (King & Persily, 2020), passive data collection methods as web-tracking and browser plugins (Keusch et al., 2019; Silber et al., 2022) or data donation (Boeschoten et al., 2022; Breuer et al., 2020, 2022). The chosen data collection approach may be influenced by the research question, the resources available, and even the skills and capabilities of the research team (Ohme et al., 2023).

Out of these solutions, this article focuses on data donation involving data download packages (DDPs, Boeschoten et al., 2022). DDPs are a collection of historical user data stored on social media platforms consisting of behavioral (e.g., likes), textual (messages), media (photos, videos), or location data. Social media users have been able to download their DDPs from major platforms because GDPR obliged platforms to provide access to their users to the data collected about them. Given the unique scientific value of this data, researchers have recently shown great interest in data donation. In data donation studies participants are typically recruited with standard survey sampling techniques, then asked to download their own social media data through DDPs and provide them to the researchers for analysis with informed consent (Boeschoten et al., 2022, Breuer et al., 2022). One key advantage of the approach is that it helps overcome some of the privacy concerns of the API approach, such as consent giving. Whilst in the case of platform-centric data collection users' ability to monitor and understand who uses their data and how they use it is limited (Halavais, 2019), user-centric approaches offer users greater transparency and the chance to provide consent under clear research terms (Breuer et al., 2022). Users' data archives, moreover, provide a richer set of data compared to API or scraping approaches, which can support various analytical goals or can be easily linked to other (such as survey) data. Nevertheless, data donation is more burdensome compared to other, passive modes of data sharing such as web-browser plugins (Keusch et al., 2019; Silber et al., 2022), since participants are required to actively download their data and then share it with the researchers. While data donation holds great promise for studying online human behavior and some recent studies reported promising results

(Baumgartner et al., 2022; Breuer et al., 2022; Kmetty & Bozsonyi, 2022; Kmetty & Németh, 2022), little is known about how to best optimize such approaches.

This study aimed to understand the underlying mechanisms of willingness to participate in academic research as a social media data donor. Earlier research has shown that digital data-sharing behavior in a broader sense can be a function of, for instance, the offered incentives, the study's sponsor and various other factors (see e.g., Pfiffner & Friemel, 2023; Silber et al., 2021; Struminskaya et al., 2020; Struminskaya et al. 2021). Our research expands existing knowledge by examining willingness to share DDPs and addressing previously uncovered factors (such as the type of social media data). We further contribute to the literature by running the same vignette experiments in two countries (the United States and Hungary) allowing for a cross-national comparison. Our results can add to the understanding of the characteristics of nonresponse bias in the case of data donation, which can raise awareness of the dimensions, thereby limiting the generalization of the results (Breuer et al., 2020). These results also provide grounds for nonresponse adjustments. Additionally, uncovering the determinants of willingness to donate such data can directly help researchers improve their recruitment strategies and research design.

Determinants of data sharing behavior and hypotheses

Incentives

Social exchange theory (Thibaut & Kelley, 2017) implies that research participants' willingness of engagement is affected by their assessment of the rewards and costs of the action (Dillman, 2000; Keusch, 2015). Survey respondents are similarly expected to weigh the pros and cons of participation as proposed by both survey participation- (Groves & Couper, 1998) and leverage-saliency theory (Groves et al., 2000). Perceived rewards include internal rewards (e.g., feeling satisfied from scientific contribution) and external rewards (e.g, getting incentives such as price rewards). Monetary incentives are one of the most common approaches to motivate participation (Keusch, 2015).

Many empirical findings have proven the effectiveness of monetary incentives in motivating people in research participation, such as responding to web surveys (see Göritz, 2006 for a meta-analysis or Keusch, 2015 for a review), and sharing digital trace data (Keusch et al., 2019; Silber et al., 2022). The results of Silber et al. (2022) showed that providing monetary incentives can have a positive effect on the willingness to share social media data. Similar effects have been found regarding willingness to share active and passive mobile data (Haas et al., 2020; Keusch et al., 2019; Ságvári et al., 2021). Additionally, earlier studies (Keusch et al., 2019; Revilla et al., 2019) have found that incentives that are too low were among the two main reasons for not participating in digital data sharing[1]. By contrast, Jäckle et al. (2019) and Beuthner et al. (2023) found no effect of incentives on downloading an app or sharing different types of data. Silber et al. (2022) also reported that incentives do not necessarily

[1] The other main reason was privacy concern.

motivate all types of data-sharing (e.g., they do not work in the case of sharing highly sensitive health data).

In the case of data donation, little is known about the optimal amount and method of offering incentives. Too high incentives may make respondents suspicious and assume that their data are highly valuable, and eventually decreasing their likelihood of participation (Silber et al., 2022). The challenge here is to estimate the optimal amount of incentive. In the context of data donation, relatively high effort is required from participants. To match with this high cost, we assume that a higher monetary incentive will make respondents more willing to donate their data. As we employed realistic incentive levels, we did not expect any non-linear or counterproductive effects associated with the highest incentive amounts.

H1a. Higher monetary incentives increase the willingness to donate data.

In addition to monetary incentives, we were also interested in whether obtaining a(n automated) summary report on the participant's own data affects their willingness to donate their data. Summary reports may contain interesting findings such as contrasts between participant-specific data and the overall sample. Offering summary reports has been successfully used to motivate participation in business-to-business (Keusch, 2015) and medical web surveys (Bietz et al., 2019; Edwards et al., 2009). However, for other web surveys, such non-monetary incentives had no or even a negative impact (see Keusch, 2015 for a review). Despite the mixed evidence in the literature (which may be due to the diverse range of outcomes examined in the literature), theoretically it is more plausible to expect a positive association. Thus, we hypothesize that offering summary reports will have a positive impact on willingness to donate data.

H1b. Offering summary reports increases willingness to donate data.

The number of platforms and the time required to download and upload data

When deciding on participation, respondents also take into account the costs that derive from the perceived *difficulty* and *burden* of the data-sharing process. Both difficulty and burden can be higher in the case of data donation compared to other passive digital data collection methods (Keusch, 2015; Silber et al., 2022). Bradburn (1978) distinguished four elements of respondents' burden regarding surveys: length, frequency, required effort, and caused stress. Following Bradburn's dimensions, the length of the task increases with the number of platforms involved in the request (clicks, download and upload time), as well as the required effort. Therefore, the more platforms involved in the request, the greater burden the participants may have. Each platform requires a somewhat different task from that of the downloader because the download procedure varies by platform. Moreover, the more platforms

involved, the more likely respondents' privacy concerns emerge (see more about privacy issues in Section 2.4).

Time can also be a type of burden. In the case of data donation, the downloading, processing, and uploading time can last for hours or days, which is longer than a usual survey participation.

According to these results, we expect that willingness to donate will decrease with the increase in the number of requested platforms, and that a declared longer download/upload time will also hinder cooperation.

H2a. Asking for more platforms decreases the willingness to donate data.

H2b. A longer download/upload time of the data decreases the willingness to donate data.

Types of data

On Social media platforms, users can share different types of user-generated content, such as textual content (tweets, posts, comments) or audio-visual materials (pictures and videos), which can be valuable data sources for research (Breuer et al., 2021). When people are asked about their participation in a data donation study, the type of data requested may affect how they decide. The level of sensitivity of the data to be donated is a particularly relevant aspect of these decisions. One of Bradburn's (1978) respondent burdens is the stress caused by the task, i. e. the discomfort the respondents feel while participating in the study. Such stress may derive from privacy issues: participants can be reluctant to be involved in the data collection because, for instance, the data to be shared is sensitive or confidential (Breuer et al., 2022). Pfiffner and Friemel (2023) examined willingness to donate data in non-experimental hypothetical settings and found that higher perceived data sensitivity was associated with lower willingness to donate; moreover, the level of perceived sensitivity of data was the most influential factor in determining the willingness to donate.

Nevertheless, it is not completely clear which types of digital data are considered sensitive by the users. Photos and videos can be sensitive because people can be directly or indirectly identified (Christen et al., 2020), which can lead to users' (and others') exposure to several risks (Bioglio & Pensa, 2022). Geolocation data can similarly uncover personal information, such as home, work, or school addresses, therefore likely evoking privacy concerns (Wenz et al., 2019).

Silber et al. (2022) reported the highest willingness to share Spotify data (59.1 percent, musical data), and the lowest for Facebook (31.2 percent, various types of data) with Twitter in the middle (41.4 percent, mostly textual content). In their other survey, 24 percent of the users shared their Twitter data, whereas less than 10 percent shared their health app data. Beuthner (2023) compared consent rates for seven types of domains and found that willingness to share Facebook data was relatively high (above 50%) compared to other administrative domains such as bank account or health insurance data sharing.

Pfiffner & Friemel (2023) found that compared to their activities on Google, people demonstrate a higher willingness to share what they encounter on social media. Additionally, individuals are least inclined to donate data that includes details about their personal social network, encompassing friends and followers.

Research on sensor-based data sharing also shows that willingness to share is partly a function of the requested data type. Wenz et al. (2019) found that respondents were less willing to share the GPS location of their smartphone compared to other tasks involving less confidential data (e.g., completing a questionnaire, or downloading an app). Revilla et al. (2016, 2019) reported that respondents' willingness to take and share pictures was higher compared to providing access to their Facebook account or sharing geolocation data. In contrast, in the study of Struminskaya et al. (2021) tasks that involved photographing or video-taking one's surroundings at home yielded far lower willingness to share rates compared to providing geolocation data with a sensor (see Struminskaya et al. (2020) for a similar approach with somewhat different results). Nevertheless, Kreuter et al. (2020) and Ságvári et al. (2021) did not find significant differences between willingness to share different, passively collected data.

To sum up, earlier findings suggest that the type of data matters in the decision on participation, and respondents perceive geolocation data as less private compared to photos and videos. Nevertheless, reluctance to share audio-visual materials strongly depends on the content that is being recorded. Additionally, most of these studies requested respondents to actively take pictures or videos for the current study, thus participants had more control over what they shared. In the case of a social media data donation, screening all content before the donation is more burdensome. For this very reason, in our hypotheses, we did not differentiate between willingness to share audio-visual materials and geolocation data but expected that willingness to share would be lower in both cases, so when the request involves pictures and videos, as well as geolocation data.

H3a. Asking for pictures/videos of the user decreases the willingness to donate data.

H3b. Asking for geolocation data of the user decreases the willingness to donate data.

Respondents' characteristics

Finally, we summarize respondent characteristics that might affect willingness to donate data. Given that the main focus of this paper was the impact of the details of a hypothetical request on participation willingness, no particular hypotheses were developed for the following factors.

Participating in a data donation study is an active task that requires a basic level of digital literacy (Baumgartner, 2022), and therefore can place more burden on people with lower technical skills. Self-assessed smartphone skills in earlier studies, however, show contradictory results: Wenz et al. (2019) and Ohme et al. (2021) found that participants with higher phone skills were more willing to

participate, but in the study of Keusch et al. (2019) and Struminskaya et al. (2020) these skills were not associated with the willingness to participate.

As we discussed earlier, privacy concerns are likely to evoke when sensitive and private data is involved in data donation. Previous studies show that privacy and security concerns lower the willingness to participate (Jäckle et al., 2019; Pfiffner & Friemel, 2023; Revilla et al., 2019; Struminskaya et al., 2020, 2021; Wenz et al., 2019, but see Elevelt et al., 2019 and Silber et al. 2022 for null results). Indeed, they are the most mentioned reasons for not being willing to participate in digital data sharing (Jäckle et al., 2019; Keusch et al., 2019; Revilla, et al., 2019).

Psychological traits can also play a role. Earlier research examined the effect of the Big Five traits on item/unit nonresponse, or attrition in surveys (Brüggen & Dholakia, 2010; Cheng et al., 2020; Kmetty & Stefkovics, 2021; Lugtig, 2014), and on downloading an app and sharing passive (Elevelt et al., 2019) or social media data (Silber et al., 2022). The results of these studies are inconsistent. Some studies found that conscientiousness, one of the Big Five traits, can increase participation while sharing GPS data is higher among introverts (Elevelt et al., 2019). However, Silber et al. (2022) did not find such positive effects regarding these psychological traits on sharing social media or health data.

Willingness to share data can also be a function of usage of the platform or device. Generally, it can be assumed that heavy users may be more open to data sharing (Breuer et al., 2022). Silber et al. (2022), for instance, found that a higher platform usage increased the likelihood of sharing social media data, but not health app data.

Finally, among sociodemographic factors, age is one of the major determinants of digital technology use and attitudes toward digital technologies as well (Ságvári et al., 2021). Age likely correlates with the willingness to participate in digital data collection, as participation willingness is typically higher in the younger generation (Pinter, 2015; Elevelt et al., 2019; Silber et al., 2019; Ságvári et al., 2021), and it decreases after age 50 (Mulder, & Bruinje, 2019). Other socioeconomic factors, like education, settlement type, and region, are not clearly associated with the willingness to participate (Pinter, 2015; Ságvári et al., 2021).

Data and methods

Data

We collected two datasets for this study and conducted one survey experiment in Hungary and another one with a similar design in the U.S. In Hungary, the survey was administered by an online polling company, NRC, on a non-probability access panel. The NRC panel consists of more than 140,000 people. Compared to the general population of Hungary, individuals with a high level of education and from bigger cities are overrepresented in the panel. We used a quota sampling method (with quotas for gender, age, and geographical region) to ensure equal representation. The company gives regular incentives for the respondents of their studies. Altogether 1,000 respondents participated in the Hungarian study. The fieldwork was carried out between 11-25, May 2022.

The U.S. dataset is a based on a panel of Harvard University called Harvard Digital Lab for Social Science (DLABSS). DLABSS is a pool of survey respondents primarily recruited using social media and other free sources. Respondents do not get incentives for filling out surveys in the panel. The size of DLABSS's pool is growing rapidly, currently counting nearly 30,000 volunteers. A study by Strange et al. (2019) found that such volunteer panels can replicate classic and contemporary social science findings and produce high levels of overall response quality comparable to paid subjects. In the end, 844 respondents participated in the U.S. study. The fieldwork was carried out at the end of 2022, between October 14 and November 8.

Both data collections and studies were pre-registered before the data collection. The anonymised pre-registration for the Hungarian study is available here: https://osf.io/r4kxm?view_only=af4b7da2aba14495b2e5df280b68a37d ; and for the U.S. study is here: https://osf.io/tvefj/?view_only=728b0fa3b66a4c47abcebc30dd07b08e

Design of the survey experiment

We built up the survey experiments similarly in the two countries. We applied a mixed factorial vignette design. Respondents had to evaluate multiple situations (called vignettes), in which we manipulated various dimensions of a fictional research. At each vignette, they had to provide the likelihood of their willingness to donate their digital data in such research on a 0 to 10 scale. These manipulated dimensions of the fictional research were the following (an example of the vignettes is available in the appendix).

Table 1

Table title

Table 1 here

Altogether we had 192 possible combinations of the dimensions in the following way: 4 [platform] * 4 [range of data] * 2 [time] * 3 [incentive] * 2 [report]. In the Hungarian study, we created 16 decks (packages of vignettes) and assigned 12 vignettes to each deck (16*12=192). Thus, one respondent had to evaluate 12 different situations. With the expected (and then realized) 1,000 respondents we had around 67 respondents per deck. The twelve vignettes per respondent is a relatively high number. To assess the validity of the results, we also conducted a robustness check (see the results section for more details).

The U.S. data comes from a volunteer panel, where we expected a higher dropout rate in a repetitive task than among paid panel members, like the Hungarian ones. To overcome the possible bias and high dropout rate caused by the large number of vignettes per respondent, we followed a slightly different strategy in the U.S. study. In this study, we created 16 decks as well, but only assigned five vignettes to each deck. We used the optBlock function of the AlgDesign package (Wheeler &

Braun, 2019) in R to find the best design. This function uses the D criterium to optimize vignette allocation. With this design, we had around 52 respondents per deck.

According to our previous power calculation, with this design 700 respondents would have been enough to achieve a 0.95 power with a 0.1 (small) effect size in both countries. The final sample sizes were over this limit.

Variables

We used the following independent variables in the analysis: gender, highest level of education, age, subjective wealth, frequency of social media usage, number of social media platforms used by the respondent, Internet Users' Information Privacy Concerns (IUIPC) scale, privacy concerns, Affinity for Technology Interaction Scale big five (BF) inventory.

There were no missing values in the dependent variable, only in the independent ones. In 26 and 28 percent of the cases in the Hungarian and U.S. study, there was at least one missing variable. To handle these missing values in the dataset in order not to lose too many cases, we applied multiple imputations. We included all the independent variables in the imputation process and used predictive-mean-matching (PMM) for the procedure. We created five imputed datasets and calculated the pool results in the regression models. We used the 'mice' package of R (van Buuren et al., 2015) for these calculations.

Analytical strategy

As the first step of our analysis, we applied a variance component model to understand how much of the variation in the response variable – willingness to donate digital footprint data – is explained by vignette level and respondent level characteristics.

As a next step, a set of multilevel regressions were performed as we had two levels in the data: one for the vignettes and another one for the respondents (because one respondent evaluated multiple vignettes). In the regression models, we allowed for random intercepts by the respondents, first regressing only vignette-level variables on willingness to donate data, then we added respondent-level characteristics as well.

We carried out the analysis using the 'lme4' (Bates et al., 2015) and related packages in R.

Results

Study 1. Hungary

Twelve vignettes were assigned to 1000 respondents in the Hungarian study, thus we had 12,000 cases on the vignette level. Thirty percent of the vignettes, respondents indicated that it is not likely at all that they would donate their data under the given circumstances, while maximal willingness was shown in 14 percent of the vignette cases. The mean value of the willingness questions was 4.2 on

the 0 to 10 scale[2]. On the respondent level, 18 percent refused any kind of data sharing, regardless of the vignette content (answered zero to all twelve vignette situations they evaluated).

In the first step of the analysis, we applied a variance component model on the Hungarian dataset (Table 2, first column). The results of this analysis showed that variation at the deck level is not significant, while variation on the respondent level explains 79.7 percent of the variation in the willingness to donate data, and the remaining 20.3 percent of the variation is explained by the vignette level.

In the next step of the analysis, we added the vignette-level variables. Regressing the outcome variable on the vignette-level explanatory variables (Table 2, second column) showed that *incentive*, *platform*, and *data type* have significant effects on the outcome variable (with *incentive* having the strongest effect), while the effect of *report* provision and the *time* to download/upload data are not significant (see the relative strength of effects in Figure 1). The effect of the *incentive* variable on the willingness to provide data is positive, with a 0.25-point increase in the expected value of the outcome variable with each additional amount of HUF worth 10 USD at Purchasing Power Parity (PPP). The effect of the *platform* variable means that compared to donating digital footprint data from the respondent's Facebook account only, the more platforms the respondent is required to provide digital footprint data from, the less likely they are to do so. The effect of the *type* variable means that as compared to providing all data except private messages, respondents are on average less likely to provide their data if private messages and photos and videos are excluded, as well as if private messages, photos, videos, and location are excluded. As these results are quite counter-intuitive, we will get back to their explanation in the discussion[3].

Figure 1

Figure title

Figure 1 here

Next, we added a set of respondent-level explanatory variables to the regression, allowing for random intercept by respondents (Table 2, third column). The results show that with the inclusion of control variables, the same vignette-level variables remain significant as in the previous setup, i.e. H1a, H2a, H3a, and H3b remain confirmed, and H1b, H2b remain contradicted. The strength of the vignette-

[2] For a robustness check of the results, we calculated the standard deviation of the willingness probability for the first and second six vignettes. A smaller standard deviation might have been a sign of fatigue for the respondent. Based on Barlett's test, we did not find differences between the standard deviations (p =.39) of the two sets.

[3] For robustness check, we re-ran this multilevel model with the first six and second six vignettes separately (see Table A1 in the supplementary). There were some differences between the evaluations of the first and second six vignettes. Still, the incentive had the most pronounced positive effect in the regressions fitted to both vignette groups. For the second six vignettes, fewer variables have a significant impact which may indicate fatigue and less attentive evaluation.

level variables does not change significantly either as compared to the regression with only vignette-level variables. Moreover, *gender* and *age*, as well as *education* are significant in explaining willingness to donate data. Female respondents are on average less likely to donate data, the older the respondent the less likely they are to donate, and the higher the level of the respondent's education the more likely they are to be willing to provide their digital footprint data. *Subjective wealth* showed no significant effect on the outcome variable.

Of the other control variables, only the number of platforms visited had a significant effect on the likelihood of sharing data. Those using multiple platforms were more likely to share their data. After including the control variables, the explanatory power of the model went up to 7.3 percent.

Table 2

Table title

Table 2 here

Study 2. USA

We had 844 respondents in the U.S. study with 5 vignette evaluations. resulting in 4,174 vignette evaluations. 59 percent of the cases on the vignette level respondents answered that it is not likely at all that they would donate their data under the given circumstances, and we observed the highest level of willingness in only 5 percent of the cases. The mean value of willingness was 2.1 on a 0 to 10 scale, where higher values mean higher willingness. On the respondent level, 52 percent mentioned that it is not likely at all that they would share their data regardless of the vignette content (answered zero to all 5 vignette situations). Overall, the willingness rate was much lower in the U.S. sample, than in the Hungarian one.

Similarly to the analysis of the Hungarian dataset, in the first step of the analysis, we applied a variance component model on the U.S. dataset (Table 3, first column). The results of this analysis show that in the U.S. dataset, the individual level explains 85.2 percent of the variation of the outcome variable, the deck level explains 0.6 percent of the variation, while the remaining 14.2 percent of the outcome variable's variation is explained by the vignette level.

When we added the vignette-level variables (Table 3, second column), we found that with the exception of *data type*, all vignette-level variables have significant effects on the willingness to donate data. Increasing the *incentive* has a positive effect on willingness to donate, 0.35-point increase in willingness with every additional ten USD, the same effect as in the Hungarian case. Offering a *report* has a significant positive effect on the outcome variable. Asking for data from more *platforms* affects willingness positively – contrary to the effect found in the Hungarian data. The effect of the *time* of

download/upload is negative, and while pointing in the same direction, it is an order of magnitude larger than in the Hungarian data. As the results of the regressions with standardized variables show (Figure 2), similarly to the Hungarian dataset, the effect of *incentive* is the strongest among the vignette-level variables. These results confirm hypotheses H1a, H1b, H2b, and H3b, and contradict hypotheses H2a and H3a in the U.S. dataset. H1a was therefore confirmed by both the Hungarian and U.S. datasets, while the evaluation of the rest of the hypotheses varied across the two datasets.

Figure 2

Figure title

Figure 2 here

Next, introducing respondent-level variables in the regression, while allowing for random intercepts by respondents (Table 3, third column), we found that the significance of the effects of vignette-level variables do not change, and the magnitudes of coefficients change only slightly compared to the previous model, which only included vignette-level independent variables. Gender has a significant effect, such that female respondents are more likely to donate their data (opposite as in Hungary). The effect of one of the *IUIPC* indicators is significant, specifically, having a more positive opinion about how one's personal data is generally collected affects willingness to donate data positively. The number of used *platforms* is positively associated with the outcome variable, and also the frequency of social media usage. After including the control variables, the model's explanatory power went up to 14,3, which is higher than the Hungarian case.

Table 3

Table title

Table 3 here

Discussion

This study aimed at understanding the mechanisms underlying the respondents' willingness to participate in an academic study as a social media data donor. To this end, we designed two vignette experiments embedded in two online surveys conducted in Hungary and in the U.S. In hypothetical requests for donating social media via DDPs, we manipulated the amount of monetary incentives (1), the presence or lack of non-monetary incentives (2), the number of platforms to which one is requested to donate (3), the estimated upload/download time (4), and the type of data to be donated (5). The results revealed that data-sharing attitudes are subject to the parameters of the actual study, and some respondent characteristics.

Monetary incentives were the strongest motivators of willingness to donate data in both countries, although the effect was stronger in the Hungarian sample. This finding is consistent with earlier results (Haas et al., 2020; Keusch et al., 2019; Ságvári et al., 2021; Silber et al., 2022).

Non-monetary incentives had a positive effect on willingness in the U.S. sample but not in the Hungarian one. This difference can be linked to the differences between the two online panels. While the Hungarian panel is a standard access panel where panel members regularly receive points and even monetary incentives for completion, the DLABSS panel is fully volunteer based. Receiving a summary report of the participant's digital behavior compared to others can be more motivating for volunteer panel members than for members who normally answer surveys for monetary incentives.

The perceived cognitive burden of the task inconsistently influenced the willingness to donate in the two countries. In line with our hypotheses, the more platforms were included in the request the less likely Hungarian respondents would have participated in the study, while we found an opposite effect in the U.S. sample. Our findings do not provide a clear explanation for these contradictory findings thus further research is needed. In line with earlier studies (Keusch et al., 2019; Ságvári et al., 2021) longer download and upload time was strongly associated with lower willingness in the U.S. sample, but not in the Hungarian one.

Earlier research suggested that participation can be a function of the requested data type, and especially depends on the sensitivity of the data type. Our results are not clearly in line with these results. The type of data had no effect on the responses in the U.S. sample, while in the Hungarian sample, respondents were somewhat more likely to share their data when more data types (including sensitive data) were asked from them. A possible explanation of this result is in the phrasing of the situation: when respondents saw the vignettes where the excluded data types were explicitly listed after each other, it decreased their willingness to share data compared to the condition that asked for the most data, so did not specify, and list the types of excluded data. This suggests that detailed information about the different types of data included in the DDPs may decrease willingness.

Lastly, some respondent characteristics influenced willingness to donate in a significant way. For instance, older and highly educated respondents were more likely to share their data in the Hungarian sample. The results reinforced that privacy and security concerns lower the willingness to participate (Jäckle et al., 2019; Revilla et al., 2019; Struminskaya et al., 2020, 2021; Wenz et al., 2019), at least among U.S. respondents. Consistently in the two samples, participants with multiple platform usage were more likely to donate their data, but for instance, the affinity for technology or personality traits (e.g., openness) of respondents did not influence willingness significantly. Nevertheless, to the extent that these self-reports overlap with actual sharing behavior, our findings altogether suggest that data donation studies should expect strong and systematic selection bias.

The differences in the observed mechanisms found between the two countries may have several reasons. The relatively large cultural differences between American and Hungarian society might cause varying levels of trust in technology companies, cultural norms surrounding data sharing, and awareness

and understanding of the benefits and implications of data donation. Moreover, even if these differences are small, the two online panels were somewhat different in their nature, as we explained earlier. To enhance our comprehension of cross-country differences, further research is needed. However, our results indicate that extrapolating findings obtained from one country to another may be limited when investigating participation in data donation.

Our study has multiple practical implications as well. First, the use of monetary incentives for social media data donation requests is advised given the perceived high cost of the study from the perspective of the respondents. We did not find that a too high incentive would backfire (Silber et al., 2022), although our level of incentives was moderate[4]. Future research could explore the effect of higher incentives to identify the point at which incentives begin to have a counterproductive impact. Second, non-monetary incentives such as personal reports on the results may also be worth considering, although they may not boost willingness in every context or culture. Third, the burden of the respondents should be kept as low as possible. DDP requests are not routine tasks to most of the platform users. Pfiffner & Friemel (2023) reported that only 7.75% of their study participants previously undertook such a request. High download and upload times can deter people from participating. Choosing tasks with the lowest burden and providing helping materials can help researchers to reduce participants' burden. To better understand our inconsistent findings about the number of platforms included in the request, further research is needed. Fourth, although our results do not suggest strong evidence against asking for specific data types, data donation requests should be carefully designed in this regard. Several earlier studies suggested that the sensitivity of the data can lower the willingness to donate (e.g., Pfiffner & Friemel, 2023). Fifth, reinforcing earlier findings, privacy concerns were strongly associated with willingness in our U.S. sample, therefore, addressing privacy concerns is key in social media data donation projects. Future studies should be conducted to explore how privacy concerns can be mitigated, for instance, by using different framing, providing more information about data protection, etc. Lastly, our study highlighted that nonresponse is expected to be high in such data donation requests, and as we can assume that this nonresponse is not random, this likely translates to strong selection bias. Future data donation studies need to develop strategies to handle different types of nonresponses.

One limitation of this study is the use of convenient samples, which limits the generalizability of our findings. Generally, non-panel member internet users are expected to be less likely to participate in data donation studies (Silber et al., 2022), although the extent to which the underlying mechanisms differ between panel- and non-panel members is unclear. Nevertheless, an advantage of using these panels can be that data donation research tends to be based on similar platforms and sampling frames. This population, compared to a random population sample, is more open to a data donation study and, due to its higher digital capabilities, is more likely to be able to retrieve and deliver its data. Another

[4] Our models were also tested by including the variable measuring the incentive as a factor in the model rather than continuously. The explanatory power of the models was not higher in the alternative runs, and the B values of the incentive variable indicated that the effect of the variable of interest was linear.

potential limitation is that we relied on self-reports and do not know to what extent willingness transfers to actual donation. A possible direction for further research is to develop similar experimental designs in which self-reports and real participation can be contrasted (see e.g., Struminskaya et al., 2021). Also, while our data collection could "only" examine a hypothetical situation and not a "real" situation, it can still provide important information on how to design "real" research. What platforms to ask for and what data to ask for within those platforms is essential when designing a survey, as is the "minimum" amount of money to ask for. In the context of willingness to participate, it is clear that the textual framing of the research is necessary because participants do not have a clear understanding of what data is available on these platforms. There are platforms such as Instagram and TikTok, where users have no option to select which specific data they want to download, so users with lower digital literacy may unknowingly provide data for research purposes they would not otherwise want to. Thus, framing research on these platforms significantly impacts the actual donation. Although the detailed description of the data requested will reduce the willingness to participate, ethical considerations should override data collection efficiency, and it is vital to be as clear and precise as possible in telling participants what data we ask them for and what this data will be used for.

Collecting individual-level social media data through DDPs with informed consent and linking this data with survey data is a promising area of research (Boeschoten et al., 2022; Breuer et al., 2022). Our study contributes to the understanding of the circumstances under which individuals are more likely to share their data, and to the assessment of the self-selection bias in data donation studies.

Acknowledgements

Author Contributions

Conceptualization Z.K., Á.S., J.S., E.O., J.K.; theoretical background: Á.S., D.D., A.K., E.P.; methodology Z.K., Á.S., J.S. and J.K..; formal analysis, Z.K., Á.S., J.S.; data curation, Z.K. and J.S..; writing—original draft preparation Z.K., Á.S., J.S., D.D., A.K., E.P., E.O., J.K..; writing—review and editing, Z.K., Á.S., J.S., D.D., A.K., E.P., E.O., J.K.; visualization, J.S..; funding acquisition, Z.K. and S.A.

Funding

The research was funded by the Eötvös Loránd Research Network within the framework of Flagship Research Projects: KÖ-32/2021. The work of Julia Koltai was funded by the Lendület "Momentum" grant of the Hungarian Academy of Sciences.

References

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumgartner, S. E., Sumter, S. R., Petkevič, V., & Wiradhany, W. (2022). A Novel iOS Data Donation Approach: Automatic Processing, Compliance, and Reactivity in a Longitudinal Study. *Social Science Computer Review*, 08944393211071068. https://doi.org/10.1177/08944393211071068

Beuthner, C., Weiß, B., Silber, H., Keusch, F., & Schröder, J. (2023). Consent to data linkage for different data domains – the role of question order, question wording, and incentives. *International Journal of Social Research Methodology*, 1–14. https://doi.org/10.1080/13645579.2023.2173847

Bietz, M., Patrick, K., & Bloss, C. (2019). Data Donation as a Model for Citizen Science Health Research. *Citizen Science: Theory and Practice*, *4*(1), Article 1. https://doi.org/10.5334/cstp.178

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *4*(2), 388-423.

Bradburn, N. M. (1978). Respondent Burden. *Proceedings of the American Statistical Association, Survey Research Methods Section.* 35-40.

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2022). User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, 1–20. https://doi.org/10.1080/1369118X.2022.2097015

Brüggen, E., & Dholakia, U. M. (2010). Determinants of Participation and Response Effort in Web Panel Surveys. *Journal of Interactive Marketing*, *24*(3), 239–250. https://doi.org/10.1016/j.intmar.2010.04.004

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Cheng, A., Zamarro, G., & Orriens, B. (2020). Personality as a predictor of unit nonresponse in an internet panel. *Sociological Methods & Research*, *49*(3), Article 3.

Christner, C., Urman, A., Adam, S., & Maier, M. (2021). Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations. *Communication Methods and Measures*, *0*(0), 1–17. https://doi.org/10.1080/19312458.2021.1907841

Dillman, D. A. (2000). Mail and web-based survey: The tailored design method. *NY: John Wiley & Sons*.

Edwards, P. J., Roberts, I., Clarke, M. J., Diguiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L. M., & Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *The Cochrane Database of Systematic Reviews*, *3*, MR000008. https://doi.org/10.1002/14651858.MR000008.pub4

Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process? *Survey Research Methods*, *13*(2), Article 2. https://doi.org/10.18148/srm/2019.v13i2.7385

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, *1*(1), 58–70.

Groves, R. M., & Couper, M. P. (1998). A conceptual framework for survey participation. *Nonresponse in Household Interview Surveys*, 25–46.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *The Public Opinion Quarterly*, *64*(3), 299–308.

Haas, G.-C., Kreuter, F., Keusch, F., Trappmann, M., & Bähr, S. (2020). Effects of Incentives in Smartphone Data Collection. In *Big Data Meets Survey Science* (pp. 387–414). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118976357.ch13

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Jäckle, A., Burton, J., Couper, M. P., & Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: Coverage and participation rates and biases. *Survey Research Methods*, *13*(1). https://doi.org/10.18148/srm/2019.v1i1.7297

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.) *Handbook of Personality: Theory and Research*, pp. 114–158. The Guilford Press.

Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, *65*(3), 183–216. https://doi.org/10.1007/s11301-014-0111-y

Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to Participate in Passive Mobile Data Collection. *Public Opinion Quarterly*, *83*(S1), 210–235. https://doi.org/10.1093/poq/nfz007

King, G., & Persily, N. (2020). A New Model for Industry–Academic Partnerships. *PS: Political Science & Politics*, *53*(4), 703–709. https://doi.org/10.1017/S1049096519001021

Kmetty, Z., & Bozsonyi, K. (2022). Identifying Depression-Related Behavior on Facebook—An Experimental Study. *Social Sciences*, *11*(3), Article 3. https://doi.org/10.3390/socsci11030135

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? A validity comparison of Facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*. https://doi.org/10.1177/07591063211061754

Kmetty, Z., & Stefkovics, Á. (2021). Assessing the effect of questionnaire design on unit and item-nonresponse: Evidence from an online experiment. *International Journal of Social Research Methodology*. *25*(5), 659-672. https://doi.org/10.1080/13645579.2021.1929714

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained Variance Measures for Multilevel Models. *Organizational Research Methods*, *17*(4), 433–451. https://doi.org/10.1177/1094428114541701

Lugtig, P. (2014). Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers. *Sociological Methods & Research*, *43*(4), 699–723. https://doi.org/10.1177/0049124113520305

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, *15*(4), 336–355. https://doi.org/10.1287/isre.1040.0032

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, *6*(3), 2056305120940703. https://doi.org/10.1177/2056305120940703

Mulder, J., & Bruijne, M. de. (2019). Willingness of Online Respondents to Participate in Alternative Modes of Data Collection. *Survey Practice*, *12*(1). https://doi.org/10.29115/SP-2019-0001

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication*, *9*(2), 293–313. https://doi.org/10.1177/2050157920959106

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking. *Communication Methods and Measures*, 1-18. https://doi.org/10.1080/19312458.2023.2181319

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, *5*(11), Article 11. https://doi.org/10.1038/s41562-021-01117-5

Pinter, R. (2015). Willingness of Online Access Panel Members to Participate in Smartphone Application-Based Research. In R. Pinter, D. Toninelli, & P. de Pedraza (Eds.), *Mobile Research Methods* (pp. 141–156). Ubiquity Press. https://www.jstor.org/stable/j.ctv3t5r9n.14

Pfiffner, N., & Friemel, T. N. (2023). Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate. *Communication Methods and Measures*, 1-23.

Revilla, M., Couper, M. P., & Ochoa, C. (2019). Willingness of Online Panelists to Perform Additional Tasks. *Methods, Data, Analyses*, *13*(2), Article 2. https://doi.org/10.12758/mda.2018.01Ságvári, B., Gulyás, A., & Koltai, J. (2021). Attitudes towards Participation in a Passive Data Collection Experiment. *Sensors*, *21*(18), https://doi.org/10.3390/s21186085

Scharkow, M. (2016). The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data. *Communication Methods and Measures*, *10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P., Stier, S., & Weiss, B. (2022). Linking Surveys and Digital Trace Data: Insights From two Studies on Determinants

of Data Sharing Behaviour, *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(2), 387–407. https://doi.org/10.1111/rssa.12954

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*. *38*(5), 503–516.

Strange, A. M., Enos, R. D., Hill, M., & Lakeman, A. (2019). Online volunteer laboratories for human subjects research. *PLOS ONE*, *14*(8), e0221676. https://doi.org/10.1371/journal.pone.0221676

Struminskaya, B., Lugtig, P., Toepoel, V., Schouten, B., Giesen, D., & Dolmans, R. (2021). Sharing Data Collected with Smartphone Sensors: Willingness, Participation, and Nonparticipation Bias. *Public Opinion Quarterly*, *85*(S1), 423–462. https://doi.org/10.1093/poq/nfab025

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2020). Understanding Willingness to Share Smartphone-Sensor Data. *Public Opinion Quarterly*, *84*(3), 725–759. https://doi.org/10.1093/poq/nfaa044

Thibaut, J. W., & Kelley, H. H. (2017). *The Social Psychology of Groups*. Routledge. https://doi.org/10.4324/9781315135007

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). Package 'mice.' *Computer Software*.

Wenz, A., Jäckle, A., & Couper, M. P. (2019). Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods*, *13*(1), Article 1. https://doi.org/10.18148/srm/2019.v1i1.7298

Wheeler, B., & Braun, M. J. (2019). Package 'AlgDesign.' *R Proj. Stat. Comput*, *1*(0), 1–25.

Appendix

A. Detailed operationalization of independent variables

Gender is a binary variable that assigns 0 to males and 1 to females. We measured the highest level of education with six-category: from "primary" to "university diploma" in Hungary and from "some high school" to "Ph.D." in the US. We operationalized age with the year of birth. Subjective wealth was measured with five categories, where the highest means that they live without financial problems, and the lowest means that they live in deprivation.

To control social media usage, we used two variables. The first variable was the frequency of Facebook usage on a 1 to 5 scale, where the lowest means never, and the highest means daily. The second variable was the number of social media platforms, on which the respondent is active. Here we asked about the following platforms: Facebook, Instagram, Twitter, Youtube, Linkedin, TikTok, and Spotify.

To measure respondents' privacy concerns, we used the Internet Users' Information Privacy Concerns (IUIPC) scale (Malhotra et al., 2004). We applied a confirmatory factor model to extract the three latent dimensions behind the eight validated items. According to our analyses, the model with the three latent variables fit the data well (Hungary CFA:0.99, RMSEA: 0.068; U.S. CFA: 0.99, RMSEA: 0.038). Out of these three dimensions, in the analysis, we only used the 'control' and 'collection' dimensions of the scale and omitted the 'awareness' one, as it highly correlates with the 'control' in both samples. High values of these dimensions mean high control over personal information and concerns about the collection of personal data by companies.

For measuring privacy concerns, we calculated the principal component of the following two variables (measured on a 1 to 7 scale):

- "Most businesses handle the personal information they collect about consumers in a proper and confidential way."
- "Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today."

High values here mean high trust in how businesses and organizations protect consumer data.

To measure the respondents' affinity for technology, we used the 9-item version of the Affinity for Technology Interaction Scale (ATI – Franke et al. (2019). We calculated the mean of the items after reverse coding the needed items. The value of the Cronbach alpha was 0.84 in the Hungarian study and 0.89 in the U.S. study. High values here mean a high affinity for technology.

The last group of independent variables was the big five (BF) inventory (John et al., 2008): Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. We used the 15-item version of the scale and applied a confirmatory factor model to extract the dimensions. In both countries, the tested factor model fit the data (Hungary CFA: 0.99, RMSEA: 0.034; US: CFA: 0.938, RMSEA: 0.062). We had to drop out the reversed coded items from the models because of their poor fit in both the Hungarian and U.S. dataset.

*B. Manipulated question (*The bold parts were the manipulated part of the experiment.*):*

„The various social media sites and platforms (Facebook, Instagram, Twitter, Google) allow users to view and even download information and data about themselves stored on the site. This data is very valuable from a scientific point of view since it captures behavioral patterns not observed elsewhere. Imagine a situation in which you are asked by the Social Science Research Centre to participate in a survey. You are invited to fill

in a questionnaire, and you are asked to share your **Facebook** data, **excluding your private messages and pictures/videos**. Downloading the data to your computer and uploading it to the research page would take **less than 1 hour**. For participating in the research, you would receive **3000 HUF** and a **personalized report on your social media usage compared to the rest of the Hungarian internet population**. Once uploaded, the data would be anonymized and only analyzed for research purposes.

Please indicate on a scale of 0-10 how likely would you be to participate in such research! 0 indicates not likely at all, and 10 indicates very likely."

Table A1

Results of the Vignette Experiment split by the first and second six vignettes – *Hungary (*multilevel mixed-effects linear regression*)*

	Vignettes 1-6			Vignettes 7-12		
(Intercept)	3.77	0.13	<0.001	3.17	0.14	<0.001
Incentive	0.24	0.01	<0.001	0.27	0.01	<0.001
Report	0.01	0.06	0.83	0.03	0.06	0.61
Platform: FB + Google	-0.32	0.06	<0.001	0.01	0.06	0.80
Platform: FB + other	-0.21	0.07	<0.001	-0.10	0.05	0.06
Platform: FB + Google + Other	-0.25	0.06	<0.001	-0.10	0.06	0.10
Time	-0.04	0.05	0.37	-0.09	0.04	0.02
Type of data: no PM/loc	-0.13	0.05	0.01	0.01	0.07	0.90
Type of data: no PM/vid	-0.21	0.07	<0.001	0.02	0.06	0.79
Type of data: no PM/loc/vid	-0.36	0.07	<0.001	0.03	0.06	0.58
Variances of random effects						
Variance: constant		11.08			11.44	
Variance: residual		2.76			2.13	
Proportion of Level 1 variance		19.9%			15.7%	
Proportion of Level 2 variance		80.1%			84.3%	
Model fit						
Variance explained (Level 1)		9.2%			11.6%	
Variance explained (overall)		1.7%			1.6%	

Tables

Table 1
Manipulated dimensions and their levels in the survey experiment

Dimension	Levels	Explanation
Platform	- Facebook - Facebook and Google - Facebook and other social media sites you use (Instagram, Twitter, Spotify) - Facebook, Google and other social media sites you use (Instagram, Twitter, Spotify)	Our research was the first step in a more extensive data donation project. Facebook data plays a major role in our data donation research, so we wanted to keep this platform as a reference.
Range of data	- all, except: private messages - all, except: private messages and location - all, except: private messages and photo, videos - all, except: private messages, and photo, videos, and location	Private messages include messages from the user and their conversation partners, so we did not consider this to be shareable data, despite the participant's consent.
Time to download/ upload data	- Less than an hour - More than an hour	DDP data is never made immediately available by the platforms and would have to wait hours or even days to become available for download. With a download/upload time of more than one hour, we wanted to explore whether it makes a difference if the respondent cannot resolve the request within one session.
Incentives	- 3000 HUF/ $10 - 5000 HUF / $20 - 10 000 HUF /$30	3000 HUF was around 8 U.S. dollars during the data collection. In order to standardize the money incentives in the two surveys, we converted the Hungarian Forint to U.S. dollars and adjusted it with purchasing power parities (see: https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm)
Additional report	- Yes - No	Additional reports mean feedback and summaries on social media usage, such as activity patterns, closest friends based on activity, or the number of friends over time.

Table 2

Results about willingness to donate data – Hungary (multilevel mixed-effects linear regression)

	Null model			Model with vignette dimensions			Model with vignette dimensions and controls		
(Intercept)	4.25	0.11	<0.001	3.55	0.12	<0.001	4.93	1.17	<0.001
Incentive				0.25	0.01	<0.001	0.25	0.01	<0.001
Report				0.01	0.03	0.66	0.01	0.03	0.68
Platform: FB + Google				-0.19	0.04	<0.001	-0.19	0.04	<0.001
Platform: FB + other				-0.21	0.04	<0.001	-0.21	0.04	<0.001
Platform: FB + Google + Other				-0.18	0.04	<0.001	-0.18	0.04	<0.001
Time				-0.03	0.03	0.34	-0.03	0.03	0.34
Type of data: no PM/loc				-0.06	0.04	0.14	-0.06	0.04	0.13
Type of data: no PM/vid				-0.19	0.04	<0.001	-0.2	0.04	<0.001
Type of data: no PM/loc/vid				-0.22	0.04	<0.001	-0.22	0.04	<0.001
Controls									
Gender							-0.59	0.22	0.01
Age							-0.03	0.01	<0.001
Education							-0.25	0.1	0.01
Subjective wealth							-0.11	0.12	0.36
IUIPC_control							0.07	0.1	0.46
IUIPC_collect							-0.12	0.08	0.2
Privacy beliefs							0.12	0.09	0.23
Tech attitudes							0.12	0.11	0.27
BF: openness							0.09	0.16	0.57
BF: conscientiousness							-0.05	0.17	0.77
BF: extroversion							0	0.1	0.98
BF: agreeability							0.06	0.17	0.73
BF: neuroticism							-0.01	0.09	0.88
Social Media usage frequency							0.19	0.14	0.18
No of platforms							0.14	0.06	0.03
Variances of random effects									
Variance: constant		11.14			11.18			10.37	
Variance: residual		2.84			2.59			2.59	
Proportion of Level 1 variance		20.3%			18.8%			20.0%	
Proportion of Level 2 variance		79.7%			81.2%			80.0%	
Model fit									
Variance explained (Level 1)					8.8%			8.8%	
Variance explained (Level 2)					0.0%			6.9%	
Variance explained (overall)					1.5%			7.3%	

Table 3

Results about willingness to donate data – US (multilevel mixed-effects linear regression)

	Null model			Model with vignette dimensions			Model with vignette dimensions and controls		
(Intercept)	2.07	0.12	<0.001	1.39	0.12	<0.001	0.47	0.96	0.62
Incentive				0.35	0.02	<0.001	0.35	0.02	<0.001
Report				0.15	0.05	0.01	0.15	0.05	<0.001
Platform: FB + Google				0.14	0.05	0.01	0.14	0.06	0.01
Platform: FB + other				0.1	0.05	0.07	0.1	0.05	0.06
Platform: FB + Google + Other				0.18	0.06	<0.001	0.19	0.06	<0.001
Time				-0.38	0.04	<0.001	-0.38	0.04	<0.001
Type of data: no PM/loc				-0.01	0.06	0.8	-0.01	0.06	0.82
Type of data: no PM/vid				-0.04	0.05	0.46	-0.03	0.06	0.53
Type of data: no PM/loc/vid				0.06	0.05	0.3	0.06	0.06	0.3
Controls									
Gender							0.35	0.16	0.03
Age							-0.01	0.01	0.08
Education							-0.01	0.06	0.84
Subjective wealth							-0.14	0.1	0.17
IUIPC_control							0.18	0.17	0.32
IUIPC_collect							-0.54	0.13	<0.001
Privacy beliefs							0.12	0.13	0.39
Tech attitudes							0.06	0.12	0.62
BF: openness							0.07	0.18	0.7
BF: conscientiousness							-0.36	0.18	0.05
BF: extroversion							-0.02	0.1	0.8
BF: agreeability							0.24	0.19	0.23
BF: neuroticism							0.03	0.05	0.57
Social Media usage frequency							0.27	0.12	0.03
No of platforms							0.24	0.09	0.01
Variances of random effects									
Variance: constant		8.27			8.28			6.99	
Variance: residual		1.44			1.32			1.33	
Proportion of Level 1 variance		14.8%			13.8%			16.0%	
Proportion of Level 2 variance		85.2%			86.3%			84.0%	
Model fit									
Variance explained (Level 1)					8.3%			7.6%	
Variance explained (Level 2)					0.0%			15.5%	
Variance explained (overall)					1.1%			14.3%	

Figures

Figure 1

Relative effects of vignette level characteristics on willingness to donate data; only vignette level independent variables; allowing for random intercept. Standardized regression coefficients. Hungary.

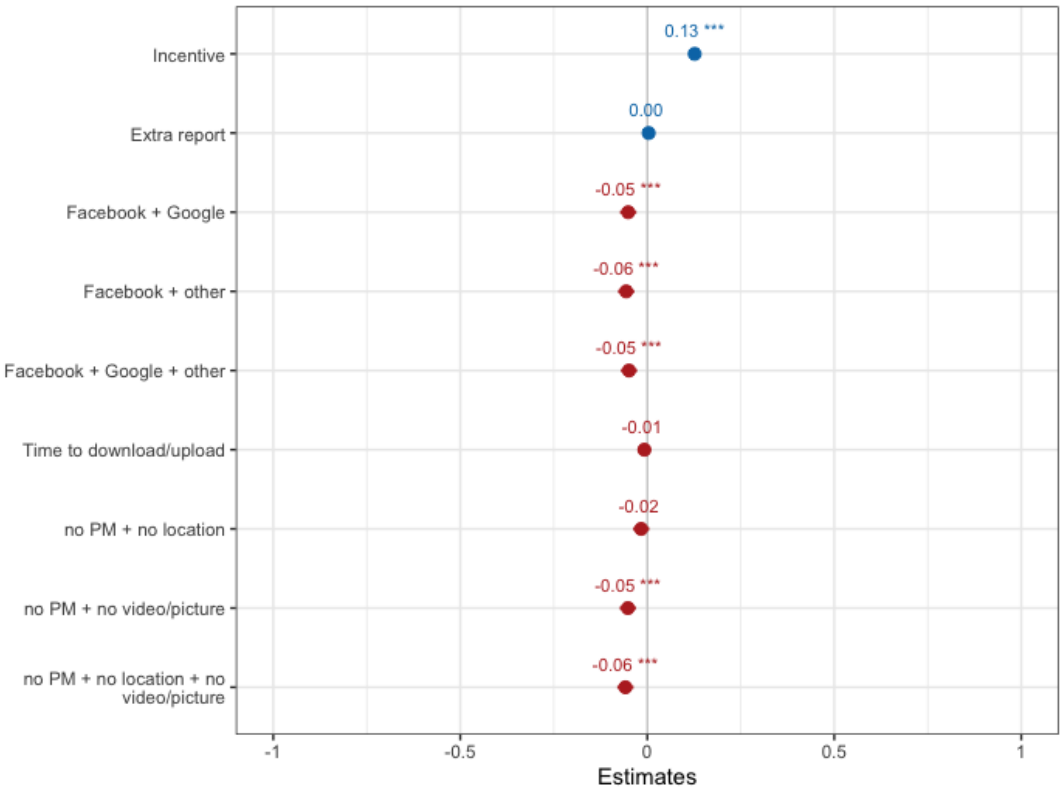

Figure 2

Relative effects of vignette level characteristics on willingness to donate data; only vignette level independent variables; allowing for random intercept (Model 2). Standardized regression coefficients. US.