# ENAGRAM: An App to Evaluate Preventative Nudges for Instagram

Nicolás E. Díaz Ferreyra
Hamburg University of Technology
Hamburg, Germany
nicolas.diaz-ferreyra@tuhh.de

Sina Ostendorf
University of Duisburg–Essen
Duisburg, Germany
sina.ostendorf@uni-due.de

Esma Aïmeur
University of Montréal
Montréal, Canada
aimeur@iro.umontreal.ca

Maritta Heisel
University of Duisburg-Essen
Duisburg, Germany
maritta.heisel@uni-due.de

Matthias Brand
University of Duisburg-Essen
Duisburg, Germany
matthias.brand@uni-due.de

## ABSTRACT

Online self-disclosure is perhaps one of the last decade's most studied communication processes, thanks to the introduction of Online Social Networks (OSNs) like Facebook. Self-disclosure research has contributed significantly to the design of *preventative nudges* seeking to support and guide users when revealing private information in OSNs. Still, assessing the effectiveness of these solutions is often challenging since changing or modifying the choice architecture of OSN platforms is practically unfeasible. In turn, the effectiveness of numerous nudging designs is supported primarily by self-reported data instead of actual behavioral information. **Objective**: This work presents ENAGRAM, an app for evaluating preventative nudges, and reports the first results of an empirical study conducted with it. Such a study aims to showcase how the app (and the data collected with it) can be leveraged to assess the effectiveness of a particular nudging approach. **Method**: We used ENAGRAM as a vehicle to test a risk-based strategy for nudging the self-disclosure decisions of Instagram users. For this, we created two variations of the same nudge (i.e., with and without risk information) and tested it in a between-subjects experimental setting. Study participants (N=22) were recruited via Prolific and asked to use the app regularly for 7 days. An online survey was distributed at the end of the experiment to measure some privacy-related constructs. **Results**: From the data collected with ENAGRAM, we observed lower (though non-significant) self-disclosure levels when applying risk-based interventions. The constructs measured with the survey were not significant either, except for participants' External Information Privacy Concerns (EIPC). **Implications**: Our results suggest that (i) ENAGRAM is a suitable alternative for conducting longitudinal experiments in a privacy-friendly way, and (ii) it provides a flexible framework for the evaluation of a broad spectrum of nudging solutions.


*EuroUSEC 2022, September 29–30, 2022, Karlsruhe, Germany*

ACM ISBN 978-1-4503-9700-1/22/09…$15.00
https://doi.org/10.1145/3549015.3555674

## CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; **Usability in security and privacy**; • **Human-centered computing → HCI design and evaluation methods**.

## KEYWORDS

privacy nudges, risk awareness, usability, personalization, online social networks

**ACM Reference Format:**
Nicolás E. Díaz Ferreyra, Sina Ostendorf, Esma Aïmeur, Maritta Heisel, and Matthias Brand. 2022. ENAGRAM: An App to Evaluate Preventative Nudges for Instagram. In *2022 European Symposium on Usable Security (EuroUSEC 2022), September 29–30, 2022, Karlsruhe, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3549015.3555674

## 1 INTRODUCTION

Over the last decades, the use of information and communication technologies has widely extended across different segments of everyday life. From making bank transactions to finding a life partner, online services have acquired an increasingly important role in the dynamics of modern societies [40]. Nevertheless, living in a digitalized world also introduces threats and challenges related to privacy and security since online services are fed and operate over large amounts of personal data. At the same time, technology must create adequate cybersecurity conditions to safeguard the privacy rights and the integrity of its users. For this, legal frameworks —such as the EU General Data Protection Regulation (GDPR)— were introduced to enforce tech companies to comply with a set of data protection principles. Overall, this contributes to ensure a secure processing and storage of personal data, preventing its non-consensual exploitation, and avoiding unfair discrimination of data subjects, among others [14]. However, to a large extent, the privacy decisions and practices of individuals have also been challenged by the affordances of media and communication technologies [6]. Particularly, Online Social Networks (OSNs) like Facebook or Instagram have redefined and blurred people's privacy boundaries by creating spaces in which they can connect seamlessly and share personal information with large (and sometimes untrusted) audiences [25].

Like in the real world, individuals disclose private information in OSNs to create and maintain social relationships with others [37, 53]. Thereby, the strength of such relationships tends to increase, and so does people's social capital [17]. However, deciding whether or not

to disclose personal information to others is not always straightforward (even more so in online environments) [25]. To a large extent, this may because online self-disclosure decisions are likely driven by short-term gratifications (e.g., likes, comments, or number of followers) instead of long-term privacy risks [35]. In turn, users are prone to disclose sensitive information unseemly to untrusted recipients and becoming victims of privacy threats such as reputation damage, social engineering, and even financial fraud [2]. Moreover, OSNs often hinder individuals' self-presentation decisions as they place different audiences (e.g., work colleagues and family) in a common communication plane [50]. Consequently, users frequently experience regret —along with unwanted incidents— after sharing personal information with unintended recipients [52].

All in all, interaction in OSNs can lead to unwanted incidents even on platforms exhibiting secure backend infrastructures and compliant with data protection principles [3, 14]. To a great extent, information and cues about the potential risks of unrestrained self-disclosures can help users regulate their exposure levels and mitigate their chances of experiencing negative consequences [16, 20, 34]. Nonetheless, current layouts and graphical interfaces of OSNs do not provide any means (i.e., cues or information) that may help users determine the potential risks and hazardous outcomes of their disclosures [48]. Conversely, platforms showcase many cues not only related to immediate gratification (e.g., like buttons) but also related to their reputation (e.g., their size) or recognition (e.g., their market presence) that contribute to larger self-disclosure levels, since such cues increase trust in the platform, and larger self-disclosure levels, in turn, serve their business model. In some cases, individuals can even develop problematic/addictive usage behaviors. Although more research is needed, a problematic/addictive usage of OSNs may (among other factors) arise due to application-specific features and affordances that especially enable the experience of immediate gratification while potentially hindering self-control and further reflective processes [11]. Moreover, the privacy policies of OSNs are also devoid of information related to potential privacy threats and leave rational risk estimations to the individual discretion of each user [14, 48]. Hence, there is an urgent call for technological affordances that promote safe and more reflective information-sharing decisions among the users of OSNs, so that the risks of online-self disclosure are mitigated or prevented.

## Motivation

Over recent years, privacy scholars have introduced a wide range of technological approaches that aim to improve people's online privacy decisions [7, 31, 33, 43, 44]. In particular, the use of nudges has gained popularity due to their capacity for assisting and guiding individuals towards safer privacy practices [1]. At their core, nudges are interventions that encourage a certain behaviour which, in turn, tends to maximize people's welfare [49]. Such interventions are the means for behavioural change and consist of small modifications in the context within which decisions are made [28]. For instance, displaying cues related to the targeted audience of a post can motivate users to employ custom friend lists [52]. Given the close relation existing between risk perception and privacy behaviour, it is not surprising that interventions portraying risk information are deemed adequate for motivating safer self-disclosure decisions in OSNs [25, 29, 42]. In particular, such interventions can prevent users from sharing posts with personal data by rendering information about the risks of unsafe self-disclosure practices [20].

Despite researchers' increasing interest in nudges and their applications to cyber-security, many solutions have remained theoretical or at early design stages (c.f., [7, 16, 27, 44]). This is because nudges are frequently conceived as interventions that should be integrated into preexisting choice architectures. That is, into current OSNs platforms or services. However, the most popular platforms do not offer integration mechanisms that would allow researchers to assess nudges' effectiveness within their intended operational environment. In turn, many approaches receive partial evaluation from end-users through mock-ups and self-reports. Hence, there is a call for evaluation approaches in which preventative nudges can be tested and assessed under more realistic working conditions.

## Contribution

In this work we present ENAGRAM, an app to evaluate preventative nudges for Instagram. ENAGRAM is an independent 3rd-party Android application that acts as an Instagram proxy. As with the original Instagram app, it allows users to elaborate posts using free text and pictures (e.g., photos taken with their phones). However, it also incorporates a nudging mechanism consisting of interventions (or pop-up messages) that are triggered at the moment of sharing the posts. Users' reactions to these interventions (i.e., whether they ignored them or not) along with some supplementary information (e.g., a hashed version of the post) are recorded by the app in a dedicated server for later analysis.

We conducted a pilot study via Prolific (N=22) to showcase ENAGRAM's evaluation features and functionalities. For this, we used the app to test a risk-based strategy in which information about common OSN incidents (e.g., reputation damage, identity theft, etc.) is applied to nudge users towards safer self-disclosure decisions. We chose this particular nudging approach as a running example since it is part of our prior work (see [16]). Nevertheless, ENAGRAM's interventions can be tailored for assessing other nudging solutions alike. All in all, the contribution of this paper is two-fold: (i) it presents and discusses ENAGRAM's affordances for the evaluation of preventative nudges, and (ii) reports the results of a preliminary study on the effectiveness of the implemented nudging strategy.

The remainder of this paper is organized as follows. In the next section we provide the paper's background and discuss related work concerning the design and evaluation of preventative nudges for OSNs. Next, in Section 3 we introduce ENAGRAM's main architectural components, whereas in Section 4 we describe the methodology applied for its empirical assessment (7-day between-subjects approach). The results of our preliminary study are presented in Section 5 and then discussed in Section 6. Particularly, we analyze ENAGRAM's benefits and drawbacks based on a set of constructs and performance metrics collected by the end of the experiment. Finally, we present the limitations of our approach in Section 7 and conclude with some prospective directions for future work in Section 8.

## 2 BACKGROUND AND RELATED WORK

Since its introduction by Nobel Prize winners Richard Thaler and Cass Sunstein, the "nudge" theory has been extensively investigated and applied repeatedly to the design of privacy-enhancing technologies. When it comes to OSNs, there is a wide variety of privacy nudges in the current literature whose goal is to support users' online self-disclosure decisions [7, 9, 10, 16, 30, 38]. For instance, Raber et al. [38] introduced *PrivacyWedges*, a visualization strategy for aiding the audience selection of social media publications. Under the premise that "close" contacts are often the most trustworthy ones, *PrivacyWedges* displays network members based on their interpersonal distance to the targeted user (i.e., well-known friends are prioritized over the rest). Thereby, the user is encouraged to keep sensitive posts within her inner circle of friends and thus away from unintended recipients. Other nudging solutions precisely seek to warn users about publications containing sensitive information. Such is the case of Botti-Cebriá et al. [9], who applied a compound of Natural Language Processing (NLP) techniques for determining whether a post contains information about one's location, health, and personal identifiers, among others. In line with this, Bracamonte et al. [10] empirically assessed users' perception of warnings about the presence of personal information in their posts. All in all, most study participants considered such nudges useful but were agnostic regarding their future adoption.

Wang et al. [51] conducted perhaps one of the most groundbreaking studies on the use of privacy nudges in OSNs. They implemented three different nudges that intervened when users were about to post something on Facebook: (i) an *audience* nudge showed visual cues about the potential recipients of the post (i.e., pictures of friends), (ii) a *sentiment* nudge displayed the sentiment of the text being posted, and (iii) a *timing* nudge delayed the actual publication of the post for some minutes. These interventions were designed to give users the chance to re-think their disclosures, edit, or even withdraw them before publication. The findings of this experiment not only yielded valuable evidence on the effectiveness of preventative nudges but also served as a reference for later contributions elaborating on the same (or closely-related) intervention strategies. Such is the case of Masaki et al. [30], who incorporate information about frequent online privacy harms to the nudge's design in a very similar way. Other solutions like the ones of Díaz Ferreyra et al. [16] and Ben Salem et al. [7] also stress the importance of personalization to increase nudges' efficacy. Particularly, they seek to overcome the limitations of one-size-fits-all approaches by adjusting the frequency and content of interventions to the individual privacy goals of each user.

Despite the great attention these nudges have received in recent years, only a reduced number of the proposed solutions have been implemented and tested under realistic conditions. That is, through evaluation instruments other than mock-ups and self-reports. As pointed out by Gómez-Barroso [21], "*...experiments are but a few when compared to surveys and theoretical approaches, even when adopting a broad definition of experiment*". When looking closer at some of the few implemented solutions, it is clear that integration is a major challenge as none of them is heavily embedded into a commercial OSN platform (e.g., Facebook, Twitter, or Instagram). For instance, the nudges of Wang et al. [51] were implemented as browser extensions scrapping Facebook's web interface, whereas a recent approach by Alemany et al. [4] was developed and tested in a non-commercial platform. In the case of browser extensions, one limitation relates to the evolving nature of OSNs, which makes them obsolete after significant changes on the site's interface are made. Furthermore, as users gravitate toward mobile apps, browser-based experimental conditions may become less engaging for research participants and thus unfeasible for conducting longitudinal studies. On the other hand, while non-commercial OSNs offer suitable nudge integration means, they may be alien to most study participants and thus fail to recreate interaction patterns like the ones emerging within commercial sites. Therefore, there is a need for evaluation methods and tools that help researchers overcome said integration limitations of proprietary OSNs.

## 3 ENAGRAM ARCHITECTURE

ENAGRAM is an Android app created to support the research and evaluation of preventative nudges on Instagram. It incorporates features for capturing users' reactions towards behavioral interventions while offering a flexible framework for testing different variants of such interventions.

As shown in Figure 1-a, ENAGRAM provides the basic functionality for creating Instagram publications, namely picture selection and caption composition features. At its core, ENAGRAM acts as an Instagram proxy: it forwards the post information (i.e., photo and caption text) to the Instagram app via Android intents. However, before passing the control over to Instagram, ENAGRAM intervenes with a pop-up message. Such a pop-up can, for instance, display the legend "Ready to share?" and a "fact of the day" describing a self-disclosure privacy threat (Figure 1-b). Users then have the chance to proceed and post their messages on Instagram or to go back to the composition screen (Figure 1-a) and edit them. If a user decides to continue, she can choose whether to post the message on her Instagram feed, create a story, or send it via direct message (Figure 1-c). This nudging strategy proposed by Díaz Ferreyra et al. [16] aims to promote reflective self-disclosure practices among OSNs users by means of risk information and awareness. However, ENAGRAM could be extended/tailored for testing other nudging solutions alike (see Section 6.2).

Figure 2 (left) illustrates ENAGRAM's main architectural components. It follows a REpresentational State Transfer (REST) communication schema where client (Android app) and server (HTML web server) interact via HTTP methods (e.g., GET, POST, and DELETE) using a lightweight data-interchange format (i.e., JSON). Communication is done through an Application Programming Interface (API) consisting of a collection of methods and operations such as *Login*, *Logout*, *Register*, and *Post*. Such API methods are PHP implemented and allow retrieving (pushing) content from (to) an SQL database (`EventsDB`) containing information about users' interactions within ENAGRAM. As shown in Figure 2 (right), the `EventsDB` consists of 5 tables:

i **`users_table`**: Contains the login credentials (username and password) of registered users, their current app version, and app language. As we describe later in Section 4, we created two versions of ENAGRAM, both in English and German (i.e., 4 variants in total).

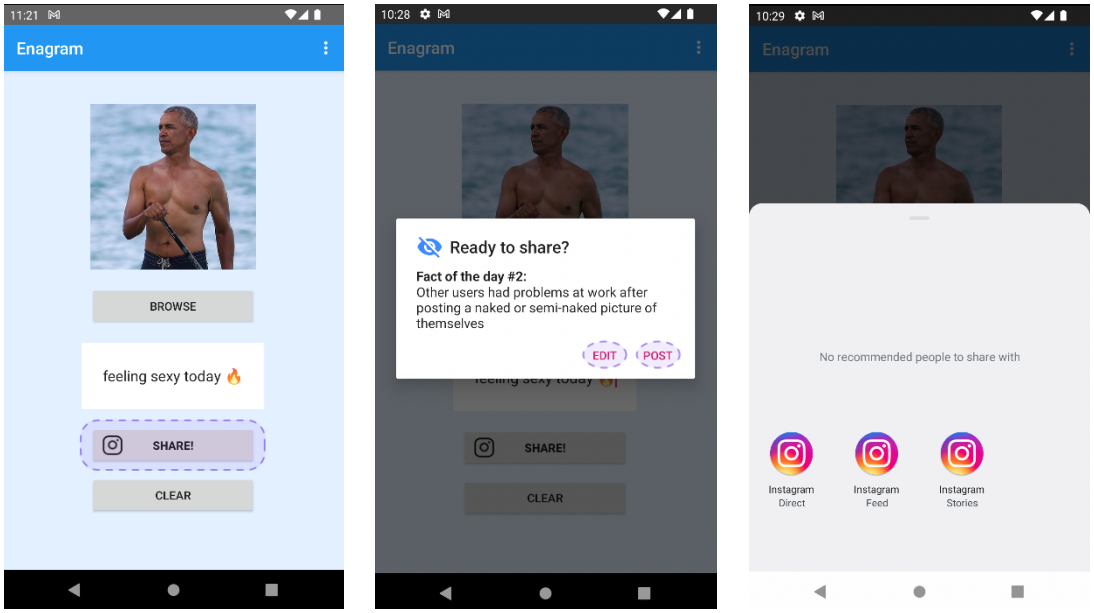


**Figure 1: ENAGRAM interfaces for Group 2: Post composition (a), Risk-based intervention (b), and posting selection (c).**

ii **`interventions`**: Contains the intervention messages shown within the app and some additional information (e.g., a risk value) that could be leveraged to adapt the display frequency of the corresponding privacy prompts (c.f., [16]). ENAGRAM includes a total of 26 different intervention messages.

iii **`intervention_categories`**: Intervention messages are grouped around 6 large categories (i) *drugs and alcohol use*, (ii) *sex*, (iii) *religion and politics*, (iv) *strong sentiment*, (v) *location*, and (vi) *personal identifiers*. Hence, each entry in the `interventions` table belongs to one of these categories (e.g., the message displayed in Figure 1-b belongs to the "sex" category). Both intervention messages and categories correspond to the ones curated by Díaz Ferreyra et al. [16].

iv **`popup_actions`**: Defines the type of actions a user can take when interacting with the nudge pop-up:

- *action_id* = 0: The user clicked "edit" after receiving an intervention.
- *action_id* = 1: The user clicked "post" after receiving an intervention.

To minimize the chances of habituation biases and annoyance, the current version of ENAGRAM intervenes at most 5 times a day with a time interval of 60 min. between interventions. Hence, it may happen that a user may not receive an intervention after clicking on "SHARE!" (e.g., if she wishes to share a picture 10 min. after being nudged for the first time). To keep track of all the sharing events within the app, we included the following action type:

- *action_id* = 2: The user clicked "SHARE!" in the main window but did not receive an intervention afterwards.

v **`user_activity`**: This table contains all the events recorded by ENAGRAM for all its users. Such events correspond to the actions listed in the `popup_actions` table characterized by the following contextual information:

- *popup_action*: The type of action being recorded (i.e., 0, 1, or 2).
- *user_id*: The id of the user whose action is being recorded.
- *msg_id*: A number between 1 and 26 pointing to the id of the warning message being displayed in the pop-up. This field is `null` when *popup_action* is equal to 2.
- *post_lenght*: Number of characters in the picture caption.
- *post_hash*: A hashed version (i.e., a pseudonym consisting of a fixed-size sequence of hexadecimal characters) of the picture caption. This can be used to check whether the user changed the caption after receiving an intervention.
- *image_hash*: A hashed version of the picture file path. Like the previous one, it can be used to check whether the user selected a different picture after receiving an intervention.
- *timestamp*: The time at which the event occurred.

The `user_activity` table is populated every time the user clicks on "SHARE!", "EDIT", or "POST". Hence, it can be seen as a collection of snapshots describing the user's self-disclosure behavior over time.

## 4 EXPERIMENTAL DESIGN

We conducted a preliminary study on a particular nudging approach to explore ENAGRAM's evaluation features. As mentioned in Section 1, we decided to test the strategy proposed in [16] since it is part of our prior work. Thus, the outcome of this study provides (i) actionable information about the effectiveness of such a strategy and (ii) empirical evidence about ENAGRAM's benefits and drawbacks. Whereas the remaining of this paper focus mainly on the former point, the latter is discussed thoroughly in Section 6.2.

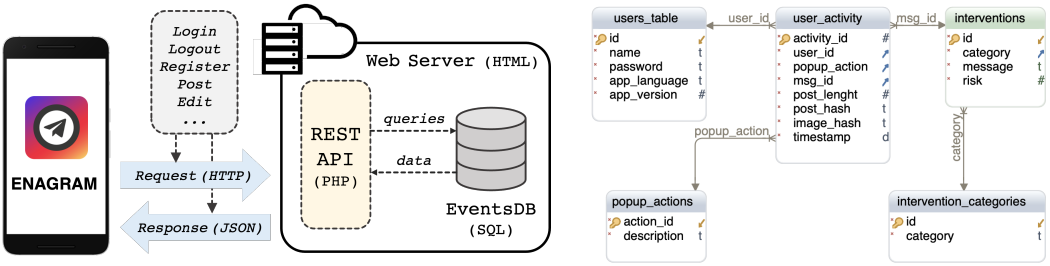


**Figure 2: ENAGRAM architecture (left) and EventsDB schema (right).**

We created 2 versions of the app and tested them in a between-groups experimental setting. The two versions only differed in the intervention pop-up: *version 1* (v1) only displayed the legend "Ready to share?" (i.e., without showing any risk information) and *version 2* (v2) included also the "fact of the day". As mentioned in Section 3, users were nudged at most 5 times a day with a minimum gap of 60 min. between interventions. In addition, v2 users did not receive the same threat description twice in the same day.

*Recruitment.* The study participants were recruited via Prolific[1] and assigned to one of the two experimental conditions: participants in Group 1 tested ENAGRAM v1 and participants of Group 2 tested ENAGRAM v2. The group assignment was done pseudo-randomly to create a gender balance within each experimental condition. Participants had to be active Instagram users, at least 18 years old, and had to have an Android phone with OS version 9.0 or higher (API 28). We asked them to have at least 1GB free space in their devices and the latest Instagram version installed. Having a computer/laptop/notebook was also a requirement as we included some questionnaires as part of the study.

*Study approach.* The study consisted of three subsequent stages, namely *briefing*, *assessment*, and *debriefing* (Figure 3). During the *briefing*, participants were asked to install ENAGRAM on their phones and answer some demographic questions (e.g., gender, age, and average time spent on Instagram). We used the following *cover story* to avoid behavioral biases: participants were told that ENAGRAM was built following a software development method created by university researchers, and that their job was to test the app for **7 days** and report possible implementation flaws (e.g., glitches, errors, missing requirements). As part of the briefing, each participant received a short tutorial describing the installation steps and instructions for creating an ENAGRAM username and password. A registration code was generated by ENAGRAM which participants had to provide to show they actually installed the app. We explicitly asked them to use a pseudonym as username to track their performance during the whole study. Good command of the German language was also required as we tested the app on its German version.

After 7 days we conducted an *assessment* and *debriefing* of the study participants. The *assessment* consisted of a sort survey asking the participants if they used ENAGRAM regularly in the last 7 days or not. Those who reported not having used the app were then debriefed and fully informed about the actual purpose of the study (viz., test an app containing a nudge mechanism for online self-disclosure). Otherwise, they were asked to complete another survey containing questions about the *performance* of the app and on the following privacy-related constructs: *perceived risks* (RSK), *perceived control* (CTRL), *perceived benefits* (BEN), and *external information privacy concerns* (EIPC). All constructs were previously elaborated and validated by other authors (i.e., EIPC by Morlok [32] and the rest by Krasnova et al. [26]) and measured using a 7-point Likert scale ranging from 1 = "strongly disagree" to 7 = "strongly agree". A summary of all employed constructs can be found in the Appendix.

*Ethical considerations.* The study was conducted in accordance with the Declaration of Helsinki and approved by the local Ethics Committee of the Department of Computer Science and Applied Cognitive Science of the University of Duisburg-Essen. All participants received information about the study procedure (including data privacy statements) and were asked to give their informed consent before moving forward in the different experimental stages. They

[1]https://prolific.co

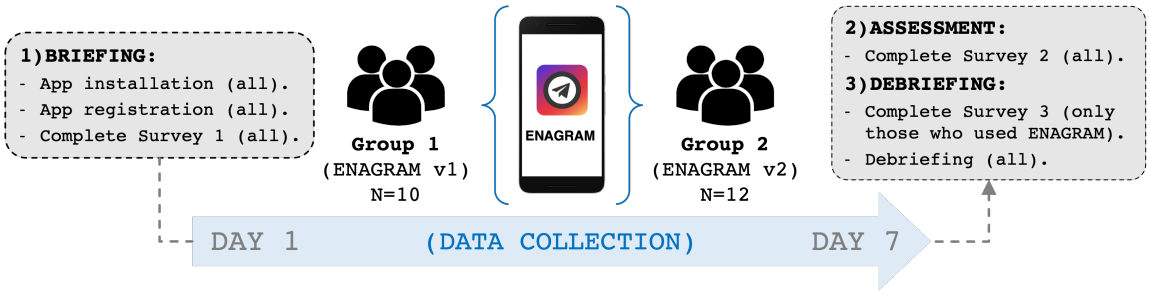


**Figure 3: Study design workflow (group sizes refer to the number of participants who completed the experiment).**

could withdraw at any time receiving a compensation for each completed part of the study: 6€ after the briefing, another €6 after the assessment, plus €5 for the final performance questionnaire. In all cases (i.e., after withdrawing or completing the study) participants were debriefed accordingly and then asked whether we could still use their data for research purposes. Survey instruments, software, and study results are available as **Supplementary Material**.

## 5 RESULTS

We recruited 24 participants for the study (12 for each group condition). One participant from Group 1 was discarded after the assessment stage (not having used the app) and another one by the end of the study (not having answered the control questions correctly). Hence, we considered and analyzed the data gathered from 22 subjects: Group 1 consisted of five females, four males, and one non-binary person (19-34 years, $M$ = 23.30, $SD$ = 5.27), and Group 2 of eight females and four males (18-33 years, $M$ = 23.33, $SD$ = 5.40).

### 5.1 Behavioral insights

We conducted a manual inspection of the data collected by ENAGRAM stored in the `EventsDB`. After a sanity check, we observed some duplicated entries in the *user_activity* table probably due to connection issues on the client side (e.g., the participant's phone lost connection at the moment of sending the data to ENAGRAM' web server). Such duplicated entries were removed resulting on a collection of 137 events: 19 edits (#EDITS) and 118 publications (#PUBLICATIONS) from which 85 (#POSTS) correspond to post actions performed after an intervention (i.e., within the intervention pop-up) and 33 to share actions that were not followed by an intervention (#SHARES). All in all, we registered 53 interventions across all participants in Group 1 and 51 across all participants in Group 2.

Edit events are of special interest for measuring the effectiveness of the nudging approach. Particularly, cases in which users change the picture or the caption after receiving an intervention could help us determine whether the nudge has indeed an impact on privacy behavior. Such cases can be identified through the *post_length*, *post_hash*, and *image_hash* values of EDIT events (*action_id*=0) that are closely followed (i.e., regarding *timestamp*) by SHARE! events (*action_id*=2). If any of these values change from one event to the other, then such a change can be interpreted as an effect of the nudge on the user's self-disclosure behavior. Nevertheless, we identified only 2 cases in which a participant changed either the picture or the caption after clicking on edit (one from Group 1 and the other from Group 2). The rest of the editing cases did not include any changes neither in the selected picture nor in the caption.

### 5.2 Group comparisons

Figure 4 shows the average #EDITS, #POSTS, #SHARES, and #PUBLICATIONS per study group. As one can observe, all of these values are higher in Group 1 than in Group 2. To determine whether such differences are statistically significant, we conducted an independent sample $t$-Test (Table 2). Since Levene's test for equality of variances was non-significant in all cases ($p > 0.05$), the corresponding $t$ statistics were computed assuming homogeneity of variances. Overall, we found no significant differences between any of the values obtained for Group 1 and Group 2. The effect sizes we obtained were in general "small" according to Cohen's convention [13].

We repeated this analysis with the constructs elicited by the end of the experiment (i.e., RSK, CTRL, BEN, EPIC). From Table 1, we can see that, with the exception of CTRL, all construct values are higher for Group 2 than for Group 1. Once again, we assumed homogeneity of variances when conducting the $t$-Tests as Levene's test was non-significant in all cases ($p > 0.05$). As shown in Table 2, the differences between the values obtained for Group 1 and Group 2 were only significant for the EPIC construct. For the rest of the constructs, such differences were not statistically significant. These results yielded "large" effect sizes for CTRL and EIPC, "medium" for BEN, and "small" in the case of RSK.

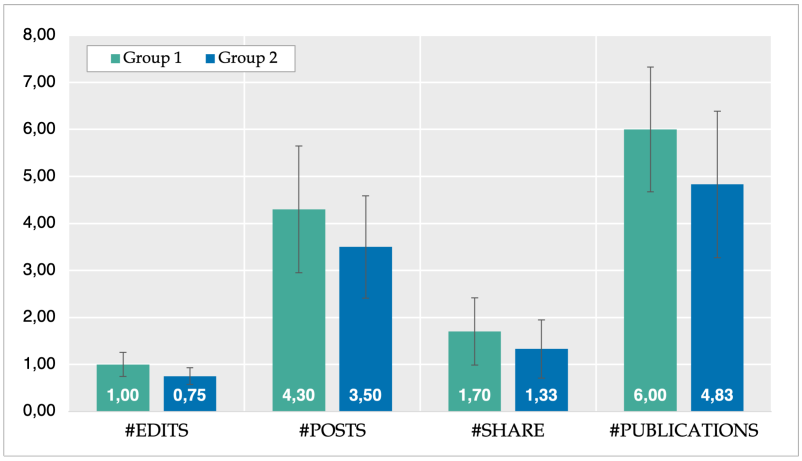


**Figure 4: Average number of events recorded by the end of the experiment per group.**

**Table 1: Descriptive group statistics.**

| Dependent Variable | Group | N | Mean | SD | SE |
|---|---|---|---|---|---|
| Number of "edits" after intervention (**#EDITS**) | 1 | 10 | 1.000 | 0.816 | 0.258 |
| | 2 | 12 | 0.750 | 0.622 | 0.179 |
| Number of "post" after intervention (**#POSTS**) | 1 | 10 | 4.300 | 4.270 | 1.350 |
| | 2 | 12 | 3.500 | 3.778 | 1.091 |
| Total number of "shares" (**#SHARES**) | 1 | 10 | 1.700 | 2.263 | 0.716 |
| | 2 | 12 | 1.333 | 2.146 | 0.620 |
| Total number of "publications" (**#PUBLICATIONS**) | 1 | 10 | 6.000 | 4.190 | 1.325 |
| | 2 | 12 | 4.833 | 5.408 | 1.561 |
| Perceived Risk (**RSK**) | 1 | 10 | 4.025 | 1.003 | 0.317 |
| | 2 | 12 | 4.396 | 1.281 | 0.370 |
| Perceived Control (**CTRL**) | 1 | 10 | 4.667 | 1.432 | 0.453 |
| | 2 | 12 | 3.389 | 1.441 | 0.416 |
| Perceived Benefits (**BEN**) | 1 | 10 | 4.850 | 0.727 | 0.230 |
| | 2 | 12 | 5.313 | 0.765 | 0.221 |
| External Information Privacy Concerns (**EIPC**) | 1 | 10 | 2.900 | 1.233 | 0.390 |
| | 2 | 12 | 4.111 | 1.072 | 0.309 |

**Table 2: Independent samples t-Test.**

| Dependent Variable | t | d.f. | Sig. | Mean diff. | $SE_{DM}$ | 95% CI | Cohen's d |
|---|---|---|---|---|---|---|---|
| **#EDITS** | 0.816 | 20 | 0.424 | 0.250 | 0.307 | (-0.389,0.889) | 0.345 |
| **#POSTS** | 0.466 | 20 | 0.646 | 0.800 | 1.716 | (-2.779, 4.379) | 0.280 |
| **#SHARES** | 0.389 | 20 | 0.701 | 0.367 | 0.942 | (-1.598, 2.331) | 0.168 |
| **#PUBLICATIONS** | 0.556 | 20 | 0.584 | 1.167 | 2.097 | (-3.207, 5.541) | 0.241 |
| **RSK** | -0.744 | 20 | 0.466 | -0.371 | 0.499 | (-1.411, 0.669) | -0.322 |
| **CTRL** | 2.077 | 20 | 0.051 | 1.278 | 0.615 | (-0.006, 2.561) | 0.890 |
| **BEN** | -1.444 | 20 | 0.164 | -0.463 | 0.320 | (-1.131, 0.206) | -0.620 |
| **EIPC** | -2.466 | 20 | 0.023* | -1.211 | 0.491 | (-2.236, -0.187) | -1.048 |

Note: (*) The mean difference is significant for $\alpha = 5\%$.

## 6 DISCUSSION

The results of this preliminary study provide not only insights about the effects of risk-based interventions, but also on the benefits (and limitations) of ENAGRAM when evaluating preventative nudges. Based on our experience, we discuss the implications of the analysis presented in Section 5 along with some important aspects that should be taken into consideration when conducting studies using ENAGRAM. Particularly, with regard to (i) the instrumentation of self-disclosure metrics and constructs, and (ii) the technical benefits and drawbacks of the app.

### 6.1 Self-Disclosure Metrics and Constructs

As mentioned in subsection 5.1, four self-disclosure metrics were considered for this study: #EDITS, #POSTS, #SHARES, and #PUBLICATIONS. Nevertheless, several other metrics could be elaborated with the data collected through ENEGRAM. For instance, each of these metrics could be expressed per time unit (e.g., per day or per week) or in a relative way (e.g., #EDITS/#PUBLICATIONS or #EDITS/#POSTS). The upper limit of interventions (5 per day in this case) could also be leveraged for the elaboration of metrics. That is, by dividing the number of interventions a user received by the end of the experiment (i.e., #EDITS + #POSTS) over the maximum number of interventions the app can generate within the experimental period (7 days x 5 interventions/day = 35 interventions).

Due to the small amount of data collected in the 7 days of experiment, we decided to analyze the self-disclosure behavior of the study participants only through absolute metrics. We observed that participants in Group 1 interacted more with the app than the ones in Group 2 (Figure 4). Prior work has emphasized the role that risk cues play in the self-disclosure behavior of OSNs users (c.f., [20, 42]). Particularly, that users' perceived risk of information sharing is one of the most important factors influencing such a behavior. Hence, the risk information displayed on the second version of the app may have (i) lessened the sharing frequency of the study participants within that group, and (ii) increased their perception of privacy risks (#RSK) at the end of the experiment (i.e.,

at day 7). Still, none of these differences were found significant and thus should be further investigated and analyzed.

Differences in participants' perceived control (CTRL), benefits (BEN), and external information privacy concerns (EIPC) were also observed at the end of the study. Regarding the former, our results differ from the ones of Kroll and Stieglitz [27] who observed higher (though marginal) levels of perceived control on those users aware of the presence of privacy nudges in OSNs. However, the nudges tested in such a study did not render any risk information, which may be determining for users' perception of control [22]. On the other hand, a vast amount of literature has emphasized that higher levels of risk awareness can negatively impact the perceived benefits of online self-disclosure and increase users' privacy concerns [22, 25, 35]. Hence, a lower BEN and a higher EPIC among Group 2 participants can be also related to the presence of risk information in the corresponding app interventions. This is particularly interesting as EIPC encompasses *social* privacy concerns towards other users, especially with regard to organizational practices affecting other people's privacy. Such concerns play an important role in OSN platforms like Instagram since users can easily compromise the privacy of other individuals when sharing group pictures. Hence, it is to expect that those with higher EIPC would be more reluctant to share information or pictures portraying others [32]. Nevertheless, our results are still preliminary and call for additional research efforts.

## 6.2 ENAGRAM's Benefits and Drawbacks

*Extensibility.* Overall, the behavioral data collected through ENAGRAM helped us to gain insight in the self-disclosure practices of the study participants. In this particular case, we took the approach proposed by Díaz Ferreyra et al. [16] and embedded it into the app for its evaluation. However, other nudging strategies (e.g., social norms, pop-out policies, or defaults [12]) could also be easily implemented as many of ENAGRAM's building blocks can be customized with just a few lines of Java code. Such is the case of the time spent between interventions or the content displayed within them. For the latter, additional changes in the `EventsDB` may be necessary, particularly in the `interventions` table as it contains the text placed inside ENAGRAM's pop-up window. Other graphical elements displayed in this window (e.g., the legends "Ready to share?" and "Fact of the day #N") can also be adjusted and adapted to the specific needs of each intervention strategy. Furthermore, users' reactions to ENAGRAM's interventions could be leveraged to achieve personalization. That is, by regulating the frequency and the content of each warning according to the number of *edits* and *shares* performed by each user in a given time frame [16].

*Pending Features.* Some aspects and functionalities of ENAGRAM still require further development. Such is the case of the interventions displayed by the app which, at the moment, are not content-dependent. Hence, a user sharing a picture (or caption) about her drinking beer may not necessarily receive a warning message referring to alcohol consumption (e.g., "Other users had problems at work after posting about their alcohol consumption"). This issue could be addressed by integrating a machine learning solution capable of classifying the content being disclosed by the user (i.e., picture, caption, or both). There is prior and ongoing research in this realm that could be leveraged for this purpose (e.g., [9, 18, 45]) and even commercial platforms offering services for the automatic classification of multimedia content (e.g., Microsoft Azure Computer Vision[2] and Google Vision AI[3]). Hence, content-aware interventions could be (in principle) shaped by integrating such off-the-shelve solutions into ENAGRAM's architecture. Still, this may not be a straight forward task as the integration of third party software often demands changes and adaptations in the targeted architecture to overcome compatibility issues.

*Users' Privacy.* As shown in Section 5.1, the information collected by ENAGRAM (e.g, the post length and the hashed version of the image path) is useful to spot changes in participants' self-disclosure behavior while preserving their privacy. In principle, such information is enough to (i) identify changes in text or (ii) determine whether a picture has been replaced after an EDIT event. However, it is insufficient to determine whether such changes entail more or less information self-disclosure. For instance, a post like "I live in New York City" is shorter but more precise than another one saying "I live in the United States of America". Likewise, two different picture paths can tell us that both images are different but not if one is more or less sensitive than the other. Methods like the ones proposed in the previous point can address this issue by collecting metadata (e.g., text sentiment, named entities, or picture explicitness) from the content disclosed within ENAGRAM. That is, by pre-processing the posts (i.e., text and image) and storing the corresponding metadata in the EventsDB for later analysis. Thereby, the effects of ENAGRAM's interventions could be better assessed without having to record participants' raw data.

*Technical Issues.* Participants also had the chance to report any technical problem they may have experienced while using the app. Some of them mentioned that the caption was not directly transferred from ENAGRAM to Instagram when sharing their posts. We have also experienced this particular issue when testing the app ourselves, so we made this limitation explicit from the beginning (i.e., at the briefing). Still, some participants seem to have missed that point and thought it was an unexpected glitch in the software. Problems were also encountered when users attempted to create Instagram stories. Many of them said that their pictures were not properly forwarded to the Instagram app and ended up having multiple publications of the same kind. We did not experience such an issue ourselves during testing but it may be the reason why we observed duplicated records inside the `EventsDB`. In line with this, some participants reported delays after clicking on "SHARE!" or "POST" forcing them to click more than once. This may have also contributed to the duplication of entries inside the database and should be then addressed in future ENAGRAM releases.

# 7 STUDY LIMITATIONS

The adoption of crowdsourcing platforms has become widespread in privacy and security research as they facilitate (to a great extent) the recruitment of study participants and the collection of large amounts of empirical data. Prior work has shown that Prolific samples provide good quality data for conducting survey research on

[2] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/
[3] https://cloud.google.com/vision/

usable privacy and security [47]. Such is also the case for longitudinal experiments like ours carried out over several days or weeks [24]. Nevertheless, conducting out-of-the-lab experiments also entails a loss of control over participants, giving room to certain types of dishonest practices. For instance, some may claim to meet the eligibility criteria for taking part in the experiment when, in fact, they do not; or may even lack extrinsic motivation (i.e., due to the absence of peer pressure) for completing their tasks [19]. Moreover, because of the limited environmental control, online study subjects are prone to get distracted and thus compromise the quality of their answers.

To minimize the effect of these perils we introduced some quality controls throughout the experiment. In particular we included attention questions in all survey instruments and placed a registration code in the app that helped us to assess participants' engagement at the beginning of the study. As described in Section 4, we also included an *assessment* stage by the end of the experiment to exclude those who did not use ENAGRAM from completing the final survey. Such an assessment does not offer any guarantee as it relies on participants' self-reports. However, after inspecting the data collected in the `EventsDB`, we observed that subjects who reported having used the app did use it at least once. Hence, we believe that the control question introduced in the *assessment* stage is a good practice in this type of experiments as it can help in the early detection of loose participants.

The pre-screening of study subjects is often regarded as a best practice when conducting online studies [41]. Hence, we used Prolific's built-in qualification features to recruit participants based on their gender (i.e., to ensure balance within study groups), social media usage (Instagram), and language skills (German). In addition, we targeted users who already took part in at least 10 other studies as these are usually more committed and less likely to drop from experiments [41]. Last but not least, we also tried to keep the length of both the individual surveys and the full experiment as short as possible to reduce participants' fatigue and minimize the chances of attrition.

As mentioned in Section 6, the results of this paper are preliminary and aim to give a first impression of ENAGRAM's evaluation features. Still, limitations related to the size and composition of the study sample should be acknowledged. Particularly, we have analyzed a relatively small sample composed exclusively of German-speaking participants. All in all, this means that the results yielded in our study are not representative of the population under analysis. Furthermore, small- to medium-size effects cannot be reliably identified in such a sample and call for future studies with a larger number of participants. One should also note that German-speaking countries (e.g., Germany, Austria, and Switzerland) are typically Western, Educated, Industrialized, Rich, and Democratic (WEIRD) nations [23]. Hence, our sample (and so the study results) are not representative of other populations with different socioeconomic and demographic characteristics. Future work should not only seek to investigate the effects of preventative nudges on larger samples, but also take into account that non-WEIRD individuals (as pointed by Dev et al. [15]) may exhibit different privacy behaviors and concerns. Moreover, prior research has also stressed-out that most studies on digital privacy are based typically on WEIRD samples providing a very narrow view on these matters [5]. Therefore, there is a call for cross-cultural studies that help us better understand the self-disclosure practices, preferences, and concerns across WEIRD and non-WEIRD populations.

## 8 CONCLUSION

Gathering behavioral evidence on the effectiveness of preventative nudges has become a struggle for privacy and security researchers. The integration barriers imposed by commercial OSNs have limited (to a large extent) the type and amount of empirical data available within the current literature. In turn, nudges targeting online self-disclosure are often evaluated through mock-ups and self-reports but not under realistic conditions. ENAGRAM seeks to aid ongoing investigations on the performance of such nudges by providing a more adequate evaluation environment suitable for conducting longitudinal experiments. The results gathered in our 7-day study show that the app could be leveraged for embedding different nudging strategies and collect insights about their effects on peoples' privacy behavior. Furthermore, the collected data can be aggregated into different performance metrics that, despite being related to the content disclosed by the users, are computed in a privacy-friendly way. That is due to the fact that ENAGRAM only stores hashed versions of the pictures and captions users share, which is adequate to identify behavior changes linked to the presence of nudges (as shown in Section 5.1).

Despite the excitement that nudges arise among privacy researchers, many still see them as a threat to people's autonomy [46]. That is because nudges often leverage well-known behavioral biases and heuristics to persuade humans toward wiser decisions [54]. Because of the fine line existing between persuasion, manipulation, and coercion, it is not surprising that many have raised concerns over potential unethical uses of nudges. Hence, it is essential to analyze the ethical implications of the approaches tested with ENAGRAM to ensure they do not jeopardize the agency and autonomy of study subjects. Renaud and Zimmermann [39] have outlined a set of ethical guidelines applicable to the design of nudges in the context of cybersecurity. We strongly advise those using ENAGRAM as an evaluation framework to adopt these (and other guidelines alike) to ensure their solutions meet ethical requirements from the very beginning.

Throughout this study, we have identified several areas of improvement and directions for future work. One is related to the feedback we received from the participants and the quality of the collected data. As mentioned in Section 6.2, some glitches in the current version of ENAGRAM are hindering its usage and may be causing duplicated entries in the `EventsDB`. Hence, we plan to have a closer look into these issues and apply the corresponding fixes to improve the overall performance of the app. On the other hand, we also aim at embedding off-the shelve machine learning solutions to support the generation of context-aware interventions. This would not only enhance the overall user experience of ENAGRAM but also open new opportunities for the evaluation of preventative nudges. Besides, an assessment with a larger and diverse sample should be conducted over a longer period of time (e.g., 4 weeks) in order to yield more significant and representative results. Particularly, especial attention shall be draw into the cultural differences between

WEIRD and non-WEIRD communities as these may considerably impact the effectiveness of preventative nudges in the practice.

Adapting nudges' content and frequency to the individual privacy goals and expectations of each user is still an ongoing research challenge [8, 12, 36]. The current version of ENAGRAM follows a one-size-fits-all approach in this regard delivering a maximum of 5 interventions per day with a minimum lapse of 60 min. between them. It would be interesting to tailor these parameters to the particular requirements of each user to increase the effectiveness of ENAGRAM's interventions (i.e., of the nudging solution under evaluation). Hence, we also plan to conduct further empirical studies with ENAGRAM to understand the role that frequency and content play in the acceptance of different nudging strategies. Particularly, to determine maximum and minimum intervention thresholds along with personalized strategies for adjusting the content of behavioral interventions.

## ACKNOWLEDGMENTS

This work was partly supported by Canada's Natural Sciences and Engineering Research Council (NSERC).

## SUPPLEMENTARY MATERIAL

Survey instruments, ENAGRAM software, and study artifacts are available at https://doi.org/10.5281/zenodo.6974704

## APPENDIX: EMPLOYED CONSTRUCTS

The reliability of the employed scales was assessed through the Cronbach's Alpha coefficient. In all cases the coefficient was higher than 0.70, which suggests that the items of each construct scale have a relatively high internal consistency (values higher than 0.7 are usually considered "acceptable"). As mentioned in Section 4, all constructs were originally introduced by Krasnova et al. [26], except for External Information Privacy Concerns (EIPC) which was elaborated by Morlok [32]. It should be noted that the Perceived Benefits (BEN) construct encompasses *Convenience* (CON), *Relationship Building* (RB), *Self-Representation* (SR), and *Enjoyment* (EN) (we have aggregated these benefits into a single BEN score).

**Perceived Benefits (BEN)**

- CON1: OSNs are convenient to inform all my friends about my ongoing activities.
- CON2: OSNs allow me to save time when I want to share something new with my friends.
- CON3: I find OSNs efficient in sharing information with my friends.
- RB1: Through OSNs I get connected to new people who share my interests.
- RB2: OSNs helps me to expand my network.
- RB3: I get to know new people through OSNs.
- SR1: I try to make a good impression on others on OSNs.
- SR2: I try to present myself in a favorable way on OSNs.
- EN1: When I am bored I often log-in to OSNs.
- EN2: I find OSNs entertaining.
- EN3: I spend enjoyable and relaxing time on OSNs.

**Perceived Privacy Risks (RSK)**

- RSK1: Overall, I see no real threat to my privacy due to my presence on the OSN (*Reversed*).
- RSK2: I feel safe publishing my personal information on the OSN (*Reversed*).
- RSK3: Please rate your overall perception of privacy risk involved when using the OSN (*very safe – very risky*).

**Perceived Control (CTRL)**

- PC1: I feel in control over the information I provide on the OSN.
- PC2: Privacy settings allow me to have full control over the information I provide on the OSN.
- PC3: I feel in control of who can view my information on the OSN.

**External Information Privacy Concerns (EIPC)**

- EIPC1: It usually bothers me to share pictures of my friends on OSNs.
- EIPC2: I am concerned that OSNs collect too many pictures of my friends.
- EIPC3: I am concerned that unauthorized people may access the pictures of my friends that I shared on OSNs.
- EIPC4: I am concerned that the pictures I share on OSNs may be kept in a non-accurate manner.
- EIPC5: I am concerned that OSNs may use the pictures of my friends I shared for other purposes without notifying me or getting my authorization.
- EIPC6: I am concerned that OSNs may sell friends' pictures I shared to other companies.

## REFERENCES

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 44.

[2] Esma Aïmeur, Hicham Hage, and Sabrine Amri. 2018. The Scourge of Online Deception in Social Networks. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 1266–1271.

[3] Samar Albladi and George R. S. Weir. 2016. Vulnerability to Social Engineering in Social Networks: A Proposed User-Centric Framework. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*. IEEE, 1–6.

[4] Jose Alemany, Victor Botti-Cebriá, Elena del Val, and Ana García-Fornes. 2022. Detection and Nudge-Intervention on Sensitive Information in Social Networks. *Logic Journal of the IGPL* (February 2022). https://doi.org/10.1093/jigpal/jzac004

[5] Frode Alfnes and Ole Christian Wasenden. 2022. Your privacy for a discount? Exploring the willingness to share personal data for personalized offers. *Telecommunications Policy* 46, 7 (2022), 102308.

[6] Anas Azzouz and Samuel Sambasivam. 2019. Strategies Needed by Social Network Builders to Develop Information Privacy. In *Proceedings of the Conference on Information Systems Applied Research ISSN*, Vol. 2167. 1508.

[7] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. 2020. A Nudge-based Recommender System Towards Responsible Online Socializing. In *Workshop on Online Misinformation- and Harm-Aware Recommender Systems (OHARS '20 (CEUR Workshop Proceedings, 2758)*, Daniela Godoy Antonela Tommasel and Arkaitz Zubiaga (Eds.). Aachen, 23–39. http://ceur-ws.org/Vol-2758/OHARS-paper2.pdf

[8] Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. 2022. The Digital Landscape of Nudging: A Systematic Literature Review of Empirical Research on Digital Nudges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 62, 16 pages. https://doi.org/10.1145/3491102.3517638

[9] Víctor Botti-Cebriá, Elena del Val, and Ana García-Fornes. 2021. Automatic Detection of Sensitive Information in Educative Social Networks. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, Álvaro Herrero, Carlos Cambra, Daniel Urda, Javier Sedano, Héctor

Quintián, and Emilio Corchado (Eds.). Springer International Publishing, Cham, 184–194.

[10] Vanessa Bracamonte, Welderufael Tesfay, and Shinsaku Kiyomoto. 2021. Towards Exploring User Perception of a Privacy Sensitive Information Detection Tool. In *Proceedings of the 7th International Conference on Information Systems Security and Privacy (ICISSP)*. INSTICC, SciTePress, 628–634. https://doi.org/10.5220/0010319706280634

[11] Matthias Brand. 2022. Can internet use become addictive? *Science* 376, 6595 (2022), 798–799.

[12] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[13] Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

[14] Sourya Joyee De and Abdessamad Imine. 2019. On Consent in Online Social Networks: Privacy Impacts and Research Directions (Short Paper). In *Risks and Security of Internet and Systems*, Akka Zemmari, Mohamed Mosbah, Nora Cuppens-Boulahia, and Frédéric Cuppens (Eds.). Springer International Publishing, 128–135.

[15] Jayati Dev, Pablo Moriano, and L Jean Camp. 2020. Lessons Learnt from Comparing WhatsApp Privacy Concerns Across Saudi and Indian Populations. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. 81–97.

[16] Nicolás E. Díaz Ferreyra, Tobias Kroll, Esma Aïmeur, Stefan Stieglitz, and Maritta Heisel. 2020. Preventative Nudges: Introducing Risk Cues for Supporting Online Self-Disclosure Decisions. *Information* 11, 8 (2020), 399. https://doi.org/10.3390/info11080399

[17] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2011. Connection strategies: Social capital implications of Facebook-enabled communication practices. *New media & society* 13, 6 (2011), 873–892.

[18] Bruce Ferwerda and Marko Tkalcic. 2018. You are what you post: What the content of Instagram pictures tells about users' personality. In *The 23rd International on Intelligent User Interfaces, March 7-11, Tokyo, Japan*. CEUR-WS.

[19] Nathan Gagné and Léon Franzen. 2021. How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience. (2021). https://doi.org/10.31234/osf.io/nt67j

[20] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. 2019. Investigating People's Privacy Risk Perception. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 267–288.

[21] José Luis Gómez-Barroso. 2018. Experiments on personal information disclosure: Past and future avenues. *Telematics and Informatics* 35, 5 (2018), 1473–1490.

[22] Nick Hajli and Xiaolin Lin. 2016. Exploring the security of information sharing on social networking sites: The role of perceived control of information. *Journal of Business Ethics* 133, 1 (2016), 111–123.

[23] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences* 33, 2-3 (2010), 61–83.

[24] Emily J. Kothe and Mathew Ling. 2019. Retention of participants recruited to a one-year longitudinal study via Prolific. (2019). https://doi.org/10.31234/osf.io/5yv2u

[25] Nicole C Krämer and Johanna Schäwel. 2020. Mastering the challenge of balancing self-disclosure and privacy in social media. *Current Opinion in Psychology* 31 (February 2020).

[26] Hanna Krasnova, Oliver Günther, Sarah Spiekermann, and Ksenia Koroleva. 2009. Privacy concerns and identity in online social networks. *Identity in the Information Society* 2, 1 (01 Dec 2009), 39–63.

[27] Tobias Kroll and Stefan Stieglitz. 2021. Digital nudging and privacy: improving decisions about self-disclosure in social networks. *Behaviour & Information Technology* 40, 1 (2021), 1–19.

[28] Yiling Lin, Magda Osman, and Richard Ashcroft. 2017. Nudge Concept, Effectiveness, and Ethics. *Basic and Applied Social Psychology* 39, 6 (2017), 293–306.

[29] Vincent Marmion, Felicity Bishop, David E. Millard, and Sarah V. Stevenage. 2017. The Cognitive Heuristics Behind Disclosure Decisions. In *International Conference on Social Informatics*. Springer, 591–607.

[30] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring nudge designs to help adolescent sns users avoid privacy and safety threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[31] Gaurav Misra and Jose M Such. 2017. PACMAN: Personal Agent for Access Control in Social Media. *IEEE Internet Computing* 21, 6 (2017), 18–26.

[32] Tina Morlok. 2016. Sharing is not Caring: The Role of External Privacy in Users' Information Disclosure Behaviors on Social Network Sites. In *Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS 2016)*. AIESeL.

[33] Francesca Mosca, Stefan Sarkadi, Jose M. Such, and Peter McBurney. 2020. Agent EXPRI: Licence to Explain. In *2nd International Workshop on Explainable Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*.

[34] Sina Ostendorf, Yannic Meier, and Matthias Brand. 2022. Self-disclosure on social networks - More than a rational decision-making process. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* (2022). in press.

[35] Sina Ostendorf, Silke M Müller, and Matthias Brand. 2020. Neglecting Long-Term Risks: Self-Disclosure on Social Media and Its Relation to Individual Decision-Making Tendencies and Problematic Social-Networks-Use. *Frontiers in Psychology* 11 (2020), 543388. https://doi.org/10.3389/fpsyg.2020.543388

[36] Eyal Peer, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. 2020. Nudge me right: Personalizing online security nudges to people's decision-making styles. *Computers in Human Behavior* 109 (2020), 106347.

[37] Janice Penni. 2017. The Future of Online Social Networks (OSN): A Measurement Analysis Using Social Media Tools and Application. *Telematics and Informatics* 34, 5 (2017), 498–517.

[38] Frederic Raber, Alexander De Luca, and Moritz Graus. 2016. Privacy Wedges: Area-Based Audience Selection for Social Network Posts. In *12th Symposium on Usable Privacy and Security (SOUPS 2016)*.

[39] Karen Renaud and Verena Zimmermann. 2018. Ethical Guidelines for Nudging in Information Security & Privacy. *International Journal of Human-Computer Studies* 120 (2018), 22–35. https://doi.org/10.1016/j.ijhcs.2018.05.011

[40] Lambèr Royakkers, Jelte Timmer, Linda Kool, and Rinie van Est. 2018. Societal and ethical issues of digitization. *Ethics and Information Technology* 20, 2 (2018), 127–142.

[41] Joni Salminen, Soon-gyo Jung, and Bernard J Jansen. 2021. Suggestions for Online User Studies. In *International Conference on Human-Computer Interaction*. Springer, 127–146.

[42] Sonam Samat and Alessandro Acquisti. 2017. Format vs. Content: The Impact of Risk and Presentation on Disclosure Decisions. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, 377–384.

[43] David Sánchez, Josep Domingo-Ferrer, and Sergio Martínez. 2018. Co-utile Disclosure of Private Data in Social Networks. *Information Sciences* 441 (2018), 50–65.

[44] Sofia Schöbel, Torben Barev, Andreas Janson, Felix Hupfeld, and Jan Marco Leimeister. 2020. Understanding user preferences of digital privacy nudges–a best-worst scaling Approach. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[45] Junho Song, Kyungsik Han, Dongwon Lee, and Sang-Wook Kim. 2018. "Is a picture really worth a thousand words?": A case study on classifying user attributes on Instagram. *PloS one* 13, 10 (2018), e0204938.

[46] Cass R. Sunstein. 2018. Misconceptions about nudges. *Journal of Behavioral Economics for Policy* 2, 1 (2018), 61–67.

[47] Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. *arXiv preprint arXiv:2202.14036* (2022).

[48] Arnout Terpstra, Alexander P Schouten, Alwin de Rooij, and Ronald E Leenes. 2019. Improving privacy choice through design: How designing for reflection could support privacy self-management. *First Monday* 24, 7 (June 2019).

[49] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, New Haven, CT and London.

[50] Jessica Vitak. 2012. The Impact of Context Collapse and Privacy on Social Network Site Disclosures. *Journal of Broadcasting & Electronic Media* 56, 4 (2012), 451–470.

[51] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web*. 763–770.

[52] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the 7th Symposium on Usable Privacy and Security, SOUPS 2011*. ACM, 1–16.

[53] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling Self-Disclosure in Social Networking Sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 74–85. https://doi.org/10.1145/2818048.2820010

[54] Verena Zimmermann and Karen Renaud. 2021. The nudge puzzle: matching nudge interventions to cybersecurity decisions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 1 (2021), 1–45.