information

MDPI

Article

Preventative Nudges: Introducing Risk Cues for Supporting Online Self-Disclosure Decisions

Nicolás E. Díaz Ferreyra [1,*], Tobias Kroll [1], Esma Aïmeur [2], Stefan Stieglitz [1] and Maritta Heisel [1]

[1] Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, 47057 Duisburg, Germany; tobias.kroll@uni-due.de (T.K.); stefan.stieglitz@uni-due.de (S.S.); maritta.heisel@uni-due.de (M.H.)

[2] Department of Computer Science and Operations Research, University of Montréal, PO Box 6128, Montréal, QC H3C 3J7, Canada; aimeur@iro.umontreal.ca

* Correspondence: nicolas.diaz-ferreyra@uni-due.de; Tel.: +49-203-379-1075

Received: 7 August 2020; Accepted: 16 August 2020; Published: 18 August 2020

check for updates

Abstract: Like in the real world, perceptions of risk can influence the behavior and decisions that people make in online platforms. Users of Social Network Sites (SNSs) like Facebook make continuous decisions about their privacy since these are spaces designed to share private information with large and diverse audiences. In particular, deciding whether or not to disclose such information will depend largely on each individual's ability to assess the corresponding privacy risks. However, SNSs often lack awareness instruments that inform users about the consequences of unrestrained self-disclosure practices. Such an absence of risk information can lead to poor assessments and, consequently, undermine users' privacy behavior. This work elaborates on the use of risk scenarios as a strategy for promoting safer privacy decisions in SNSs. In particular, we investigate, through an online survey, the effects of communicating those risks associated with online self-disclosure. Furthermore, we analyze the users' perceived severity of privacy threats and its importance for the definition of personalized risk awareness mechanisms. Based on our findings, we introduce the design of preventative nudges as an approach for providing individual privacy support and guidance in SNSs.

Keywords: adaptive privacy; digital nudging; social network sites; risk perception; human-computer interaction; artificial intelligence

1. Introduction

Risk is a standing component of everyday life since there is always some uncertainty associated with the outcome of people's decisions. Moreover, whether consciously or unconsciously, people often assess the potential consequences of their actions guided by their perception of risk [1]. In particular, an individual's sense of risk is likely to influence the evaluation of available choices, and therefore have a certain impact on her final decision [2–4]. Nonetheless, humans' cognitive capacity is limited and cannot take into consideration many risk factors [5]. Given this limitation, providing risk information is key for helping people to avoid misjudgments, unseemly behavior, and, ultimately, to safeguard them from unwanted incidents. Hence, risk information should not only be *accessible* but also *explicit* and *adequate* to effectively enhance individuals' sense of awareness [6,7].

Deciding whether or not to disclose personal information to others is a daily exercise that is essential for establishing and maintaining social relationships [8]. Furthermore, it is a practice that is done under uncertainty conditions related to its potential benefits and privacy costs [9]. With the introduction of Social Network Sites (SNSs) like Twitter or Instagram, privacy decisions became more frequent since these are spaces in which people constantly share information with large and diverse audiences. In particular, people are liable to reveal more personal information in SNSs than what they normally would in a traditional offline context [10,11]. This makes SNSs a gateway for accessing large amounts of sensitive data and, consequently, a target for social engineering attacks [12,13]. However, average users find it hard to properly estimate the risks associated with their disclosures and, in turn, often become victims of privacy threats such as scamming, grooming, or identity theft. In addition, SNSs hinder individual's decisions related to self-presentation since their affordances place different audiences (e.g., work colleagues and family) in the same communication plane [14]. Consequently, users frequently experience regret—along with unwanted incidents—after having shared personal information with an unintended audience [15].

Despite its importance, rational estimations of self-disclosure risks and benefits are frequently replaced by heuristic judgements related to trust and credibility [16,17]. For instance, users tend to reveal more or less information about themselves according to their individual perception of the platform's legitimacy and trustworthiness as a data controller [17]. Such an assessment is normally based on cues related to the platform's reputation (e.g., its size) or recognition (e.g., its market presence), among others. This, on one hand, tends to simplify complex self-disclosure decisions. However, it also undermines people's privacy-preserving behavior since SNSs portray many trust-related cues, yet scarce risk information [17]. Furthermore, privacy policies are also devoid of risk cues which, in turn, hinder users' decisions related to consent on data processing activities [18]. In consequence, even users who attempt a considered approach to self-disclosure lack adequate means for conducting a well-grounded privacy calculus [19].

1.1. Motivation

Over recent years, privacy scholars have introduced a wide range of technological approaches that aim to improve people's online privacy decisions [20–22]. In particular, the use of *nudges* has gained popularity due to their capacity for assisting and guiding individuals towards safer privacy practices [23]. At their core, nudges are interventions that encourage a certain behavior which, in turn, tends to maximize people's welfare. Such interventions are the means for behavioral change and consist of small modifications in the context within which decisions are made [24]. For instance, displaying cues related to the targeted audience of a post can motivate users to employ custom friend lists. Given the close relation existing between risk perception and privacy behavior, it is not surprising that interventions portraying risk information are deemed adequate for motivating safer self-disclosure decisions in SNSs [17,19,25]. In particular, such interventions can prevent users from sharing posts with personal data by rendering information about the risks of unsafe self-disclosure practices [26].

One approach to risk-based interventions is through the elaboration of self-disclosure scenarios or patterns. Basically, these are descriptions of privacy harms that may occur when certain pieces of personal data are revealed to untrusted audiences in SNSs [26–29]. For instance, a scenario describing *"Revealing bank account details can increase the chances of financial fraud"* can be leveraged for motivating a user to keep her financial information away from public disclosure. Nevertheless, risky events are often perceived differently across individuals. Consequently, a certain incident such as *financial fraud* may be assessed as "catastrophic" by one user and "minor" by another one. Therefore, understanding the nuances across users' perception of self-disclosure patterns becomes crucial for the design of effective nudging applications. Nonetheless, the preventative effects of these patterns and the perceived severity of the incidents they portray have not been extensively explored and documented within the current literature. Hence, further insights should be gained regarding these aspects to develop successful risk-based interventions and nudges.

1.2. Contribution

This work elaborates on the use of risk cues as a strategy for supporting self-disclosure decisions in SNSs. In particular, it investigates the nudging effects of risk-based interventions through an online survey conducted via Amazon Mechanical Turk (N = 281). Overall, the contributions of this study are two-fold. First, it provides insights into the preventative effects of risk-based interventions and the perceived severity of certain patterns of information disclosure. In line with prior work, our findings suggest that individuals are less willing to reveal private information when they are aware of the negative consequences it may bring to them. On the other hand, the paper investigates the role that the perceived severity of privacy threats has for the development of adaptive nudging mechanisms. In particular, recent findings suggest that nudges can be more effective if they are tailored to the individual privacy goals and expectations of each user [30,31]. Based on this premise, this work introduces an approach that leverages the nuances in the perceived severity of unwanted incidents for adapting the frequency and content of self-disclosure interventions.

The rest of this paper is organized as follows. In the next section, related work in online self-disclosure and preventative nudges is discussed and analyzed. Following, Section 3 introduces the theoretical foundations of this paper. In particular, the notion of Self-disclosure Patterns (SDPs) is presented and discussed along with its relevance for the design of preventative nudging solutions. Sections 4 and 5 introduce the design of our online survey and its findings, respectively. Next, in Section 6, we address the elaboration of personalized interventions taking into consideration the insights obtained from the survey. The strengths and limitations of this approach are analyzed in Section 7, whereas the ones of the survey method are discussed in Section 8. Finally, in Section 9, we outline the conclusions of this paper and introduce directions for future work.

2. Related Work

Privacy scholars have studied self-disclosure practices in SNSs through the lens of multiple theories and disciplines. Moreover, such studies have served as foundations for the development of several preventative technologies and the application of persuasive mechanisms for behavioral change. This section discusses findings related to the use of risk cues in state-of-the-art solutions. In line with this, we analyze related work that elaborates on the importance of such cues for promoting safer self-disclosure decisions in SNSs.

2.1. Self-Disclosure Behavior in SNSs

Since the advent of SNSs, much effort has been dedicated to understanding the factors behind self-disclosure decisions in online platforms. One of the earliest contributions in the field of psychology is the so-called *privacy paradox* which describes an offset phenomenon between peoples' privacy concerns and behavior [32]. In particular, such paradox manifests when individuals who claim to be highly concerned about their privacy end up revealing personal information in SNSs [33]. Due to its high impact, the privacy paradox has paved the way for subsequent studies and the development of new theories about self-disclosure behavior. Among them, the *privacy calculus* has become one of the most plausible theories in privacy research [19]. Basically, the calculus posits that self-disclosure decisions are the result of an estimation of the potential risks and gratifications of revealing private information to others [34–36]. However, despite its novelty, it has been argued that estimations about the consequences of a self-disclosure act can be affected by personal characteristics, emotions, or missing knowledge [23]. Hence, average users often fail on conducting a rational assessment of their privacy behavior due to optimistic biases or false estimations [19].

To reduce the cognitive effort of privacy decision-making, users often apply *cognitive heuristics* instead of a rigorous uncertainty calculus [17,37]. Basically, heuristics are mental shortcuts that allow people to make quick judgements and reduce the complexity of their decisions. Overall, self-disclosure heuristics are triggered by cues that are tied to the different affordances of SNSs [16]. Such heuristics can be classified into *positive* or *negative* depending on whether they foster self-disclosure behavior or not [38]. For instance, *bandwagon* is a positive heuristic which is often triggered by the networking affordances of SNSs, and consists of disclosing more personal information as others are seen doing it so [37]. Conversely, *expectancy violation* is a negative heuristic which consists of diminishing the amount of self-disclosure if the credibility of the platform is perceived as low [39], i.e., if the platform exhibits an inferior design, bad navigability, or poor visual appearance. Overall, self-disclosure heuristics lead to snap judgements and reduce the complexity of self-disclosure decisions. However, SNSs portray many cues associated with positive heuristics, yet scarce cues that would ease the application of negative heuristics [17]. In consequence, privacy-preserving behavior is often impaired by heuristic judgements which, instead of discouraging, tend to foster self-disclosure processes.

2.2. Preventative Nudges

In recent years, several efforts have been made to understand and assist people's privacy decisions in SNSs. In particular, scholars have introduced a wide range of technical solutions for countering the biases and cognitive limitations associated with privacy decision-making [23]. Many of these technologies are grounded on findings in the area of behavioral economics and, to a large extent, on soft paternalistic principles. Among them, the *nudge* theory by Richard Thaler and Cass Sunstein has been widely adopted among researchers, and its application closely explored and documented [40]. Essentially, nudges represent small changes introduced in a *choice architecture* (i.e., the context in which decisions are made) with the purpose of influencing people's behavior in a predictable way. In particular, nudges aim to promote decisions that are considered beneficial for the users to maximize their welfare.

Overall, nudges can be applied for supporting human decisions across different domains and scenarios including online shopping [41], education [42], and health [43]. However, a growing body of literature has focused on developing *preventative nudges* which aim to guide individuals towards safer cybersecurity practices [23]. For example, Wang et al. [15] designed three nudges for Facebook consisting of (i) introducing a 30 seconds delay before a message is posted, (ii) displaying visual cues related to the post's audience, and (iii) showing information about the sentiment of the post. These interventions gave users the chance to re-consider the information disclosed inside their posts and reflect on the potential privacy threats. However, under this and other approaches alike, the forecast of threats and the estimation of risks associated with them remains a task that the users must conduct on their own.

To prevent people from making misjudgments about the consequences of their privacy decisions, some nudging solutions incorporate explicit risk information to their design. For instance, De et al. [44] introduced a nudge that informs about the privacy risks associated with the use of lax privacy settings in SNSs (e.g., the risks of having a public profile). Their approach combines empirical evidence and attack trees to estimate the risk associated with particular combinations of settings. In line with this, Díaz Ferreyra et al. [28] elaborated on risk scenarios that can be used for nudging textual publications in SNSs. Such scenarios consist of empirical risk evidence which is captured in patterns of information disclosure and used for the generation of personalized interventions. These interventions are warning messages describing unwanted incidents that may occur when particular pieces of private information reach the hands of untrusted audiences in online platforms. Nonetheless, empirical evidence supporting the effectiveness of these scenarios for nudging online self-disclosure behavior has not been documented so far in the literature.

3. Theoretical Background

As discussed in Section 2, risk cues play a key role in people's online self-disclosure decisions. Hence, preventative nudges should, in principle, elaborate on behavioral interventions that inform about the privacy risks that may occur when revealing private information in SNSs. Under this premise, this work explores the use of self-disclosure patterns for the generation of such interventions and their effect on users' information-sharing behavior. Furthermore, it investigates the role that the perceived severity of privacy risks has for the personalization of preventative nudges. In the following section, the theoretical foundations of this work are introduced and discussed.

3.1. Self-Disclosure Patterns

Often, self-disclosure practices across SNSs derive in unwanted incidents (e.g., identity theft, financial fraud, or harassment) after certain pieces of personal information reach an untrusted audience. For instance, it is likely that the chances of suffering from financial fraud increase if we disclose our credit card number inside a public post. Likewise, the probabilities of experiencing harassment become higher as we share our current location along with a textual publication. All in all, these are examples of events that repeat themselves over time and can be represented as Self-disclosure Patterns (SDP) [45]. In particular, a SDP can be modelled as a triple <*PAs, Audience, UIN*> where *PAs* corresponds to a set of private attributes, *Audience* to a collection of recipients (e.g., Facebook friends), and *UIN* to an *Unwanted Incident*. Under this approach, a situation in which a person gets harassed after revealing her phone number in a public post can be modelled as a SDP consisting of <*phone number, public, harassment*>.

Essentially, SDPs are abstractions of scenarios in which *UINs* take place after revealing a set of *PAs* to an untrusted *Audience*. Overall, a large body of SDPs can be extracted from the negative experiences reported by SNSs users across the literature [45]. Furthermore, it has also been suggested that SDPs could be retrieved from the content people delete from their profiles through the application of machine learning techniques [28]. Nonetheless, SDPs not only offer the possibility of representing risky self-disclosure scenarios but can also support the generation of behavioral interventions. In particular, preventative nudges could intervene whenever a user attempts to share a post with the information specified in an SDP. For this, the *Audience* and *UIN* of such SDP could be used to elaborate a warning message describing the negative consequences that may occur if the post reaches a group of untrusted recipients. For example, a message like *"Revealing a phone number in a public post may lead to situations of harassment. Do you want some hints for protecting your privacy?"* could be generated when a user is about to reveal her phone number in a public post. Such an approach introduced by Díaz Ferreyra et al. [28] aims to promote safer self-disclosure decisions by introducing risk cues in behavioral interventions. In Section 5.2, the nudging effects of SDPs are investigated and analyzed empirically.

3.2. Personalized Risk Awareness

In general, the instances of preventative nudges described in the current literature resemble a "one-size-fits-all" persuasive design, i.e., they apply the same behavioral intervention to a large group of people without acknowledging their individual goals, expectations, and nuances across their personalities. However, there is an increasing demand for personalized nudges that address individual differences in privacy decision-making and regulate their interventions, accordingly. Prior work has elaborated on adaptive approaches that capture differences across people's personality traits and privacy attitudes [29,30,46]. Nevertheless, a recent study by Warberg et al. [31] suggests that in general, personality-based interventions do not introduce significant changes in peoples' privacy decisions. On the other hand, a growing body of literature began to put emphasis on the role of risk and the perceived severity of unwanted incidents for the elaboration of personalized nudging solutions [21,26,44]. In line with this, Díaz Ferreyra et al. [28] introduced a risk-based strategy for adapting the frequency in which warning messages are displayed to the users. That strategy takes into account (i) the number of times a user accepts or rejects a warning, and (ii) the perceived severity of the *UIN* communicated by a warning. Although the former aims to address the individual privacy goals of each user, the latter seeks to acknowledge the subjective perception of SDPs. In particular, an *UIN* represented by an SDP can be perceived as insignificant by one user and catastrophic by others. Hence, such perception is crucial for the adaptation of risk-based interventions. Nevertheless, nuances in the perceived severity of SDPs have not been yet investigated or supported by empirical evidence. Therefore, the study introduced in the next section seeks to shed light on the subjective perception of SDPs. Furthermore, it set the basis for the elaboration of an approach that unlike the one described above, does not require a fine-grained analysis of user content when generating the corresponding intervention.

4. Method

All in all, the performance of privacy decisions depends largely on the availability of risk cues. Therefore, nudges should inform about the risks of unsafe self-disclosure practices to promote a preventative behavior among the users of SNSs. As mentioned earlier, this can be achieved by shaping interventions out of SDPs describing the risks of sharing personal data with untrusted audiences. Hence, we conducted an online survey to evaluate the effectiveness of this strategy and analyze the impact of risk-based interventions in online self-disclosure decisions. Moreover, the study explores the perceived severity of unwanted incidents and its importance for the development of personalized nudging solutions. This last point is further elaborated in Section 6.

4.1. Survey Design

To evaluate the effectiveness of interventions generated out of SDPs we followed a simple three-step, before-and-after design. First, a set of response variables related to online self-disclosure behavior were measured using well-established constructs and scales. In particular, questions related to *(i) self-disclosure*, *(ii) perceived control*, *(iii) trust in other platform members*, *(iv) trust in the SNS provider*, and *(v) perceived privacy risk* were asked to the participants at the beginning of the survey. Next, participants were requested to rate the severity of the unwanted incidents described in a set of self-disclosure scenarios. Such scenarios represented the type of behavioral interventions that can be elaborated out of the information contained in SDPs. Finally, the response variables were measured again, and their values compared against the ones obtained prior to the scenario assessment task. For this, a paired sample *t*-Test was conducted to determine whether exposing participants to risk-based interventions can significantly modify their immediate information-sharing behavior.

For assessing the proposed scenarios, participants were asked to rate the corresponding unwanted incidents as *insignificant*, *minor*, *moderate*, *major* or *catastrophic* based on their perceived severity. For instance, given the scenario *"You post a picture of you drunk at a party. You feel embarrassed after your work colleagues forward the picture to your boss"*, participants were requested to evaluate the severity of the incident "feeling embarrassed". In addition, participants were also asked whether they had experienced the scenarios themselves and, if so, if they deleted the corresponding post afterwards. The complete list of scenarios and a detailed description of the instruments employed in the survey can be found in Appendix D. As shown in Table 1, a total of 26 scenarios were defined and grouped around 6 information categories:

i *Drugs and alcohol use:* These scenarios correspond to situations in which people may suffer unwanted incidents after posting information related to their consumption habits of alcohol or drugs.

ii *Sex:* Scenarios defined under this category represent cases where people are liable to experience negative consequences after sharing details about their sexual life in SNSs.

iii *Religion and politics:* These scenarios describe negative consequences that may occur when sharing a political statement or disclosing one's religious affiliation in online platforms.

iv *Strong sentiment:* This category groups together scenarios in which unwanted incidents can take place as a result of sharing content with a strong or negative sentiment.

v *Location:* These scenarios describe unwanted incidents that are likely to occur when people reveal their current location or places they frequently visit inside their posts.

vi *Personal identifiers:* Scenarios defined under this category portray situations in which negative consequences can occur after sharing information containing personal identifiers such as one's credit card or social security numbers.

Both scenarios and their respective categories were derived from prior studies that analyze the type of information people usually regret having shared in SNSs [15,47,48]. Furthermore, many of them are considered categories of personal data under Article 9 of the European Union's General Data Protection Regulation (GDPR) [49].

In sum, each participant evaluated the incidents of 9 randomly selected scenarios. In particular, 2 random scenarios were picked from each of the first three categories (i.e., *drugs and alcohol use*, *sex*, and *religion and politics*), and 1 from each of the three last ones (i.e., *strong sentiment*, *location*, and *personal identifiers*). Therefore, we ensure a fair proportion of assessments across the different information categories while preserving the individual evaluation likelihood of each scenario. Furthermore, this criterion allowed us to maximize the amount of evidence collected on each information category.

Table 1. Categories of self-disclosure scenarios.

No.	Category	Scn. IDs	Example	SDP <PAs, Audience, UIN>
I	Drugs and alcohol use	1–6	Scn. 6: *"You post a picture of you drunk at a party. You feel embarrassed after your work colleagues forward the picture to your boss"*	SDP_6 <alcohol consumption, work colleagues, embarrassment>
II	Sex	7–12	Scn. 8: *"You post a naked or semi-naked picture of you. You get a wake-up call from your superior after a colleague shows it to her"*	SDP_8 <nudity, work colleagues, employer warning>
III	Religion and politics	13–18	Scn. 15: *"You share a post giving your opinion about a religious issue or statement. Some of your friends decide to end up their relationship with you because they found your post offensive"*	SDP_{15} <religious beliefs, close friends, end up friendship>
IV	Strong sentiment	19–21	Scn. 21: *"You share a post with a negative comment about your employer. You lose your job after a work colleague forwards the post to your boss"*	SDP_{21} <employer judgement, work colleagues, job joss>
V	Location	22–23	Scn. 22: *"You share a post and include the location where you are at the moment. You get stalked by a person who sees your post and is at the same place as you are"*	SDP_{22} <location, public, stalking>
VI	Personal identifiers	24–26	Scn. 24: *"You share a post including your new phone number. You get messages and calls from a person who was not supposed to reach you"*	SDP_{24} <phone number, public, harassment>

4.2. Population and Sampling

The survey was conducted in September of 2019 through Amazon's Mechanical Turk (Mturk) https://www.mturk.com, a crowdsourcing marketplace where *requesters* can allocate Human-Intelligence Tasks (HITs) that are completed by the platform's *workers* [50]. Mturk is a platform frequently used by researchers in the area of usable privacy and security for conducting experiments with human subjects [51]. The HIT in this case was the survey described in Section 4.1 and, for ensuring good quality responses, workers were required a HIT approval rate $\geq$ 95% and several approved HITs $\geq$ 1000 [52]. A remuneration of $1.25 was offered to each worker/participant considering an average completion time of 15 min per survey and the payment standards of the Mturk community. In sum, a total of 289 responses from participants across the United States and Germany were collected from which 281 were considered for the analysis and the rest rejected or discarded due to inconsistencies. In particular, responses from participants who either (i) failed on answering the control questions, (ii) answered the survey in a very short time, or (iii) provided an invalid completion code where not taken into consideration for the analysis. Table A1 in Appendix A summarizes the self-reported demographic characteristics of the study sample.

5. Results and Findings

Following, we summarize the results of our online survey. In particular, we analyze how individuals assess the severity of different self-disclosure scenarios for each of the information categories introduced in Section 4.1. For this, descriptive metrics were elaborated to identify those categories with the highest level of perceived severity. Moreover, a hypothesis test was conducted to determine the immediate effect of such scenarios upon participants' self-disclosure behavior. The findings reported below are leveraged in Section 6 for the elaboration of a personalized nudging solution.

5.1. Assessment of Self-disclosure Scenarios

Table 2 summarizes the participants' severity assessment of the proposed self-disclosure scenarios. As mentioned in Section 4.1, each participant was asked to evaluate 9 scenarios, so a total of 2529 evaluations were carried out by the end of the survey. In particular, categories I, II, and III received 562 observations each, while 281 responses were obtained for categories IV, V, and VI, correspondingly. In other words, the first three categories received two evaluations per participant, whereas the last three received one. Figure 1 illustrates the perceived severity of the proposed scenarios aggregated per information category (average values were considered for categories I, II, and III since they were assessed twice by each participant). As it can be observed, the category with the highest mean severity is *location* ($\overline{X}_V = 4.19 \pm 0.85$) and the one with the lowest is *religion and politics* ($\overline{X}_{III} = 3.29 \pm 0.78$) together with *personal identifiers* ($\overline{X}_{VI} = 3.30 \pm 1.05$). As already mentioned, severity can be measured using a 5-point ordinal scale ranging from *insignificant* (1) to *catastrophic* (5). Hence, the severity of *location* scenarios is perceived overall as "major", whereas scenarios defined under the categories *religion and politics* and *personal identifiers* approach "moderate" severity values. For the remaining categories *drugs and alcohol use* ($\overline{X}_I = 3.65 \pm 0.73$), *sex* ($\overline{X}_{II} = 3.81 \pm 0.74$), and *strong sentiment* ($\overline{X}_V = 3.68 \pm 0.95$), the severity assessment is, on average, between "moderate" and "high".

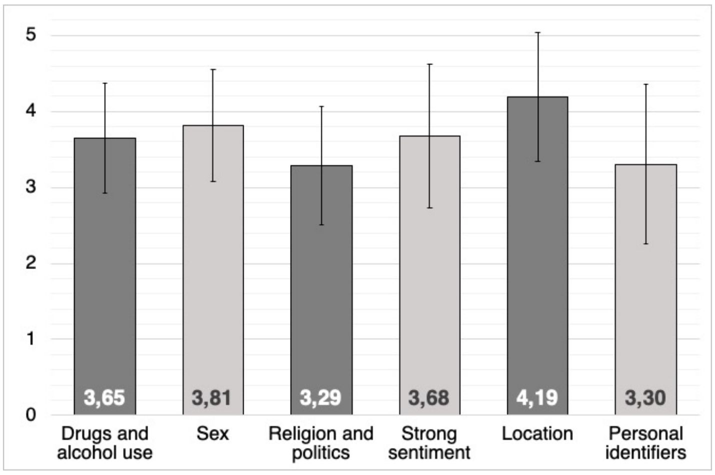


Figure 1. Perceived severity of each category of scenarios.

At a glance, one can observe differences in the perceived severity of the scenarios across information categories. To determine whether such differences are significant or not we conducted a one-way ANOVA test followed by a posthoc non-parametric test (see Appendix B). The results of the ANOVA test show a significant difference in the perceived severity across all information categories with $F_{5,1680} = 43.075$, $p < 0.05$ (Table A3). Since the studied sample violates the *homogeneity of variances* condition, a Games-Howell test was conducted to analyze where these differences occur. In particular, to determine which pairs of information categories show significant differences in their perceived severity. The results of this analysis are summarized in Table A4 (Appendix B) and reveal significant differences for the majority of the paired groups ($p < 0.05$). In particular, the following paired categories did not show significant differences in their means: *(drugs and alcohol use—sex)*, *(drugs and alcohol use—strong sentiment)*, *(sex—strong sentiment)*, and *(religion and politics—personal identifiers)*.

Along with the severity assessment, participants were asked if they had experienced the scenarios themselves and, if so, whether they deleted the corresponding publication afterwards. In sum, from a total of 2529 evaluations, only 132 correspond to scenarios that were experienced by the participants. However, out of this amount, 61 were reported as situations in which the corresponding post was deleted afterwards. Hence, the ratio of deleted posts over experienced self-disclosure scenarios is of 46.21% in our sample. Furthermore, when this analysis is conducted per information category, one can observe that the highest deleted/experienced ratio is 80% for *personal identifiers*, whereas the lowest one corresponds to *religion and politics* with 16%. The remaining categories *drugs and alcohol use*, *sex*, *strong sentiment* and *location* show deleted/experienced ratios of 37%, 58%, 44% and 54%, respectively.

Table 2. Participant's severity evaluation of the scenarios.

Category	Scn.	N	Mean	SD	Experienced	Deleted
Drugs and alcohol use	1	100	3.75	0.94	0	0
	2	93	2.74	0.85	19	9
	3	83	4.30	0.89	0	0
	4	144	4.34	0.66	2	1
	5	50	3.00	1.07	1	0
	6	92	3.14	0.92	5	0
Sex	7	140	4.50	0.73	2	1
	8	100	4.03	0.85	0	0
	9	100	3.32	1.09	4	2
	10	96	3.44	0.93	0	0
	11	41	4.02	0.94	3	2
	12	85	3.32	0.97	3	2
Religion and politics	13	102	4.10	0.91	4	1
	14	92	2.79	0.90	5	1
	15	92	2.91	1.01	4	0
	16	95	2.78	0.92	17	1
	17	87	2.74	0.90	1	1
	18	94	4.29	0.77	1	1
Strong sentiment	19	95	3.21	0.86	12	5
	20	97	3.63	0.96	4	2
	21	89	4.22	0.73	2	1
Location	22	129	3.90	0.89	8	4
	23	152	4.43	0.73	5	3
Personal identifiers	24	99	3.03	0.87	9	8
	25	92	4.27	0.63	1	0
	26	90	2.61	0.83	20	16

5.2. *Effects of Risk-Based Interventions*

As described in Section 4.1, a statistical test was conducted to analyze the effects that risk-based interventions may have on peoples' self-disclosure behavior. In particular, we ran a paired sample *t*-Test to determine if, following the scenarios' assessment, the means of the response variables showed significant differences.

The pre-assessment (PRE) and post-assessment (POS) values of the response variables are summarized in Table 3, whereas the outcome of the *t*-Tests can be found in Table 4. As it can be observed, most variables show higher means in the pre-assessment than after the post-assessment except for *perceived risk* whose mean increased from 3.543 to 3.630 (Mean diff. = −0.093). As for the rest, *perceived control* is the variable with the highest decrease (Mean dif. = 0.235) and *self-disclosure* the one with the lowest (Mean diff. = 0.093). In the case of *trust in member*, this value decreases 0.165 whereas for *trust in provider* in 0.123 points. As shown in Table 4, all these differences are statistically significant (i.e., *p*-values bellow 0.05) for a confidence level of 95%. The corresponding effect sizes of all variables are between "medium" and "low" except for *self-disclosure* which is below "low" according to Cohen's convention [53].

Table 3. Paired samples statistics.

Variable		Mean	N	SD	SE
Self-Disclosure	PRE	3.751	281	1.480	0.088
	POS	3.648	281	1.560	0.093
Perceived Control	PRE	4.268	281	1.483	0.088
	POS	4.033	281	1.558	0.093
Trust in Member	PRE	3.855	281	1.205	0.072
	POS	3.689	281	1.251	0.075
Trust in Provider	PRE	3.532	281	1.372	0.082
	POS	3.409	281	1.379	0.082
Perceived Risk	PRE	3.543	281	0.902	0.054
	POS	3.630	281	0.913	0.054

Overall, the results of the *t*-Test suggest that risk-based interventions can promote a preventative behavior among the users of SNSs. In particular, results show that *self-disclosure* intentions tend to decrease after participants are informed about unwanted incidents that may occur when personal information is revealed in online platforms ($t_{280} = 3.468, p < 0.001, d = 0.156$). Furthermore, as expected, these cues also increase people's immediate perception of *privacy risks* ($t_{280} = -3.481, p < 0.001, d = 0.202$). However, results also show a negative impact on the participants' trust in both the *service provider* ($t_{280} = 3.468, p < 0.001, d = 0.207$) and in *other SNS members* ($t_{280} = 5.018, p < 0.001, d = 0.300$). Moreover, this is also the case for *perceived control* ($t_{280} = 3.468, p < 0.001, d = 0.304$). Nevertheless, given the current experimental setting, these should only be considered to be short-term effects since measurements were taken right after the scenarios' assessment without capturing any mid- or long-term consequences.

Table 4. Paired sample *t*-Tests.

Pair	Mean diff.	SD	SE	d.f.	t	*p*	Cohen's d
(PRE-POS) Self-disclosure	0.103 *	0.614	0.037	280	3.468	0.001	0.156
(PRE-POS) Perceived Control	0.235 *	0.773	0.046	280	3.468	0.001	0.304
(PRE-POS) Trust in Member	0.165 *	0.553	0.033	280	5.018	0.000	0.300
(PRE-POS) Trust in Provider	0.123 *	0.593	0.035	280	3.468	0.001	0.207
(PRE-POS) Perceived Risk	−0.093 *	0.446	0.027	280	−3.481	0.001	0.202

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

6. Personalized Risk-Based Interventions

The results of our study show that risk-based interventions can have a preventative effect on people's self-disclosure behavior. However, as described in Section 3.2, results also indicate that the perceived severity of a particular SDP can vary among participants. Given the close relation existing between risk perception and privacy behavior, such nuances become crucial for the design of personalized nudging solutions. One way to personalize behavioral interventions is by tailoring their frequency and content to the particular goals of each user. Following this approach, an adaptation strategy driven by the risk estimation of SDPs is introduced in this section along with an application example.

6.1. Content, Frequency and Timing

At their core, SDPs represent events of a certain *likelihood* and *severity* which occur when particular pieces of private information are revealed to untrusted audiences. Hence, it is possible to determine their corresponding *risk* value based on the frequency and severity of the Unwanted Incidents (UINs) they describe. In particular, such estimation can be carried out through a normalized index which assigns values closer to 1 when the risk of the UIN is high, and closer to 0 when it is low (see

Appendix C for an extended description of the risk metric and its estimation approach). Likewise, an index of similar characteristics can be applied for representing the *risk threshold* of each user by assigning values closer to 1 to individuals with high risk tolerance, and closer to 0 otherwise. Then, an approach for adjusting the *content* of interventions may consist of just communicating those UINs whose risk values are above the user's threshold. Furthermore, a similar approach can be applied to regulate the intervention's *frequency*, i.e., by adjusting periodically the user's threshold depending on how often she accepts or ignores the generated interventions [28]. For instance, if the number of ignored interventions is higher than the accepted ones in a certain time period, then the user's threshold is increased. Conversely, the threshold is reduced when the number of rejected interventions exceeds the accepted ones.

A central aspect of nudging design is the definition of an adequate *timing* [54], i.e., the moment in which interventions are generated and applied for promoting behavioral change. As a general rule, interventions should be applied at the time when users are likely to be more receptive to privacy advice [23]. In the case of preventative nudges addressing self-disclosure decisions, current approaches normally intervene when the user is about to publish a post with personal or sensitive data [22,23,29]. Hence, their timing is realized through the identification of personal information inside the user's post. In particular, this task can be executed through different methods and techniques for Natural Language Processing (NLP) such as regular expressions, named entity recognition, or sentiment analysis [55]. Nevertheless, these techniques cannot attain in isolation the identification of a wide spectrum of personal information. Moreover, custom solutions to personal data detection should also differentiate between *self-referencing* posts such as "I love working at the University of Oxford" and more general ones like "The University of Oxford looks like an amazing place for working" [56]. Consequently, the analysis of textual publications is a complex challenge that often requires crafted solutions and the orchestration of multiple NLP approaches.

6.2. Intervention Approach

An alternative *timing* approach could consist of intervening when posts are assessed as "regrettable". In particular, it has been shown that users who regret having posted something in SNSs often delete such content afterwards [15,48]. Furthermore, this notion is supported by the results of our online survey. Hence, a regret classifier for textual publications could be elaborated in principle out of a corpus of deleted posts, i.e., by labelling deleted content as "regrettable" and using it for training a machine learning model (e.g., Artificial Neural Networks or Support Vector Machines) capable of classifying a particular post into regrettable or not.

Algorithm 1 integrates the approach described above with the adaptation strategies for *content* and *frequency* introduced in Section 6.1. Essentially, this algorithm is triggered when the user attempts to share a post P through her SNSs account. First, the function $IsRegrettable$ determines whether the content disclosed inside the post is likely to be regretted later. The result of this assessment is assigned to the variable $regrettable$ (line 4) and used thereafter to determine if an intervention must be elaborated or not (line 5). For instance, the post of Figure 2 is quite likely to be assessed as "regrettable" by the classifier. In that case, an SDP is selected from a *knowledge base* (KB) taking into consideration the risk threshold φ of the end-user. In particular, such a KB comprises a collection of SDPs and their corresponding perceived severity frequencies (as in Table 5). Consequently, function $SelectSDP$ selects a random SDP with a risk value higher than φ and assigns it to the variable $inSDP$ (line 6). If no SDP of such characteristics is found inside the KB, then the algorithm terminates and no intervention is generated (line 7). Considering $\varphi = 0.5$ and Table 5 as our KB (i.e., built upon the severity assessments obtained from the online survey), the following scenarios/SDPs could be selected by the function $SelectSDP$: 1, 3, 4, 6-13, and 18-25. Therefore, under such conditions, the algorithm can proceed and elaborate on a proper intervention.

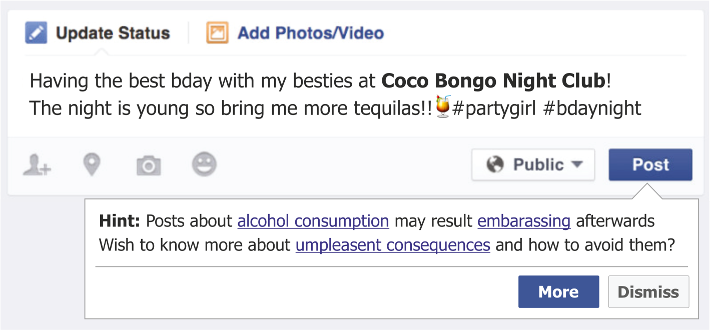
Algorithm 1: Personalized interventions

```
Function GenerateIntervention(P):
    InterventionMSG inMSG;
    Action usrAction;
    Boolean regrettable := IsRegrettable(P);
    if $regrettable = true$ then
        SDP inSDP:= SelectSDP(KB, $\varphi$);
        if $inSDP \neq null$ then
            inMSG.SetMsg(inSDP);
            ApplyIntervention(inMSG);
            usrAction := WaitForUsrAction();
            UpdateRiskThreshold(usrAction);
        end if
    end if
return
```

Table 5. Criticality index and severity frequencies of each scenario.

Category	Scn.	N	1	2	3	4	5	$\hat{I}$	SE
Drugs and alcohol use	1	100	2	9	20	50	19	0.688	0.047
	2	93	5	31	42	13	2	0.435	0.044
	3	83	2	2	6	32	41	0.825	0.049
	4	144	0	2	9	71	62	0.835	0.027
	5	50	4	12	18	12	4	0.500	0.075
	6	92	2	21	37	26	6	0.535	0.048
Sex	7	140	1	4	2	50	83	0.875	0.031
	8	100	0	4	22	41	33	0.758	0.042
	9	100	5	20	26	36	13	0.580	0.054
	10	96	2	11	38	33	12	0.609	0.047
	11	41	0	2	11	12	16	0.756	0.072
	12	85	2	15	31	28	9	0.579	0.052
Religion and politics	13	102	1	7	10	47	37	0.775	0.045
	14	92	6	29	36	20	1	0.448	0.046
	15	92	9	21	34	25	3	0.478	0.052
	16	95	8	27	40	18	2	0.445	0.047
	17	87	8	24	39	15	1	0.434	0.048
	18	94	1	1	9	42	41	0.822	0.040
Strong sentiment	19	95	3	16	36	38	2	0.553	0.044
	20	97	4	5	30	42	16	0.657	0.049
	21	89	0	3	7	46	33	0.806	0.039
Location	22	129	1	10	22	64	32	0.725	0.039
	23	152	1	3	7	60	81	0.857	0.030
Personal identifiers	24	99	2	26	42	25	4	0.508	0.044
	25	92	0	1	6	52	33	0.818	0.033
	26	90	7	33	39	10	1	0.403	0.044

Note: Severity level 1 = *insignificant*, 2 = *minor*, 3 = *moderate*, 4 = *major*, 5 = *catastrophic*; 🟦 Scenarios with risk index $\hat{I}$ higher than 0.5.



Having the best bday with my besties at **Coco Bongo Night Club**!
The night is young so bring me more tequilas!!🍹#partygirl #bdaynight

Hint: Posts about alcohol consumption may result embarassing afterwards
Wish to know more about unpleasent consequences and how to avoid them?

Figure 2. Envisioned Interface.

Once a valid SDP has been chosen and assigned to *inSDP*, an intervention can be elaborated by setting the content of the warning message *inMSG*. In particular, the method *setMsg* instantiates *inMSG* with the *unwanted incident* and *personal information* described by the SDP allocated in *inSDP* (line 8). For example, if such SDP corresponds to scenario No 6, then *inMSG* is instantiated with the values *embarrassment* and *alcohol consumption* (Table 1). Once the content of the message is defined, it is communicated to the user through the function *ApplyIntervention* (line 9). At this point, the user has the chance to re-think the content of her post or proceed with its publication (Figure 2). As mentioned in Section 6.1, prior work suggests that such a user feedback can be leveraged for adapting the frequencies of the interventions. In particular, Díaz Ferreyra et al. [28] propose to regulate this parameter according to the amount of *accepted*/*dismissed* interventions observed in a given period of time. For this, the function *WaitForUsrAction* waits for the user's input and forwards it to the function *UpdateRiskThreshold* which takes care of updating the value of φ (lines 10 and 11). This function monitors of the number of times the user has dismissed/followed the warnings within a τ period of time. After each τ period, if $\#dismiss > \#accept$, then the value of φ is increased in δ (i.e., $\varphi_{\tau+1} := \varphi_\tau + \delta$). Conversely, if $\#dismiss < \#accept$, then φ is decreased in δ (i.e., $\varphi_{\tau+1} := \varphi_\tau - \delta$).

7. Discussion

Overall, the results of our online survey reveal nuances in the perception of privacy risk among the users of SNSs. Furthermore, such nuances become salient not only across different self-disclosure scenarios but also throughout their corresponding information categories. On one hand, these findings are aligned with prior work in risk perception and privacy behavior. In particular, a recent study by Gerber et al. [26] revealed contrasts across the perceived severity of unwanted incidents related to data collection and manipulation. Furthermore, their findings suggest that in terms of risk awareness, scenarios that solely describe a general probability of harm are not as effective as the ones who illustrate specific negative consequences. In line with this, the analysis conducted in Section 5.2 further supports the application of self-disclosure scenarios for nudging purposes. In particular, our results indicate that interventions elaborated out of SDPs have a short-term negative effect on participants' self-disclosure, trust, and control but a positive one on their perception of risk. Similar effects were observed by Kroll and Stieglitz [57] in an experiment that tested the impact of current persuasive elements on Facebook (i.e., privacy wizards and profile-visibility check-ups). Furthermore, the negative impact of privacy risks over users' trust in SNSs has also been confirmed by Nemec Zlatolas et al. [58] through a structural equation model. Nevertheless, our results have yielded intervention effects of medium-low size on all response variables. Hence, a further analysis employing larger samples may be necessary for gaining further insights on the effectiveness of risk-based interventions.

Besides the insights on the persuasive effects of SDPs, the results summarized in Section 5 have also provided valuable input for the design of preventative nudges. In particular, the strategy described in Section 6.2 leverages the information gathered from the online survey not only for estimating the risk impact of SDPs but also for adapting the timing of behavioral interventions. One distinctive aspect of this approach is the machine learning module which allows the identification of potentially regrettable posts. As already mentioned, an approach of such characteristics would, in principle, reduce the amount of NLP craftsmanship necessary for analyzing the content of a post. Nevertheless, up to some extent, it compromises the usability of the nudging strategy since it hinders the elaboration of content-specific interventions. On one hand, this raises less privacy concerns since it performs a *coarse-grained classification* of the post (i.e., into "regrettable" or not). However, a *fine-grained analysis* (i.e., the identification of specific patterns of private information) would allow the selection of more relevant SDPs [28]. For instance, if the information disclosed by the user is *location*, a fine-grained analysis would allow choosing a SDP from the KB whose information component (PAs as described in Section 3.1) corresponds to *location*. However, a coarse-grained classification does not provide the adequate level of abstraction which is necessary for performing such a content-aware selection. Consequently, Algorithm 1 may generate interventions decoupled from the content of the post since these are elaborated out of SDPs which are randomly chosen. This is a central design aspect that must be taken into consideration and further investigated.

8. Limitations

Although the approach employed in this work has yielded interesting results, there are some limitations that should be acknowledged. First, the results summarized in Section 5 were obtained from the assessment of hypothetical self-disclosure scenarios. Hence, this approach does not ensure that participants' actual behavior (i.e., in a real case scenario) would be consistent with the one they reported during the online survey. Furthermore, although the assessed scenarios portray the three main components of an SDP (i.e., personal information, untrusted audience, and unwanted incident), they were not framed as the interventions they suppose to generate. On one hand, this strategy remains adequate for the estimation of the privacy risks associated with each SDP. However, such an approach may not be suitable enough for capturing the nudging effectiveness of SDPs to a large extent.

Another aspect to be considered is the number of scenarios and information categories that were defined for this study. In particular, the proposed scenarios cover just a fraction of the cases reported in the literature. Hence, some categories may not be extensively represented and require the definition of additional scenarios for better assessment. Moreover, such an assessment may be influenced by the cultural background of the participants which, depending on its collectivist (or individualistic) nature, can predispose them to a more (or less) privacy risk-avoidance behavior [35]. Therefore, a closer look into risk perception across cultures should be taken to unveil significant differences in the severity values assigned to particular SDPs.

Another limitation point is related to the use of crowdsourcing platforms for carrying out online surveys. In particular, conducting surveys over Mturk often supposes a loss of control over the experimental setting [50,59]. In particular, participants may compromise the quality of their answers due to distractions present in their physical environment. Moreover, workers sometimes provide fast or nonsense answers to optimize their profit. Nevertheless, it has been shown that studies conducted over Mturk can provide results as relevant as those obtained using traditional survey methods [50]. This can be achieved by applying some good practices such as (i) controlling the time workers actually spend in the task, (ii) filtering-out workers with a low HIT approval rate, and (iii) adding control questions [60,61]. These practices were followed to ensure good quality results.

9. Conclusions and Future Work

Privacy-preserving behavior in SNSs is often impaired by heuristic judgements and optimistic biases. In particular, instant gratifications and short-term benefits tend to outweigh the risks associated with unrestrained self-disclosure practices and increase, thereby, the chances of regret. Hence, it is of great importance to promote a reflective thinking among the users of social media platforms to preserve their privacy and contextual integrity. For this, individuals must have the ability to reflect on the consequences of their privacy choices and understand what is at stake when interacting through these technological means [62]. Under this premise, this work has studied the role of risk cues in online self-disclosure behavior and their importance for the design of preventative nudges. The results of our online survey suggest that behavioral interventions generated from SDPs can increase users' perceived severity of privacy risks and reduce their self-disclosure intentions. Furthermore, results also reveal nuances in individual's perception of unwanted incidents and thus provide a valuable insight for the elaboration of personalized nudging solutions.

There are several questions and research directions that can be drawn from the results obtained in this work. The most salient one is the implementation and evaluation of the approach described in Section 6.2. In particular, such task introduces challenges related to (i) the technical realization of Algorithm 1, and (ii) the definition of an experimental setting for investigating the long-term effects of the proposed nudging strategy. On the technical side, a critical point is the timing mechanism embedded in Algorithm 1 which must be trained from a corpus of "regrettable" posts. Overall, prior research concerning the analysis of deleted content in SNSs can provide well-grounded theoretical and practical foundations to this challenge. For instance, Tinati et al. [63] and Gazizullina et al. [64] elaborated on machine learning solutions of similar characteristics for the classification of deleted publications on Instagram and Twitter, respectively. On the other hand, the analysis of long-term effects is expected to be performed through an application for mobile phones that implements the proposed nudging strategy, i.e., an app from which users can post messages using their SNS accounts (e.g., via the corresponding API) and intervenes according to the approach described in Section 6.2. Thus, we plan to evaluate the preventative effects of the generated interventions by testing this app with a group of users and analyzing the outcome of their self-disclosure decisions. In line with this, we also plan to further elaborate on the estimation of the parameters employed by Algorithm 1 (i.e., φ, τ and δ) in order to maximize users' privacy awareness while minimizing the potential negative effects of behavioral interventions (e.g., annoyance, irritation or distraction).

Another future research direction is related to the ethical issues that certain designs of choice architectures may introduce. In the recent years, nudges and their applications have become quite popular across different disciplines and research areas. Furthermore, they have also caught the attention of governments across the world for addressing societal issues at large. For instance, the British government has recently considered nudging as a solution for controlling the spread of COVID-19 across the UK [65]. However, despite the optimism of many, nudges are often considered to be a threat to individuals' agency and autonomy. Essentially, this is due to the fine line existing between persuasion, manipulation, and coercion. For example, citizens may be encouraged to share their smartphone's location with the excuse of ensuring social distancing measures when, in fact, the main objective is monitoring their movements. Such applications introduce several ethical questions including *"who should benefit from nudges?"*, *"should users be informed of the presence of a nudge?"*, and *"how nudges should (not) influence the users?"*. So far, these questions have been addressed through ethical guidelines and principles that should be included into the nudges' design. In particular, Renaud et al. [66] have introduced guidelines and checklists for the ethical design of nudges in the areas of privacy and security. However, it remains unclear how these guidelines should be incorporated into the development process of preventative nudges and how can they be promoted among the designers of these technologies. In our future work, we expect to elaborate on these ethical aspects and contribute thereby to an ethical-by-design approach for preventative nudges.

Author Contributions: Conceptualization, N.E.D.F.; methodology, N.E.D.F. and T.K.; validation, T.K., E.A., S.S. and M.H.; formal analysis, N.E.D.F. and T.K.; investigation, N.E.D.F. and T.K.; data curation, N.E.D.F.; writing—original draft preparation, N.E.D.F.; writing—review and editing, T.K., E.A., S.S. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media", the H2020 European Project No. 787034 "PDP4E: Privacy and Data Protection Methods for Engineering", and Canada's Natural Sciences and Engineering Research Council (NSERC).

Acknowledgments: The authors acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

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

Abbreviations

The following abbreviations are used in this manuscript:

SNS	Social Network Site
SDP	Self-Disclosure Pattern
UIN	Unwanted Incident
PA	Private Attribute

Appendix A. Studied Sample

Table A1. Demographic characteristics of the studied sample.

Demographic	Ranges	Freq.	Responses (%)
Age	18-25 years	29	10.3
	26–35 years	135	48
	36–45 years	66	23.5
	46–55 years	31	11
	$>$ 56 years	20	7.12
Gender	Male	156	55.5
	Female	123	43.8
	Non-binary	2	0.7
Occupation	Employed full time	205	73
	Employed part time	34	12.1
	Unemployed and not searching for work	13	4.6
	Unemployed searching for work	8	2.8
	Disabled or retired	7	2.5
	Student	14	5
Education	Graduate degree (MSc, PhD)	44	15.7
	Undergraduate degree (BSc, BA)	104	37
	Some college	87	31
	High school or less	43	15.3
	Primary school or less	3	1.1

Appendix B. ANOVA Posthoc Test

Table A2. Descriptive statistics of each information category.

Category	N	Mean	SD	SE
(I) Drugs and alcohol use	281	3.649	0.728	0.043
(II) Sex	281	3.811	0.737	0.044
(III) Religion and politics	281	3.288	0.782	0.047
(IV) Strong sentiment	281	3.676	0.951	0.057
(V) Location	281	4.185	0.850	0.051
(VI) Personal identifiers	281	3.302	1.051	1.051

Table A3. One-way ANOVA between information categories.

	SS	d.f.	MS	F	*p*
Between groups	158.654	5	31.731	43.075	0.000
Within groups	1237.573	1680	0.737		
Total	1396.227	1685			

Table A4. Games-Howell Simultaneous Tests for Differences of Means.

Difference of Levels	Difference of Means	SE	p	95% CI
drugs and alcohol use—sex	−0.162	0.062	0.094	(−0.339, 0.015)
drugs and alcohol use—religion and politics	0.361 *	0.064	0.000	(0.179, 0.544)
drugs and alcohol use—strong sentiment	−0.027	0.072	0.999	(−0.231, 0.178)
drugs and alcohol use—location	−0.536 *	0.067	0.000	(−0.727, −0.345)
drugs and alcohol use—personal identifiers	−0.347 *	0.076	0.000	(0.129, 0.365)
sex—religion and politics	0.523 *	0.064	0.000	(0.340, 0.707)
sex—strong sentiment	0.135	0.072	0.413	(−0.070, 0.341)
sex—location	−0.374 *	0.067	0.000	(−0.566, −0.182)
sex—personal identifiers	0.509 *	0.077	0.000	(0.290, 0.728)
religion and politics—strong sentiment	−0.388 *	0.074	0.000	(−0.598, −0.178)
religion and politics—location	−0.897 *	0.069	0.000	(−1.109, −0.700)
religion and politics—personal identifiers	−0.014	0.078	1.000	(-0.238, 0.209)
strong sentiment—location	−0.509 *	0.076	0.000	(−0.727, −0.291)
strong sentiment—personal identifiers	0.374 *	0.085	0.000	(0.132, 0.616)
location—personal identifiers	0.883 *	0.081	0.000	(0.652, 1.113)

Note: (*) The mean difference is significant for $\alpha = 5\%$. Welch's $F_{5,782.05} = 42.83$, $p < 0.05$.

Appendix C. Risk Criticality Index

A risk criticality index is an instrument for measuring the impact of an unwanted incident given its severity and frequency. In particular, Facchinetti et al. [67] introduced an approach in which severity is assumed to be measured through an ordinal scale such as *insignificant*, *minor*, *moderate*, *major*, and *catastrophic*. Under this premise, such index I draws upon a categorical random variable X with ordered categories x_k which represent decreasing severity levels $k = 1, 2, ..., K$. Therefore, an *estimator* of I can be obtained out of a sample of size n of the categorical variable X with the following equation [67]:

$$\hat{I} = \frac{\sum_{k=1}^{K} \tilde{F}_k - 1}{K - 1}$$

where for a severity scale of K levels, the values $k = 1$ and $k = K$ correspond to the highest and lowest severity values of an unwanted incident, respectively. Likewise, $\tilde{F}_k$ corresponds to the *empirical* distribution function of the random variable X, which for a category x_k is computed as the number of observations r_l in the sample with consequence levels between 1 and k:

$$\tilde{F}_k = \sum_{l=1}^{k} \frac{r_l}{n} \quad \text{for} \quad k = 1, 2, ..., K$$

Table 5, summarizes the values of $\hat{I}$ for each of the 26 scenarios/SDPs included in the survey described in Section 4.1 according to the participants' severity assessments. In particular, values of $\hat{I}$ closer to 0 suggest that the impact of an unwanted incident is likely to be low whereas values closer to 1 indicate that such impact is likely to be high. The corresponding variance of $\hat{I}$ is given by the following equation:

$$Var(\hat{I}) = \frac{1}{n(K-1)}\left[\sum_{k=1}^{K-1}(K-k)^2 p_k(1-p_k) - 2\sum_{k=1}^{K-1}(K-k)p_k\sum_{l=1}^{k-1}(K-l)p_l\right]$$

where p_k is the proportion of observations in the sample corresponding to the severity level k. From this equation, a confidence interval for $\hat{I}$ can be obtained as:

$$\hat{I} - Z_{\alpha/2}\cdot S(\hat{I}) \leq I \leq \hat{I} + Z_{\alpha/2}\cdot S(\hat{I})$$

where $S(\hat{I})$ is the standard deviation of $\hat{I}$, α the significance level, and $Z_{\alpha/2}$ the standard score for $\alpha/2$.

Appendix D. Survey Instruments

As described in Section 4.1, participants were asked to evaluate 9 randomly selected scenarios by answering the following questions:

- **Q1**: *"Please indicate how severe is for you the consequence described in this scenario"*. **Options**: *insignificant*, *minor*, *moderate*, *major*, or *catastrophic*.
- **Q2**: *"Have you experienced a situation similar to that before?"*. **Options**: *yes*, or *no*.
- **Q3**: (if the answer to Q2 was *yes*) *"Have you deleted such content afterwards?"*. **Options**: *yes*, or *no*.

Constructs were measured before and after the assessment of the scenarios. For the post-assessment measurement, we used differently phrased questions to minimize habituation biases. The employed constructs correspond to the ones introduced by Krasnova et al. [68] and were measured using a 7-point Likert scale ranging from *entirely disagree* to *entirely agree*. The reliability of the employed scales was assessed through the Cronbach's Alpha coefficient which in all cases was higher than 0.80. Since values higher than 0.70 are considered "acceptable", this suggests that the items of each construct scale have a relatively high internal consistency.

All participants gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of the University of Duisburg-Essen (Req. No. 1907PKKT1415). All the questions of the survey were deemed mandatory.

Appendix D.1. Employed Constructs

Trust in Provider (TP)

i Social media is open and receptive to the needs of its members.
ii Social media makes good-faith efforts to address most member concerns.
iii Social media is also interested in the well-being of its members, not just its own.
iv Social media is honest in its dealings with me.
v Social media keeps its commitments to its members.
vi Social media is trustworthy.

Trust in Other Members (TM)

i Other members on social media will do their best to help me.
ii Other members on social media do care about the well-being of others.
iii Other members on social media are open and receptive to the needs of each other.
iv Other members on social media are honest in dealing with each other.
v Other members on social media keep their promises.
vi Other members on social media are trustworthy.

Perceived Control (PC)

i I feel in control over the information I provide on social media.
ii Privacy settings allow me to have full control over the information I provide on social media.
iii I feel in control of who can view my information on social media.

Self-disclosure (SD)

i I have a comprehensive profile on social media.
ii I find time to keep my profile up to date.
iii I keep my friends updated about what is going on in my life through social media.
iv When I have something to say, I like to share it on social media.

Perceived Privacy Risk (RISK)

i (R) Overall, I see no real threat to my privacy due to my presence on social media.
ii I fear that something unpleasant can happen to me due to my presence on social media.
iii (R) I feel safe publishing my personal information on social media.
iv Overall, I find it risky to publish my personal information on social media.
v Please rate your overall perception of privacy risk involved when using social media.

Appendix D.2. Self-Disclosure Scenarios

Alcohol and illegal drug use

1. You share a post describing your experience with drugs. You get a wake-up call from your superior after a colleague forwards this post to him.
2. You post a picture in which you are drunk at a party. You feel embarrassed after you realize this picture was seen by all your contacts including close friends, family and acquaintances.
3. You share a post describing your experience with drugs. You lose your job after your work colleagues forward this post to your boss.
4. You post a picture in which you are drunk at a party. You lose your job after your work colleagues forward this picture to your boss.
5. You share a post describing your experience with drugs. You feel embarrassed after you realize this post was seen by all your contacts including close friends, family and acquaintances.
6. You post a picture in which you are drunk at a party. You get a wake-up call from your superior after a colleague forwards this picture to him.

Sex

7. You post a naked or semi-naked picture of you. You lose your job after your work colleagues forward this picture to your boss.
8. You post a naked or semi-naked picture of you. You get a wake-up call from your superior after a colleague forwards this picture to him.
9. You share a post describing a personal sexual encounter or experience. You feel embarrassed after you realize this post was seen by all your contacts including close friends, family and acquaintances.
10. You share a post describing a personal sexual encounter or experience. You get a wake-up call from your superior after a colleague forwards this post to him.
11. You post a naked or semi-naked picture of you. You feel embarrassed after you realize this picture was seen by all your contacts including close friends, family and acquaintances.
12. You share a post describing a personal sexual encounter or experience. You feel embarrassed after you realize this post was seen by all your contacts including close friends, family and acquaintances.

Religion and Politics

13. You share a post giving your opinion about a religious issue or statement. You lose your job after your work colleagues forward this post to your boss.
14. You share a post giving your opinion about a political issue or statement. You get a wake-up call from your superior after a colleague forwards this post to him.
15. You share a post giving your opinion about a religious issue or statement. Some of your friends decide to end up their relationship with you because they found your post offensive.
16. You share a post giving your opinion about a political issue or statement. Some of your friends decide to end up their relationship with you because they disagree with what you wrote.
17. You share a post giving your opinion about a religious issue or statement. You get a wake-up call from your superior after a colleague forwards this post to him.
18. You share a post giving your opinion about a political issue or statement. You lose your job after your work colleagues forward this post to your boss.

Strong Sentiment

19. You share a post with a negative comment about someone else. Friends in common decide to end up their relationship with you after seeing what you wrote.
20. You share a post with a negative comment about your employer. You get a wake-up call from your superior after a colleague forwards this post to him.
21. You share a post with a negative comment about your employer. You lose your job after your work colleagues forward this post to your boss.

Location

22. You share a post and include the location where you are at the moment. You get stalked by a person who saw your post and is at the same place as you are.
23. You share a post including your new home address. Someone who saw your post breaks into your house to rob your belongings.

Personal Identifiers

24. You share a post including your new phone number. You get messages and calls from a person who was not supposed to see your post.
25. You share a picture of your brand-new credit card. Some days later you realize that someone has been buying stuff on your behalf.
26. You share a post including your new email address. Thereafter, you start getting spam messages from someone you don’t know.

References

1. Williams, D.J.; Noyes, J.M. How does our perception of risk influence decision-making? Implications for the design of risk information. *Theor. Issues Ergon. Sci.* **2007**, *8*, 1–35. [CrossRef]
2. Ashby, N.J.S.; Glöckner, A.; Dickert, S. Conscious and unconscious thought in risky choice: Testing the capacity principle and the appropriate weighting principle of unconscious thought theory. *Front. Psychol.* **2011**, *2*, 261. [CrossRef]
3. Slovic, P.; Peters, E. Risk Perception and Affect. *Curr. Dir. Psychol. Sci.* **2006**, *15*, 322–325. [CrossRef]
4. Loewenstein, G.F.; Weber, E.U.; Hsee, C.K.; Welch, N. Risk as feelings. *Psychol. Bull.* **2001**, *127*, 267. [CrossRef] [PubMed]
5. Fischer, A.R.H. Perception of Product Risks. In *Consumer Perception of Product Risks and Benefits*; Emilien, G., Weitkunat, R., Lüdicke, F., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 175–190.
6. Kim, H.K. Risk Communication. In *Consumer Perception of Product Risks and Benefits*; Emilien, G., Weitkunat, R., Lüdicke, F., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 125–149.

7. Yang, Z.J.; Aloe, A.M.; Feeley, T.H. Risk Information Seeking and Processing Model: A Meta-Analysis. *J. Commun.* **2014**, *64*, 20–41. [CrossRef]
8. Wang, Y.C.; Burke, M.; Kraut, R. Modeling Self-Disclosure in Social Networking Sites. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16, San Francisco, CA, USA, 27 February–2 March 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 74–85. [CrossRef]
9. Acquisti, A.; Brandimarte, L.; Loewenstein, G. Privacy and human behavior in the age of information. *Science* **2015**, *347*, 509–514. [CrossRef]
10. Ampong, G.; Mensah, A.; Adu, A.; Addae, J.; Omoregie, O.; Ofori, K. Examining Self-Disclosure on Social Networking Sites: A Flow Theory and Privacy Perspective. *Behav. Sci.* **2018**, *8*, 58. [CrossRef]
11. Such, J.M.; Criado, N. Multiparty Privacy in Social Media. *Commun. ACM* **2018**, *61*, 74–81. [CrossRef]
12. Albladi, S.; Weir, G.R.S. Vulnerability to Social Engineering in Social Networks: A Proposed User-Centric Framework. In Proceedings of the 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), Vancouver, BC, Canada, 12–14 June 2016; pp. 1–6.
13. Krombholz, K.; Hobel, H.; Huber, M.; Weippl, E. Advanced social engineering attacks. *J. Inf. Secur. Appl.* **2015**, *22*, 113–122. [CrossRef]
14. Vitak, J. The Impact of Context Collapse and Privacy on Social Network Site Disclosures. *J. Broadcast. Electron. Media* **2012**, *56*, 451–470. [CrossRef]
15. Wang, Y.; Norcie, G.; Komanduri, S.; Acquisti, A.; Leon, P.G.; Cranor, L.F. "I regretted the minute I pressed share": A Qualitative Study of Regrets on Facebook. In Proceedings of the ACM 7th Symposium on Usable Privacy and Security, SOUPS 2011, Pittsburgh, PA, USA, 20–22 July 2011; pp. 1–16. ANSWER: Confirmed
16. Sundar, S.S.; Kang, H.; Wu, M.; Go, E.; Zhang, B. Unlocking the Privacy Paradox: Do Cognitive Heuristics Hold the Key? In Proceedings of the ACM CHI '13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 811–816.
17. Marmion, V.; Bishop, F.; Millard, D.E.; Stevenage, S.V. The Cognitive Heuristics Behind Disclosure Decisions. In *Social Informatics. SocInfo 2017*; Lecture Notes in Computer Science Series; Ciampaglia, G., Mashhadi, A., Yasseri, T., Eds.; Springer: Cham, Switzerland, 2017; Volume 10539, pp. 591–607. ISBN 978-3-319-67216-8.
18. De, S.J.; Imine, A. On Consent in Online Social Networks: Privacy Impacts and Research Directions (Short Paper). In *Risks and Security of Internet and Systems*; Zemmari, A., Mosbah, M., Cuppens-Boulahia, N., Cuppens, F., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 128–135.
19. Krämer, N.C.; Schäwel, J. Mastering the challenge of balancing self-disclosure and privacy in social media. *Curr. Opin. Psychol.* **2020**, *31*, 67–71. [CrossRef] [PubMed]
20. Mosca, F.; Sarkadi, S.; Such, J.M.; McBurney, P. Agent EXPRI: Licence to Explain. In Proceedings of the 2nd International Workshop on Explainable Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS), Auckland, New Zealand, 9–13 May 2020.
21. Sánchez, D.; Domingo-Ferrer, J.; Martínez, S. Co-utile Disclosure of Private Data in Social Networks. *Inf. Sci.* **2018**, *441*, 50–65. [CrossRef]
22. Misra, G.; Such, J.M. PACMAN: Personal Agent for Access Control in Social Media. *IEEE Internet Comput.* **2017**, *21*, 18–26. [CrossRef]
23. Acquisti, A.; Adjerid, I.; Balebako, R.; Brandimarte, L.; Cranor, L.F.; Komanduri, S.; Leon, P.G.; Sadeh, N.; Schaub, F.; Sleeper, M.; et al. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 44. [CrossRef]
24. Lin, Y.; Osman, M.; Ashcroft, R. Nudge: Concept, Effectiveness, and Ethics. *Basic Appl. Soc. Psychol.* **2017**, *39*, 293–306. [CrossRef]
25. Samat, S.; Acquisti, A. Format vs. Content: The Impact of Risk and Presentation on Disclosure Decisions. In Proceedings of the USENIX Association Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017), Santa Clara, CA, USA, 12–14 July 2017; pp. 377–384.
26. Gerber, N.; Reinheimer, B.; Volkamer, M. Investigating People's Privacy Risk Perception. *Proc. Priv. Enhanc. Technol.* **2019**, *2019*, 267–288. [CrossRef]
27. Aimeur, E.; Diaz Ferreyra, N.E.; Hage, H. Manipulation and Malicious Personalization: Exploring the Self-Disclosure Biases Exploited by Deceptive Attackers on Social Media. *Front. Artif. Intell.* **2019**, *2*, 26. [CrossRef]

28. Díaz Ferreyra, N.E.; Meis, R.; Heisel, M. Learning from Online Regrets: From Deleted Posts to Risk Awareness in Social Network Sites. In Proceedings of the ACM 27th Conference on User Modeling, Adaptation and Personalization, Larnaca, Cyprus, 9–12 June 2019; pp. 117–125.
29. Masaki, H.; Shibata, K.; Hoshino, S.; Ishihama, T.; Saito, N.; Yatani, K. 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*; ACM: New York, NY, USA, 2020; pp. 1–11.
30. Peer, E.; Egelman, S.; Harbach, M.; Malkin, N.; Mathur, A.; Frik, A. Nudge Me Right: Personalizing Online Nudges to People's Decision-Making Styles. *Comput. Hum. Behav.* **2020**, *109*, 106347. [CrossRef]
31. Warberg, L.; Acquisti, A.; Sicker, D. Can Privacy Nudges be Tailored to Individuals' Decision Making and Personality Traits? In Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society, WPES'19, London, UK, 11 November 2019; pp. 175–197.
32. Kokolakis, S. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Comput. Secur.* **2017**, *64*, 122–134. [CrossRef]
33. Barnes, S.B. A Privacy Paradox: Social Networking in the United States. *First Monday* **2006**, *11*. [CrossRef]
34. Dienlin, T.; Metzger, M.J. An Extended Privacy Calculus Model for SNSs: Analyzing self-disclosure and Self-Withdrawal in a Representative U.S. Sample. *J. Comput. Mediat. Commun.* **2016**, *21*, 368–383. [CrossRef]
35. Trepte, S.; Reinecke, L.; Ellison, N.B.; Quiring, O.; Yao, M.Z.; Ziegele, M. A Cross-Cultural Perspective on the Privacy Calculus. *Soc. Media Soc.* **2017**, *3*, 1–13. [CrossRef]
36. Chen, H.T. Revisiting the Privacy Paradox on Social Media With an Extended Privacy Calculus Model: The Effect of Privacy Concerns, Privacy Self-Efficacy, and Social Capital on Privacy Management. *Am. Behav. Sci.* **2018**, *62*, 1392–1412. [CrossRef]
37. Spottswood, E.L.; Hancock, J.T. Should I Share That? Prompting Social Norms That Influence Privacy Behaviors on a Social Networking Site. *J. Comput. Mediat. Commun.* **2017**, *22*, 55–70. [CrossRef]
38. Gambino, A.; Kim, J.; Sundar, S.S.; Ge, J.; Rosson, M.B. User Disbelief in Privacy Paradox: Heuristics That Determine Disclosure. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, San Jose, CA, USA, 7–12 May 2016; pp. 2837–2843.
39. Metzger, M.J.; Flanagin, A.J. Credibility and trust of information in online environments: The use of cognitive heuristics. *J. Pragmat.* **2013**, *59*, 210–220. [CrossRef]
40. Weinmann, M.; Schneider, C.; vom Brocke, J. Digital Nudging. *Bus. Inf. Syst. Eng.* **2016**, *58*, 433–436. [CrossRef]
41. Esposito, G.; Hernández, P.; van Bavel, R.; Vila, J. Nudging to prevent the purchase of incompatible digital products online: An experimental study. *PLoS ONE* **2017**, *12*, e0173333. [CrossRef]
42. Damgaard, M.T.; Nielsen, H.S. The use of nudges and other behavioural approaches in education. *Anal. Rep. Eur. Expert Netw. Econ. Educ. (EENEE)* **2017**, *29*, 52.
43. Shaffer, V.A. Nudges for Health Policy: Effectiveness and Limitations. *Mo. Law Rev.* **2017**, *82*, 727.
44. De, S.J.; Le Métayer, D. Privacy Risk Analysis to Enable Informed Privacy Settings. In Proceedings of the 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), London, UK, 23–27 April 2018; pp. 95–102.
45. Díaz Ferreyra, N.E. Instructional Awareness: A User-centred Approach for Risk Communication in Social Network Sites. Ph.D. Thesis, University of Duisburg-Essen, Duisburg Germany, 2019. [CrossRef]
46. Malkin, N.; Mathur, A.; Harbach, M.; Egelman, S. Personalized Security Messaging: Nudges for Compliance With Browser Warnings. In *2nd European Workshop on Usable Security (EuroUSEC)*; Internet Society: Paris, France, 2017.
47. Guha, S.; Baumer, E.P.S.; Gay, G.K. Regrets, I've Had a Few: When Regretful Experiences Do (and Don't) Compel Users to Leave Facebook. In Proceedings of the 2018 ACM Conference on Supporting Groupwork, ACM, Sanibel Island, FL, USA, 7–10 January 2018; pp. 166–177.
48. Zhou, L.; Wang, W.; Chen, K. Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones. In Proceedings of the 25th International Conference on World Wide Web, Montréal, QC, Canada, 11–15 April 2016; pp. 603–612.
49. European Parliament and Council of European Union. *Regulation (EU) 2016/679*, 2016. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN (accessed on 12 August 2020).

50. Paolacci, G.; Chandler, J.; Ipeirotis, P.G. Running Experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* **2010**, *5*, 411–419.
51. Kelley, P.G. Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk. In Proceedings of the Symposium on Usable Privacy and Security (SOUPS), Redmond, WA, USA, 14–16 July 2010.
52. Tips for Academic Requesters on Mturk. Available online: https://bit.ly/3dUAI0y (accessed on 7 August 2020).
53. Cohen, J. Statistical Power Analysis. *Curr. Dir. Psychol. Sci.* **1992**, *1*, 98–101. [CrossRef]
54. Schaub, F.; Balebako, R.; Cranor, L.F. Designing Effective Privacy Notices and Controls. *IEEE Internet Comput.* **2017**. [CrossRef]
55. Tesfay, W.B.; Serna, J.; Pape, S. Challenges in Detecting Privacy Revealing Information in Unstructured Text. In Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web—Policy and Technology (PrivOn), Kobe, Japan, 18 October 2016; Brewster, C., Cheatham, M., d'Aquin, M., Decker, S., Kirrane, S., Eds.; CEUR Workshop Proceedings, CEUR-WS.org: Aachen, Germany, 2016; Volume 1750.
56. Nguyen-Son, H.Q.; Tran, M.T.; Yoshiura, H.; Sonehara, N.; Echizen, I. Anonymizing Personal Text Messages Posted in Online Social Networks and Detecting Disclosures of Personal Information. *IEICE Trans. Inf. Syst.* **2015**, *98*, 78–88. [CrossRef]
57. Kroll, T.; Stieglitz, S. Digital nudging and privacy: improving decisions about self-disclosure in social networks. *Behav. Inf. Technol.* **2019**, 1–19. Available online: https://www.tandfonline.com/doi/abs/10.1080/0144929X.2019.1584644 (accessed on 7 August 2020). [CrossRef]
58. Nemec Zlatolas, L.; Welzer, T.; Hölbl, M.; Heričko, M.; Kamišalić, A. A Model of Perception of Privacy, Trust, and Self-Disclosure on Online Social Networks. *Entropy* **2019**, *21*, 772. [CrossRef]
59. Kittur, A.; Chi, E.H.; Suh, B. Crowdsourcing user studies with Mechanical Turk. In Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Florence, Italy, 5–10 April 2008; pp. 453–456.
60. Oh, J.; Wang, G. Evaluating Crowdsourcing Through Amazon Mechanical Turk as a Technique for Conducting Music Perception Experiments. In Proceedings of the 12th International Conference on Music Perception and Cognition, Thessaloniki, Greece, 23–28 July 2012; pp. 1–6.
61. Amazon. *Mechanical Turk: Requester Best Practices Guide*; Technical Report; Amazon Inc.: Bellevue, WA, USA, 2011.
62. Terpstra, A.; Schouten, A.P.; de Rooij, A.; Leenes, R.E. Improving privacy choice through design: How designing for reflection could support privacy self-management. *First Monday* **2019**, *24*. [CrossRef]
63. Tinati, R.; Madaan, A.; Hall, W. InstaCan: Examining Deleted Content on Instagram. In Proceedings of the 2017 ACM on Web Science Conference, ACM, Troy, NY, USA, 25–28 June 2017; pp. 267–271.
64. Gazizullina, A.; Mazzara, M. Prediction of Twitter Message Deletion. In Proceedings of the IEEE 2019 12th International Conference on Developments in eSystems Engineering (DeSE), Kazan, Russia, 7–10 October 2019; pp. 117–122.
65. Yates, T. Why is the government relying on nudge theory to fight Coronavirus? *The Guardian*, 13 March 2020. Available online: https://bit.ly/2WYEQGf (accessed on 7 August 2020).
66. Renaud, K.; Zimmermann, V. Ethical guidelines for nudging in information security & privacy. *Int. J. Hum. Comput. Stud.* **2018**, *120*, 22–35.
67. Facchinetti, S.; Osmetti, S.A. A Risk Index for Ordinal Variables and its Statistical Properties: A Priority of Intervention Indicator in Quality Control Framework. *Qual. Reliab. Eng. Int.* **2018**, *34*, 265–275. [CrossRef]
68. Krasnova, H.; Spiekermann, S.; Koroleva, K.; Hildebrand, T. Online social networks: Why we disclose. *J. Inf. Technol.* **2010**, *25*, 109–125. [CrossRef]

Sample Availability: All datasets generated for this study are can be found at the following online repository: https://uni-duisburg-essen.sciebo.de/s/ISyoWPgwEFuIxSE.