Truth or Dare: Understanding and Predicting How Users Lie and Provide Untruthful Data Online

Kopo M Ramokapane
marvin.ramokapane@bristol.ac.uk
Univeristy of Bristol
Bristol, UK

Gaurav Misra
gm@youtility.co.uk
Youtility Limited
London, UK

Jose Such
jose.such@kcl.ac.uk
King's College London
London, UK

Sören Preibusch
acm@soeren-preibusch.de
preibusch.de
Mountain View, USA

ABSTRACT

Individuals are known to lie and/or provide untruthful data when providing information online as a way to protect their privacy. Prior studies have attempted to explain when and why individuals lie online. However, no work has examined into *how* people lie or provide untruthful data online, i.e. the specific strategies they follow to provide untruthful data, or attempted to predict whether people would be truthful or not depending on the specific question/data. To close this gap, we present a large-scale study with over 800 participants. Based on it, we show that it is possible to predict whether users are truthful or not using machine learning with very high accuracy (89.7%). We also identify four main strategies people employ to provide untruthful data and show the factors that influence the choices of their strategies. We discuss the implications of findings and argue that understanding privacy lies at this level can help both users and data collectors.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; **Usability in security and privacy**; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Untruthful data, Privacy lies, Privacy Protective Behaviors, False Information

ACM Reference Format:
Kopo M Ramokapane, Gaurav Misra, Jose Such, and Sören Preibusch. 2021. Truth or Dare: Understanding and Predicting How Users Lie and Provide Untruthful Data Online. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445625

CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445625

1 INTRODUCTION

Personal data is important for various purposes: personalization of services, product development, targeted advertising, revenue predictions, market intelligence, and improved services. However, illegitimate use of personal data has become a threat to data collection, and as a result, users often find themselves having to devise ways of protecting their data. One standard method of protecting personal data is falsification, providing data that is not true [43]. This may be problematic for both users and data processors, as users are being asked to provide data they would not like to, and data processors need to deal with data that may be false. A win/win situation would then be that users are not asked, where possible[1], about data they would not like to provide, and that data processors can have assurances about the validity of the data that is actually provided when asked.

Previous literature [11, 39, 43, 52] on falsification as a privacy protective behavior focuses on the reasons why people lie, the impact of lies, frequency of lying, and encouraging people to tell the truth. Also, there is research looking at whether certain factors like the data being asked [7, 22, 25] and the context in which it is being asked [15, 34, 39], influence providing truthful information. However, the question is, can we really predict, in practice, for a given context and data attribute whether a particular individual is going to provide truthful information or not? In this paper, we show that predicting whether an individual will provide truthful information given a particular type of data and a context is possible using machine learning techniques with high accuracy. The results in this paper quantify the effect of the context on the prediction of truthfulness, as well as the trade-offs involved for service providers to implement such a classifier in practice.

Another limitation of previous literature is that it considers all falsified information the same. However, there is little understanding of exactly what *strategies* people follow to falsify information. Everyday anecdotal evidence suggests that false information may differ, or at least individuals provide false information that may be different. For instance, when requested to provide a home address, an individual may provide partially true information, providing their correct zip or postcode instead of home address.

Mainstream research in this vein refers to data that is not true or falsified as *privacy lies.* However, privacy lies may appear to

[1] Note: Here, we are not talking about data needed for authentication or age-checking purposes for instance when privacy-preserving options are not possible.

judge individuals' intentions when providing such information (e.g., intention to deceive), and such intentions may not always be clear. Therefore, we use in this paper *untruthful data* in general to refer to data that is not true or falsified, which includes privacy lies and we may refer to them as such at some points. Previous studies that have examined privacy lies have investigated why or when people provide such lies; none have explored how these lies are made. Nevertheless, such findings would be of considerable interest. Understanding privacy with this regard is beneficial, as lies can be just half-truth, i.e., users just revealing the part of the requested info they would like to. In this paper, we characterize and show strategies that individuals use to provide false information. We also show that the strategies used depend upon contextual and personal factors. Our main *research questions* are:

RQ1 Is it possible to predict with high accuracy whether an individual is going to provide truthful information when asked about a particular type of personal information?

RQ2 What contextual and personal factors among those used to make the predictions influence the most truthful information providing behaviour?

RQ3 What strategies do users employ to avoid providing truthful information?

To answer these questions, we conducted an empirical investigation in which we asked participants to provide personal information in the context of a web form to purchase discounted movie tickets. This allowed us to vary contextual factors (i.e., users' perceptions) regarding the sensitivity, relevance, and effort in providing information, and we were able to observe how participants reported truthful or untruthful information. In addition, we chose the discounted movie ticket scenario, as it had already been used for privacy experiments [14], yielding high purchase rates, which indicates universal appeal and relevance, and may help avoid bias. To analyze the data, we used both quantitative (machine learning and regression analysis) and qualitative (thematic analysis) methods. Our results suggest that a machine learning classifier can predict with very high accuracy (89%) whether an individual may provide truthful information or not in this environment. Regardless of the participant's general tendency to provide truthful information or not, we also investigated and showed how untruthful information could be classified into four distinct types in terms of the strategy users follow when falsified information. Specifically, we make the following contributions:

- We study factors that affect the disclosure of truthful information in the context of purchasing movie tickets at a discount. We ask participants various questions and their feedback about the data item requested regarding relevance, effort, comfort, and truthfulness. We evaluate these to determine each factor's robustness in determining how likely someone will provide truthful information.
- We develop a truthfulness prediction model using machine learning techniques based on these factors. We show that the prediction model can help determine which questions or requested data items are likely to receive truthful responses. Our study also shows how each model determinant contributes to the prediction model.
- We classify strategies to provide untruthful information into four categories, responses that: do not follow any pattern or format (invalid information), follow a format or pattern but are completely untrue, follow a format or pattern but are partially true, and lastly, those that suggest an unwillingness to answer.

2 RELATED WORK

2.1 Privacy Lies

Several previous studies have attempted to understand *Privacy lies*, i.e. why people lie, how often they lie, and the media they use to lie the most. Lying behaviors often emerge when individuals' privacy concerns rise [39, 46]. Users may also be motivated to lie when they want to avoid unsolicited communication, offline repercussions, harm in general, and data misuse or secondary data use. Some express their strong customer dissatisfaction by falsifying information or perceived anonymity [11, 18, 39, 43, 46]. Poddar et al. [39] found that users may falsify information to distance themselves from the data collector. Moreover, Sannon et al. [43] found that users can lie when they have negative perceptions about the data requester, or feel that the requested information is not needed. Regarding privacy-protective behaviours, Son and Kim [46] found that users may provide false information to misrepresent themselves. While previous studies yield quite interesting findings (e.g., Miltgen and Smith [32] found that falsification can be explained by a trade-off among perceptions of risks, benefits and trust), they viewed falsifications as a general privacy-protective construct.

Hancock et al. [11] found that the highest proportion of lies occurred over the phone and the least on emails, while Sannon et al. [43] found that people provide false information to systems in order to get access to services or avoid being left out. People are more likely to tell self-serving lies in every media to individuals not well known to them [52] but less likely to lie to people who know them, as they are likely to know a lot about them already [39, 43]. These studies also argue that context and relevance contribute significantly to why people lie, but not to how they lie. In this study, we adopt a movie purchasing context to understand privacy lies, and in fact, also ask participants how relevant the requested data is to purchasing movie tickets online. Moreover, these previous studies had also not explored the possibilities of predicting whether individuals may provide truthful or false information given a particular context.

Sannon et al. found that users usually refrain from lying when they find no reason to lie, or, if they view lying as morally and practically wrong [43]. Poddar et al. [39] suggest that users may provide truthful information if they expect the requested information to enhance their future interactions with the service provider. More specifically to questionnaire based scenarios, prior studies [18, 20] argue that giving participants an option not to disclose information (e.g. "prefer not to say") may reduce the number of falsified responses, non-responses, and default selections. For example, Joinson et al. [18] got a low deception rate (less than 1%) for the date of birth when they gave respondents an option not to disclose. With our objective of understanding lying behaviors, we ask respondents how truthful they were when providing answers. This way, we do not make assumptions about their disclosure but use their actual responses to characterize their behaviors.

Regarding social networks, literature has been mainly concerned with issues of trust and reputation, for instance, individuals providing information if they trust the recipient [53]. Others (e.g., Abraham [1] and Krombholz [24]) showed how misrepresentation in social networks can be used for social engineering purposes to steal information from other social media users. Finally, Church et al. [5] provided evidence that most misrepresentation intentions generalize across the socio-economic spectrum and other far-reaching aspects of social network behavior. They also found that social media users expressing greater altruistic desires were less likely to engage in misrepresentation, while competitive desires were associated with increased tendencies towards misrepresentation.

Overall, we find that previous studies attempt to understand disclosure and privacy lies, but they assume privacy lies are all the same. Our work in this paper shows that there are various ways of untruthful strategies, and that they are affected by various factors. Our work complements other prior studies by particularly looking at the different, practical methods people employ when falsifying their information online.

2.2 Information Disclosure Attitudes

Prior studies on data disclosure attitudes suggest that such attitudes are multidimensional; they cannot be defined by a single construct but compose of a relationship between aspects such as sensitivity, trust, context, specific information sought, and individual personality.

Several studies [7, 22, 25] have found that if users perceive data as sensitive, they are more likely to alter it than providing specific information. Moreover, users are likely to provide deliberately vague responses to requests for information they think can personally identify them [8, 22, 50]. Regarding trust, prior studies [19, 35] show that the data recipient's trustworthiness influences data disclosure attitudes. Norberg et al. [35] tested the effect of trust on the intention to disclose information and actual disclosures and found that regardless of how trustworthy the data receiver is, participants had lower intention to disclose than their actual disclosure level. Prior experience with a particular data collector may also lead to more openness towards disclosure [39].

Context and relevance have also been found to play a significant role in users' disclosing behaviors. Based on Contextual Integrity theory [34], many scholars argue that when users perceive data collection or use as appropriate, they are likely to share information. Others [39] found that colors, presentation of information, and the presence of adverts determined how likely participants were to provide information. John et al. [15] investigated the effect of contextual cues when requesting sensitive data and found that disclosures are sometimes responsive to cues that are not connected or even inversely related. For instance, users may disclose more information in unprofessional settings when their privacy concerns are assuaged. They also suggest that collecting data may be successful if data collectors make fewer promises to protect consumer privacy. In non-pecuniary contexts, trust and reduced perception of risks have been found to be the most important factors in reducing the withholding of information [32]. Sundar et al. [47] investigated whether cognitive heuristics can predict information disclosures in contexts that contain cues related to those heuristics. They conclude that contextual cues compel users to disclose more information.

Our work aims to understand how individuals provide untruthful data or lies based on individual data items and how their personal characteristics affect this. However, unlike prior works, we do not attempt to investigate the reasons behind this behavior nor try to classify users' attitudes towards lying. We argue that classifying untruthful behavior is more accurate since one group may lie in a particular way for a specific data item, while another group may adopt several ways to answer the same question.

3 METHODOLOGY

Our goal was to understand how users tell privacy lies and provide untruthful data online using the answers provided when asked for particular types of personal data. For this, we conducted a survey-based study focusing on purchasing movies at a discounted price. This allowed us to have a practical scenario with the possibility to instantiate the contextual variables considered, particularly the relevance of the data asked for. Also, this scenario had already been successfully used for privacy experiments before [14], yielding high purchase rates, universal appeal and relevance, which may help avoid bias.

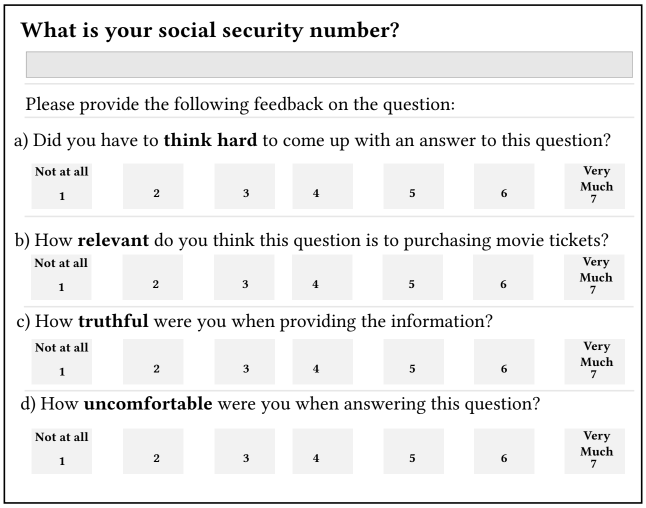
3.1 Statement of ethics

This study was reviewed and approved by our institutional review board (IRB). We did not employ deception nor ask participants to lie or tell the truth while taking the survey. However, we applied a mild form of incomplete disclosure, as we did not completely reveal the study's purpose beforehand. We only revealed who we were and that the study was about the best way to design and test web forms for purchasing discounted movie tickets. We plainly explained in detail how the raw data from the study would be used and handled, e.g., data not being shared or used by anyone else other than us for study (plus other provisions like data retention period).

The only part we did not entirely reveal was that the study was actually about privacy and that we wanted to observe participants' behavior regarding truthfulness. This was clarified at the end of the survey in a debrief statement that explained the study's real intent, as is usual in studies that require initial incomplete disclosure. Since some questions asked for personal information, we contacted and discussed this with Prolific, who approved it. For this, we disclosed to Prolific the study's real aim and shared with them our survey and the debrief statement. We also disclosed that we were not interested in the data itself but respondents' behavior regarding truthfulness.

3.2 Study Design

We designed a survey consisting of six sections: (1) movie purchasing questions; (2) personality questions; (3) reciprocity questions; (4) IUIPC; (5) demographics; and (6) online presence. Our study created an online shopping context (using movie tickets as a representative example) where participants were offered tickets at a discounted price provided they answered some questions, inspired by [41]. There were 50 questions which asked respondents for their personal information (e.g., passport number), information about other people (e.g., employer's full name), and movies (e.g., the last movie they watched at a cinema), which we selected and refined



What is your social security number?

Please provide the following feedback on the question:

a) Did you have to **think hard** to come up with an answer to this question?

Not at all 1	2	3	4	5	6	Very Much 7

b) How **relevant** do you think this question is to purchasing movie tickets?

Not at all 1	2	3	4	5	6	Very Much 7

c) How **truthful** were you when providing the information?

Not at all 1	2	3	4	5	6	Very Much 7

d) How **uncomfortable** were you when answering this question?

Not at all 1	2	3	4	5	6	Very Much 7

Figure 1: Sample question (social security number). Note: i) a response was mandatory for the data asked; ii) questions a-d were presented in random order each time.

with the help of pre-survey and pre-tests, as detailed later in Section 3.3. The 50 questions are in the Supplementary Material for this paper. In addition to each question, we asked about four contextual variables: relevance, truthfulness, effort, and comfort using a 7-point Likert scale "Not at all" (1) to "very much" (7), as shown in Fig. 1. Our goal was to give respondents the chance to answer questions perceived relevant to online purchasing in general and those that are not. Contextual variables (i.e., effort, relevance, comfort) were aimed at sourcing respondents' perceptions regarding each requested data item. We asked about truthfulness so that we could analyze responses reported as untruthful by participants. Prior studies investigating privacy lies do not analyze actual reported data.

Regarding personality, we used the 10-item Big Five Inventory (BFI-10) to measure respondents' "big five" personality characteristics: extraversion, agreeableness, conscientiousness, neuroticism, and Openness [16, 42]. We also used the IUIPC scale [27] to measure privacy concern, and the scale used by DIW [49] to measure reciprocity and personal stability. These were all measured using a 7-point Likert scale. By measuring all these, we aimed to investigate their influence on truthfulness. With regards to demographics, we requested for age, gender, education, and profession. Lastly, concerning online presence, we asked how much time they spent being online, which website they use to purchase online, products they regularly buy, and the number of close friends.

We designed our study so that each participant would be presented randomly with 20 movie purchasing questions and five questions: personality, reciprocity, IUIPC, demographics, and online presence. All the 25 questions were mandatory, but some questions had extra validations or rules to respect. For instance, all the email address questions required the answer to meet the email requirements. Furthermore, not all the questions were open-ended; some questions were closed or with a varying number of options to select from and another option to add a preferred response. For instance, for gender, we provided three options, male, female, and "other" where a respondent can add their preferred answer. All questions were mandatory because we did not want people to skip questions, another privacy-preserving strategy reported by other studies. We aimed to understand how people will respond to answering questions that they would generally prefer not to answer in online forms.

3.3 Procedure

To host and administer our study, we used two platforms, Qualtrics survey platform, to host the study and Prolific platform to recruit participants. We conducted two pre-tests with a sample of 50 participants each to optimize the survey design. We also asked participants to share their views on the sensitivity, relevance, and effort of each data item we requested. This aimed to balance the number of questions perceived as sensitive, relevant, and demanding, as previous works [26] provided some evidence that when users perceive data items to be sensitive, non-relevant and demanding, they are more likely to choose to withhold or provide false information. As a result of the pre-tests, we increased the number of questions perceived as sensitive and added questions about movies (replacing/editing 14 questions in total after the two pre-tests). We excluded all the data collected from both pre-tests studies from the analysis.

Participants willing to take our study were first presented with an information sheet that explained what the study was about and asked for their consent. We then presented 20 questions in a randomized order, with attention checks interspersed after questions 7 and 14. This was to minimize any systematic effects caused by being presented with questions of the same nature. To conclude the survey, we randomly served them with the last 6 questions before asking them personality, character, and attitude questions. The last section was a debrief information sheet concerning their submissions and our data practices.

To estimate the required sample size targeting 99% confidence level in our regression findings, we used an adaptation of Cochran's formula for online surveys [23] and Green's method of estimating sample size based on the number of independent variables in the regression analysis [10]. We estimated that we required at least 670 participants to obtain statistically significant effects from our regression models. The average completion time was 18 minutes, and the participants who completed the study were compensated $3.25 for their time.

3.4 Data Quality and Participants

To ensure data quality, our survey employed popular quality assurance methods [13, 21, 28, 36, 38], attention check questions asking less questions to prevent fatigue. Submissions with at least one failed attention check questions were excluded from the analysis. We also removed submissions that contained inconsistent responses and the second submissions for those who took the study twice. We blacklisted all the participants who took part in the pre-tests not to take the main study.

In the end, 875 participants completed the survey. Thirty-nine (39) participants failed at least one attention checking question and were removed from the study. Six (6) participants took the study twice. We kept their first attempt and removed their second response. We also removed one participant with inconsistent responses, mismatch information regarding their demographics from Prolific platform. Two (2) more participants were removed because they did not meet our criteria (i.e., 18 years of age or above). In total, 48 participants were removed to ensure data quality, and we ended up with 827 participants in total.

Of the 827 participants, 379 and 436 participants identified as females and males, respectively. Twelve (12) identified as other. The average age was 31.6 years (std=10.7, median=29). The mean and the standard deviation for reciprocity were 4.02 and 0.82, while personal stability was 4.20 and 0.87. Moreover, the mean and the standard deviation for IUIPC were as follows: Collection (std=1.44, mean=2.28), awareness (std=1.53, mean=1.88) and control (std=1.39, mean=2.33).

3.5 Analysis

We performed both quantitative and qualitative analyses to address our research questions. First, we use machine learning techniques to predict whether an individual is likely to provide false information (**RQ1**). Then, we used regression analysis to measure the effect of participants' perception of the question and their personal characteristics on truthfulness (**RQ2**). To identify the strategies that users employ when providing false information (**RQ3**), we adopted thematic analysis [6]. These methods are explained in detail in the following sections.

4 PREDICTING TRUTHFULNESS

4.1 Machine Learning for Predicting Truthfulness

We wanted to evaluate whether we can train a machine learning model to predict whether a survey response is truthful or not, based on the data we collected in our study (*RQ1*).

4.1.1 Data Preprocessing. In order to create machine learning models to predict truthfulness, we first decided to binarize the class variable (truthfulness) into a boolean variable. As mentioned earlier in section 3.2, we used a 7-point Likert scale for all variables in the survey, including truthfulness. For classification purposes, we decided to convert this to a boolean variable ("yes"/"no") and create a binary classifier instead of a multi-class classifier. We treated any value equal or higher than 4 on the 7-point Likert scale as having more truth than not, and assigned it a class label of '1'. All values lower than 4 were considered not to be truthful and were assigned a class label '0'.

We had a total of 16,540 instances (i.e., 827 participants answering 20 questions each), out of which a large majority (85.6%) were encoded as the "truthful" class, based on the self-reported truthfulness value provided by the participants. Our classification model included the three contextual variables (i.e., *Effort, Relevance* and *Uncomfortable*) as well as all the variables representing the personal characteristics, introduced in section 3.2. We divided the dataset into training and test (evaluation) sets, using a 'Stratified Sampling' approach [31], which ensures identical class distributions (i.e., truthfulness values of '0' and '1') in both the training and test sets. We created an 80:20 training vs. validation split (i.e., training set is 80% of the total data). Table 1 shows the class distribution for the test and training sets, respectively.

Class	Value	Full Dataset		Training		Validation	
		Count	**%**	**Count**	**%**	**Count**	**%**
Untruthful	**0**	2382	14.4%	1905	14.4%	431	14.4%
Truthful	**1**	14158	85.6%	11327	85.6%	2877	85.6%
Total		**16540**		**13232**		**3308**	

Table 1: No. of instances in the training and validation datasets for predicting truthfulness

Experiment	Instances	Accuracy	AUC	Recall	Precision	F1
Training	13232	89.2%	0.891	0.962	0.917	0.939
Validation	3308	89.7%	0.900	0.962	0.921	0.941

Table 2: Classifier 1 : Results for predicting truthfulness (Light Gradient Boosting Classifier) using all variables

4.1.2 Model Selection and Tuning. We used the Python machine learning framework *PyCaret* [3] to conduct our experiments. We tried several classification algorithms and performed 10-fold cross-validation (using the training set of 13232 instances) for each of them. The results of our experiments (shown in Table 8 in Appendix B) indicate that all algorithms performed at a high level. The high F1-score and AUC (Area under ROC curve, which compares "true positive rate" and "true negative rate") denote that the models produced by all algorithms were of very high quality, despite the large imbalance in the class variable (85.6% "true response" class as shown in Table 1). The best performing algorithm with our data in the cross-validation stage was **Light Gradient Boosting Classifier**, which produced an average accuracy (across 10 folds) of 89.2% and an F1-score of 0.939. We then followed the standard practice of tuning the hyper-parameters of this classifier [12] using the 'Bayesian optimization' technique [45], which is used to find the most appropriate set of hyper-parameters for each classification algorithm based on the data provided, and is faster than a simple grid-search based method, hence resulting in better performance [45]. Once the hyper-parameters were tuned, and the optimum configuration was identified, this configuration was then used to generate predictions on the validation set.

We also tuned the hyper-parameters of all the other algorithms and experimented with the training data (shown in Table 9 in Appendix B) to verify that our choice of algorithm for validation was correct. We found that there was negligible difference between the performance of all algorithms in the optimal configuration (tuned hyper-parameters), which highlights that the choice of algorithm for validation is not central to our solution and practitioners who implement our solution are free to choose any of the other algorithms discussed in Appendix B based on their preference.

4.1.3 Classifier Results. As shown in Table 1, we had a total of **3308** instances to validate our model. The results in Table 2 show that the model performs slightly better during validation (on previously unseen data) than the training phase in all the metrics, which shows that our model did not overfit the data during the training phase, despite the large class imbalance. The high recall value (0.962), also known as "sensitivity" of the model, indicates that we can identify the vast majority of the truthful responses and minimize the

Variable	Model 1: All variables ($R^2 = 0.301$)		Model 2: Contextual variables ($R^2 = 0.296$)	
	Coefficient	p value	Coefficient	p value
Effort	-0.101	0.000	-0.103	0.000
Uncomfortable	-0.447	0.000	-0.448	0.000
Relevance	0.052	0.000	0.054	0.000
Gender	-0.042	0.026		
Age	-0.007	0.000		
Extraversion	0.016	0.093		
Agreeableness	0.070	0.000		
Conscientousness	-0.024	0.108		
Neuroticism	0.027	0.100		
Openness	0.018	0.103		
Reciprocity	0.049	0.004		
Personal Stability	0.055	0.001		
IUIPC-Control	0.035	0.024		
IUIPC-Awareness	-0.048	0.002		
IUIPC-Collection	0.000	0.983		
Online Presence	-0.004	0.305		

Table 3: Regression models to study the effect of contextual and personal information on truthfulness

Experiment	Instances	Accuracy	AUC	Recall	Precision	F1
Training	13232	87.7%	0.863	0.946	0.913	0.929
Validation	3308	87.6%	0.853	0.948	0.910	0.929

Table 4: Classifier 2 : Results for predicting truthfulness (Light Gradient Boosting Classifier) using Effort, Comfort and Relevance

"false negatives" (i.e., predicting a truthful response as untruthful) as a result. The high precision value (0.921) indicates that "false positives" (i.e., an untruthful response being predicted as truthful) were also very low for our model. Overall, it can be seen from the results that our classification model is very effective in predicting whether someone is answering truthfully or not. When analyzing the predictions made by the model, we found it able to detect truthful responses better than untruthful ones, i.e. it has more "false positive" errors than "false negatives". This is due to the class imbalance in our data (85.6% truthful instances, see Table 1) which can be mitigated using class balancing techniques such as synthetic oversampling (SMOTE). Practitioners can decide how they manage the trade-off between prioritizing truthful or untruthful responses in their implementation and can oversample the training data accordingly. We have included an example usage of SMOTE oversampling to mitigate class imbalance in Appendix C.

4.2 Factor Influence on Truthfulness

We wanted to evaluate the effect of the participants' perception of the question as well as some measurable personal characteristics on the (self-reported) truthfulness of their responses (*RQ2*).

4.2.1 Regression Analysis. To answer this question, we created two regression models to study the effects of participants' perception of each question's context, measured by self-reported values for the contextual variables in the survey and their personal characteristics, on the truthfulness of their answers. We used an "Ordinal Logit" regression model where *Truthfulness*, measured on a 7-point Likert scale as explained earlier, was considered the dependent variable. We implemented the model using the "mord" library of Python [37] and conducted two experiments to study the effects of the variables on truthfulness.

In the first iteration of the regression model (Model 1 - All variables), we considered all contextual variables and personal characteristics of participants and examined their effect on the truthfulness of responses ($R^2 = 0.301$). Table 3 shows the value of the β coefficient for each of the variables. It can be seen that being uncomfortable in answering ($\beta = -0.447, p < 0.001$), perceived effort ($\beta = -0.101, p < 0.001$) as well as relevance ($\beta = +0.052, p < 0.001$) play a big role in the user answering a question truthfully. Out of the three contextual variables, comfort has the biggest effect on participants' truthfulness, and the high negative coefficient indicates that participants are more likely to answer untruthfully if they feel uncomfortable in answering it. Some personal characteristics of the user, such as agreeableness ($\beta = 0.07, p < 0.001$), reciprocity ($\beta = 0.05, p < 0.01$), personal stability ($\beta = 0.055, p < 0.01$) and IUIPC-Awareness ($\beta = -0.048, p < 0.01$) also have some effect on the user's truthfulness.

Having observed the relatively higher effect of the contextual variables, i.e., effort, comfort, and relevance, on truthfulness, we created a second model (Model 2 - Effort, Comfort, and Relevance), which only considered these three variables as independent and measured the effect that they had on truthfulness. Table 3 shows that the difference in the values of the β coefficients is minimal and that the overall R^2 value is slightly lower (0.296) than model 1. This suggests that the contextual variables seem to have substantially more influence and contribute more to the model fit than the personal variables.

4.2.2 Machine learning with only Context Variables. The results with Model 2, showing the dominant role contextual variables, i.e., effort, comfort, and relevance, have on the truthfulness value, prompted us to also consider how effective an ML classification model using only these features would be in predicting truthfulness. Therefore, we followed the same method already reported in Section 4.1 but only using comfort, effort, and relevance and attributes for the ML model.

We can see from the results in Table 4 that classifier:2, which only uses the three context variables as features, performs only slightly worse than classifier:1 (Table 2). This is a significant result, as it means that one can predict truthfulness with still a very good performance by just looking at contextual factors associated with the data being asked for, rather than personal factors. This means that data processors have it easier to control what information they will ask for, which may determine how uncomfortable this is for users or how relevant they perceive the information to be, which may impact truthfulness.

4.2.3 Cross-question Influence on Truthfulness. Finally, we sought to explore relationships between questions. We randomized both the subset of questions we would show to each participant. We also randomize the order in which the subset of questions selected for each participant would be presented. Still, it could be possible that, even in these circumstances, there was some influence between questions so that the answers a participant gave to one question would influence how they would answer other questions. In particular, we wanted to explore any relationships between the questions of the survey. The first step was to see if we could gain insight into the participants' attitudes towards answering questions truthfully based on correlation. The mutual correlation coefficients between truthfulness values for the 50 questions is shown in Appendix D. This pairwise correlation coefficients indicate whether a participant is likely to answer two questions with the same degree of truthfulness. For example, a negative correlation coefficient indicates that a participant is likely to answer in the opposite manner (truthful for one and untruthful for the other). In contrast, a positive correlation coefficient indicates that if a participant answers one of the two questions truthfully, they are likely to be truthful in answering the other question.

As shown in Table 5, there are only three pairs of questions for which the correlation of truthfulness was statistically significant. The first pair in the table suggests that a participant is likely to answer the two questions (i.e., social security/national insurance number and country of residence) with opposite truthfulness levels. So, if a participant is known to answer one of those two questions truthfully, there is a good possibility that they would answer the other one untruthfully. For the second pair (sexual orientation and relationship status), the correlation coefficient is positive, indicating that if a participant answers one of them truthfully, they are likely to answer the other one truthfully, similarly, for question 21 and 23.

5 UNTRUTHFUL STRATEGIES

To understand the strategies that participants used to avoid providing truthful information, we coded all the responses which scored less than four (4) for truthfulness. We first extracted and open-coded all the responses that scored three or less for truthfulness to identify recurring themes. First, the lead researcher developed the initial codebook, and then two researchers coded some of the data independently using the same codebook. The first step towards developing our codebook involved verifying some responses which could be verified for existence, e.g., place names, or famous actors. We used various online databases to verify participants' responses where possible, mainly place names, because our sample contained people from Europe and North America, where most place names are on record and available in various databases such as Google Maps. Likewise, when the question concerned movies, actors, musicians, and songs, we searched various databases to verify existence. However, we were careful not to dismiss or mark some responses as non-existence, since we acknowledged that not all items are popular to be on the databases we used. When encountering such instance, we would check whether the response represented a meaningful text or random text with no meaning. This process helped us identify various lying strategies and eliminate the chances of labeling item non-existence just because it was not available in various online databases.

We performed data triangulation using demographics data from Prolific regarding other responses such as age, date of birth, and gender. We compared these responses to see how they differed from the data we got from Prolific, which we assumed it was truthful. This way, we could tell how they lied about their age or gender. This method's primary purpose was not to verify whether participants were truthful or not, but rather how they were untruthful. All the matching (i.e., respondents' answers matching with data from Prolific) were excluded from coding. In the end, our Cohen's Kappa intercoder agreement between the coders was 0.71, which is a high degree of agreement [30].

Ultimately, we all validated the coding scheme through a series of discussions (i.e., arguing to consensus [17]) to resolve discrepancies between the coders. Open coding was followed by an iterating process of identifying relationships and patterns between the codes. This process continued until no more new codes or relationships were observed.

From our analysis and considering all the validations and rules imposed on various questions, we identified three possible privacy untruthful strategies (See Table 6) that people adapt to falsify information when they feel the need to refrain from sharing the truth. The first strategy concerned providing an invalid response to the question being asked. The second is about responses that follow a format or pattern but partially true or completely false. The last strategy discusses responses where the respondents refused to provide an answer to the question being asked. We explain and discuss these strategies in detail below, and we also demonstrate our findings using anchor examples from our data.

5.1 Invalid information

First, we found that when participants provide untruthful data, they may choose to provide information that holds no meaning or follows any pattern or format. Their responses did not conform to any rule, and at times, they were made up of a mixture of letters and numbers. These responses ignored the type of data that was being requested. For example, when asked for the last four digits of their credit number, some participants would provide characters of any length, while when asked for their sexual orientation, they would provide a random string of text containing both numbers and characters. Invalid information also included characters close to each other on the keyboard, e.g., "qwerty" or "asdfgh". Furthermore, we also observed invalid information when participants were given an option to add another response that was not listed in the question, such as gender and ethnicity. Invalid information was not tied to any particular question but to all the questions that did not enforce a format or pattern. However, concerning questions that enforced numeric values only, e.g., phone number, "0" was considered invalid information, considering that we did not enforce maximum length and phone numbers are made up of more than one digit.

5.2 Valid format or pattern

The second way participants provide untruthful data is by giving responses that follow the pattern, format or type of the information being requested. Thus, sometimes untruthful information follows a pattern or format to appear as truthful information. We found that respondents crafted their responses to match what was being asked to match with validations or contexts (without validations). For instance, respondents would not provide a long number for Zip Code but were likely to do so for telephone numbers. We found that lies that conform to patterns and format were either completely untrue or partly true.

Q	Q	Coefficient	p-value
4 - Social Security/National Insurance Number	5 - Country of Residence	-0.249	< 0.001
19 - Sexual Orientation	20 - Relationship Status	+0.195	< 0.001
21 - Receiving a fine	23 - Languages spoken	+0.205	< 0.001

Table 5: Pearson correlation between question truthfulness

Strategy	# Description
Invalid information	Providing an answer which does not have any meaning or follow any pattern or format.
Valid format and pattern	Providing an answer which adheres to a format or pattern but may be partially or completely untrue: • Format and Pattern valid but completely untrue: a response that follows pattern or format but does not carry any meaning or it is completely not true. • Format and Pattern valid but partially untrue: a response that follows a pattern or format but is partially untrue. The respondent can choose to share part of the truth, not whole truth (i.e., suppression) or adding noise to part of information that is true.
Refusing to answer	This is where respondents stated that they will not answer the question or gave an answer that suggested they are not answering the question.

Table 6: Summary of Untruthful Strategies.

5.2.1 Valid and Untrue. Valid and completely untrue is the information that follows a pattern or format, but it is completely false—for instance, someone declaring "fake.email@email.com" as their personal email. We also observed responses that included a run of numbers (e.g., 1234 - the last 4 digits of a credit card), repdigit (e.g., 5555), random text, and phrases like "qwerty" which follows a keyboard pattern or valid for alphanumeric questions but does not carry a meaning with regards to our questions. Other valid responses were existing and common words or names like cancer for illnesses, Windows Explorer for favorite browser. Moreover, we found that some of the responses reported as untruthful conformed to a format and pattern that were related to the context of our study—movies; actors' names, fictional places, and languages—for example, when asked which language they spoke, one participant said "Elvish and Vulcanian", which are fictional languages used in the star trek movie.

5.2.2 Valid and Partially Untrue. Our analysis also revealed that some data reported as untruthful might contain partially true information. A participant may decide to provide part of the true information and withhold the rest (i.e., suppression) or add some false information to the true information (i.e., noise). When suppressing truthful information, some respondents provided broad or vague answers about what was being requested. This was common for addresses or place names. For instance, participants providing their city's name instead of their home address or providing a region/country for a city of residence. This was also common concerning questions that asked for full names; respondents provided one part of the name and left out the rest.

On the other hand, responses with added noise may contain the truth at the beginning, in the middle, or at the end. The noise was mostly added to responses for questions that had two or more parts like date of birth, full names, place names, and addresses. The first part of some of the postcodes provided by participants was valid, while the last parts were not. This way of lying was also common regarding the date of birth where most responses contained the year of birth that was true but with the wrong day and month, e.g., a correct year with "01/01". We also observed the possibilities of noise adding for questions that did not require two-part answers such as phone number and passport numbers. Participants were starting with correct information and ending with lies.

5.3 Refusing to answer

Some responses that were marked as untruthful were actually participants refusing to answer some questions. These answers included responses like "no", "not answering", "prefer not to say", and "not applicable". These responses were commonly used for questions that allowed alphanumeric values and did not require any striking pattern or format—for example, addresses. Nevertheless, despite validations in some questions, such as email addresses, we also found responses about emails that suggested that participants did not want to answer the questions. For instance, "notproviding@email.com" as a personal email and "prefernottosay@survey.com" as work email.

5.4 Use of Untruthful Strategies

Considering the categorization described above, and as shown in Table 7, overall, the most adopted strategy was providing completely untrue information, followed by refusing to answer. The least adopted strategy was providing partially untrue information. Participants mostly refused to answer and provided invalid information when asked for their *social security number* and the *last 4 digits of their credit card number*, respectively. We also observed high occurrences of "refusal to answer" for various questions requesting personally identifiable information, e.g., username for streaming services. Participants rather refused to answer than making up responses int these cases.

Our coding also revealed that participants do not always adopt all the available strategies; their choices may be influenced by other factors such as how they perceive the questions (i.e., comfort, relevance, and effort). When they are uncomfortable answering a question, they favor providing completely untrue information than employing other strategies. For instance, when asked for their passport numbers (i.e., highest average score uncomfortable, shown in

					Valid	
Question	**Total**	**Untruthful**	**R**	**I**	**P**	**U**
Full Name	311	73 (23.5%)	20	7	9	37
Gender	348	4 (1.1%)	0	0	▨	4
Date of Birth	342	51 (14.9%)	▨	▨	21	30
Social Security No.	341	231 (67.7%)	102	22	1	106
Resident Country	302	7 (2.3%)	0	0	3	4
City of Residence	333	20(6.0%)	5	4	0	11
Home Postcode	309	68 (22.0%)	7	8	3	50
Home Address	342	155 (45.3%)	54	10	42	49
Employer Name	313	52 (16.6%)	15	3	10	24
Workplace Zip	338	78 (23.1%)	10	10	3	55
Work Address	307	103 (33.6%)	30	9	20	44
Passport Number	338	224 (66.3%)	43	50	1	130
Personal Email	333	137 (41.1%)	▨	▨	19	118
Professional Email	339	157 (46.3%)	▨	▨	19	138
Phone Number	308	167 (54.2%)	▨	49	1	117
Ethnicity	353	4 (1.1%)	0	0	0	4
Politics	334	6 (1.8%)	1	0	0	5
Religion	342	5 (1.5%)	1	3	0	1
Sexual Orientation	342	11 (3.2%)	4	1	0	6
Relationship Status	310	5 (1.6%)	1	0	0	4
Received fine	318	11 (3.5%)	1	0	0	10
Memorable Event	346	23 (6.6%)	2	1	15	5
Languages	312	4 (1.3%)	0	2	0	2
Serious Illnesses	344	27 (7.8%)	7	1	0	19
Education	341	5 (1.5%)	0	0	0	5

					Valid	
Question	**Total**	**Untruthful**	**R**	**I**	**P**	**U**
Annual Income	309	45 (14.6%)	5	2	0	38
Hobby/Pastime	342	10 (2.9%)	0	1	0	9
Drunk	347	2 (0.6%)	▨	▨	▨	2
Shared "X" movies	309	9 (2.9%)	▨	▨	▨	9
Father’s Full Name	338	112 (33.1%)	21	3	32	56
Sentiments hurt	345	28 (8.1%)	2	0	0	26
Holiday Destination	340	15 (4.4%)	1	1	1	12
Lied about Age	318	5 (1.6%)	▨	▨	▨	5
Fav. Director	337	15 (4.5%)	2	3	0	10
Movie Soundtrack	338	16 (4.7%)	2	0	0	14
Musician/Band	303	14 (4.6%)	2	0	0	12
Emotional Movie	341	12 (3.5%)	0	0	1	11
Web Browser	337	6 (1.8%)	1	0	0	5
Age for Adult movie	342	38 (11.1%)	▨	▨	0	38
Fav. Actor/Actress	340	14 (4.1%)	0	3	0	11
Fav. Movie	341	6 (1.8%)	1	0	0	5
Fav. Movie Genre	311	3 (1.0%)	0	0	0	3
Fav. Song	313	15 (4.8%)	3	1	0	11
Last cinema movie	334	8 (2.4%)	0	0	0	8
Money on cinema	341	4 (1.2%)	0	1	0	3
Online Username	309	95 (30.7%)	50	1	1	43
Illegal streaming	344	11 (3.2%)	▨	▨	▨	11
Cinema Visits	338	5 (1.5%)	0	0	0	5
Adult movies freq.	340	18 (5.3%)	2	2	0	14
Card last 4 digits	337	248 (73.6%)	18	68	0	162

Table 7: Prevalence of strategies for each question; R – Refusing to answer, I – Invalid information, P – Partial untrue, and U – Completely Untrue. Greyed area denotes that the strategy is not applicable because of the enforced validation. Green represents the area where the strategy was applicable but was not adopted.

Appendix A), they avoided partially true responses. Instead, they would choose to provide invalid data. This was also observed when they responded to the *last 4 digit of credit card number* question (i.e., comfort - low).

Completely untrue was also prevalent when responding to questions that contained validations. For this type of question, respondents preferred to provide completely untrue information, such as responses for email addresses. When given a list and a chance to add another option (e.g., gender and ethnicity), respondents preferred not to add extra options or any information suggesting that they are not willing to answer the questions. They preferred to provide completely untrue information. Questions with only two options, *Yes* and *No*, e.g., ‘Lied about age,’ had all the responses completely untrue, since participants could only tell the truth or lie. When the numeric-only format was enforced, we found that most participants preferred the completely untrue strategy over providing invalid responses. Most participants provided “0” as their phone number, and only four participants stated that they were “zero” years old when they first watched an adult movie. Our analysis revealed that questions that permitted alphanumeric responses attracted both completely untrue responses and refusal to answer.

Regarding questions that were perceived to be relevant, respondents falsified less information, but they preferred to be completely untrue rather than adopt other strategies when they did. Analyzing their responses, e.g., movie genre, participants provided a valid or existing movie genre, even though they had a chance to make up their responses. However, for questions they considered irrelevant like ‘fathers’ full name,’ participants provided information that appeared to be completely untrue, such as famous male actors’ names.

Our analysis also revealed that the partially true information strategy was mostly employed when it concerned responses that had two or more parts (e.g., Full names) or could be vague (e.g., addresses). For names, respondents provided responses that appeared to be part of the name or initials. For addresses/zip codes, respondents provided names of cities/countries when asked for home addresses. We also observed partially true (i.e., existing) postcodes for home and work addresses. This was also true for employer names, with participants providing company names or employment sector. Partially true information was also used more often than other strategies regarding questions like memorable events. In this case, participants preferred to share part of the information that is true rather than providing completely untrue information. Their responses covered four themes: (1) celebrations, (2) traveling, (3) mourning moments, and (4) family occasions such as weddings or giving birth. Regarding the completely untrue strategy, one participant stated that they did not know.

Our analysis suggests that the date of birth information was very often partially untrue. Comparing user responses with the demographics data (age from Prolific), we found that some participants may have provided information that is partly true, mainly, the year of birth (count = 21) but with the wrong date or month. The most common day and month combination was "01/01". They mostly said they were younger 16 times (i.e., providing a younger year than the one from Prolific) and older year only six times.

5.4.1 Factor Influence on Untruthful Strategies. We finally sought to explore the influence of each factor on the adopted strategies in a more systematic way to formulate or suggest hypotheses that could be confirmed by future studies around false information. That is, even if our coding process was rigorous with a high inter-rater agreement, as explained before, we did not take the resulting coded data as sufficiently objective or representative for a confirmatory statistical test. Hence, the intent and interpretation of the method described next were purely exploratory and speculative in nature. For this, we created a multi-nominal regression (MN Logit) model where the strategy used was the dependent variable, and the perception of the question and personal characteristics were factors. We used the "statsmodels" library in Python [44] to run the regression experiments.

We present the resulting model in figure 2 showing the regression coefficients which were found to be statistically significant at the 95% confidence level ($p < 0.05$) and the "pseudo R2" value of 0.223 denotes the quality of our model, which is deemed to be a good quality model (values between 0.2 and 0.4 are considered good quality according to McFadden's evaluation of logistic regression models [29]). Still, the results should be interpreted with caution due to the coded untruthful strategies used as the dependent (categorical) variable. Comfort seems to have the most significant influence on the participant's choice of untruthful strategy. As we observed earlier (Section 4.2), participants are more likely to provide an untruthful response when they feel uncomfortable while answering a question. Moreover, these results seem to suggest that participants are also less likely to provide partially true information when they have an option to refuse to answer or can provide invalid information. For example, most respondents provided untrue information when asked to provide their passport number (i.e., low comfort), followed by invalid information and refusing to answer, as shown in Table 7. The model also suggests that when participants perceive the question to be irrelevant to the context, they are more likely to refuse to answer the question. A person's characteristics such as "reciprocity" and "personal stability" also influence the choice of strategy, particularly whether to provide information (valid or invalid) or refuse to answer. For instance, participants with higher personal stability appeared more likely to provide a valid response to a question, even if they deem it to be uncomfortable and/or irrelevant.

6 DISCUSSION

This paper aimed to understand *Privacy lies* and, in general, untruthful data reporting in a practical domain of application—a web form to buy discounted movie tickets. Our result advances the state of the art with regards to untruthful information reporting, as previous research had mainly investigated through regression analysis the effect of different variables in lying behaviour [15, 22, 25, 39, 40] but had not demonstrated the actual high accuracy with which, given a particular individual, data to be asked, and domain of application, it may be possible to predict whether users are going to provide untruthful information or not. And if not, how they will provide untruthful information.

Demonstrating the variation of untruthful strategies itself is but a foundation to the real challenge of discovering the antecedent conditions that fully explain why this phenomenon exists and how the differences between truthful and falsified information disclosures can be addressed. This work provides evidence that, under certain conditions, it is possible to characterize untruthful information and, under some conditions, predict how users may respond to specific questions.

This paper also provides evidence that a machine learning model can be trained to predict truthful and untruthful responses by users. The overall high accuracy of the model, regardless of the classifier used, demonstrates the practicality of implementing such a solution in real-world systems. We also found that a model trained on data excluding any personal characteristics (i.e., considering only comfort, relevance, and effort) provides very good performance, almost as good as the one with personal characteristics. From a practical standpoint, this suggests that service providers could predict whether an individual will answer truthfully without knowing their personal characteristics, but only with their perception of the data being asked for. Service providers could particularly benefit from such predictions without involving personal data, as it would allow them to design systems that collect personal data in a responsible, non-invasive way more easily. Moreover, service providers can identify which questions to ask or avoid to increase their chance of getting better truthful responses. Finally, service providers can control these variables, e.g., making them more or less relevant to the context. An interesting future line of research would be to study the extent to which this control may actually lead to intentional manipulation to increase truthfulness in data reporting, as it has been shown in other privacy-related dimensions, such as information disclosure [2].

Our study showed that participants who reported falsifying information adopted different various untruthful strategies. However, they always had a predominant strategy. We posit that this is because participants make rules for what they can disclose and cannot under any circumstances. We believe this change all the time as users' disclosure behavior changes. This finding supports the idea that users also make rules of engagement online all the time [39] when it comes to providing truthful data or not.

Our results also support the notion that individuals' ability or options to protect their data online is significantly limited, particularly when given no chance not to engage (i.e., prefer not to answer). Moreover, an option not to provide truthful information, partially true information, may still provide enough data to make inferences—engagement may protect one's data to some extent but still with severe consequences. However, for service providers, our findings suggest that falsified information may not all be useless. Specific questions and validations may still encourage some truthful disclosures; thus inferences can still be made from users' responses

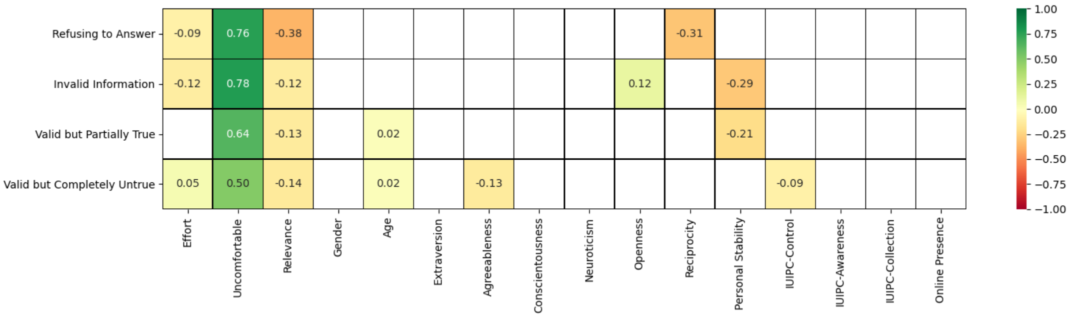


Figure 2: Statistically significant coefficients for Multinomial Logistic Regression model (pseudo R2 = 0.223)

combined with other data. Nevertheless, our findings suggest that to receive more truthful information, service providers may still need to offer users a chance not to disclose information, which is supported by previous studies [18, 51].

Our analysis also revealed that some lies may contain partially true information. A participant may decide to provide part of the information that is true and withhold the rest (i.e., suppression/generalization) or add some false information to the true information (i.e., noise). Interestingly, these strategies seem to directly map to existing Privacy-Enhancing Technologies (PETs), such as the techniques used to achieve k-anonymity [48] (suppression/generalization) and differential privacy [9] (noise addition). Although there are, of course, differences in the way participants used these strategies and their technical counterparts, which are more systematic, the basic intuition behind them is pretty much the same. This has implications in terms of the usability of these PETs, and the understanding that users may have of how they work to protect their privacy, which is known to be challenging [4, 33]. In particular, our results provide evidence that in some cases, users may naturally tend to act similarly as these PETS, and so this could potentially be leveraged to better educate and explain how these PETS work, as a more systematic and safe way to achieve a desired level of privacy.

Our results provide evidence to suggest that validations may influence the choice of untruthful strategies. Future studies, therefore, could investigate how prevalent this is with regards to privacy lies. Furthermore, to identify useful online form designs that could help increase data disclosures, more research needs to be conducted exploring heuristics cues that influence untruthful strategies and which cues instantiate specific strategy.

6.1 Limitations

The machine learning model with the highest predictive power uses variables that web forms may not include or may not be practical to include (e.g. the 10-item IUIPC scale). However, we also show that a model with just three questions about perceptions of data renders almost the same very high accuracy. Another potential limitation is that the design of our study may have encouraged participants to provide falsified information, since the initial information sheet made it clear that honest feedback about questions was sought (and the feedback questions in the survey included truthfulness). This could have suggested to participants that lying was permitted, and they would not be penalized for it. Also, it is possible that some participants could have reported lying while they were truthful or vice-versa. We, however, also believe that by giving participants the chance to disclose whether they were truthful or not is a methodological consideration necessary to understand truthfulness when reporting data, avoid biases, maintain authenticity and improve data quality. We did analyze a very substantial dataset of 2,382 untruthful information items (as reported by the participants) as part of RQ3 to observe the untruthful information reported, which, in most cases, we were able to check their untruthfulness triangulating with other information (e.g. demographics verified by Prolific per participant, online databases of addresses and movies, etc.), which suggests both the predictions and the analysis of untruthful strategies is valid and meaningful. This seems in line with previous research on voluntary self-disclosure [40], where, with some ground truth verification, accuracy was generally high as well. In general, it is worth noting the challenge to study information falsification in the wild, especially as some ground truth may not be available to compare against the information provided.

7 CONCLUSION

This paper explores how people provide untruthful data (including so-called privacy lies), the strategies they use, the factors that influence their choices, and attempted to predict whether an individual would provide truthful data or not. We ran a large-scale study (n > 800), identified four ways individuals use to falsify information, and applied regression and machine learning techniques to predict if/how they lie. Our classifier can predict with an accuracy of 89% whether an individual will provide false information. Moreover, our findings suggest that some false information, depending on the strategy that has been used to falsify it, may still be useful in revealing part of the data.

ACKNOWLEDGMENTS

This research was partially funded by Google through the project awarded to Dr Jose Such "Context and Personality in Privacy Choices".

REFERENCES

[1] Sherly Abraham and InduShobha Chengalur-Smith. 2010. An overview of social engineering malware: Trends, tactics, and implications. *Technology in Society* 32, 3 (2010), 183–196.

[2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[3] Moez Ali. 2020. *PyCaret: An open source, low-code machine learning library in Python.* https://www.pycaret.org PyCaret version 2.1.

[4] Brooke Bullek, Stephanie Garboski, Darakhshan J Mir, and Evan M Peck. 2017. Towards Understanding Differential Privacy: When Do People Trust Randomized Response Technique?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3833–3837.

[5] Mitchell Church, Ravi Thambusamy, and Hamid Nemati. 2020. User misrepresentation in online social networks: how competition and altruism impact online disclosure behaviours. *Behaviour & Information Technology* 39, 12 (2020), 1320–1340.

[6] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* (2015), 222–248.

[7] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. 2005. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 81–90.

[8] Cailing Dong, Hongxia Jin, and Bart P Knijnenburg. 2016. Ppm: A privacy prediction model for online social networks. In *International Conference on Social Informatics*. Springer, 400–420.

[9] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[10] Samuel B Green. 1991. How many subjects does it take to do a regression analysis. *Multivariate behavioral research* 26, 3 (1991), 499–510.

[11] Jeffrey T Hancock, Jennifer Thom-Santelli, and Thompson Ritchie. 2004. Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 129–134.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[13] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48, 1 (2016), 400–407.

[14] Nicola Jentzsch, Sören Preibusch, and Andreas Harasser. 2012. Study on monetising privacy: An economic model for pricing personal information. *ENISA, Feb* 1, 1 (2012).

[15] Leslie K John, Alessandro Acquisti, and George Loewenstein. 2011. Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of consumer research* 37, 5 (2011), 858–873.

[16] Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.

[17] Barbara Johnstone. 2017. *Discourse analysis*. John Wiley & Sons.

[18] Adam N Joinson, Carina Paine, Tom Buchanan, and Ulf-Dietrich Reips. 2008. Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. *Computers in Human Behavior* 24, 5 (2008), 2158–2171.

[19] Adam N Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. 2010. Privacy, trust, and self-disclosure online. *Human–Computer Interaction* 25, 1 (2010), 1–24.

[20] Adam N Joinson, Alan Woodley, and Ulf-Dietrich Reips. 2007. Personalization, authentication and self-disclosure in self-administered Internet surveys. *Computers in Human Behavior* 23, 1 (2007), 275–285.

[21] Yujin Kim, Jennifer Dykema, John Stevenson, Penny Black, and D Paul Moberg. 2019. Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review* 37, 2 (2019), 214–233.

[22] Bart P Knijnenburg, Alfred Kobsa, and Hongxia Jin. 2013. Dimensionality of information disclosure behavior. *International Journal of Human-Computer Studies* 71, 12 (2013), 1144–1162.

[23] JWKJW Kotrlik and CCHCC Higgins. 2001. Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research. *Information technology, learning, and performance journal* 19, 1 (2001), 43.

[24] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. 2015. Advanced social engineering attacks. *Journal of Information Security and applications* 22 (2015), 113–122.

[25] Scott Lederer, Jennifer Mankoff, and Anind K Dey. 2003. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI'03 extended abstracts on Human factors in computing systems*. 724–725.

[26] Miguel Malheiros, Sören Preibusch, and M Angela Sasse. 2013. "Fairly truthful": The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In *International Conference on Trust and Trustworthy Computing*. Springer, 250–266.

[27] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.

[28] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[29] Daniel McFadden. 1974. The measurement of urban travel demand. *Journal of public economics* 3, 4 (1974), 303–328.

[30] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[31] Ross J Micheals and Terrance E Boult. 2001. Efficient evaluation of classification and recognition systems. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Vol. 1. IEEE, I–I.

[32] Caroline Lancelot Miltgen and H Jeff Smith. 2019. Falsifying and withholding: exploring individuals' contextual privacy-related decision-making. *Information & management* 56, 5 (2019), 696–717.

[33] Jack Murtagh, Kathryn Taylor, George Kellaris, and Salil Vadhan. 2018. Usable differential privacy: A case study with psi. *arXiv preprint arXiv:1809.04103* (2018).

[34] Helen Nissenbaum. 2009. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

[35] Patricia A Norberg, Daniel R Horne, and David A Horne. 2007. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs* 41, 1 (2007), 100–126.

[36] Leonard J Paas and Meike Morren. 2018. Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters* 29, 1 (2018), 13–21.

[37] Fabian Pedregosa-Izquierdo. 2015. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses. Université Pierre et Marie Curie - Paris VI. https://tel.archives-ouvertes.fr/tel-01100921

[38] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[39] Amit Poddar, Jill Mosteller, and Pam Scholder Ellen. 2009. Consumers' rules of engagement in online information exchanges. *Journal of Consumer Affairs* 43, 3 (2009), 419–448.

[40] Sören Preibusch, Kat Krol, and Alastair R Beresford. 2013. The privacy economics of voluntary over-disclosure in Web forms. In *The Economics of Information Security and Privacy*. Springer, 183–209.

[41] Sören Preibusch, Dorothea Kübler, and Alastair R Beresford. 2013. Price versus privacy: an experiment into the competitive advantage of collecting less personal information. *Electronic Commerce Research* 13, 4 (2013), 423–455.

[42] Estrella Romero, Paula Villar, J Antonio Gómez-Fraguela, and Laura López-Romero. 2012. Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory (TIPI) in a Spanish sample. *Personality and Individual Differences* 53, 3 (2012), 289–293.

[43] Shruti Sannon, Natalya N Bazarova, and Dan Cosley. 2018. Privacy lies: Understanding how, when, and why people lie to protect their privacy in multiple online contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 52.

[44] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[45] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.

[46] Jai-Yeol Son and Sung S Kim. 2008. Internet users' information privacy-protective responses: A taxonomy and a nomological model. *MIS quarterly* (2008), 503–529.

[47] S Shyam Sundar, Jinyoung Kim, Mary Beth Rosson, and Maria D Molina. 2020. Online Privacy Heuristics that Predict Information Disclosure. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[48] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[49] TNS Infratest Sozialforschung. 2012. SOEP 2010 - Methodenbericht zum Befragungsjahr 2010 (Welle 27) des Sozio-oekonomischen Panels. SOEP Survey Papers 75: Series B. Berlin: DIW/SOEP.

[50] Eran Toch, Justin Cranshaw, Paul Hankes Drielsma, Janice Y Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong, and Norman Sadeh. 2010. Empirical models of privacy in location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 129–138.

[51] Mark Warner, Agnieszka Kitkowska, Jo Gibbs, Juan F Maestre, and Ann Blandford. 2020. Evaluating'Prefer not to say'Around Sensitive Disclosures. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[52] Monica T Whitty and Siobhan E Carville. 2008. Would I lie to you? Self-serving lies and other-oriented lies told across different media. *Computers in Human Behavior* 24, 3 (2008), 1021–1031.

[53] Jyh-Jeng Wu and Alex SL Tsang. 2008. Factors affecting members' trust belief and behaviour intention in virtual communities. *Behaviour & Information Technology* 27, 2 (2008), 115–125.

A CONTEXTUAL DESCRIPTIVE STATISTICS

As discussed in the Method Section, for each question that each of our participants encountered, in addition to their answer to the question (e.g. "What is your relationship status?"), they also had to

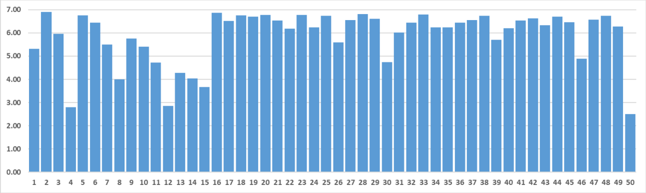


Figure 3: Truthfulness (mean=5.87, sd=1.15)

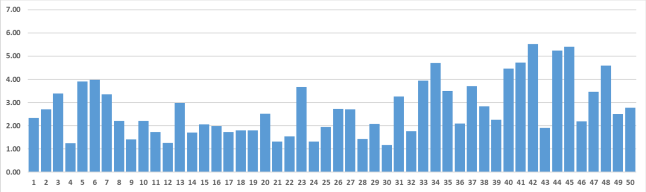


Figure 4: Relevance (mean=2.74, sd=1.19)

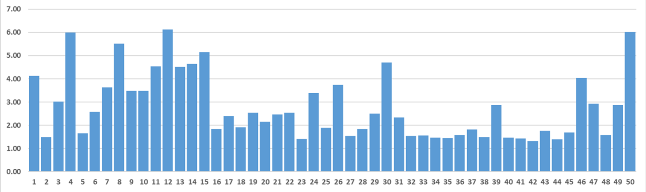


Figure 5: Uncomfortable (mean=2.79, sd=1.41)

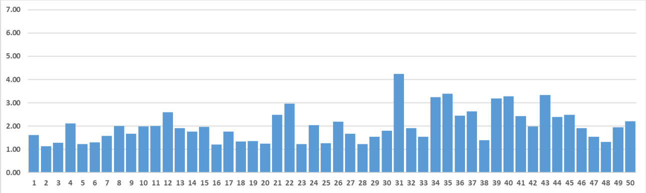


Figure 6: Effort (mean=2.01, sd=0.71)

select values (from a 7-point likert scale) for 4 contextual variables, namely, *Truthfulness, Effort, Uncomfortable & Relevance.* The participants were instructed to select values for these variables based on their perception of the particular data item being requested in that question. The options for all the questions ranged from *"Not at all (1)"* to *"Very much (7)"*.

Truthfulness. The participants were asked to report how truthful they were in answering a particular question - *"How truthful were you when providing the information?"*. Figure 3 shows the distribution of average truthfulness value for each of the 50 questions. It can be seen that most of the questions had an average truthfulness value of more than 5 (mean=5.87, sd=1.15). Question 50 (*"What are the last 4 digits of your credit card number that you use for online shopping (e.g. purchasing movie tickets)?"* had the lowest overall average truthfulness value (2.51).

Relevance. The participants were asked *How relevant do you think this question is to purchasing movie tickets?*. Figure 4 shows the distribution of average relevance values reported by participants for each question. It can be seen that several of the 50 questions in the questionnaire were considered irrelevant by the participants (mean=2.74, sd=1.19) to the given context of the survey, i.e. purchasing online movie tickets. Question 30 (*"What is your father's full name?"*) had the lowest overall average relevance value (1.17).

Uncomfortable. We asked the participants - *"How uncomfortable were you in answering this question?"* for each question they faced in the survey. Note that the 7 point likert scale measures increasing level of discomfort (i.e. low score means the participant was not uncomfortable, and hence comfortable, in answering the question). From figure 5 we can see that most questions in the survey had low values for uncomfortable (mean=2.79, sd=1.41) and that users were mostly comfortable in answering them. We found that the participants were most uncomfortable (6.01) when answering question 50 (*"What are the last 4 digits of your credit card number that you use for online shopping (e.g. purchasing movie tickets)?"*.

Effort. We wanted the participants to reflect on their effort in answering each question they faced in the survey and hence asked them - *"Did you have to think hard to come up with an answer to this question?"*. Figure 6 shows that the participants could answer most of the questions without too much effort (mean=2.01, sd=0.71). Question 31 - *"Tell us the name of one movie that hurt your beliefs or feelings (religious, political, sexual, etc)."* had the highest overall average value for effort (4.25).

B CROSS VALIDATION RESULTS

Table 8 shows the 10-fold cross validation results of all 8 classification algorithms which were experimented with for the truthfulness prediction machine learning model. It can be seen from the figures in the table that all the eight algorithms produce very high performance according to all the metrics, including *F1 score*, which shows high model quality. In terms of comparison of the algorithms, it can be seen that *Light Gradient Boosting Machine (Light GBM)* marginally outperforms the other algorithms for all the evaluation metrics and therefore was chosen for the final model to predict truthfulness.

Algorithm	Accuracy	AUC	Recall	Precision	F1 Score
Light GBM	89.2%	0.891	0.962	0.917	0.939
CatBoost	88.6%	0.881	0.960	0.912	0.935
Extra Trees	88.4%	0.837	0.952	0.916	0.933
XgBoost	88.2%	0.867	0.961	0.907	0.933
Gradient Boosting	88.2%	0.869	0.960	0.907	0.933
Ada Boost	87.8%	0.853	0.954	0.908	0.930
Random Forest	87.8%	0.842	0.943	0.916	0.930
Logistic Regression	87.2%	0.847	0.960	0.900	0.928

Table 8: Cross Validation Results for all algorithms with default parameters (ordered by F1 score)

Table 9 shows the performance of all the algorithms once their hyper-parameters have been tuned while prioritizing higher accuracy. We can see from the numbers that the performance of the algorithms is even closer to each other once all of them are at optimum configuration of hyper-parameters and the difference between them is very low.

Algorithm	Accuracy	AUC	Recall	Precision	F1 Score
Light GBM	89.2%	0.896	0.955	0.919	0.937
CatBoost	89.2%	0.895	0.964	0.915	0.939
Extra Trees	89.2%	0.888	0.965	0.915	0.939
XgBoost	89.2%	0.891	0.970	0.910	0.939
Gradient Boosting	89.2%	0.885	0.964	0.915	0.939
Ada Boost	87.4%	0.864	0.951	0.907	0.928
Random Forest	89.2%	0.891	0.964	0.915	0.939
Logistic Regression	87.3%	0.853	0.957	0.900	0.928

Table 9: Cross Validation Results after tuning hyperparameters (optimizing accuracy) of all algorithms

C OVERSAMPLING

As shown in the paper, the collected dataset had class imbalance, as 85.6% of the instances in our data were coded as being truthful. This resulted in the classifier being able to detect truthful responses better than untruthful ones. We used SMOTE oversampling technique to demonstrate the effect it has on the classifier in this section.

Experiment	Instances	Accuracy	AUC	Recall	Precision	F1
Validation	3308	89.9%	0.899	0.949	0.934	0.942

Table 10: Results for Light Gradient Boosting with SMOTE

True/Predicted	Class 0	Class 1
Class 0	272	190
Class 1	144	2702

Table 11: Confusion Matrix with SMOTE

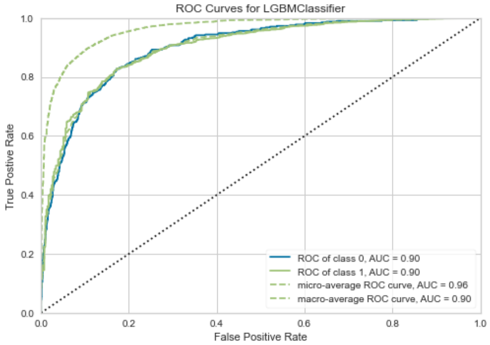


Figure 7: ROC Curve with SMOTE

While we found SMOTE oversampling to be an effective mitigation to the class imbalance problem in our data, resulting in better overall performance (see Tables 10 and 11, and Figure 7), practitioners are encouraged to evaluate the balance of their data and experiment with multiple balancing techniques before using it in practice.

D CORRELATION BETWEEN QUESTIONS

Figure 8 shows a heatmap of pairwise correlation coefficients between questions, which indicates whether a participant is likely to answer two questions with the same degree of truthfulness. For example, a negative correlation coefficient indicates that a participant is likely to answer in opposite manner (truthful for one and untruthful for the other) whereas a positive correlation coefficient indicates that if a participant answers one of the two questions truthfully, they are likely to be truthful in answering the other question as well.

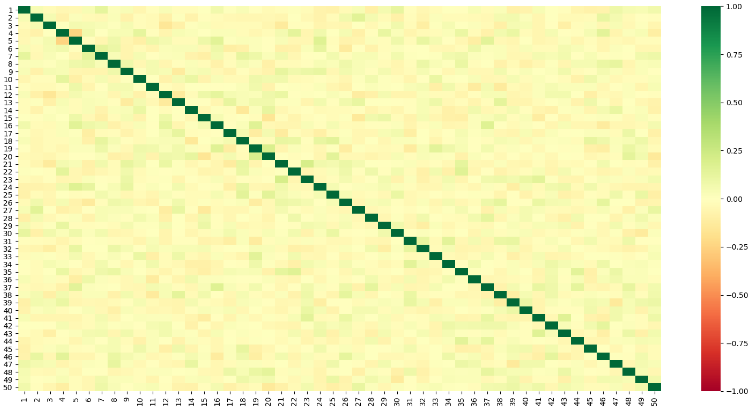


Figure 8: Correlation between truthfulness values of questions