community (Chassang 2017; Mostert et al. 2016; Mourby et al. 2019; Rumbold and Pierscionek 2017; Veale et al. 2018).

As such, obtaining the informed consent of research participants is an important ethical (and potentially legal) milestone to strive for. Yet, there is *further* ongoing debate about whether consent should be one-off and broad, or continuous and specific (Kaye et al. 2014; Luger and Rodden 2013; Morrison et al. 2014; Steinsbekk et al. 2013). With the former, the participant lacks granular control and autonomy over what data is appropriate for sharing with the researcher, and for what purposes. The latter, in contrast, grants more control to the participant – though the cost of this control is an increased burden placed on the participant to manage their data based on what they find appropriate. In short, a trade-off exists between configurability and convenience of consent mechanisms. Baarslag et al. (2017) argue that "there is a pressing need for automating privacy negotiation that can make meaningful decisions on the user's behalf while minimizing their burden".

In this paper, we explore machine learning as a mechanism for automating dynamic consent decisions – striving for high predictive accuracy, while minimising the burden associated with repeated requests. We evaluate a number of algorithms and configurations, and outline potential implications of such an approach. Our work aims to be a case study into how the problem could be technically approached—warts and all—before then deconstructing and critiquing the approach taken in order to provide lessons and insights into this subject.

As such, our findings should be interpreted carefully, and with full hindsight of the intrinsic challenges of predicting consent as outlined in the Discussion section. In this regard, the limitations and biases attempt to serve a cautionary tale for those with an interest in this topic. We believe this work is of particular relevance to the research community, especially those with an interest in research ethics, user studies, social media, privacy, and fair and accountable machine learning.

In this paper, we firstly outline and present the results of a user study (n = 67) to predict dynamic consent decisions with regard to social media data in health research. We focus on this context given the potentially sensitive nature of medical data and its prevalence as an outcome variable in research involving social media data. Secondly, we outline several considerations for model optimisation in a consent prediction context. And finally, we discuss some of the ethical, technical, and practical implications of using such a technique to predict consent decisions based on observations from our study. Our intention is not only to explore whether we *can* predict participant consent decisions, but also under what circumstances we *should* do so.

Background

Informed consent is outlined by the American Psychological Association (2014) as "the process by which researchers working with human participants describe their research project and obtain the subjects' consent to participate in the research based on the subjects' understanding of the project's methods and goals". It is "widely seen as fundamental to medical and research ethics" (Manson and O'Neill 2007), and is a way in which the researcher can fulfil ethical responsibilities with regards to the protection of data, privacy, and the autonomy of the participant (Morrison et al. 2014).

Recent debates have questioned the suitability and practicality of traditional approaches to participant consent in the case of studying online communications. A common argument is that gaining consent from each participant studied in such an environment (which may include many thousands of unique accounts) may be impractical, or even impossible (Hudson and Bruckman 2004; Solberg 2010; Willis 2017), yet research has suggested that individuals in online environments generally do not approve of being studied without their consent (Fiesler and Proferes 2018; Hudson and Bruckman 2004). Reviews of the literature into user attitudes toward the analysis of social media data for research have resulted in "equivocal findings" (Mikal et al. 2016). Despite this, the literature is increasingly reaching the consensus that consent should be sought in such cases (boyd and Crawford 2012; Chiauzzi and Wicks 2019; Conway and OConnor 2016; Hunter et al. 2018; Hutton and Henderson 2015; Rothstein 2015; Zimmer 2010). This is particularly the case when health-related data is involved (Chiauzzi and Wicks 2019; Conway and OConnor 2016; Fiesler and Proferes 2018; Hunter et al. 2018; Norval and Henderson 2017; Rothstein 2015).

Broad, dynamic, and contextual consent

Different approaches for obtaining consent have been defined. The de facto standard in research, referred to as 'broad consent' (Kaye et al. 2014),[6] is typically a one-off, catch-all request conducted at the outset of a study. This approach has come under criticism in the literature for being unsuitable for research using online communications due to its lack of flexibility, transparency, and ongoing participant control (Kaye et al. 2014; Luger and Rodden 2013; Morrison et al. 2014).

As a result, many have argued for a more fluid approach, often referred to as 'dynamic consent' (Kaye et al. 2014). This approach involves giving the participant increased control over their data, including the ability to grant or revoke access to certain data for certain research (Kaye et al. 2014; Luger and Rodden 2013; Steinsbekk et al. 2013). It promotes data re-use (with the knowledge and consent of the individual), and preferences can be modified over time on an ongoing basis (Kaye et al. 2014). It has been described as superior with regard to autonomy, information, increased engagement, control, social robustness, and reciprocity (Steinsbekk et al. 2013). Dynamic consent is not without its own criticisms, however. Steinsbekk et al. (2013) have argued that "broad consent combined with competent ethics review and an active information strategy is a more sustainable solution". One issue of dynamic consent is that repeated requests may lead to 'consent fatigue', potentially risking attrition (Hutton and Henderson 2015; Kaye et al. 2014; Morrison et al. 2014; Steinsbekk et al. 2013).

Some have subsequently sought a middle-ground approach. Hutton and Henderson (2015) have outlined 'contextual integrity consent', based on Nissenbaum's model of contextual integrity (Nissenbaum 2004). This approach looks to consider the contextual nature of appropriate data flow, taking into account factors such

as the type of data, who is requesting it, and why it is being requested. Hutton and Henderson also put forward a statistical approach to inferring when this decision might be automated, so as to not over-burden the participant.

Automated consent procedures

Predictive algorithms have been suggested as a potential solution to the problem of participant burden and informed consent in large-scale observational research studies (Baarslag et al. 2017; Hutton and Henderson 2015; Jones et al. 2018; Norval and Henderson 2017). For example, Hutton and Henderson investigated an approach involving the collection of consent decisions as the participants agreed or disagreed with sharing data for research purposes. As the participant continually answered these requests, the distribution of their answers for each data type was compared to the distribution of other participants. If the participant conformed with these "norms", the decision was automated, reducing participant burden. Hutton and Henderson's results suggest that such an approach may be an effective method of automating consent for those who are "norm-conformant", with such an approach defaulting to a 'Dynamic Consent' approach when conformance could not be assumed.

Gomer et al. (2014) outline a proposal for a semi-autonomous agent using machine learning or a rule-based system for predicting consent decisions. Similar work has looked into automated negotiation agents for incentivised data sharing requests (Aydoğan et al. 2017; Baarslag et al. 2017), suggesting that such an approach can result in statistically higher accuracy compared to random chance (Baarslag et al. 2017). Further calls for such an agent-based approach have been made with regard to consent in the age of the Internet of Things (schraefel et al. 2017).

While using statistical models to predict participant consent decisions may be possible, challenges with such an approach have been raised (Jones et al. 2018; Norval and Henderson 2017). So far, much of the work in automated consent procedures have called for further research to be undertaken in order to understand the predictive capabilities of such approaches. In line with this, our work attempts to build on these findings by exploring the practical and technical implications of such an approach – while simultaneously contributing to the debate on some of the wider ethical questions that such techniques could raise.

User Study

To explore how consent prediction might work in practice, we outline a scenario where we wish to predict whether an individual would find it appropriate for their social data (e.g. photos, status updates, page likes) to be shared with a given audience. More specifically, we are interested in those who use Facebook (due to its relative popularity) for health purposes (be it for information retrieval, participating in online support groups, discussing medical conditions, etc.). This offers an interesting and pertinent use-case, given the interest in social media for health research (outlined previously). Our research hypothesis is that we can predict whether or not a given bit of social data should be shared (dependent variable) based on attributes of the social data itself and of its author (independent variable). This prediction might factor in, for example, the kind of social content in question (e.g. a photo, a status update, a page like), with whom it would be shared (i.e. the audience), and how willingly the individual has approved similar requests in the past.

To create and evaluate a predictive consent model for the above task, we designed a web-based study to collect a corpus of data. Participants were repeatedly asked whether they would find it appropriate to have different types of their social data shared with different hypothetical audiences within a medical context. From this dataset, we went on to train models to predict the appropriate flow of such data, and explored how the different factors of the model each influenced the consent decision.

Application Development

We developed a web application to conduct this data collection study remotely. This application made use of the Facebook API to present participants with social data from their Facebook profile, attempting to tap into the motivation to adequately protect their own social data (Madejski et al. 2012). However, this raised a number of considerations to overcome in order to conduct the study in an ethically aware way.

Given the potentially sensitive nature of the social data, we chose not to collect or analyse the social data itself. Rather, we collected contextual metadata about (i) the social data, (ii) the participant, and (iii) the request in question. While collecting and analysing the data itself (e.g. through image recognition on pictures, sentiment analysis on status updates) may have led to improvements in classification accuracy over metadata alone, such an approach would have a separate set of ethical and practical implications. As such, we consider such approaches to be outside the scope of this particular study.

To mitigate concerns over working with the Facebook data of participants, we used the PRISONER framework, described as an "architecture for ethical and privacy-sensitive social network experiments" (Hutton and Henderson 2013). This framework acted as middleware between our web application and the Facebook API, handling authentication along with the sanitation and collection of data during the study. PRISONER Configuration files specified which data should be temporarily accessible to the application, and which data should be stored for later retrieval by the researcher; The former allowed our application to present the Facebook data in question to the participants, and the latter meant that only metadata was collected and accessible to the researchers. Constraints of the Facebook API meant that only data created by the authenticated participant (i.e. no social data produced by their friends) was accessible to the application, and this did not include content posted to private groups.

We identified a number of potential contextual factors which could be of importance to the decision of whether data sharing was appropriate. This included the participant (who was being asked), the data type (what kind of data was being requested), the audience (with whom it would be shared, and for what purpose). Each item of social data could be categorised as one of the following Facebook data types, based on prior work (Hutton and Henderson 2015) in this

area:[7] (i) 'Liked' Facebook pages, (ii) Status updates, (iii) Location check-ins, (iv) Photos, (v) Photo albums.

We also outlined four hypothetical audiences within an online medical context to explore how this would impact the decision to share or withhold the social data. These audiences included: (i) Researchers, (ii) Clinicians, (iii) Medical support group members (iv) Members of the general public. Explanations and examples of each of these hypothetical audiences were outlined to participants in an information sheet prior to starting the study, in an attempt to mitigate the abstract nature of the request. While we recognise that this study is a narrow representation of consent decisions, we argue that it offers construct validity within a particular scope which has precedence in the literature (i.e. the work by Hutton and Henderson 2015).

Recruitment

The recruitment of participants proved to be one of the major challenges of this study. We began by identifying 50 UK-based medical support groups on Facebook.[8] We contacted the administrators of each of these groups, outlining the aims of our research, the nature of the study, measures taken to protect the privacy of participants, and finally asking if they would be willing to distribute information about the study to their members. This approach was not particularly effective, and even resulted in a few negative responses – further elaborated in the Discussion section of this paper. As such, we also created Facebook adverts targeting those with an interest in health conditions, distributed posters and leaflets at a digital health conference (to people who worked with patients), distributed an invitation to participate on Twitter, and used an academic recruitment website to promote our study. We specifically did not record the recruitment approach for any participant, and so cannot tie any specific respondent to a recruitment method or Facebook group. The study was open to all who met the recruitment criteria, and restrictions were thus not in place to control for demographic attributes. Further implications of our sample are critiqued in the Discussion section, and readers should consider this section carefully before extrapolating the implications of our analysis.

Method

In the first instance, our research proposal and study plan were scrutinised by the ethical review committee of the authors' institution, and approval to proceed with the research was granted. Following approval, we recruited 100 UK-based adults over 18, who self-identified as using social media for health purposes, to participate in the research. 67 completed the study, with the remaining 33 participants either not finishing or actively withdrawing during the process. Participants who completed the study received a £10 amazon.co.uk gift card.

The study consisted of three stages:

Stage 1: After reading the information page and granting consent, participants 'signed in' to the experiment application using the Facebook API, granting the application access to their Facebook data. Participants then completed a demographic questionnaire, collecting information about their age (bucketed, e.g. 18 – 24, 25 – 34), gender, level of education, Facebook privacy setting, Facebook friend count (rounded to the nearest 50 to obfuscate identity), and nationality. All of these values were optional. Participants were also asked for their email address, which was used both as a measure of uniqueness and for distributing the gift cards.

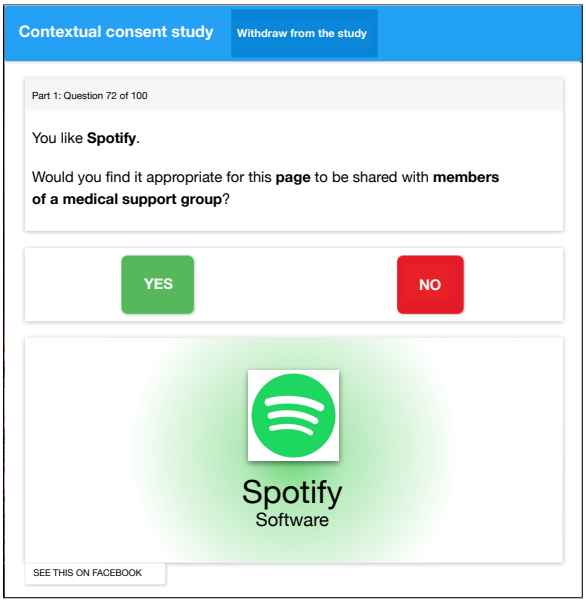

Figure 1. A question from Stage 2 of the study. Participants answer whether they would find it appropriate for a given social item from their Facebook profile to be shared with a hypothetical audience.

Stage 2: Participants were presented with a social item from their Facebook profile, and were asked if they would find it appropriate to have that social data shared with a randomised hypothetical audience (see Figure 1). This was repeated with a different combination of social data and audience, up to 100 times per participant. The combination of data type and audience were pseudo-randomised to achieve a roughly even distribution of combinations (≤ 20 questions for each of the five data types). If the participant did not have enough social data for a given data type (e.g. they only had 10 check-ins), the rest of the questions for that data type were omitted. Data collected from each response included the hypothetical audience, the type of social data in question, the number of likes it received, the number of comments it received, the date and time published (rounded to the nearest hour to obfuscate identity), and the data's privacy setting.

Stage 3: A response from Stage 2 was randomly selected and presented to the participant, and they were asked if that social data contained anything health-related. If the participant had specified that the data should *not* have been shared, they would also be asked to specify their reasons for their answer via pre-defined check-boxes and an open text area (see Figure 2). The reasons given in the pre-defined check-boxes were partly adapted from work investigating reasons for self-censorship (offensive, uninteresting, self presentation) (Das and Kramer 2013; Sleeper et al. 2013), and partly informed by contextual integrity (the content, the data type, the audience). This was repeated for up to 20 of the participant's prior answers.

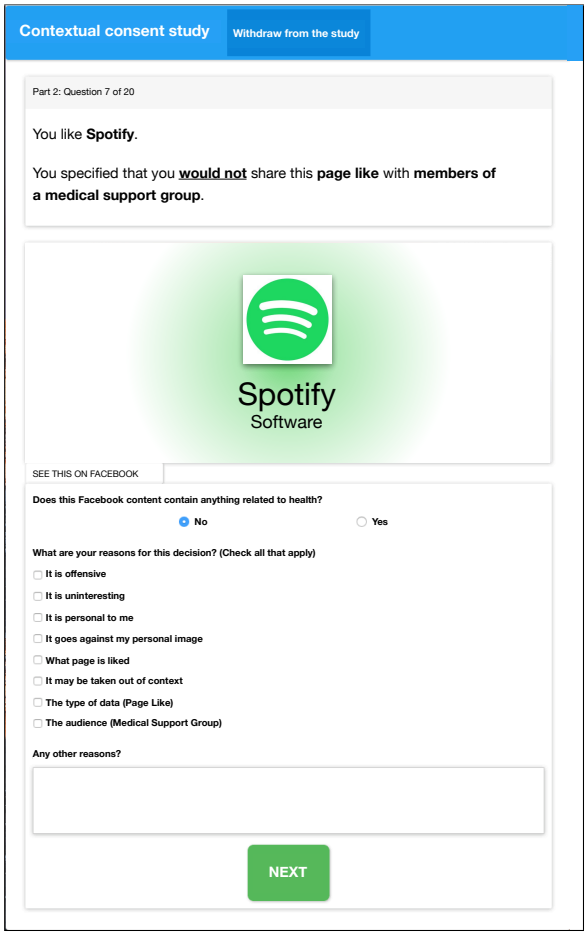

Figure 2. A question from Stage 3 of the study. Participants can specify the reasons why they chose not to share a social item with an audience.

Data Processing and Feature Selection

This study resulted in a dataset of 4,660 consent decisions (stage 2)—of which, 1,027 had contextual reasoning (Stage 3)—from 67 participants. 2,474 (53.1%) of these consent decisions were seen as an acceptable flow of data (i.e. consent was granted). A link to our dataset and analysis code is included at the end of this paper.

First, participants were partitioned into three distinct groups: a training group (n_{pcpt} = 43), a testing group (n_{pcpt} = 12), and a validation group (n_{pcpt} = 12), controlling for the proportion of questions where consent was given per participant. The response data from these groups of participants made up a training set ($n_{\text{responses}}$ = 3,031), a testing set ($n_{\text{responses}}$ = 830), and a validation set ($n_{\text{responses}}$ = 799). This ensured that each dataset consisted of the responses from a distinct group of participants, therefore allowing us to evaluate the generalisability of our approach when testing. In each dataset, responses were 'inner joined' with the demographic information of the participant who answered, so that each row of data contained all relevant information (dependent variables). Missing data for the optional demographic questions were re-assigned using the most frequent responses for each variable.

Second, we computed aggregate data for each participant. We hypothesised that those who consented to a high proportion of data being shared with a particular audience would likely continue to share a similar proportion with that audience in the future. To evaluate this as a potential predictor, we randomly sampled 20% of responses for each participant and used that data to calculate an overall estimated share proportion per participant, and an estimated share proportion for each audience type per participant. The data used to calculate these values was then discarded (since retaining this data for analysis would have raised issues over rigour). Finally, these computed proportions were 'inner joined' with the remaining 80% of response data per participant-audience combination such that each row of data contained (i) the share proportion for the participant in question and (ii) the share proportion for the participant-audience combination (in addition to all other dependent variables).

While these share proportion variables loosely attempted to proxy sharing behaviours over time, we recognise that our data collection was not longitudinal. Further research exploring how sharing behaviours change over long and short terms would therefore be highly complementary. The calculation of this potential predictor also comes at the cost of two downsides. First, we lose a fifth of the response data for all of our participants. And second, any model which makes use of these variables would need to calculate (or estimate) these share proportions as a prior value, known as the 'Cold Start' problem (Park and Chu 2009).

Results: Can we Predict Consent?

These results are broken down into four stages. First, we present the output from a multilevel logistic regression analysis for the purposes of variable inference. Second, we train a number of established binary classification algorithms, evaluating the predictive performance of each. Next, we explore how some model optimisation techniques impact the performance of the models, discussing each within the context of automated participant consent. Finally, we select one model based on the above criteria and validate it with a set of previously unseen participants (the validation set), therefore evaluating how this model might generalise to new participants in the real world.

Sampled sharing proportions appear to be a highly significant predictor of consent decisions. We present the results of a multilevel logistic regression model, using backward stepwise elimination for variable selection,[9] in Tables 1 and 2. The first step contains only variables which relate to the context of the request. The second step then includes demographic data of the participant. The third step then includes the sampled sharing proportions. These two variables are a highly significant predictor, increasing the effect size (McFadden's R^2) from .061 to .286. We present all three steps for the purpose of inference, showing how the inclusion of different predictor variables influenced the model's efficacy.

Following the multilevel logistic regression results, we present a mathematical formula in Figure 3 for predicting whether social media data should be shared. This formula includes all of the dependent variables which were found to be significant predictors in the logistic regression model, and is the set of dependent variables used when training subsequent models in this analysis.

Considerations for Predictive Consent Models

There are a number of factors that should be considered when predicting consent decisions. One such consideration is how different machine learning algorithms affect the predictive efficacy of the models on our testing dataset.

To explore this, a selection of machine learning algorithms were trained and evaluated using the training and testing datasets respectively (using the formula outlined in Figure 3). 10-fold cross-validation was used to reduce the risk of over-fitting, and hyperparameters were optimised via grid search. All analysis was performed in R using the 'caret' package. The results of these evaluations are presented in Table 3, and their ROC curves are illustrated in Figure 4.

Table 1. The first two steps of the multilevel logistic regression model to predict consent decisions. With only contextual and demographic factors, Step 2 does not capture much variance (McFadden's R^2 = .06).

Coefficient	Value	SE	p
Step 1: Contextual			
Intercept	-0.64	0.13	0.000 ***
Data type			
Checkin	0.47	0.13	0.000 ***
Like	1.11	0.12	0.000 ***
Note	0.67	0.12	0.000 ***
Photo	0.40	0.12	0.001 ***
Audience type			
Group	0.06	0.10	0.569
Public	0.02	0.11	0.813
Researcher	0.32	0.11	0.003 **
Published Time			
Evening	0.04	0.11	0.728
Morning	0.19	0.10	0.069
Night	0.21	0.10	0.038 *
	McFadden's R^2 = .027		
Step 2: Contextual & Demographic			
Intercept	-2.45	0.33	0.000 ***
Data type			
Checkin	0.46	0.13	0.000 ***
Like	1.14	0.12	0.000 ***
Note	0.67	0.12	0.000 ***
Photo	0.41	0.12	0.001 ***
Audience type			
Group	0.04	0.11	0.689
Public	-0.01	0.11	0.907
Researcher	0.33	0.11	0.003 **
Published Time			
Evening	0.02	0.11	0.843
Morning	0.18	0.11	0.080
Night	0.22	0.10	0.037 *
Education			
High School	0.88	0.12	0.000 ***
Undergraduate Degree	0.74	0.10	0.000 ***
Postgraduate Degree	0.72	0.13	0.000 ***
Profile Visibility			
Friends Only	1.01	0.30	0.001 ***
Other	0.26	0.39	0.514
Public	2.57	0.58	0.000 ***
Number of Friends	0.00	0.00	0.000 ***
	McFadden's R^2 = .061		

* p < .05; ** p < .01; *** p < .001

$$is\ shared \approx social\ data\ type + participant's\ education + number\ of\ friends + sampled\ overall\ share\ proportion + sampled\ share\ proportion\ with\ that\ audience$$

Figure 3. The formula showing the selection of predictor variables used in follow-up analysis.

A trade-off exists between false positives and false negatives, and these can be re-balanced. A further consideration to binary classification models is that the importance of different performance metrics may often depend heavily on the context of what is being predicted. For example, a medical professional making clinical diagnoses may weigh a false negative (a sick person incorrectly classified as healthy) to be significantly more 'costly' than a false positive (a healthy person incorrectly classified as sick). In line with this, we can therefore explore how our models might perform when taking optimisation steps to minimise the number of false positives (data is shared when it should have been withheld) at the expense of false negatives (data is withheld when it should have been shared).

One method of achieving this cost-sensitive classification involves adjusting the probability threshold for where one class should be selected over the other (Kuhn and Johnson 2013; Sinha and May 2004; Zhao 2008). Given that we wish to lower false positives, we can adjust the probability threshold of each of our models such that the calculated probability of a *Share* outcome must be higher than this threshold value in order for *Share* to be predicted. This could be increased from 50% to 95%, resulting in a more conservative model, with fewer instances where data is leaked undesirably at the cost of lower model sensitivity (i.e. correctly identifying fewer instances where the data should be shared).

The predictions generated during the cross-validation training process were used to retrieve the probability threshold value where specificity was approximately .95.

Table 2. The third step of the multilevel logistic regression model, adding in past share behaviours. This increases the amount of variance that the model can account for from ˜6% to ˜29% (McFadden's R^2 = .286).

Coefficient	Value	SE	p
Step 3: Contextual, Demographic & Share Proportions			
Intercept	-3.30	0.18	0.000 ***
Data type			
Checkin	0.57	0.16	0.000 ***
Like	1.41	0.15	0.000 ***
Note	0.80	0.15	0.000 ***
Photo	0.37	0.14	0.010 *
Education			
High School	0.05	0.14	0.719
Undergraduate Degree	-0.09	0.12	0.477
Postgraduate Degree	0.47	0.14	0.001 **
Number of Friends	0.00	0.00	0.007 **
Total Share Proportion	2.19	0.26	0.000 ***
Audience Share Proportion	2.63	0.20	0.000 ***
	McFadden's R^2 = .286		

* p < .05; ** p < .01; *** p < .001

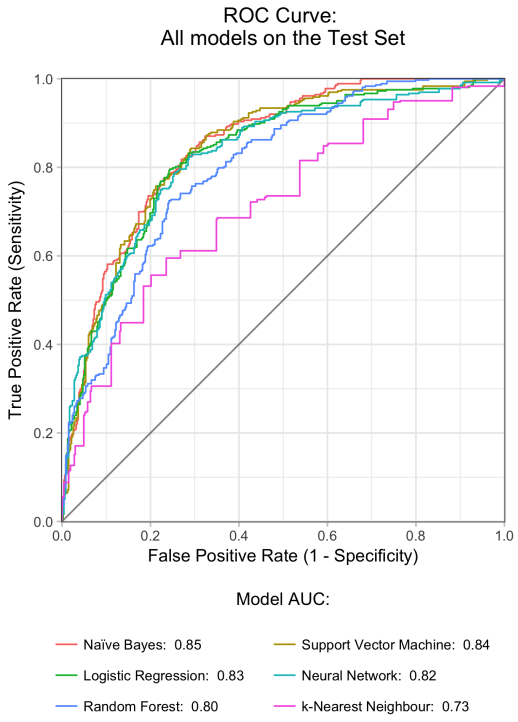

Figure 4. A collection of ROC curves of our models on the test set. Many of these curves appear very similar, with little variance in shape or skew.

This represents the point at which approximately 95% of data that should not be shared would be correctly withheld. We then evaluated each of our models' ability to predict on the testing set when this threshold was used, illustrated in Figure 6. We found that using this threshold adjustment technique reduced the False Positive Rate to approximately 5% for many of our models, though lowering the sensitivity in the process.

Data Leaks can be eliminated at the cost of burdening the participant. If false positives are seen as entirely unacceptable, the participant could simply be consulted every time a 'Share' outcome is predicted – essentially eliminating data leaks through erroneous share predictions. Taking this approach would mean that the proportion of requests requiring participant input would be dependent on how frequently the model predicts that data should be shared. More conservative models would result in lower burden, but also lower sensitivity (true positives). In the worst case (i.e. all data is predicted as 'Share'), this approach would match the burden associated with non-automated dynamic consent (the participant is asked whether each item should be shared).

A model with fewer predictors may have minimal performance impacts, while being more privacy-aware. We can investigate the performance impact associated with following 'data minimisation'[10] principles—collecting and processing less data—by removing some of the predictors from our models. These 'Minimised' models may be more acceptable in situations where participants are concerned about their data being accessed or processed to make a consent prediction. Given that the Share Proportion variables were highly significant predictors of consent decisions (Table 2), we define a 'minimised' formula comprising of the share proportions as predictors, as shown in Figure 5.

$$is\ shared \approx proportion\ shared\ total + proportion\ shared\ with\ that\ audience$$

Figure 5. The 'Minimised' formula.

Given its previous performance (Table 3), we compare the impact of using this formula on the naïve Bayes models for both threshold-adjusted and non-adjusted approaches. Results are presented in Table 4. Given the reduction of information, one might expect that the minimised formula would perform notably worse than the 'Full' formula, outlined in Figure 3. Results of this comparison, however, suggest that this did not appear to be the case. This raises follow-up questions about situations in which participants might prefer the minimised model, and what drop in predictive performance might be seen as acceptable. The use of share proportions as the only predictors, however, does have technical limitations – as will be outlined in the Discussion section.

Selecting and Evaluating a Model

So far, we have evaluated multiple machine learning models (various algorithms; with and without adjusting the probability threshold; two mathematical formulas). However, performing multiple evaluations on the test set increases the likelihood of finding good results by chance. Selecting one 'best-performing' model and evaluating it with the (previously unseen) validation set will give us more confidence in the generalisability of our findings. Yet, what constitutes a 'best-performing' model is heavily dependent on the task in question, and requires us to determine a context under which we would opt for certain attributes of the model (threshold adjusted, full or minimal formula, etc.).

Table 3. Models evaluated on the test set. Many of the models perform comparatively when using the ROC curve (AUC) or the F_1 score as a performance metric, however, the naïve Bayes classifier has the highest specificity – suggesting fewer data leaks.

Model	Accuracy	Precision	Sensitivity	Specificity	F_1 Score	AUC
Naïve Bayes Classifier	0.765	0.827	0.737	0.802	0.779	0.849
Support Vector Machine (RBF)	0.773	0.807	0.786	0.758	0.796	0.840
Logistic Regression	0.755	0.776	0.794	0.705	0.785	0.830
Neural Network (MLP)	0.770	0.804	0.782	0.755	0.793	0.827
Random Forest	0.730	0.786	0.715	0.749	0.749	0.803
k-Nearest Neighbour	0.690	0.708	0.764	0.595	0.735	0.732

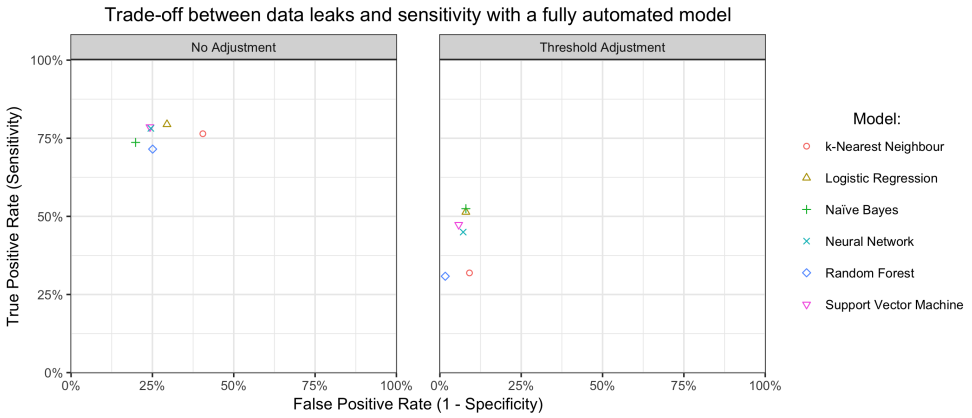

Figure 6. A comparison of the trade-offs between the sensitivity and the proportion of False Positives for different models, both with and without threshold adjustment. The threshold adjusted models attempt to reduce the instances of data leaks (false positives) at the cost of lower sensitivity — which may be an acceptable sacrifice in some contexts.

To evaluate a single model with the validation dataset, we select one that fits the criteria based on what we believe is a promising use-case of this consent prediction technique: Automating the dynamic consent process in a research databank containing social media data. For this hypothetical use-case, participants can answer a few example questions to configure their sharing preferences (i.e. calculate their share proportion variables), and then opt-in to having their social data shared with researchers or clinicians automatically. The system determines which specific data should be accessible to different audiences as requests for research participants are made. We select a model which (i) Minimises data leaks to approximately 5% of requests where consent should not be given, (ii) minimises participant burden beyond the initial calculation of sharing preferences, and (iii) uses the full set of predictor variables that the databank would already have access to.

Based on this criteria, we choose the threshold adjusted naïve Bayes model using the set of predictors outlined in Figure 3 to evaluate our validation dataset of previously unseen participants. The confusion matrix of this evaluation is presented in Table 5. Based on these results, 97.0% of consent requests which should not have been granted were correctly refused. This, however, comes at the cost of sensitivity – only 32.7% of requests where consent would have been granted were correctly predicted as such. Overall accuracy was 65.3%.

Discussion: Should we Predict Consent?

The question of whether or not algorithms *should* predict participant consent decisions is, of course, complex. While an answer to this question is beyond the reach of a single paper, we do believe that it is a question that those with an interest in research ethics need to start thinking about – particularly as researchers have already began to suggest such an approach (Baarslag et al. 2017; Hutton and Henderson 2015; Jones et al. 2018; Norval and Henderson 2017). Automated consent prediction requires a substantial discussion within the research community, along with further research, in order to better understand some of the potential implications. In this section, we attempt to further this topic by outlining some observations raised from our study, and discuss a few considerations which we believe are pertinent to this topic.

There are *very* serious differences between consent prediction with & without the participant's permission. We selected a model for validation based on a hypothetical databank use case. In this scenario, the participant would opt-in (broad consent) to a system which automated their granular sharing decisions (dynamic consent) – with their full oversight. This is a very different scenario to predicting whether someone would consent to participation without their prior knowledge, and without first obtaining their permission. Predicting 'overall' consent (i.e. deciding what social profiles to scrape) without that participant's permission would not absolve a researcher of the ethical and legal implications outlined in the Introduction section of this paper. A clear distinction should be made between instances where overall consent is, and is not, obtained. As such, we argue that *any form of consent prediction should not be used without the participant's prior knowledge and informed permission*.

Prediction need not necessarily mean automation. Of course, also relevant is the purposes for which consent prediction is performed. There are differences between predicting consent decisions and then seeking confirmation from the participant (in the form of active and ongoing dynamic consent) as opposed to outright automating the consent procedure. Of course, regardless, the participants in question should be aware, consenting, and involved in this discussion prior to any predictions or automation taking place.

Most support groups were not willing to share information about this study with their members. Participant recruitment was a significant challenge for our study. Firstly, while the study was designed with careful consideration of the ethical and privacy-related implications of handling participants' social data, we received a low response rate from moderators of the UK-based medical support groups on Facebook. 50 groups were contacted, 15 replied (30%), 9 of which were willing to allow the study to be shared with their members. The 6 groups who responded negatively either expressed the desire to protect their members, specified that their group was not meant for research, or raised specific concerns over the study accessing

Table 5. Confusion matrix for the threshold adjusted naïve Bayes model using the full set of predictors, evaluated using the validation set of participants.

		Predicted value: Share	Predicted value: Do not share	total
Actual value	Share	129	265	394
	Do not share	12	393	405
	total	141	658	

the members' Facebook data.[11] As a result, the number of participants recruited for this study was less than we had originally intended.

Requests for publicly available data were not always granted, confronting the argument of 'implied consent'. We have previously discussed the argument that just because social media data might be publicly available does not mean that it is appropriate to scrape and analyse without consent (boyd and Crawford 2012; Conway and OConnor 2016; Hutton and Henderson 2015; Zimmer 2010). Our data appears to corroborate this viewpoint. Of data which included privacy/visibility settings (only Albums and Notes, as per Facebook's API), 15.4% was publicly avaiable. Of this publicly available data, 41.2% did not receive participant consent deeming it appropriate for sharing. When looking only at instances where the question requested sharing with researchers, this value was 35.7%. In other words, this Facebook data were publicly available, however, the participants did not find it appropriate for that data to be accessed and used by researchers in a healthcare context.

31% of data not shared with researchers was, in part, due to being perceived as uninteresting. We can look at data from Stage 3 of the study to explore reasons why consent was not given, and particularly we can look at cases where data was not shared with researchers. Of data which was withheld, 'It is uninteresting' was checked one third of the time. This could indicate a misconception by participants that data they deem to be irrelevant isn't valuable to researchers. Among other major reasons, 33% was due to the data being personal, 22% because it would be shared with a researcher, 16% due to what the social data contained, and 13% due to the type of data in question.

Requiring consent can lead to selection bias, and this may be an unavoidable cost of ethical research. There is inherent selection bias in any data collected with consent (Fiesler and Proferes 2018). There is irony in that a recognised limitation of our study is selection bias – a consequence which may not have been as much of an issue if less scrupulous methods for data collection were utilised. As an example, the vast majority of our respondents were female (86.6%), though it is not distinguishable if this is due to females being more likely to engage with medical content on social media, or more likely to participate in research, or any other reasons leading to the gender disparity. As such, this raises questions over how well our model would generalise if deployed to different types of individuals.

By asking people if they are willing to participate in a data collection study, the subsequent models generated from that dataset are based solely around the types of people who are likely to agree to take part in such research – those who were uncomfortable in participating are therefore not represented. Nevertheless, researchers need to be aware of the implications, develop alternative recruitment strategies which mitigate this bias, and argue the case that ethical approaches are still worth pursuing.

Performance metrics are only part of the picture, and accurate models may lead to further selection bias. During the model comparison process, we tested a model which used the minimised formula with a threshold adjustment, outlined in Table 4. This model identified 57% of data that should be shared while reducing data leaks to 6.9% of data which should not. From a purely metric-oriented perspective, it appeared to perform respectively. However, in line with the previous subsection, the generalisability of any such approach to consent prediction should be carefully considered and scrutinised. Further investigation into our predictive model identified the potential for unintended consequences, which could have significant research implications, if it were put into practice in the real-world.

Since the sole predictor of our model was the participant's share proportions, and the model was threshold adjusted, the decision of whether consent was granted ended up depending entirely on the participant having previously shared a very high proportion of requests (e.g. $>\tilde{}95\%$). A very small number of participants with a high share proportion had all of their data shared, whereas participants with a lower share proportion ($<\tilde{}95\%$) had none of theirs shared. This led to a model which appeared to have relatively good performance metrics (and few data leaks), however, any research which utilised a dataset generated from such a model would be at risk of severe selection bias. Further, this share proportion

Table 4. Naïve Bayes models evaluated on the test set. Models either use the full or minimised formulas, either with or without threshold adjustments. The threshold-adjusted models have slightly lower accuracy for the full and minimised models. See the Discussion section for recognised limitations of the Minimised + Threshold Ajusted model.

Formula	Cost Adjustment	Accuracy	Precision	Sensitivity	Specificity	F_1 Score	AUC
Full	None	0.765	0.827	0.737	0.802	0.779	0.849
Full	Threshold	0.698	0.894	0.525	0.920	0.661	0.849
Minimised	None	0.745	0.737	0.848	0.612	0.789	0.817
Minimised	Threshold	0.728	0.914	0.570	0.931	0.702	0.817

was calculated from a fixed point in time – no longitudinal data was collected. As a result, it would be easy to assume that it was a proxy of sharing behaviours over time. However, given that it is perfectly feasible for sharing behaviours to change over time, this could lead to incorrect predictions being made if deployed in the real world.

In short, these observations illustrate how conducting solely metric-driven approaches to predicting consent decisions could lead to widely different findings if deployed in the real world. Our model performed seemingly well at a tertiary glance – though upon further scrutiny, we have identified several points of consideration. We raise these points to emphasise that performance metrics for consent prediction aren't everything, and further investigation is vital before deploying a model with such consequences. We intend for these insights to act as a cautionary tale, outlining some of the considerations that those developing consent prediction systems must consider in order to ensure accurate and ethical deployments going forward.

Best Practices

Machine learning practitioners (whether researchers, data scientists, or hobbyists) are facing increasing calls for ethical, transparent, and accountable practices (Singh et al. 2019). We have argued that this is particularly true for predicting when social media data should be used for health research, and so-called 'implied consent' cannot be assumed. More widely, we argue that any form of consent prediction should not be used without the participant's prior knowledge and informed permission. Further, predicting granular consent decisions (i.e. where permission has been given for consent prediction to take place) raises a number of considerations, which we have outlined in the Discussion section.

In all, researchers must consider the wider context in which these models are deployed. While performance metrics may tell a part of the story about the efficacy of a predictive model, it may not accurately reflect the challenges it will face when deployed in practice. Seemingly high-performing models may predict poorly under particular circumstances, or with certain cohorts, and this may not become apparent until data leaks occur and harms result. This paper intends to lay the foundations for discussing some of these considerations, although it is not an exhaustive list, and careful circumspection is therefore advised.

Research Agenda

The present work is outlined as one exploratory case study into consent prediction for social media data in health research. As such, there are many intriguing areas for follow-on research. Firstly, we are careful to stress that our study contains no longitudinal elements. Given our finding that the proportion of social media data that an individual shares is a highly significant predictor of granular consent decisions (Table 2), whether and how this predictor changes over time is an outstanding research question. Indeed, we believe that this is of paramount importance to the concept of consent prediction – consent, after all, is fluid. A greater understanding of the longitudinal implications of this predictor could therefore help prevent the risks of data leakage if such an approach were ever deployed.

Additionally, as discussed, our research specifically explored predictions based on contextual factors – loosely based on Nissenbaum's model of contextual integrity (Nissenbaum 2004). Follow-up work which explores consent prediction using more involved means of data mining (e.g. through image recognition on pictures, sentiment analysis on status updates)—providing it can be done in an ethically appropriate way—may lead to significant increases in accuracy. This would have strong implications on the degree to which such consent mechanisms can be relied upon. Based on our findings, we have argued that our approach should be looked at tentatively. However, alternate approaches which boast exceptionally high predictive accuracy may offer a way forward for consent automation in certain situations.

Other platforms also offer further opportunities for research, given that health information online is not constrained to Facebook. Our study could easily be replicated on other social media platforms with health-related communities, such as Twitter and Reddit. Given the contextual nature of requests, this could go on to help provide researchers with a richer understanding of what participants deem appropriate, where differences exist, and what accuracy can be achieved across these platforms.

Educational Implications

Raising awareness of this topic, along with some of the pitfalls and challenges which may not have otherwise been apparent, will better ensure that those serving on ethical committees can better scrutinise large-scale social media research with automated consent mechanisms. We believe that this is good for researchers and participants alike.

Additionally, we found that one third of publicly available data was deemed by the participants to be inappropriate for sharing with researchers—an empirical confrontation to the belief that publicly available data are fair game—which corroborates the viewpoint that making data public does not equate to implied consent. This has serious implications for researchers, both established and in training, about what participants expect of them.

We also found that 31% of share requests with the 'Researcher' audience-type were refused because they were deemed to be 'uninteresting' by the participant. This could indicate potentially lost opportunities for researchers to recruit participants who have misconceptions, reservations, and/or concerns about how their data are used. We believe that this highlights the importance of ensuring that (to the greatest extent possible) participants are made aware of how their data is used and processed by researchers. This may include providing illustrative examples of the types of data analysed by the researcher (e.g. demonstrating that participant data is processed in aggregate, as opposed to directly identifiable), or indicating what the researchers hope to gain from access to their data (e.g. looking at patterns in sharing behaviour, as opposed to content analysis).

Conclusion

Social media platforms have continued to grow in popularity, making them a valuable resource for researchers. Yet, such social data is often scraped and analysed without the explicit consent of the participant(s) in question, raising significant ethical and legal implications about such research. Even when consent is obtained, current mechanisms are either broad and inflexible, or put the burden of continual data management onto the participant. We have *tenatively* explored a mechanism which attempts to combine broad and dynamic consent approaches by using machine learning to predict appropriate data flow, and we present results of several predictive models, finding respectable accuracy.

Yet, consent prediction remains a highly contextual and complex task, with pertinent ethical and legal implications. We highlight several considerations that those exploring consent prediction systems should consider. For example, while possible to obtain reasonable performance metrics in studies, obtaining representative samples and evaluating performance over time are vital – though pitfalls associated with these biases may not be overly obvious. More widely, we want to raise awareness of such considerations, so that consent prediction—if and when it ever becomes commonplace—might better represent the intentions of our participants going forward.

Acknowledgements

We would like to greatly thank the administrators of the groups who helped us by sharing our study with their members, and all participants who agreed to take part. This work was supported by the Wellcome Trust [UNS19427]. The first author has since received funding from Microsoft through the Microsoft Cloud Computing Research Centre (MCCRC).

Dataset

Our dataset and analysis code are available at: https://github.com/cnorval/automating-dynamic-consent-dataset

Notes

1. GDPR, Arts 13 and 14.
2. Personal data is defined as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;" GDPR, Art 4(1).
3. Note that pseudonymised personal data is still considered personal data, due to the potential for future re-identification. GDPR, Recital 26.
4. The entities responsible for determining the purposes and means of processing the personal data. GDPR, Art 4(7).
5. While Article 14 does outline an exception where "provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes", GDPR, Art 14(5)(b), the onus will be on the data controller to demonstrate the disproportionality with the supervisory authority, should a complaint be made.
6. It is sometimes referred to in the literature simply as 'informed consent', though this is likely more to do with it being the go-to default approach to 'informed participant consent', rather than any statement that it leaves participants any more informed than other consent approaches we will discuss (such as dynamic consent).
7. A sixth data type, identifying the name of a friend of the participant, was omitted due to restrictions of the Facebook API introduced since Hutton and Henderson's original study.
8. Note that despite the health-oriented focus, we make no assumptions about any medical conditions that any participants may have.
9. This removed any variables that were found to be non-significant predictors of consent decisions in each step.
10. GDPR, Art 5(1)(c).
11. It may be worth noting that this study was conducted before the Cambridge Analytica revelations involving allegations of Facebook data misuse through 'personality test' Facebook apps designed to harvest data (Cadwalladr and Graham-Harrison 2018); the evidence is mixed as to how well such studies are addressed by research ethics committees (Schneble et al. 2018) and indeed the personality test studies did not have approval (Weaver 2018). Our study did, however, take place after the earlier 'emotional contagion' Facebook research controversy (Flick 2016).

References

American Psychological Association (2014) APA ethics code addresses when obtaining informed consent from research participants is necessary. Available at: http://www.apa.org/news/press/releases/2014/06/informed-consent.aspx (accessed: 2019-10-11).

Aydoğan R, Øzturk P and Razeghi Y (2017) Negotiation for incentive driven privacy-preserving information sharing. In: An B, Bazzan A, Leite J, Villata S and van der Torre L (eds.) *PRIMA 2017: Principles and Practice of Multi-Agent Systems*. Cham: Springer International Publishing. ISBN 978-3-319-69131-2, pp. 486–494.

Baarslag T, Alan AT, Gomer R, Alam M, Perera C, Gerding EH and schraefel mc (2017) An automated negotiation agent for permission management. In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, pp. 380–390.

boyd d and Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15(5): 662–679. DOI: 10.1080/1369118X.2012.678878.

Cadwalladr C and Graham-Harrison E (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Available at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (accessed: 2019-10-11).

Chassang G (2017) The impact of the EU General Data Protection Regulation on scientific research. *ecancermedicalscience* 11(709): 1–12. DOI:10.3332/ecancer.2017.709.

Chiauzzi E and Wicks P (2019) Digital trespass: Ethical and terms-of-use violations by researchers accessing data from an online patient community. *Journal of Medical Internet Research* 21(2): 1–12. DOI:10.2196/11985.

Conway M and OConnor D (2016) Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology* 9: 77–82. DOI:https://doi.org/10.1016/j.copsyc.2016.01.004.

Das S and Kramer ADI (2013) Self-censorship on Facebook. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

De Choudhury M, Kiciman E, Dredze M, Coppersmith G and Kumar M (2016) Discovering shifts to suicidal ideation from mental health content in social media. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16. New York, NY, USA: ACM. ISBN 978-1-4503-3362-7, pp. 2098–2110. DOI:10.1145/2858036.2858207.

European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Fiesler C and Proferes N (2018) "Participant" perceptions of Twitter research ethics. *Social Media + Society* 4(1): 1–14. DOI:10.1177/2056305118763366.

Flick C (2016) Informed consent and the Facebook emotional manipulation study. *Research Ethics* 12(1): 14–28. DOI:10.1177/1747016115599568.

Gomer R, schraefel mc and Gerding E (2014) Consenting agents: Semi-autonomous interactions for ubiquitous consent. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct. New York, NY, USA: ACM. ISBN 978-1-4503-3047-3, pp. 653–658. DOI:10.1145/2638728.2641682.

Hudson JM and Bruckman A (2004) "Go Away": Participant objections to being studied and the ethics of chatroom research. *The Information Society* 20(2): 127–139. DOI:10.1080/01972240490423030.

Hunter P Ruth F, Gough P Aisling, O'Kane B Niamh, McKeown P Gary, Fitzpatrick M Aine, Walker P Tom, McKinley P Michelle, Lee P Mandy and Kee M Frank (2018) Ethical issues in social media research for public health. *American Journal of Public Health* 108(3): 343–348. DOI:10.2105/AJPH.2017.304249.

Hutton L and Henderson T (2013) An architecture for ethical and privacy-sensitive social network experiments. *SIGMETRICS Perform. Eval. Rev.* 40(4): 90–95. DOI:10.1145/2479942.2479954.

Hutton L and Henderson T (2015) "I didn't sign up for this!": Informed consent in social network research. In: *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.

Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD and Argyle T (2014) Tracking suicide risk factors through Twitter in the US. *Crisis* 35(1): 51–59. DOI:10.1027/0227-5910/a000234.

Jones ML, Kaufman E and Edenberg E (2018) AI and the ethics of automating consent. *IEEE Security Privacy* 16(3): 64–72. DOI:10.1109/MSP.2018.2701155.

Kaye J, Whitley EA, Lund D, Morrison M, Teare H and Melham K (2014) Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics* 23(2): 141–146. DOI:10.1038/ejhg.2014.71.

Kuhn M and Johnson K (2013) *Applied predictive modeling*. Springer.

Lee JA, Efstratiou C and Bai L (2016) OSN mood tracking: Exploring the use of online social network activity as an indicator of mood changes. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16. New York, NY, USA: ACM. ISBN 978-1-4503-4462-3, pp. 1171–1179. DOI:10.1145/2968219.2968304.

Li J and Cardie C (2013) Early stage influenza detection from Twitter. *arXiv:1309.7340* .

Luger E and Rodden T (2013) An informed view on consent for UbiComp. In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13. New York, NY, USA: ACM. ISBN 978-1-4503-1770-2, pp. 529–538. DOI:10.1145/2493432.2493446.

Madejski M, Johnson M and Bellovin SM (2012) A study of privacy settings errors in an online social network. In: *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. pp. 340–345. DOI:10.1109/PerComW.2012.6197507.

Manson NC and O'Neill O (2007) *Rethinking Informed Consent in Bioethics*. Cambridge University Press.

Mikal J, Hurst S and Conway M (2016) Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC Medical Ethics* 17(22).

Moreno MA (2012) Associations between displayed alcohol references on Facebook and problem drinking among college students. *Archives of Pediatrics & Adolescent Medicine* 166(2): 157–163. DOI:10.1001/archpediatrics.2011.180.

Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE and Becker T (2011) Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depression and Anxiety* 28(6): 447–455. DOI:10.1002/da.20805.

Morrison A, McMillan D and Chalmers M (2014) Improving consent in large scale mobile HCI through personalised representations of data. In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14. New York, NY, USA: ACM. ISBN 978-1-4503-2542-4, pp. 471–480. DOI:10.1145/2639189.2639239.

Mostert M, Bredenoord AL, Biesaart MC and van Delden JJ (2016) Big data in medical research and EU data protection law: Challenges to the consent or anonymise approach. *European Journal of Human Genetics* 24(7): 956–960.

Mourby M, Gowans H, Aidinlis S, Smith H and Kaye J (2019) Governance of academic research data under the GDPR—lessons from the UK. *International Data Privacy Law* DOI:10.1093/idpl/ipz010.

Nissenbaum H (2004) Privacy as contextual integrity. *Washington Law Review* 79: 119–158.

Norval C and Henderson T (2017) Contextual consent: Ethical mining of social media for health research. In: *Proceedings of the 1st Workshop on Mining Online Health Reports (MOHRS)*

at the 10th ACM international conference on Web Search and Data Mining.

Ofcom (2018) Adults' media use and attitudes report 2018. Technical report, Ofcom.

Park ST and Chu W (2009) Pairwise preference regression for cold-start recommendation. In: *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09. New York, NY, USA: ACM. ISBN 978-1-60558-435-5, pp. 21–28. DOI: 10.1145/1639714.1639720.

Reece AG and Danforth CM (2017) Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6(15): 1–12. DOI:10.1140/epjds/s13688-017-0110-z.

Rothstein MA (2015) Ethical issues in big data health research: Currents in contemporary bioethics. *The Journal of Law, Medicine & Ethics* 43(2): 425–429. DOI:10.1111/jlme.12258.

Rumbold JMM and Pierscionek B (2017) The effect of the General Data Protection Regulation on medical research. *Journal of medical Internet research* 19(2): 1–6. DOI:10.2196/jmir.7108.

Samuel G, Derrick GE and van Leeuwen T (2019) The ethics ecosystem: Personal ethics, network governance and regulating actors governing the use of social media research data. *Minerva* 57(3): 317–343. DOI:10.1007/s11024-019-09368-3.

Schneble CO, Elger BS and Shaw D (2018) The Cambridge Analytica affair and Internet-mediated research. *EMBO reports* 19(8). DOI:10.15252/embr.201846579.

schraefel mc, Gomer R, Alan A, Gerding E and Maple C (2017) The Internet of Things: Interaction challenges to meaningful consent at scale. *Interactions* 24(6): 26–33. DOI:10.1145/3149025.

Singh J, Cobbe J and Norval C (2019) Decision Provenance: Harnessing data flow for accountable systems. *IEEE Access* 7: 6562–6574. DOI:10.1109/ACCESS.2018.2887201.

Sinha AP and May JH (2004) Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems* 21(3): 249–280. DOI:10.1080/07421222.2004.11045815.

Sleeper M, Balebako R, Das S, McConahy AL, Wiese J and Cranor LF (2013) The post that wasn't: Exploring self-censorship on Facebook. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13. New York, NY, USA: ACM. ISBN 978-1-4503-1331-5, pp. 793–802. DOI:10.1145/2441776.2441865.

Solberg L (2010) Data mining on Facebook: A free space for researchers or an IRB nightmare? *University of Illinois Journal of Law, Technology & Policy* 2: 311–343.

Steinsbekk KS, Kare Myskja B and Solberg B (2013) Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem? *European Journal of Human Genetics* 21: 897–902. DOI:10.1038/ejhg.2012.282.

Veale M, Binns R and Van Kleek M (2018) Some HCI priorities for GDPR-compliant machine learning. *arXiv:1803.06174* .

Vitak J, Shilton K and Ashktorab Z (2016) Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16. New York, NY, USA: ACM. ISBN 978-1-4503-3592-8, pp. 941–953. DOI:10.1145/2818048.2820078.

Weaver M (2018) Cambridge University rejected Facebook study over 'deceptive' privacy standards. *The Guardian*. Available at: https://www.theguardian.com/technology/2018/apr/24/cambridge-university-rejected-facebook-study-over-deceptive-privacy-standards (accessed: 2019-10-11).

Willis R (2017) Observations online: Finding the ethical boundaries of Facebook research. *Research Ethics* 15(1): 1–17. DOI: 10.1177/1747016117740176.

Zhao H (2008) Instance weighting versus threshold adjusting for cost-sensitive classification. *Knowledge and Information Systems* 15(3): 321–334. DOI:10.1007/s10115-007-0079-1.

Zimmer M (2010) "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology* 12(4): 313–325. DOI:10.1007/s10676-010-9227-5.