


# Assessing Trust Versus Reliance for Technology Platforms by Systematic Literature Review

Social Media + Society  
April-June 2020: 1–8  
© The Author(s) 2020  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2056305120913883  
journals.sagepub.com/home/sms  


Trevor Deley  and Elizabeth Dubois 

## Abstract

We do not trust technologies like we trust people, rather we rely on them. This article argues for an emphasis on reliance rather than trust as a concept for understanding human relationships with technology. Reliance is important because researchers can empirically measure the reliability of a given technology. We first explore two frameworks of trust and reliance. We then examine how reliance can be measured by conducting systematic literature reviews of reported success metrics for given technologies. Specifically, we examine papers which present models for predicting private traits from social media data. Of the 72 models for predicting private traits that were surveyed from 31 papers, 80% of the methods reported success rates lower than 90%, indicating a general unreliability in predicting private traits. We illustrate the current applicability of this method throughout the article by discussing the Cambridge Analytica scandal that began during the 2016 US Presidential election.

## Keywords

trust, reliance, social media, systematic literature review

## Introduction

Drawing on Annette Baier's philosophy of trust, we argue that trust can only exist between people, not people and technological entities like Facebook. Instead, most relationships between humans and technology are based on reliance, not trust. Therefore, when talking about trust, we must ask questions about relationships between people, and when talking about technology we must ask questions about reliance. It follows that "trust in social media" then has to do with how our reliance on technology shapes our beliefs about trust toward the people who control that technology.

To illustrate this argument, we use the case of the Cambridge Analytica (CA) scandal before turning to our systematic literature review of peer-reviewed articles which present models for predicting private traits based on social media data. In our review of these papers, we focus on the ways in which reliance and reliability are, or are not, assessed. We conclude with a discussion of the implications of shifting our discussion from trust to reliance and what that means for researchers aiming to use predicted private traits from social media data.

## The CA Scandal

In March of 2018, Facebook founder and CEO Mark Zuckerberg published apology letters in seven British and three American newspapers, formally apologizing for a data leak that he called a "breach of trust." The ads were in response to what is now known as the CA scandal where data collected from a third-party app was used to predict the personalities of voters. The predicted personalities were used for shaping personalized ads, but also to allegedly engage in voter suppression tactics (Solon, 2018).

In mainstream media coverage, the events surrounding CA's involvement in the 2016 US elections were covered several times before they were deemed a scandal (Davies, 2015; Grassegger & Krogerus, 2017). It was not until *The Guardian* and *The New York Times* jointly broke the story

University of Ottawa, Canada

### Corresponding Author:

Trevor Deley, Department of Communications, University of Ottawa,  
Ottawa, Ontario K1N 6N5, Canada.  
Email: tdele013@uottawa.ca



of whistleblower Christopher Wylie that it became a scandal (Cadwalladr & Graham-Harrison, 2018; Rosenberg et al., 2018). It was framed as a data breach, or data leak even though users consented to share their data through a third-party Facebook application, though most users likely did so unknowingly (Glum, 2018). After Zuckerberg's apology, Facebook implemented the European Union's (EU) General Data Protection Regulation for its users globally (Rahman, 2018).

The technical details are that CA used a third-party Facebook application developed by Aleksandr Kogan to have Facebook users fill out personality tests. The application also collected participant Facebook likes, public profile, location, and birthday. These data were used to predict personality scores of individuals based on their Facebook likes—a method partially pioneered by Michal Kosinski and his team at Cambridge (M. Kosinski et al., 2013). Once the models were trained, personalities of novel users could be predicted simply from collecting their Facebook likes. By calculating aggregate personalities for different electoral districts, CA could estimate what kind of advertising would be most effective in a given region (e.g., ads depicting breaking and entering would sell better to groups who scored high for "neuroticism" within the OCEAN personality assessment) (Concordia Summit, 2016).

### *Trust Versus Reliance*

Annette Baier argued that trust can only exist in relationships where there is a possibility for betrayal, and that we cannot truly form trust relationships with technology because technology cannot "betray" us in the exact sense of the word (Baier, 1986). We do not trust technologies like we trust people, rather we rely on them and they can succeed or fail. Technology disappoints us but does not betray us when it fails. Baier defines trust as a special case of reliance, where one party is relying specifically on the good will of the other party. Reliance, on the other hand, is defined as a continued relationship on the basis of one party's dependable habits toward the other (Baier, 1986). Insofar as technologies cannot possess a good will there is no possibility for a genuine relationship of trust in Baier's view. Technology rather produces dependable habits that we come to rely upon or not. In contrast, it is the good will of the creators and managers of a given technology that we either trust or do not trust.

In Baier's view, Facebook as a platform is not eligible to betray anyone but can only fail to fulfill dependable habits or expectations of users. It would only be individuals such as those within Facebook who could willingly betray users in some that would constitute a breach of trust. For example, in Mark Zuckerberg's apology ads, the headline read, "We have a responsibility to protect your information, if we can't we don't deserve it." His statement indicates personal responsibility for a breach of trust, but who is the "we" he is referring to? Both media framing and elements of Zuckerberg's statement are incongruous with Baier's definitions of trust. The abstraction of online privacy in this case is difficult to pin down,

especially when each user's privacy was not placed in the explicit trust of an identifiable individual but rather distributed among an anonymous "we" which in fact was a technological system that had been relied upon and failed. When looked at from Baier's perspective, the diffusion of responsibility within Facebook nullified the preconditions for a trusting relationship even though popular media and Facebook both referred to a certain breach of trust.

What we will argue is that reliability is a mediator of trust toward the makers of a technology and that relationships of reliance mediate our relationships of trust.

### *Mediators of Trust*

In a thorough literature review, McKnight and Chervany synthesize 65 articles and monographs that contained trust constructs from psychology, social psychology, sociology, communications, political science, management science, and economics to create a reconciled typology of trust (Harrison McKnight & Chervany, 2007). Trust is categorized according to "trust referent" properties, or the characteristics a trustworthy person would possess; and according to concepts of trust, which is a catalog of how different disciplines study trust.

McKnight and Chervany sorted characteristics in the trust referent category to find four overarching characteristics trustees generally possess: benevolence, integrity, competence, and predictability. Benevolence was defined as "caring and being motivated to act in one's interest rather than acting opportunistically," integrity as "making good faith agreements, telling the truth, and fulfilling promises," competence as "having the ability or power to do for one what one needs done," and predictability as "trustee actions (good or bad) . . . are consistent enough to be forecasted in a given situation" (Harrison McKnight & Chervany, 2007). Competence and predictability in this typology are very similar to Baier's notion of reliance being "one party relying on another's dependable habits." Meaning that in both frameworks, our reliance on technology is dependent on it doing what we need done, and predictably.

McKnight and Chervany grouped concepts of trust from the literature into three overarching categories: dispositional concepts, institutional concepts, and interpersonal concepts. Dispositional concepts cover the "disposition to trust" in individuals, or the reasons that cause a person to be "willing to depend on others generally." Institutional concepts refer to institution-based trust and the study of beliefs about protective structures that enable people, who otherwise might not trust each other, to trust each other. Interpersonal concepts generally describe conditions that often exist between trusting parties in three subcategories: trusting intentions, trusting beliefs, and trusting behaviors. Trusting intentions describe when a person is willing to depend on a trustee, trusting beliefs describe when a person believes a trustee has characteristics that engender their trust, and trusting behaviors describe instances where a person voluntarily depends on the trustee. Trusting intentions, beliefs, and behaviors all assume that the

trustor has a feeling of relative security despite not having control over the trustee where negative consequences are possible (Harrison McKnight & Chervany, 2007).

The McKnight–Chervany typology is very simple once definitions are in place. In this typology, trust is described by properties that make something appear trustworthy, factors that influence one's disposition to trust, and the availability of institutional trust (e.g., protective structures in society that are generally fair). The presence or absence of these elements will affect one's intention to trust others, their beliefs of others trustworthiness, and ultimately their trust-related behaviors. The strength of belief in a trustee is then modeled by perceiving the four trustee characteristics they derived from their literature review: benevolence, integrity, competence, and predictability. Therefore, issues with competence and predictability of a technology negatively impact our trust beliefs toward the makers of that technology. This is helpful because competence can be measured by looking at how successful a given technology is in its task. Meaning we can develop empirical proxies for how trustworthy an entity like Facebook may be based on how successfully it performs certain tasks that we rely on. For example, if it turned out that Company A was very bad at predicting things about users for the purpose of serving ads, but did so anyway, then they may not be very trustworthy on the basis that they are not particularly competent.

### Measuring Reliance

One way to evaluate competence is by evaluating the success of a method; for example, in predicting a trait like personality scores. Generally, there is a variable or class that is to be

predicted such as a person's personality type. After the prediction is made, the correct answer is recognized by looking at that individual's measured scores or ground truth data. In classification tasks, a comparison between the prediction and the ground truth data produces a confusion matrix with four possible outcomes (Table 1). If the trait is classified as positive and has a ground truth positive value that is a true positive or a correct classification. Conversely, if the individual is classified as negative, and has a ground truth negative value that is a true negative. The other options are false classifications or errors. A type one error occurs when an example is classified as negative but has a ground truth positive value, this is a false negative. A type two error occurs when an example is classified as positive but has a ground truth negative value, this is a false positive.

There are many ways to combine confusion matrix scores into assessments for the success and effectiveness of prediction measures, many common uses and the uses seen in this literature review have been outlined below (Table 2).

To review, Annette Baier's philosophy provides philosophical criteria to exclude ineligible trustees by distinguishing between trust and reliance. From the McKnight–Chervany

**Table 1.** Confusion Matrix for Classification Tasks.

	Positive classification	Negative classification
Ground truth positive	True positive (tp)	False negative (fn)
Ground truth negative	False positive (fp)	True negative (tn)

**Table 2.** Definitions of Evaluative Measures for Machine Learning Tasks as per (Sokolova et al, 2006).

Measures	Word definition	Equation
Accuracy	True classifications over all observations	$Acc. = \frac{tp + tn}{tp + fp + fn + tn}$
Specificity	True negatives over all actual negative observations	$Spec. = \frac{tn}{fp + tn}$
Precision	True positives over all positive classifications	$Prec. = \frac{tp}{tp + fp}$
Recall	True positives over all actual positive observations	$Rec. = \frac{tp}{tp + fn}$
F-score	The harmonic mean of precision and recall. Often weighted by a factor of $\beta$ when precision or recall is of special interest	$F_{\beta} = \frac{(\beta^2 + 1) * Prec. * Rec.}{\beta^2 * Prec. + Rec.}$
Receiver-operating characteristic (ROC)	The ratio of conditional probabilities of positive classification over negative classification, given the data entry $x$ . Found by plotting the recall against the false-positive rate ( $1 - Spec.$ ) at various thresholds	$ROC = \frac{P(x   positive)}{P(x   negative)}$
Area under the curve (AUC)/balanced accuracy	The area under the ROC curve	$AUC_b = \frac{(Rec. + Spec.)}{2}$

typology, researchers can operationalize assessing said party's trustworthiness by evaluating the competence of a given technology via a confusion matrix or similar methods. For example, in this framework "Facebook" as a platform is not an eligible trustee because "Facebook" as a technology stack has no good will and can offer no relationship with the possibility for betrayal. However, executives at Facebook, or Facebook's newsfeed team, and so on, all would count as eligible trustees. While assessing the trustworthiness of Facebook as a platform is not a reasonable task for this framework, the reliability of aspects within the Facebook platform do mediate our perceptions of benevolence, integrity, competence, and predictability of Facebook executives and engineers and whether it is reasonable to trust them or not.

### Systematic Literature Review

Competence can be assessed through systematic reviews of certain technologies that summarize confusion matrix data. For example, while it is very difficult to assess the benevolence and integrity of a potential trustee in advance, competence of a potential trustee and thereby the reliability can be assessed. For example, in the case of CA say, we know nothing about the benevolence or integrity of CA in advance, but we know they are predicting private traits using machine learning models and Facebook data. In this case we cannot directly assess CA's models, however, by assessing the reliability of similar models, we can make better judgments about the competence, and trustworthiness of actors like CA, which should help in the project of creating more trustworthy social media platforms. The next section will illustrate a systematic literature review of the prediction of private traits and their reliability.

While many literature reviews have been done, and even systematic literature reviews exist on the specific topics of Twitter analysis, sentiment analysis, and opinion mining in political contexts (Nakov, 2017), there is no systematic literature review of methods for private trait prediction using social media data in a political context, or the respective error rates of said metrics. Individual metrics in this context refers to any trait predicted about an individual user that can be estimated without that user explicitly volunteering that information. Error rate in this context refers to definitions outlined in Table 2. In our analysis, we address the following research questions:

First, we aim to examine which types of private traits are being predicted and how to map the current landscape:

*RQ1.* What personal traits about individuals can be predicted using social media data?

Next, we wanted to know if and how researchers measure the success and reliability of their metrics:

*RQ2.* How were error rates reported?

While no direct comparison can be easily performed between different measures of success without access to the original data from each experiment, it is possible to form a general sense of how well private traits can be predicted which became the basis of our third research question:

*RQ3.* What were the error rates for predicting private traits from social media data?

### Screening Methods

Four databases were queried: Scopus, Web of Science, Communications and Mass Media Collections, and Communication Source. The former two were chosen due to their comprehensive nature, and coverage of scientific fields. The latter two were chosen to specifically investigate what has been written about this topic within the communications field.

Kosinski's research papers were used as a benchmark for our search query due to its centrality to CA's work. This meant the final search results had to successfully include Kosinski's work.

### Search Query

This was the general search query used in each of the databases:

(Individual\*) AND ("Social media" OR Facebook OR Twitter OR Reddit OR "Digital Footprint") AND (predict\* OR model\* OR "machine learning") AND ("private trait" OR personality) AND (metric)

Queries were limited to peer-reviewed articles between the period 2007–present. The year was chosen because 2008 was the first major election in the United States that used sentiment analysis of social media data as part of an election strategy (Carr, 2008). English and French articles were included as well.

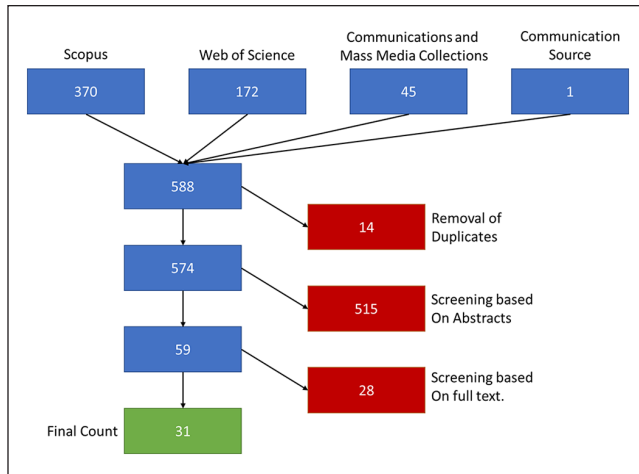
### Screening

For screening the initial publications, the titles and abstracts were read. Papers were excluded if from the abstract it could be determined that

1. There was no predictive model used;
2. The predictive model was not derived from digital footprint data;
3. The predicted measure was not an individual-level measure;
4. The predicted measure was not a private trait.

From an initial 574 publications, the pool of eligible papers was cut down to 59 papers which were read for





**Figure 1.** PRISM chart of paper screening process, and inclusion/exclusion steps.

further screening. The same criteria were applied to the 59 papers which left 31 papers that predicted an individual-level predicted private trait derived from digital footprint data (Figure 1).

## Results

### RQ1

From the 31 eligible publications, 25 different predicted private traits were described with reported measures of success ranging widely both in quality and chosen measurement. There was a heavy focus in the papers on personality prediction from Facebook and Twitter data (Table 1), with 16 of the 31 papers being about personality prediction. A large reason for this can be attributed to the use of the “My Personality” dataset that was collected and is available for research use by one of the prominent authors in this field, Dr Michal Kosinski. Another major area of interest was in predicting demographic data. Personality was grouped into the category “psychological factors” which had 10 traits including “personal well-being,” “temporal orientation,” and “intelligence.” The remaining private traits were grouped into four groups: preferences which contained six traits such as “political orientation” and “sexual orientation,” health outcomes which contained the trait “substance abuse,” economic outcomes which contained two traits “consumer behavior” and “financial outcomes,” and demographics which contained traits like “age” and “ethnicity” (Table 1).

The ability to predict private traits seems to be determined by the ability of researchers to model the desired trait in a psychometric test. Many studies that were attempting to predict complex private traits like mood or personality used a psychometric test as ground truth data for their predictive models. Examples include the OCEAN big five personality traits, the Myers–Briggs personality test, IQ evaluations, and

Beck’s Anxiety Inventory. Other studies, however, used judges to annotate text, photos which then served as ground truth data.

### RQ2

Twenty of the 31 papers only reported one measure of success, of which 14 were accuracy as defined in Table 2. Nineteen papers in total reported some measure of accuracy, six papers reported area under the curve (AUC), four papers reported precision, four papers reported recall, four papers reported an F-score, six papers reported a Pearson correlation coefficient for continuous predicted values, one paper reported precision, and several other measures were included in papers such as mean squared error and kappa statistics. Overall, the large use of accuracy as a success metric makes comparison easier, but masks certain kinds of prediction errors. For example, accuracy as a measure only relays how many true positives are predicted in relation to all other error types. We have no information if the model is prone to false positives or false negatives. Another discrepancy is that correlation coefficients are often used for regression based predictive models, where the trait being predicted is a continuous variable. This makes comparison challenging impossible because correlation coefficients are generally below .50, while measures of accuracy can often be .70 and above—both are decent indicators of success but not easily comparable.

### RQ3

All surveyed publications reported some measures of error rates. For the purposes of this study, error rate was reported as “success rate” which serves as a proxy for the various reporting measures authors chose as mentioned in RQ2. Measure of correlation coefficients was excluded from Table 3 as they are incompatible. Measures of success were sorted as greater than 90% success, between 70% and 90% success, and less than 70% success rate (Table 3). Of the 31 papers, 13 papers contained at least one private trait that was predicted with greater than 90% success. There were seven traits including Personality, Interests, Sexual Orientation, Gender, Ethnicity, Age Group, and Life Stage that were predicted in some aspect with higher than 90% success. Of 72 predicted traits in the 31 papers, only 59 were predicted with less than 90% success, meaning overall 80% of predicted traits had mediocre or poor success rates which would indicate in general terms that the prediction of private traits from social media data is unreliable.

## Discussion

We cannot form trust relationships with technologies, we can only have trust between people. We can, however, rely on technology, and measure that reliability empirically. This is taken from Annette Baier’s argument that trust can only exist

**Table 3.** List of 25 Predicted Traits by Publication, the Data Sources by Publication, and Reported Accuracy by Publication.<sup>a</sup>

Predicted trait	Data sources	Self-reported accuracy <sup>b</sup>		
		>90	>70	<70
Psychological factors				
Personality <sup>1,3,30,31,33,39–42,13,20,22–27</sup>	Mobile <sup>1,33,41</sup> , Twitter <sup>1,3,24,26,27,30</sup> , Facebook <sup>3,20,22,23,31,39,40</sup> , Weibo <sup>42</sup> , E-Learning Software <sup>13</sup> , YouTube <sup>3</sup> , Essay <sup>3</sup>	3,13,27,39,40	1,3,13,24,30,33,39,40,42	3,20,22,23,25,30,31,39,41
Personal well-being <sup>20,36</sup>	Facebook <sup>20,36</sup>			20,36
Boredom proneness <sup>33</sup>	Mobile <sup>33</sup>			
Intelligence <sup>20</sup>	Facebook <sup>20</sup>			20
Parental separation <sup>20</sup>	Facebook <sup>20</sup>			20
Intrinsic factors <sup>25</sup>	Facebook <sup>25</sup>		25	
Temporal orientation <sup>32</sup>	Facebook <sup>32</sup> , Twitter <sup>32</sup>		32	
Values <sup>28</sup>	Facebook <sup>28</sup>		28	28
Mood <sup>43,44</sup>	LifeJournal <sup>43</sup>		44	43
Aggression <sup>19</sup>	Facebook <sup>19</sup>		19	
Preferences				
Sentiment <sup>10</sup>	Twitter <sup>10</sup>			10
Interests <sup>11,19</sup>	App installation logs <sup>11</sup> , Facebook <sup>19</sup>	19	11,19	11
Sexual orientation <sup>20,21</sup>	Facebook <sup>20</sup> , US dating site <sup>21</sup>	21	20,21	
Religious views <sup>20</sup>	Facebook <sup>20</sup>		20	
Political views <sup>8,20,22</sup>	Facebook <sup>8,20,22</sup>		8,20,22	
Friendship network <sup>20</sup>	Facebook <sup>20</sup>			20
Health factors				
Substance use <sup>20</sup>	Facebook <sup>20</sup>		20	20
Economic factors				
Consumer behaviour <sup>26,33</sup>	Mobile <sup>33</sup> , Twitter <sup>26</sup>		33	26,33
Financial outcomes <sup>2</sup>	Mobile <sup>2</sup> , phone bills <sup>2</sup>		2	2
Demographics				
Education level <sup>33</sup>	Mobile <sup>33</sup>			33
Gender <sup>11,18,20,22,33</sup>	Mobile <sup>18,33</sup> , Facebook <sup>20,22</sup> , app installation logs <sup>11</sup> , phone bills <sup>18</sup>	11,20,22	11,18,25	33
Ethnicity <sup>20</sup>	Facebook <sup>20</sup>	20		
Age <sup>16,18,20,22,33</sup>	Mobile <sup>18,33</sup> , Facebook <sup>20,22</sup> , phone bills <sup>18</sup> , Twitter <sup>16</sup>	16	16,20	18,22
Life stage <sup>12</sup>	App installation logs <sup>12</sup>	12	12	
Relationship status <sup>20</sup>	Facebook <sup>20</sup>			20

<sup>a</sup>For formatting reasons, each citation is replaced with a superscript that denotes the order the paper appears in the bibliography. <sup>b</sup>“Success” was not reported uniformly across studies, this is based off the reported metric in the paper and can refer to any number of error measures: accuracy, precision, recall, area under the curve (AUC), F-score, and absolute mean-error (AME) as defined in Table 2.

in relationships where there is a possibility for betrayal, and that we cannot truly form trust relationships with technology because technology cannot “betray” us. Reliability is operationalized from the McKnight–Chervany typology of trust as competence and predictability. Competence and predictability and thereby reliability can be measured empirically through measures of error as defined in Table 2. In conducting a systematic literature review of articles offering models for predictive private traits from social media data, we show that only including a single measure is typical, and that there are high error rates across a majority of current work.

We show that 80% of methods for predicting private traits from social media data have less than a 90% success rate. Of the 80% of methods more than half are well below 70% success rates. This would indicate a general unreliability of predicting private traits from social media data at this time, and

a lack of trustworthiness based on measures of competence and predictability.

Let us now consider the implications of a lack of accuracy and lack of ability for people to rely on these models/technologies for large technology companies such as Facebook. While trust in technology platforms has become a key concern in public and policy discussions, we suggest drilling down to the specific relationship between humans and the technologies they interact with is also crucial. This is where the idea of reliance comes in. Technology companies should not predict the private traits of users or allow any similar third-party activity as long as low accuracy and a lack of reliability remain the norm.

While an inability to rely on technologies to do what we expect them to do may be enough from a base functional level, there are other reasons technology companies should

be wary of using predictive traits which have low accuracy or success. That is, if people cannot rely on a piece of technology, they may begin to question whether they can trust those who created or promoted that technology. These trusting relationships among humans are complex and involve more than simple reliance as is demonstrated in McKnight–Chervany’s focus on benevolence and integrity as the other key aspects of trust.

The approach demonstrated in this article, however, only deals with the competence aspect of trust. This means that if a company could prove their technology was 100% reliable, then on competence alone they would be “trustworthy.” However, that is where the rest of the McKnight–Chervany typology is useful and where higher levels of analysis come in. For example, if CA was untrustworthy because their goals were not considered benevolent, then it does not matter how competent they are perceived to be. This seemed to be reflected in the general public discourse surrounding the scandal and as reflected in media coverage. Much of the fallout was centered on how a “data-breach” or Facebook’s perceived lack of competence led to damage to other forms of trust. For example, institutional trust in democratic processes was scrutinized during much of the CA scandal as a result.

We therefore suggest that platforms desiring trustworthy status should not predict private traits about users, or at least must provide a transparent reason for doing it with assurances of reliability related to the activity. In the general sense, platforms wishing to be trustworthy should employ reliable technology and be transparent about the unreliability of newer technologies they are experimenting with or already employing. Overall, the competence measure of trust is a minimum evaluation that can be used when other aspects like benevolence, or integrity, are unknown or unknowable.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Trevor Deley  <https://orcid.org/0000-0002-3111-0208>

Elizabeth Dubois  <https://orcid.org/0000-0003-1323-516X>

### References

Note that the numbers in parentheses denote the no. of the reference as cited in Table 3.

- Adali, S., & Golbeck, J. (2014). Predicting personality with social behavior: A comparative study. *Social Network Analysis and Mining*, 4(1), 1–20. (1)
- Agarwal, R. R., Lin, C. C., Chen, K. T., & Singh, V. K. (2018). Predicting financial trouble using call data—On social capital, phone logs, and financial trouble. *PLOS ONE*, 13(2), Article e0191863. (2)

- Alsadhan, N., & Skillicorn, D. (2017). Estimating personality from social media posts. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 350–356). IEEE. (3)
- Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2), 231–260. <https://doi.org/10.1086/292745>
- Cadwalladr, C., & Graham-Harrison, E. (2018, March). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>
- Carr, D. (2008, November). The media equation: How Obama tapped into social network’s power. *The New York Times*. <https://www.nytimes.com/2008/11/10/business/media/10carr.html>
- Concordia Summit. (2016). Cambridge Analytica—The power of big data and psychographics. <https://www.youtube.com/watch?v=n8Dd5aVXLCc>
- David, E., Zhitomirsky-Geffet, M., Koppel, M., & Uzan, H. (2016). Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. *Online Information Review*, 40(5), 610–623. (8)
- Davies, H. (2015, December). Ted Cruz campaign using firm that harvested data on millions of unwitting Facebook users. *The Guardian*. <https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data>
- Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & González-Castaño, F. J. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Systems with Applications*, 103, 74–91. (10)
- Frey, R. M., Xu, R., Ammendola, C., Moling, O., Giglio, G., & Ilic, A. (2017). Mobile recommendations based on interest prediction from consumer’s installed apps—insights from a large-scale field study. *Information Systems*, 71, 152–163. (11)
- Frey, R. M., Xu, R., & Ilic, A. (2017). Mobile app adoption in different life stages: An empirical analysis. *Pervasive and Mobile Computing*, 40, 512–527. (12)
- Ghorbani, F., & Montazer, G. A. (2015). E-learners’ personality identifying using their network behaviors. *Computers in Human Behavior*, 51, 42–52. (13)
- Glum, J. (2018, March). Is the Facebook Cambridge Analytica case data breach? *Money*. <https://money.com/facebook-data-breach-experts/>
- Grassegger, H., & Krogerus, M. (2017, January). The data that turned the world upside down. *Vice*, pp. 1–22.
- Guimaraes, R. G., Rosa, R. L., De Gaetano, D., Rodriguez, D. Z., & Bressan, G. (2017). Age groups classification in social network using deep learning. *IEEE Access*, 5, 10805–10816. (16)
- Harrison McKnight, D., & Chervany, N. L. (2007). Trust and distrust definitions: One bite at a time. In R. Falcone, M. Singh, & Y. H. Tan (Eds.), *Trust in cyber-societies* (pp. 27–54). Springer.
- Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., ‘Sandy’ Pentland, A., & De Montjoye, Y.-A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science*, 6(6), Article 3. (18)
- Kandias, M., Gritzalis, D., Stavrou, V., & Nikoloulis, K. (2017). Stress level detection via OSN usage pattern and chronicity analysis: An OSINT threat intelligence module. *Computers and Security*, 69, 3–17. (19)
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human

- behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), 5802–5805. (20)
- Kosinski, M., & Wang, Y. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2), 246–257. (21)
- Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, 21(4), 493–506. (22)
- Laleh, A., & Shahram, R. (2017). Analyzing Facebook activities for personality recognition. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 960–964). IEEE. (23)
- Lee, T., Yoong, C., Ngatirin, R., & Zainol, Z. (2017). Personality prediction based on social media using decision tree algorithm. *Pertanika Journal of Science and Technology*, 25, 237–248. (24)
- Liu, Y., Wang, J., Jiang, Y., Sun, J., & Shang, J. (2017). Identifying impact of intrinsic factors on topic preferences in online social media: A nonparametric hierarchical Bayesian approach. *Information Sciences*, 423, 219–234. (25)
- Liu, Z., Wang, Y., Mahmud, J., Akkiraju, R., Schoudt, J., Xu, A., & Donovan, B. (2016). To buy or not to buy? Understanding the role of personality traits in predicting consumer behaviors. In E. Spiro & Y. Y. Ahn (Eds.), *Lecture notes in computer science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10047 LNCS, pp. 337–346). Springer. (26)
- Lukito, L. C., Erwin, A., Purnama, J., & Danoekeoesomo, W. (2016). Social media user personality classification using computational linguistic. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1–6). IEEE. (27)
- Mukta, M. S. H., Ali, M. E., & Mahmud, J. (2016). User generated vs. supported contents: Which one can better predict basic human values? In E. Spiro & Y. Y. Ahn (Eds.), *Lecture notes in computer science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10047 LNCS, pp. 454–470). Springer. (28)
- Nakov, P. (2017). Semantic sentiment analysis of Twitter data. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining* (pp. 1–12). Springer.
- Ngatirin, N. R., Zainol, Z., & Yoong, T. L. C. (2016). A comparative study of different classifiers for automatic personality prediction. In *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 435–440). IEEE. (30)
- Park, G., Schwartz, A. H., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–9520. (31)
- Park, G., Schwartz, H. A., Sap, M., Kern, M. L., Weingarten, E., Eichstaedt, J. C., . . . Seligman, M. E. P. (2017). Living in the past, present, and future: Measuring temporal orientation with language. *Journal of Personality*, 85(2), 270–280. (32)
- Park, S., Matic, A., Garg, K., & Oliver, N. (2017). When simpler data does not imply less information: A study of user profiling scenarios with constrained view of mobile http(s) traffic. *ACM Transactions on the Web*, 12(23), Article 9. (33)
- Rahman, M. (2018). *Amidst data scandal, Facebook will voluntarily enforce EU's new privacy rules "everywhere."* xda-developers. <https://www.xda-developers.com/facebook-voluntarily-enforce-eu-privacy-law/>
- Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018, March). How Trump consultants exploited the Facebook data of millions. *The New York Times*. <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>
- Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., & Ungar, L. H. (2016). Predicting individual well-being through the language of social media. *Biocomputing*, 2016, 516–527. (36)
- Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In A. Sattar & B. Kang (Eds.) *AI 2006: Advances in artificial intelligence* (pp. 1015–1021). Springer.
- Solon, O. (2018, May). Cambridge Analytica whistleblower says Bannon wanted to suppress voters. *The Guardian*. <https://www.theguardian.com/uk-news/2018/may/16/steve-bannon-cambridge-analytica-whistleblower-suppress-voters-testimony>
- Thilakarathne, M., Weerasinghe, R., & Perera, S. (2016). Knowledge-driven approach to predict personality traits by leveraging social media data. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 288–295). IEEE. (39)
- Wang, M., Zuo, W., & Wang, Y. (2015). A novel adaptive conditional probability-based predicting model for user's personality traits. *Mathematical Problems in Engineering*, 2015, 1–14. (40)
- Xu, R., Frey, R. M., Fleisch, E., & Ilic, A. (2016). Understanding the impact of personality traits on mobile app adoption e Insights from a large-scale field study. *Computers in Human Behavior*, 62, 244–256. (41)
- Xue, D., Hong, Z., Guo, S., Gao, L., Wu, L., Zheng, J., & Zhao, N. (2017). Personality recognition on social media with label distribution learning. *IEEE Access*, 5, 13478–13488. (42)
- Yang, Y.-H., & Liu, J.-Y. (2013). Quantitative study of music listening behavior in a social and affective context. *IEEE Transactions on Multimedia*, 15(6), 1304–1315. (43)
- Zhou, Q. (2017). Multi-layer affective computing model based on emotional psychology. *Electronic Commerce Research*, 18, 109–124. (44)

## Author Biographies

**Trevor Deley** (MSc Carleton University) is a PhD Candidate in E-Business at the University of Ottawa. His research interests include machine learning methods for the social sciences, and the impact of computational logic on democratic norms and customs.

**Elizabeth Dubois** (PhD University of Oxford) is an Assistant Professor at the Department of Communication and Faculty Member at the Center for Law, Technology and Society, University of Ottawa, Canada. Her research focuses on political uses of digital media, media manipulation, and political opinion formation.