Minds and Machines
https://doi.org/10.1007/s11023-019-09497-4

Can Machines Read our Minds?

Christopher Burr[1] · **Nello Cristianini**[1]

Received: 6 September 2018 / Accepted: 25 February 2019

Abstract
We explore the question of whether machines can infer information about our psychological traits or mental states by observing samples of our behaviour gathered from our online activities. Ongoing technical advances across a range of research communities indicate that machines are now able to access this information, but the extent to which this is possible and the consequent implications have not been well explored. We begin by highlighting the urgency of asking this question, and then explore its conceptual underpinnings, in order to help emphasise the relevant issues. To answer the question, we review a large number of empirical studies, in which samples of behaviour are used to automatically infer a range of psychological constructs, including affect and emotions, aptitudes and skills, attitudes and orientations (e.g. values and sexual orientation), personality, and disorders and conditions (e.g. depression and addiction). We also present a general perspective that can bring these disparate studies together and allow us to think clearly about their philosophical and ethical implications, such as issues related to consent, privacy, and the use of persuasive technologies for controlling human behaviour.

Keywords Machine learning · Inference · Psychometrics · Digital footprints · Social media · Intelligent systems

✉ Christopher Burr
chris.burr@bristol.ac.uk

Nello Cristianini
nello.cristianini@bristol.ac.uk

[1] Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, England BS8 1UB, UK

Published online: 27 March 2019

1 Introduction

Recent news stories have brought to the public's attention a research trend that has been developing for several years across different research communities, and which is aimed at providing machines with the capability to infer information about the mental states and psychological traits of their users.[1]

However, the controversial technology behind these announcements is representative of a wider set of research interests than is captured by any specific news story, and is carried out for very different reasons by different scientific communities. A key observation, which motivates our enquiry, is that data scientists have come to discover that people leak personal information during online interactions with intelligent systems (i.e. "digital footprints"), which can then be used to train machine learning (ML) algorithms to infer information about the mental states and psychological traits of human users (e.g. Kosinski et al. 2013; Chen et al. 2014; Yang and Srinivasan 2016). This observation has had profound effects.

In a review of how digital footprints can be used to predict personality traits, for example, Lambiotte and Kosinski (2014, p. 1934) state that the collection and analysis of human activities mediated by online platforms is "changing the paradigm in the social sciences, as it undergoes a transition from small-scale studies, typically employing questionnaires or lab-based observations and experiments, to large-scale studies, in which researchers observe the behavior of thousands or millions of individuals and search for statistical regularities and underlying principles." This is because the digital footprints left behind during our online interactions with intelligent systems can be treated as *samples of behaviour*, and in turn used to infer additional psychological information about each individual, under certain conditions outlined later in this paper (Sect. 3).[2] There are now vast datasets of such behavioural samples, which are gathered from online repositories, social media APIs, or IoT enabled devices (among other sources), and which make these studies possible.

Furthermore, in addition to their scientific interest, the types of studies that Lambiotte and Kosinski (2014) allude to, are also of interest to businesses, governments and society, more generally. For example, as Matz et al. (2017) have shown, the automated detection of personality traits by ML algorithms, can also be used to tailor persuasive messages that demonstrably increase the chance of a user clicking on an online advertisement and purchasing a product. As such, there is a clear financial incentive for businesses and organisations to implement and deploy some of the methods detailed in these studies, connecting further communities to the ongoing

[1] For example, MIT Technology Review reported on how smartphones can be used to predict scores in tests designed to assess cognitive function (Metz 2018), and The Guardian provided extensive coverage on the use of psychographic modelling for use in election campaigning and marketing (Hern 2018). This development followed on from public backlash towards the use of similar technologies by Facebook (Rosenberg et al. 2018), as identified by an article that a research team at Facebook released describing the ability to manipulate the emotional states of users (Kramer et al. 2014), and also hinted at in a patent filing (Nowak & Eckles 2014).

[2] As we discuss later (Sect. 4.3), each of these samples of behaviour could also be potentially considered as an item in a psychometric test.

research and technological developments. However, these incentives may not necessarily align with the interests of individuals and society more generally, raising important social, legal and ethical questions (Wachter and Mittelstadt, Forthcoming). An obvious example in this regards is the use of psychometric data to influence political campaigning (Hern 2018), and the continued rise of so-called 'neuropolitics' (Schreiber 2017; Svoboda 2018). Even if the effects of these techniques are sometimes overstated by companies trying to market their latest product, the potential risks involved justify the ongoing analysis and scrutiny of these technological developments.

Therefore, it is worth reflecting on what information we reveal during our online interactions, as well as how much of this information can be used by intelligent systems to 'read our minds'. This is important, because no business invests money into large-scale behaviour monitoring for the sake of merely knowing more about their users. Rather, the process of inferring psychological information is often to improve the accuracy of *consequential decisions* made by autonomous intelligent systems about how best to predict, persuade, and ultimately control the behaviour of the user.

In light of this interest, the current paper explores a central question that underlies the aforementioned technical developments and news announcements, and which may not be immediately clear to all of the communities involved:

> Can machines infer (probabilistic) information about the psychological traits and mental states of individual users, on the basis of samples of their behaviour?

This question is replete with many thorny philosophical and methodological issues, which we wish to avoid in order to focus on other matters.[3] Therefore, in Sect. 2, we begin by unpacking and clarifying what is meant by the question, before detailing two case studies of influential technologies at the heart of recent advances. In order to address this question, in Sect. 3, we present an overview of a significant portion of the scientific literature, across a range of different research communities, and identify 17 categories of psychological constructs, which can be inferred (to varying degrees) by machines on the basis of a variety of samples of behaviour or other observable quantity. We present 26 studies that have explored these various

[3] An important clarificatory note, however, is that while we refer to psychological traits and mental states in our question, our review in fact encompasses a wider range of theoretical constructs (e.g. political orientation and skills or abilities). In general, psychological traits differ from mental states in the sense that the former are typically treated as dispositions that affect behaviour but which are relatively stable over time (i.e. personality traits), whereas states tend to be more transitory (e.g. particular emotions). We sometimes refer to only one of these terms (i.e. 'trait' or 'state'), unless the context requires more specificity, in which case the relevant theoretical term is employed. In other cases, we use the more general term 'psychological construct' when we need to refer to the full set our review covers (also see footnote 6). We acknowledge that our grouping together of these constructs, and indeed our treatment of them as psychological constructs, is far from being theoretically uncontroversial. However, our primary aim in this paper is to better understand and draw attention to an emerging methodology in computer science (and related disciplines), rather than to take a substantive position on the nature of theoretical entities such as psychological traits.

constructs, and highlight the types of behavioural samples that can be used to infer information pertaining to them.

The purpose of this review is to better understand the extent to which autonomous intelligent systems can influence and shape our behaviour, but we do not attempt to offer a systematic meta-analysis of a specific literature (see Sect. 3.1). Instead, we are primarily interested in understanding what kind of psychological information can be inferred on the basis of our online activities, and whether an intelligent system could use this information to improve its ability to subsequently steer our behaviour towards its own goals. Therefore, it is sufficient for our purposes to simply note an emerging theme that has begun to appear across a wide range of studies and across a wide range of different communities.

In Sect. 4, we discuss the findings of our review, building on earlier work that presented a conceptual framework for understanding and analysing the interactions between autonomous intelligent systems and human users (Burr et al. 2018).[4] In this earlier paper, we employed the language of control theory to frame our discussion. The basic notion of control theory, the feedback loop, tells us that when a controller (e.g. an autonomous intelligent system) has access to information about the state of a controlled system (e.g. a human user), then it can choose appropriate actions to govern that state. We can break this feedback loop into two parts: a) the observational component, where a controlling agent can monitor the state (e.g. mental state) of a controlled user, and b) the action component, where the controlling agent can make decisions, conditional upon the observed state and its own goals, in order to steer the behaviour of the controlled user.

In (Burr et al. 2018), we focused on the part of the feedback loop concerned with actions taken by the controlling agent (i.e. an intelligent system). Specifically, we discussed the risks entailed in cases when the values and goals that drive the decisions of an intelligent system are misaligned with our own, and the risk of positive feedback loops emerging and leading to unintended consequences (e.g. political polarisation or behavioural addiction). This article focuses on the other component of the feedback loop: the observational component. Our review is designed to help demonstrate the types of mental states and psychological traits that intelligent systems can now detect, with the subsequent aim being to explore how the increasing ability for intelligent systems to 'read our minds' may alter the dynamics of the aforementioned feedback loop.[5]

By framing our discussion in terms of control theory and bounded rationality, we are able to highlight important philosophical and ethical questions, such as whether implied consent is sufficient in situations where it is unclear what psychological information can be inferred from our online behaviour, and how user's trust

[4] In our (2018) paper, we refer to intelligent systems as 'Intelligent Software Agents', developing on the standard definition of learning agent defined by Russell and Norvig (2010). Rather than motivating the use of the term 'agent' in this paper, we have chosen to simply adopt the former label instead.

[5] Although this work develops and extends upon earlier research, we also believe there is intrinsic value in discussing the article's central question for its own sake. Therefore, although we would encourage the reader to explore this paper's findings alongside the earlier framework, the two articles can be read independently of one another.

is impacted by the respective technological developments (Sect. 4.2). These questions are especially important given recent research findings (discussed in Sect. 4.2), which demonstrate the surprising scope of behavioural data that is collected from our smartphones during everyday activities (Schmidt 2018).

Finally, we also discuss, briefly, how the technological developments explored in this paper will likely impact the development of the behavioural sciences, most notably psychometrics (Sect. 4.3).

2 Unpacking the Question

The title of this article is informally 'can machines read our mind?', but in order for this question to be well-posed it requires some unpacking. The following definitions help clarify our framing:

- Our use of the term 'machine' refers to algorithms, and more specifically, to those machines that can learn (i.e. improve performance on a task) from data (i.e. experience). These systems are the object of study in the field of machine learning (Mitchell 1997).
- By 'mind' we mean the set of psychological constructs for any given individual, which typically fall within the remit of psychometrics, and partially determine the subject's observable behaviour.
- By 'psychological construct', we limit ourselves to the sub-case of theoretical constructs that are currently measured by various psychometric assessments, or may result from a medical diagnosis.[6]
- By 'read' we mean the ability to (probabilistically) infer or predict some information pertaining to the postulated psychological construct, based on a sample of the subject's observable behaviour.
- By 'samples of behaviour' we mean the observation of any actions of the user or their interactions with the machine.

Therefore, a more precise formulation of the question is, 'can machines infer (probabilistic) information about the psychological constructs of individual users, on the basis of samples of behaviour?'[7] Ultimately, this is a problem of *inference*: to know something without direct observation, on the basis of its effects. As such it can be modelled mathematically as an *inverse problem*, which is studied in various

[6] In psychometrics, the target of measurement is a postulated psychological construct, which is defined and delineated in relation to the process of measurement. The process of *construct validation*, including its epistemological and metaphysical assumptions (see Alexandrova 2017; Borsboom 2005 for helpful discussions), is complicated and beyond the scope of this paper—though we do say a bit about the process in Sect. 4.3. We focus on psychometrics in this paper because it is the science of psychological measurement.

[7] We are not interested in whether a machine can determine the 'content of our thoughts', though some have begun working on this (e.g. Shen et al. 2017; Wang et al. 2017).

disciplines (e.g. reconstructing a 3D shape based on a 2D projection is an example of an inverse problem commonly solved in radiography), and is a typical focus of ML.

In addressing this question, there are two further issues we wish to sidestep, but which it is worth saying something briefly about here. Firstly, by employing terms such as 'psychological trait' or 'mental state', we do not wish to take a stand on debates in related areas such as philosophy of mind about the nature or existence of such psychological constructs. For example, situationists (and to some extent interactionists) will find much to disagree with in the literature we survey, and these debates have well known consequences for related discussions in moral philosophy (Harman 1999). However, for the purpose of this paper we wish to sidestep these concerns in order to focus more specifically on uncovering an important methodology that is emerging in the computer sciences.[8]

Secondly, and relatedly, we do not discuss well-studied theoretical procedures in psychological assessment such as *construct validation* (Rust and Golombok 2009; Alexandrova and Haybron 2016). Instead, we ask if the outcome of certain psychological assessments can reliably be predicted by a machine based on samples of user behaviour, thereby *bypassing* the need for administering the original assessment. This approach was taken in a study, which administered a series of psychometric tests to a large number of Facebook users, and then used ML algorithms to learn how to map their online data to the outcome of the respective tests (Kosinski et al. 2013). Here also, the question of construct validity was bypassed, and the algorithm predicted whatever the authors of the original test considered as a 'latent psychological trait'. This study is representative of a research trend being conducted by many different communities (often independently), which collectively allows us to address the above question. To further understand the nature of this question, we explore this study in more detail, alongside a further case study that also represents an example of an emerging methodology being utilised across the aforementioned communities.[9] It is our hope that with this methodology clearly laid out, philosophers will be able to engage with the material and perhaps develop on some of the underlying theoretical assumptions that pertain to debates such as those mentioned above.

2.1 Case Study 1: MyPersonality

Social media platforms have been interested in the possibility of inferring private psychological traits from samples of users' behaviour for a while, as evidenced by a patent filed by Facebook in 2012, and subsequently granted in 2014, which

[8] As one tangential remark, however, it is interesting to note that some studies in areas such as HCI and affective computing do note the importance of situational and contextual factors in inferring psychological traits and mental states (e.g. Baras 2016; Freitas 2017), and see developments in ubiquitous computing (e.g. IoT devices) as promising developments for improving our ability to accurately incorporate this type of contextualising data.

[9] Both of the techniques discussed in the following two case studies have been influential across a wide-range of communities, as will become evident in the review Sect. (3.2). It is for this reason that they have been selected.

explored the possibility of determining user personality traits on the basis of their social media activity (Nowak & Eckles 2014). However, the techniques by which this is possible were made clear to the public following the publication of (Kosinski et al. 2013).

This paper provided details of an application (MyPersonality), developed by researchers at the University of Cambridge, which allowed Facebook users to participate in a range of psychometric tests, including: a 20-item version of the IPIP (5-factor personality) test; a 20-item version of Raven's Standard Progressive Matrices (Intelligence) test; and a 5-item Satisfaction with Life Scale test.

Following the tests, users were asked if they were happy for their profile information to be collected for research purposes. This information included, but is not limited to:

- 55,814 possible "Likes" recorded and decomposed (using Singular Value Decomposition) into a 100-component vector for each user (n = 58,466);
- The user's age, gender, sexual orientation, relationship status, political views, religion, and social network information (e.g. network density), if recorded by the user;
- Details of the users' consumption of alcohol, drugs, and cigarettes and whether a user's parents stayed together until the user was 21 years old (recorded using online surveys); and
- Visual inspection of profile pictures, in order to assign ethnicity to a randomly selected subsample of users.

In order to predict the user's psychological traits, a combination of linear regression and logistic regression algorithms were used (both with 10-fold cross validation), in order to predict numerical variables (e.g. score for 'openness' trait) and binary variables (e.g. gender) respectively. These methods enabled the researchers to predict various psychological traits and demographic information with differing degrees of accuracy (details are reported in Sect. 3).

The method and dataset that Kosinski et al. (2013) presented has subsequently been utilised by additional researchers, some of whom have used the dataset for different experiments (e.g. Boyd et al. 2015; Annalyn et al. 2018)—Sect. 3 will review some of these experiments in more detail.

An interesting point, raised by Kosinski et al. (2013), in their discussion, was that the "similarity between Facebook Likes and other widespread kinds of digital records, such as browsing histories, search queries, or purchase histories suggests that the potential to reveal users' attributes is unlikely to be limited to Likes. Moreover, the wide variety of attributes predicted in this study indicates that, given appropriate training data, it may be possible to reveal other attributes as well" (Kosinski et al. 2013, p. 5805).

The possibility of digital samples of behaviour revealing further (perhaps unknown) psychological traits of users is a primary motivation for this paper, and will be discussed further in Sect. 4.

2.2 Case Study 2: LIWC

Another influential technology is the Linguistic Inquiry and Word Count (LIWC): a popular method in computational linguistics for inferring psychological information based on an individual's language use (Pennebaker et al. 2015).

Development of LIWC began in the early 1990 s, taking advantage of modern computing and the rise of the internet (Tausczik and Pennebaker 2010). The goal was to create a program that could look for and count words that belonged to "psychology-relevant categories" at scale and across multiple text files (Tausczik and Pennebaker 2010, p. 27). After several iterations the product has evolved into a comprehensive software tool that contains over 6400 words. (Pennebaker et al. 2015).[10]

LIWC has two central features: (a) the processing component and (b) the dictionary. The processing feature is a computer program, which opens a series of text files (e.g. essays, blogs, or novels) and counts each word in the file. The dictionary is organised into categories, which serve the purpose of scoring a text file for various attributes (e.g. positive or negative emotion words; function words), as well as defining which of the target words in the file should be counted and which should be ignored. For example, 'it' is counted as an instance of a 'function word', a 'pronoun', and, more specifically, an 'impersonal pronoun'. Each category is incremented when a member of the category is detected, and at the end, a score can be given that identifies the percentage of words in a text that are included within each of the hierarchically-organised categories.

The purpose of LIWC and its categories is to capture the language correlates of psychological traits or mental states such as attentional focus, emotional state, social relationships, and thinking styles (e.g. analytic use of distinctions, degree of cognitive complexity). For example, "[t]he function and emotion words people use provide important psychological cues to their thought processes, emotional states, intentions, and motivations" (Tausczik and Pennebaker 2010, p. 37). There is now a huge amount of literature assessing the psychometric properties of LIWC.[11]

Evaluating the psychometric properties of LIWC is similar to standard psychometric questionnaire evaluation, in that *reliability* and *validity* are assessed—word counts can be treated as responses, in the sense of item response theory (IRT) (see Sect. 4.3 for discussion). However, assessing the reliability of LIWC differs from traditional questionnaires, because an individual does not tend to use the same language in multiple iterations (e.g. test-retest reliability). In terms of validation, a number of studies are worth mentioning:

- Kahn et al. (2007) assessed the construct validity of LIWC's emotion categories (e.g. positive and negative emotions), and reports that LIWC appears to be "a valid method for measuring verbal expression of emotion".

[10] For an overview and introduction to LIWC, see (Pennebaker 2011).

[11] A good starting point is (Tausczik and Pennebaker 2010), which contains a large list of references for validation studies. Pennebaker et al. (2015) also provide an overview of the psychometric properties of the most recent release of LIWC (LIWC2015).

- Alpers et al. (2005) found that LIWC ratings of positive and negative emotion words correspond with human ratings of writing excerpts.
- Mehl et al. (2006) found that, in transcripts of spoken dialogue, higher word count and use of fewer large words (for both males and females) predicted extraversion.
- Rude et al. (2004) found that individuals with depression are more likely to use an increased number of first-person singular and negative emotions words in emotional writings, than individuals who are not depressed.

LIWC is known as a 'closed-dictionary' approach, due to the fixed nature of its categories.[12] As an example, LIWC "ignores context, irony, sarcasm, and idioms", leading to codings of words such as 'mad' as instance of 'anger'. However, as LIWC is a probabilistic system, the advent of big data techniques and large-scale content analysis means that many of these weaknesses can be mitigated with sufficiently large datasets. As such, LIWC is frequently used in ML studies (e.g. De Choudhury et al. 2013; Chen et al. 2014; Hao et al. 2014), and the increasing amount of publicly available web data offers new insights for the social sciences (Lazer et al. 2009)—for example, computational methods, such as LIWC, may help to test the degree to which word use is contextual and whether particular findings hold with different groups across a wide range of domains.

Although we have focused on two case studies, it turns out that many different research communities have been interested in automating or bypassing psychological testing for a while. A non-exhaustive list would include communities such as: human-computer interaction, computational social science, digital humanities, affective computing, psychoinformatics, health informatics, and many more.[13] While each of these communities may be interested in specific mental states (e.g. emotion in the case of affective computing), the general interest in inferring psychological information from samples of behaviour is common to all. This is important to note, because as the communities become increasingly integrated, it is possible that more can be achieved than could otherwise be done in isolation. As we demonstrate in Sect. 4, the consequences of this raises important philosophical and ethical questions.

[12] See (Schwartz et al. 2013) for a discussion of closed- versus open-vocabulary approaches, including a consideration of LIWC.

[13] The HCI community routinely hold challenges for researchers, in which different teams compete to demonstrate the most effective method for automatically extracting relevant features from common datasets, across multiple modalities (e.g. extracting and predicting emotional content from audio, video etc.). Examples of these challenges include the International Workshop on Audio/Visual Emotion Challenge (component of the ACM Multimedia Conference) and the SemEval challenge. These workshops help to develop methodologies and techniques across domains such as signal detection theory, and therefore, even if one study only focuses on a specific area (e.g. predicting affective state from a video recording of an individual's gait), the techniques can also serve to advance research in wider domains (e.g. identification in surveillance systems).

3 Machine Inference of Psychological Traits

In this section, we review 26 studies, across 17 categories, which goes some way to answering the question of whether machines infer (probabilistic) information about the psychological traits and mental states of individual users, on the basis of samples of their behaviour.

As noted in the introduction, the purpose of this review is to better understand a research trend that has emerged across a wide range of communities and to explore the philosophical and ethical consequences of the techniques being developed—we see these consequences as demanding urgent attention and *ongoing* scrutiny, in order to meet the changing demands that arise from constant innovation. Therefore, although the review is non-systematic, and was not designed to meet the standards of a scientific meta-analysis or quantitative review, it is sufficient for our purposes to demonstrate the main characteristics of an emerging trend, which we aim to capture and formalise in the next section.

3.1 The General Format

The general process for these studies involves an algorithm having access both to samples of an individual's behaviour and to a *normative group* of many individuals for whom both psychometric information and observable behaviour are known.[14] It can be summarised as follows:

- A study takes the values of a measure of some theoretical construct (P) (e.g. a psychological trait). Typically, these values refer to the answers or score to a validated psychometric test. However, they may also represent a diagnosis in the case of psychopathologies (e.g. the binary classification representing the result of a diagnosis), as well as a range of additional self-reported labels (e.g. political or sexual orientation). These values represent the 'ground truth' for the subsequent experiment.
- The above values are paired with another set of values, which correspond to a measure of some set of observable behavioural samples (B).
- The set of pairs $\langle P_i, B_i \rangle$, for each subject i in the study, comprises the labelled training data that is used as input to a machine-learning algorithm (A) (e.g. support vector machine). This training set plays a role that is analogous to a normative group in psychometrics (see footnote 10).
- The model that is the output of this process (M: $B \rightarrow P$) is then used to predict, for a new subject s, their values for P_s on the basis of B_s.

[14] The concept of a normative group (or, normative population) is a fundamental notion in modern psychometrics. It enables the assessment of an individual to be compared relative to the performance of a wider population. As such, the existence of this reference group is what gives certain scales their meaning (e.g. candidate X has an average score, relative to the results of the normative group) (see Rust and Golombok 2009, for an introduction to modern psychometrics).

In a less formal manner, when an ML algorithm is trained on a set of values of psychological traits (P_i) and a set of behavioural samples (B_i), for a normative group that has undertaken a pre-existing psychological assessment, it can use this information to infer the respective information about other individuals not in the original sample, thereby bypassing the need for all individuals to take the original assessment. Although some of the studies in our review depart from this general process in specific ways, the perspective that this formal setting offers is nevertheless instructive for understanding the research being conducted and developed by many different communities.

We organise our review according to the theoretical constructs that are both (a) the object of enquiry for the original psychological assessment, and (b) the target that the ML algorithm aims to predict on the basis of some sample(s) of behaviour. The 17 categories of theoretical constructs are organised into five parent categories: affect and emotion (Sect. 3.2.1), aptitudes and skills (Sect. 3.2.2), attitudes and orientations (Sect. 3.2.3), personality (Sect. 3.2.4), and disorders and conditions (Sect. 3.2.5).[15] Across these categories, a broad range of behavioural signals were found to correlate with one or more of the subsequent constructs, including (but not limited to) visual signals (e.g. profile pictures; facial expressions), audio signals (e.g. paralinguistic features of speech), written text (e.g. social media posts, email communication), physiological signals (e.g. heart rate), and other samples of behavioural signals (e.g. computer and smartphone usage, website choice, typing patterns, and social media "likes").

By conducting this review, we do not wish to endorse or critically evaluate the studies themselves, though we present relevant metrics where possible.[16] Furthermore, we accept that many of the studies could be improved, and that many of the reported measures of accuracy are currently insufficient to allow for practical application of the relevant techniques. In spite of these limitations, some organisations have already begun trying to control user behaviour on the basis of the inferred information, which raises important ethical issues that we discuss in Sect. 4. As such, we believe it is imperative that we understand the scope of what is being researched, and the consequences of these communities increasingly converging.

3.2 The Review

3.2.1 Inferring Affect and Emotion

3.2.1.1 Discrete Emotions In affective science, we can distinguish two theories—those which categorise emotions as basic or discrete [e.g. anger, fear, sadness, enjoy-

[15] The organisation of these categories does not follow any specific taxonomy found within the existing psychological literature, but is designed to capture the broad interests of the relevant communities and studies that this review covers, while retaining an intuitively plausible grouping.

[16] Where relevant, we present the metrics in the original form of the respective study, rather than attempting to translate into a common measure. Some of these measures or techniques may be unfamiliar to the wider audience. Where this is the case, we direct the interested reader to the respective study.

ment, disgust and surprise (Ekman 1992)], and those which emphasise the affective (continuous) dimensions [e.g. valence and arousal (Russell 1980)] of emotions. Different methods are used depending on the theoretical assumptions made by the researchers conducting the study. For example, in the affective computing community, a number of techniques have been developed for automated face analysis (AFA) (Cohn and de la Torre 2015). AFA can be used to extract 'facial action units'—anatomically-based descriptors of facial activity—from images or video. These action units can then be used as input for a sign-based measurement process to infer "basic emotions" such as amusement, sadness, anger, fear, surprise, disgust, contempt, and embarrassment. This process is known as the Facial Action Coding System (FACS), and relevant manuals allow human observers to code action units and translate them into the emotional categories, such as basic (discrete) emotions (Ekman and Rosenberg 2005). However, there is also disagreement over how many distinct emotion categories should be represented by the relevant system (e.g. Du et al. 2014).

Study 1 Mavani et al. (2017) trained a convolutional neural network to bypass the FACS process, by removing the need for extracting action units. Their study found an overall test accuracy of 95.71% for their model when trained and tested on the Radboud Faces Database (Langner et al., 2010), but fell to 65.39% when attempting to generalise across datasets.[17] Angry and sad faces were most likely to be confused, with a per-class accuracy of 46.27% each. Disgusted faces achieved the highest per-class accuracy of 90.05%.

Study 2 Utilising a different method, Hu and Flaxman (2018) took user-tags (e.g. '#happy') from Tumblr, a social media site, as self-reports of emotional states, and combined these labels with corresponding images and text posted by the individual. 15 tags were selected, based on how frequently they occured in the posts and also whether they appeared in the PANAS-X psychometric scale (Watson and Clark 1999). After filtering the initial dataset to only include posts with one of the 15 emotional tags and the corresponding text and image, the authors were left with 256,897 posts. These multimodal posts were initially processed separately, using a convolutional neural network for the images and a combination of word embeddings and a long short-term memory neural network for the text. The output of these two components was then fed into a further multimodal neural network, in order to classify the posts. Their model achieved a 72% accuracy during testing.

3.2.1.2 Affective Dimensions Many of the studies in affective computing that deal with the automatic prediction of *affective dimensions* face a similar problem to the FAC system above—the extraction of relevant features from multimedia such as speech and video recordings (sometimes referred to as 'signal detection and processing').[18]

[17] Dataset was split into 70% training, 15% validation and 15% test.

[18] A number of informative review articles, discussing the automated extraction of emotion-content features from images, video, speech recordings, and text can be found in (Calvo et al. 2015). The technical details are beyond the scope of this article.

Study 3 Bone et al. (2012) present an unsupervised learning method for producing ratings of one affective dimension (arousal) through the extraction of salient prosodic features of speech recordings. They utilised four publicly available databases containing speech recordings from acted and natural emotional conversations in German and English (see article for details regarding databases used), which had been rated along the arousal dimension in order to provide ground truth. They report that the Spearman's rank correlation (and binary classification accuracy) achieved by their unsupervised learning method on the four arousal databases were: 0.62 (73%), 0.77 (86%), 0.70 (82%), and 0.65 (73%).

Study 4 Karg et al. (2010) used an optical tracking system to record the gait of actors who had been asked to "feel angry, happy, neutral, or sad and to imagine a situation in which they feel a particular affect". From these instructions, the authors split the database into two groups containing 520 strides for analyzing discrete affective states and 780 strides for analyzing affective dimensions. The gait patterns (embodied using a visually animated manikin model) were also evaluated by human raters who had to determine whether the stride expressed either a low, medium, or high level of pleasure, arousal, or dominance, on a five-item Likert scale. The study compares multiple feature extraction/reduction methods (e.g principal component analysis (PCA), linear discriminant analysis), as well as multiple classification methods (e.g. Neural Network, Naive Bayes, Support Vector Machine). Using PCA to reduce the input to 15 features, the authors achieved the following mean accuracies for detecting person-dependent, discrete affective states (i.e. predicting affective states for individuals, rather than interindividual prediction): Neural Network (92%), Naive Bayes (92%), Support Vector Machine (95%). For person-dependant affective dimensions, they achieved the following accuracies (neural network without PCA): valence 88%; arousal (97%); dominance (96%).

3.2.1.3 Subjective Well-Being Subjective well-being is a self-reported measure of how an individual evaluates their life or a specific life event (Diener 1984). Typically, it includes an affective component (i.e. frequent positive affect and infrequent negative affect) and a cognitive judgement (i.e. evaluation of life satisfaction).[19] Psychometric measures for these two components can be treated independently, or summed to produce an overall measure. There are over 1400 wellbeing and quality-of-life instruments, covering a range of sub-groups (e.g. different cultures, ages, contexts, etc.), including instruments that focus on negative aspects such as depression (see Sect. 3.2.5) (Calvo & Peters 2014).

Study 5 Hao et al. (2014) showed how sets of features extracted from Chinese microblogging service Sina Weibo could be used to predict an individual's score on these two components. The features included demographic information (e.g. gender,

[19] Although subjective well-being is widely assumed to be multidimensional, there is disagreement over just how many dimensions to include. Huppert et al. (2013), for example, argues that ten factors are needed: competence, emotional stability, engagement, meaning, optimism, positive emotion, positive relationships, resilience, self-esteem, and vitality. The interested reader can see (Alexandrova 2017) for a helpful discussion on this issue.

age, and location), behavioural signals (e.g. number of posts, privacy settings, length of nickname), and linguistic information obtained with a simplified Chinese version of LIWC (see Sect. 2.2). As with Case Study 1 (Kosinski et al. 2013), their subjects completed two questionnaires: the positive and negative affect schedule (PANAS) (Watson and Clark 1999) and the psychological well-being scale (PWBS) (Ryff and Keyes 1995). The scores from these tests formed the labels used in the training data, and a number of ML algorithms were compared, with stepwise regression performing the best. They found that by using a combination of demographic, behavioural and linguistic information, their predictions achieved a Pearson's Correlation Coefficient of 0.45 for positive affect, 0.27 for negative affect, and a mean of 0.45 for psychological wellbeing.

3.2.2 Inferring Aptitudes and Skills

3.2.2.1 General Intelligence General intelligence is a psychometric factor that summarises correlations between an individual's proficiency across a range of cognitive abilities. The factor was originally proposed by Charles Spearman in the early 20th century, and is still explored in modern psychometrics (Rust and Golombok 2009).

Study 6 In addition to the other psychological traits already discussed, Kosinski et al. (2013) also found correlations between social media "likes" and general intelligence. They measured subjects' general intelligence using a 20-item version of Raven's Standard Progressive Matrices—a nonverbal multiple choice test. Using linear regression, they found that an individual's "likes" showed a correlation of 0.39 with their scores on the above test. They also state that of these, "the best predictors of high intelligence include "Thunderstorms," "The Colbert Report," "Science," and "Curly Fries"" (Kosinski et al. 2013, p. 5804).

3.2.2.2 Writing Ability Automated assessment of educational tests has been eagerly pursued since the advent of computers, and many companies offer software that claim to be able to replace the need for human markers. In cases where the test is multiple choice, the process is relatively straightforward, but written essays pose a greater challenge, due to the more holistic manner in which human graders tend to evaluate a student's ability.

Study 7 The Education Testing Service (ETS) developed the e-rater system for automated assessment of a student's writing ability (Attali and Burnstein 2005). The system uses natural language processing techniques (see Burnstein et al. 2003) to extract features from essays, which include 'word choice' (e.g. relative occurrence of words; word length), 'grammatical conventions' (e.g. rates of errors, spelling, punctuation), 'fluency and organization' (e.g. use of passive voice, repetition of words, essay structure) and 'topical vocabulary usage' (assessed against a normative group of high-scoring essays on similar topics). These features can be used to train a linear regression model to find the optimal weights for each of the features (combined with some fixed weights), which best predict the score of trained human readers (scoring according to grade-specific rubrics). The performance metric Attali and Burnstein (2005) choose to emphasise is the *test-retest reliability* for individual essays (across multiple grades), as they were attempting to bypass the assessment

of human raters (assumed to have low inter-rater reliability). Overall, across 1987 essays, the e-rater system (0.60) outperforms individual single human raters (0.50) and a combined average from two human raters (0.58).

3.2.2.3 Verbal Fluency Verbal fluency tests aim to measure the ease with which a person can produce words, and are used in clinical batteries to diagnose cognitive disorders associated with aphasia (e.g. Alzheimer's) and guide neuropsychological investigation (e.g. possible lesions in frontal cortex impacting executive functioning).

Study 8 Jimison et al. (2008) developed a computer assessment for measuring verbal fluency, based around a simple game in which subjects are required to come up with as many words as possible from a series of letters. To test the system, they administered a neuropsychological battery to 30 elderly participants (average age 80.4) who had played their computer game over the course of 1 year.[20] This score was used as the basis for a linear regression algorithm based on derived features extracted from the game logs (e.g. average time and word complexity). They reported a correlation of 0.459 (R^2) with the original tests.

3.2.3 Inferring Attitudes and Orientations

3.2.3.1 Values According to Schwartz's theory of Basic Human Values, individuals have a set of values (i.e interlinked, abstract ideas that are judged to be desirable and important) and trans-situational goals that motivate their behaviour (Schwartz 2003). The theory postulates ten universal values across five dimensions, which are assumed to be recognisable across cultures—making it useful for intercultural research.

Study 9 A research team from IBM recruited 799 participants from the social media site Reddit (Chen et al. 2014), each of whom were required to complete the Portrait Value Questionnaire (PVQ)—a 21-item test, using a 6-point Likert scale, which measures an individual's value orientations (Schwartz 2003). Using LIWC to extract word categories from the user's posts on Reddit, the authors performed a regression analysis on the extracted categories and questionnaire scores (one per dimension), and found a range of correlations (R^2) between the regressed scores and the actual scores (as measured by the PVQ) from 0.39 (self-transcendence) to 0.41 (openness-to-change and hedonism).

Study 10 Boyd et al. (2015) tested whether values extracted using a topic-modelling technique [meaning extracting method (MEM) (Chung and Pennebaker 2008)], which allows researchers to automatically discover relevant words that repeatedly co-occur across a corpus, predicted an individual's scores on the Schwartz Value Survey (SVS) (Schwartz 1992). Participants were recruited using Amazon's Mechanical Turk,[21] and required to complete the SVS, as well as provide free-form responses to two questions asking the subject to reflect on their personal values and behaviours. 16 themes associated with values (e.g. faith, growth, indulgence) and 27 themes associated with behaviour (e.g. fiscal concerns, time awareness, relaxation)

[20] The authors do not provide details for which neuropsychological battery they used.

[21] https://www.mturk.com/.

were extracted from the texts using the above natural language processing techniques. In two studies—the second performed using a subset of the MyPersonality dataset (Kosinski et al. 2013)—the authors found mostly weak correlations between the extracted topics and the scores derived from the SVS (the majority of R^2 correlations were <0.04).[22]

3.2.3.2 Sexual Orientation As with other examples in this review, the 'ground truth' for sexual orientation is simply the self-report of the individual concerned, which may not necessarily be accurate (Kosinski et al. 2015). Nevertheless, assuming the accuracy of these self-reports, some studies have demonstrated that it may be possible to predict sexual orientation through the use of alternative digital footprints.[23]

Study 11 As we discussed in Case Study 1 Kosinski et al. (2013) predicted a range of attributes pertaining to individuals (including sexual orientation, i.e. homosexual or heterosexual), from the set of their Facebook "likes". Using logistic regression, they found that the prediction accuracy (expressed by the area under receiver operating characteristic curve (AUC) coefficient) for males was 88% and for females was 75%.

Study 12 In a more recent study with Yilun Wang (2018), Michal Kosinski has also used a deep neural network (VGG-Face) to extract facial features from a set of profile photos taken from an online dating site and convert them into 4096 variables. These variables, along with the self-reported sexual orientation of the dating site users, can then be used to train a logistic regression analysis to correctly classify sexual orientation with a similar level of accuracy to the previous study (81% for men and 71% for women, also expressed using the AUC coefficient).

3.2.3.3 Political Orientation Big data and ML have been used in election campaigns in the US since at least 2008 (Issenberg 2012), but typically the information used was restricted to traditional forms of demographic data. More recently, we have begun to see increasing interest in groups inferring political orientations on the basis of social media information, due to the value this information has for election campaigns (Rosenberg et al. 2018).

Study 13 Cohen and Ruths (2013) collected hashtags from 2496 Twitter users, segmented into three groups (and three corresponding datasets): (a) *politicians* affiliated with a political party (n = 397), where the label was obvious (i.e. 'Republican' or 'Democrat'); (b) *politically active users* with self-reported affiliation in profile (n = 1837); and (c) *politically modest users* (n = 262) who were categorised by multiple Mechanical Turk workers (for inter-rater agreement). The collected hashtags (1000 most recent for each individual) were used to construct feature vectors to train a Support Vector Machine. Average accuracies for 10-fold cross validation

[22] The authors also compared the extracted values and SVS scores with reported behaviours, in order to test whether there was a closer link between either the open- or closed-form assessments and an individual's self-report of relevant behaviours (see Boyd et al. 2015 for details).

[23] In both instances, the authors have highlighted the significant ethical implications that these technologies could pose for the privacy and safety of the individuals concerned.

were reported as 91% (politicians), 84% (politically-active), and 68% (politically modest).[24]

3.2.3.4 Brand Perception Neuromarketing uses research from neuroscience and psychology in an attempt to gain commercially valuable insights into consumer experience, and to understand how an individuals purchasing behaviour could be predicted on the basis of neuroimaging data (Ariely and Berns 2010). A fundamental aspect of this area is inferring traits related to how individuals perceive and respond to various stimuli from potential advertising campaigns.

Study 14 Wei et al. (2018) used electroencephalography (EEG) data collected from 30 male participants while watching 4–5 adverts randomly selected from a possible set of 220. The participants were also required to complete a proprietary questionnaire consisting of a mixture of Likert-based items and binary items, for each of the products advertised. The questionnaire was designed to measure attitudes related to brand perception, and was based on a consumer experience model that emphasises four relevant attributes: attention, interest, desire, and action (AIDA). Some of the questions assessed whether the subject would be likely to buy the respective product. The results of the questionnaire were converted into a format suitable for a binary classification model (i.e. Support Vector Machine). Various predictions were made for each of the different product types (e.g. car, food, technology, clothes), and multiple accuracies were reported (see full text for details). Overall, their study achieved an accuracy of 77.28% using EEG data to predict brand perception and purchasing intentions.

3.2.4 Inferring Personality

3.2.4.1 Big-5 Traits (OCEAN) In contemporary personality science, the dominant paradigm is the five-factor model, which has been shown to subsume a wide variety of other personality scales (McCrae and Costa 1987). The five traits postulated by the model are 'openness', 'conscientiousness', 'extraversion', 'agreeableness', and 'neuroticism', collectively known as the Big-5, and often referred to using the acronym OCEAN (see Nettle 2009 for an accessible introduction).

There are many studies that show how personality can be predicted from digital footprints. In a review of these studies, Lambiotte and Kosinski (2014, p. 1934) acknowledge that one of the reasons behind this recent interest in personality psychology is that the "[a]bility to automatically assess psychological profiles opens the way for improved products and services as personalized search engines, recommender systems, and targeted online marketing".[25]

[24] Part of their study was also to show how reports of performance for earlier classifiers of political orientation trained on social media (e.g. Twitter) were over optimistic because of their reliance on politically active users. Therefore, they also showed how models trained on individual datasets performed poorly when generalised to novel datasets (e.g. model of politically modest users performed with 54% accuracy when classifying politicians).

[25] Initial evidence from (Matz et al. 2017) seems to support this idea.

Study 15 We have already introduced the exemplary study produced by Kosinski et al. (2013) (see Case Study 1 for details). In this study, the authors achieved the following levels of accuracy for their regression model (measured by the Pearson correlation coefficient): openness (0.43); conscientiousness (0.29); extraversion (0.4); agreeableness (0.3); neuroticism (0.3).

Study 16 Annalyn et al. (2018) also made use of the MyPersonality dataset, but focused on those "likes" that represented books. In combination with data mined from the book review site Goodreads.com, they were able to collect user-generated tags (i.e. keywords acting as proxies for the books content) for books that Facebook users had also liked. These pairings could then be used to test whether book preferences predicted personality traits. This development allowed the authors to discover correlations between genres of books and certain personality traits (e.g. philosophical-novel and openness: $r = 0.25$).[26] Using Lasso regression on the most predictive clusters of book tags, the authors were able to predict the Big-5 traits from book preferences to the following degrees (R^2): openness (0.41); conscientiousness (0.30); extraversion (0.32); agreeableness (0.34); and neuroticism (0.38).

Study 17 Grover and Mark (2017) tested whether patterns of smartphone and computer activity (e.g. usage duration, screen switching patterns), automatically collected from logging software, could predict personality traits. Unlike, the previous two examples, their study utilised a significantly smaller dataset (76 features of smartphone usage for 62 participants, each of whom completed the NEO five-factor personality inventory). Interestingly, some of the features referred to information about the ratio of duration spent on social media to the total usage duration for the device, which the authors hypothesised were related to personality traits. Using an optimal set of features, the authors trained a Random Forest Classification model for each of the five traits using 10-fold cross validation. They reported the following average binary classification/AUC values: openness (0.80/0.82); conscientiousness (0.65/0.66); extraversion (0.72/0.78); agreeableness (0.72/0.69); and neuroticism (0.73/0.72).

Study 18 Finally, Hoppe et al. (2018) were able to demonstrate that eye movements, measured during a natural-environment exploration study, could reliably predict four of the big-five personality traits (conscientiousness, extraversion, agreeableness, neuroticism). 42 students were required to walk around campus and purchase any items of their choice from a campus shop. They were also required to complete the NEO Five-Factor Inventory (60-item questionnaire). During their time exploring the campus, gaze data was tracked and recorded using a head-mounted video-based eye tracker, with 207 features subsequently extracted from the gaze data, and used to train a Random Forests model for each of the big-five traits. The performance of the classifiers was evaluated in terms of an average F_1 score across three score ranges, and the following accuracies were achieved: neuroticism (40.3%), extraversion (48.6%), agreeableness (45.9%), conscientiousness (43.1%)—the classifier for openness (30.8%) performed below chance level (33%).

[26] There are too many individual tag-trait pairings to report here (see original article for details).

3.2.4.2 Perceptual Curiosity Perceptual curiosity refers to an individual's level of interest in and reaction to novel stimuli that involve feelings of interest or uncertainty.

Study 19 In addition to predicting four of the five personality traits, Hoppe et al. (2018) were also able to predict *perceptual curiosity* from the acquired gaze data (see above). They used the Perceptual Curiosity scale—a self-report questionnaire developed by Collins et al. (2004)—as their ground truth. Using the same methodology as above, the Random Forest classifier achieved a 37.1% accuracy for predicting perceptual curiosity scores.

3.2.5 Inferring (Diagnosing) Disorders and Conditions

3.2.5.1 Autism Diagnosis of autism spectrum disorder (ASD) often involves assessment by a qualified speech and language therapist, due to the close association between ASD and abnormal vocal prosody.

Study 20 Nakai et al. (2017) recruited 30 children diagnosed with ASD by the Kobe University Hospital Developmental Behavioral Pediatric Clinic [according to DSM-V criteria (American Psychiatric Association, 2013)] and 51 children with typical development. They were required to verbally name objects and animals on picture cards, and the subsequent audio recordings (24 extracted features) were used as the basis for training a Support Vector Machine. The results of the classification algorithm were compared against the performance of 10 speech and language therapists, and a F_1 score was used to measure their performance. For the ML algorithm and therapist, respectively, the scores were as follows: true-positive rate = 0.81, 0.54; false-negative rate = 0.19, 0.46; false-positive rate = 0.27, 0.21; true-negative rate = 0.073, 0.80. Their experiment demonstrates that a ML algorithm can achieve similar levels of accuracy to a qualified specialist, and sometimes outperform them (true-positive). However, it should be noted that vocal prosody is only one element of a holistic assessment for children with suspected ASD.

3.2.5.2 Depression The DSM-V lists a series of depressive disorders (e.g. major depressive disorder), which have the common feature of the "presence of sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's capacity to function" (American Psychiatric Association 2013). A number of psychological assessments exist to measure the severity of symptoms associated with depression, including the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff 1977) and the Beck's Depression Inventory (Beck et al. 1961).

Study 21 A research team at Microsoft (De Choudhury et al. 2013), found that *major depressive disorder* could be predicted on the basis of a range of behavioural signals collected from Twitter. These signals include attributes such as engagement (e.g. volume of posts; proportion of reply posts), network statistics (e.g. ratio of followers and followees, embeddedness within network), emotion (measured by psycholinguistic properties through LIWC, see Case Study 2), and depressive language (also using LIWC lexicon). 476 participants, recruited through Mechanical Turk, were required to complete the self-reported 20-item CES-D questionnaire, and were split into two groups based on whether they

scored above a certain threshold on the CES-D. The scores and feature vectors (derived from Twitter data) were used to train a Support Vector Machine classification algorithm, which had to correctly classify the users as belonging to one of the two classes. Their subsequent model yielded an average accuracy of ~70% and high precision of 0.74.

Study 22 Reece and Danforth (2017) extracted features from 43,950 photographs using colour analysis, metadata components, and algorithmic face detection. These photos were taken from the accounts of 166 Instagram users (recruited using Mechanical Turk), 71 of whom had a history of depression as measured using the CES-D questionnaire. Using a 100-tree Random Forest algorithm to classify depressed users from non-depressed users, they acquired the following levels of prediction accuracy: recall (0.697), specificity (0.478), precision (0.604), negative predictive value (0.579), F_1 (0.647).

3.2.5.3 Dyslexia Eye fixation studies have explored how particular patterns of eye movements reflect an individual's difficulty with reading (Hyönä & Olson 1995), which may be used to detect dyslexia. The increased presence of webcams, or front-facing cameras on smartphones, therefore, presents an opportunity for automating the detection of dyslexia.

Study 23 Rello and Ballesteros (2015) trained a Support Vector Machine to classify Spanish readers with and without dyslexia. 97 subjects were required to read 12 different texts and 48 of the subjects had been diagnosed by a human expert as having dyslexia. The readings were recorded using eye tracking technology, and a variety of features were extracted (e.g. reading time, mean of fixations, and age of the participant). Their classifier achieved 80.18% accuracy in a 10-fold cross validation experiment.

3.2.5.4 Psychopathy Psychopathy refers to a range of personality disorders, which the WHO's International Classification of Diseases (ICD-11) (World Health Organisation, 2018) defines as "problems in functioning of aspects of the self, and/or interpersonal dysfunction that have persisted over an extended period of time". As with personality more generally, psychopathy is manifest in patterns of cognition, emotional experience, emotional expression, and behaviour, and is manifest across a range of personal and social situations, but is specifically treated as maladaptive.

Study 24 Steele et al. (2017) tested incarcerated youths for psychopathic traits using the Hare Psychopath Checklist: Youth Version (PCL: YV) (Hare, 2003), administered by trained researchers. Neuroimaging data was also collected for each of the individuals, who were subsequently split into three groups based on the scores obtained in the test: incarcerated youth with high psychopathy scores (HP) (n = 71); incarcerated youth with low psychopathy scores (LP) (n = 72); and non-incarcerated youth as healthy controls (HC) (n = 21). Features extracted from the neuroimaging data, were used to train Support Vector Machines, and their binary classification models obtained the following overall accuracies (additional measures are reported in original article): HP versus LP (69.23%); HP versus HC (78.26%); LP versus HC (79.57%).

3.2.5.5 Stress There are many forms of stress, including occupational and psychological stress, as well as forms of cognitive stress experienced during demanding tasks. In mild forms, stress can play an adaptive or motivational role in responding to environmental cues (e.g. competitive sports). However, many workers will have experience with forms of stress that go beyond its milder forms.

Study 25 Koldijk et al. (2016) tested whether unobtrusive sensors could be used to detect occupational stress in offices. They performed multiple experiments and extracted various features from four modalities: computer interactions from log files (i.e. mouse movement, keyboard usage, and application usage); facial expressions from webcams (i.e. head orientation, facial movements, action units, emotion), body posture from a Kinect 3D camera (i.e. distance, joint angles, and bone orientations), and physiological data (i.e. heart rate variability from ECG and skin conductance). Three pre-existing questionnaires were used as ground truth and also compared: the NASA Task Load Index (NASA-TLX) (Hart & Staveland 1998), which measures perceived workload; the Rating Scale Mental Effort (RSME) (Zijlstra & van Doorn 1985), which measures perceived mental workload; the Self-Assessment Manikin (SAM) (Bradley & Lang 1994). An initial exploratory study found that mental effort could be best predicted, with a correlation of 0.7920. Other variables could also be predicted with varying degrees of accuracy: valence (0.7139), arousal (0.7118), frustration (0.7117), perceived stress (0.7105), task load (0.6923), temporal demand (0.6552). They were able to achieve a higher correlation with mental effort scores (0.8416), by utilising a regression tree and using the 25 best features across the various modalities—features associated with facial expressions and posture provided the most information.

Study 26 Unlike many of the above examples, Vizer et al. (2009) conducted a study that used experimentally-defined conditions as the ground truth for their ML algorithms. They set up five conditions grouped into cognitive stress (i.e. mental multiplication and number recall tasks), physical stress (cardiovascular exercise and resistance exercise) and a control situation. These task labels were used in the supervised ML task. In each of the three conditions, subjects were required to spontaneously generate text through keyboard input, and a range of features associated with typing patterns and linguistic patterns were extracted. The two best classification models for physical stress (artificial neural network) and cognitive stress (kNN), achieved accuracies (reported using the AUC measure) of 0.625 and 0.75 respectively.

4 Discussion

Our review was undertaken in order to answer the question 'can machines infer (probabilistic) information about the psychological traits or mental states of individual users, on the basis of samples of their behaviour?' The findings in the previous section support an affirmative answer to this question for a variety of psychological constructs. This demonstrates that particular samples of behaviour are sufficient, in some instances, when the machine has been trained on the data referring to the psychological values and behavioural signals of a large number of other people (i.e. the

set of pairs $\langle P_i, B_i \rangle$). It follows that some of our online behaviour, if analysed in the context of a large 'normative group' (or training set), discloses personal (sometimes private) information about our mental states and psychological traits.

As we indicated in the introduction, this raises a number of considerations about what one can and should do when they have access to the aforementioned information—specifically whether an autonomous intelligent system could utilise this information to control a user's behaviour. In Sect. 4.1 we present the following actions as relevant to this first consideration: diagnose, predict, persuade and (more speculatively) control. In principle, these actions can be taken without active participation or explicit consent of the individuals concerned—we discuss these issues in Sect. 4.2.

In addition, our review also demonstrates that samples of online behaviour can be used to segment users into groups that share some psychological trait or mental state (e.g. group of users with high levels of depression). If we assume that the algorithm could access other samples of behaviour, or combine current signals in linked datasets, it is possible that ML techniques, such as unsupervised learning, may in the future find more effective criteria for grouping subjects together than have currently been discovered. These, as of yet, unnamed traits may still have psychological reliability, and perhaps validity, without belonging to our established lexicon. Although the consequences of these technologies for the future of psychometrics is not a key aspect of this paper—we focus on traditional forms of psychological assessment primarily to simplify our discussion—it is clear that the wider research community need to address the consequences of machines reading the minds of their users, whether they are known or unknown to current psychological science. Therefore, we also briefly discuss the connection between ML and psychometrics in Sect. 4.3.

4.1 What can be done with the inferred knowledge?

Given that machines can infer information about our psychological traits and mental states, it is important to consider what can (and should) be done on the basis of this information. Four categories are useful for discussing this point: diagnosis, prediction, persuasion, and (more speculatively) control. The first two represent passive forms of knowledge acquisition, whereas the final two introduce forms of intervention or action, conditional on some information pertaining to a user's psychological traits or mental states.

4.1.1 Diagnosis

Our review explored a number of cases where diagnosis of certain psychopathologies (e.g. depression and psychopathy) and other mental disorders or conditions could be bypassed by using ML algorithms, trained on relevant data. ML-based diagnosis is of significant interest within the medical community (e.g. DeepMind Health in the UK), because of the obvious benefits that improved levels of reliability can bring. However, diagnostic information can also be valuable to other organisations, such as health insurance companies, dating or gambling websites, or in hiring

decisions made by employers (e.g. whether to offer a job to an individual with high levels of depression).[27]

In all of these cases, diagnosis is typically a first step in a larger process of consequential decision-making, and depending on the subsequent decision, particular diagnoses can have significant practical consequences for the individual concerned (e.g. 'what, if any, treatment option should be given?'; 'should a particular candidate be hired?'). Therefore, it is important to consider the reliability and validity of any diagnosis in connection with the domain in which it is used. For example, one could argue that the use of ML-based medical diagnosis for the purpose of determining treatment options should require a much higher level of accuracy than alternative applications (e.g. advertising mindfulness apps or holidays to subjects displaying high levels of stress).

4.1.2 Prediction

Prediction utilises historical data (e.g. samples of behaviour) in order to predict the outcome of future events, on the assumption that certain statistical patterns are likely to recur. For example, this could be the likelihood of a user purchasing some product (conditional on some set of past purchases), or it could be the chance of an individual voting for a political candidate (conditional on the inferred values of their political attitudes or orientation).

Machine predictions are typically probabilistic in nature, and are often connected with a corresponding risk score (e.g. risk of defaulting in the case of loans and mortgages; risk of dropping out or quitting in college and job admissions; risk of recidivism in criminal justice decisions). As such, many communities are keenly interested in whether these predictions can be improved, and whether (and how) new forms of data-driven ML can assist. However, the considerations that prediction raises for each community are not necessarily shared. For example, the tolerance for risk varies across domain (e.g. insurance versus criminal justice) and risk-weighted predictions must reflect the prevailing attitudes of the community. Secondly, it may be unethical to treat predictions concerning individuals displaying psychopathological states in the same way as those for neurotypical individuals.

4.1.3 Persuasion

Action is a key ingredient in the generation of control systems and feedback loops. An intelligent system that has access to our mental states, in the context of other valuable data, can take actions that are designed to steer an individual's behaviour towards particular goals, while also monitoring feedback from its actions (i.e. the

[27] Chamorro-Premuzic et al. (2016, 2017) discuss the growing interest in big data analytics in hiring decisions and human resource management, along with the possibility of using digital fingerprints and gamified assessments as alternative samples of behaviour to supplement traditional job assessment methods. We do not include this work in the above review because many of the techniques are proprietary and the companies involved are not required to disclose the validity or reliability of their tools.

subsequent actions taken by the human user). This process can create a feedback loop, enabling an intelligent system to update its model regarding the probability of whether some future action will be effective in reaching its goal.

In the case of persuasion, for example, an intelligent system could use information about an individual's mental states for various ends. In one instance, Matz et al. (2017) show how personality can be used to more effectively target persuasive advertising messages that are expected to increase sales. And, in another, Lin et al. (2017) developed an app that can detect problematic usage based of smartphone usage patterns (daily use/non-use frequency, and duration of usage), which could in turn enable developers to nudge users who are at risk of smartphone addiction, with reminders about their usage.[28]

4.1.4 Control

Many in the area of positive computing—an offshoot of the more general area of positive psychology—have already begun exploring whether technology could be used to make people happier by promoting psychological traits and attributes such as positive emotions, self-awareness, motivation, engagement, mindfulness, empathy, and compassion, through value-sensitive design (Calvo & Peters 2014). A final (more speculative) consideration is the possibility of directly controlling an individual's mental state, such as those explored by positive computing.

By this, we mean a machine that continuously measures an individual's mental state and takes actions that are designed to directly control the associated variable (i.e. the latent variable), rather than simply trying to steer their behaviour through unmonitored persuasive appeals (e.g. nudges). Such attempts at control could have enormous benefits to individual and social levels of well-being, and many studies have begun to explore technology- or internet-based forms of medical intervention (i.e. therapeutic or promotional efforts to improve physical or mental health) (Calvo & Peters 2014). However, another example is a study conducted by a research team at Facebook (Kramer et al. 2014), which involved attempts at controlling the emotional states of users of the social media platform. News feeds of certain users were manipulated to show a greater proportion of positive or negative emotional content, in order to test levels of emotional contagion (i.e. the degree to which emotional states are transferred to others). Some users' news feeds were filtered to only see positive or negative emotional content, and the study found that when positive expressions of emotion were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. This is problematic. As is well-understood in control theory, minor increases in the level of inaccuracy associated with the estimation of state variables (i.e. inference of latent traits) can lead to drastic variation in the variables following

[28] Although Lin et al. (2017) considered their app-derived parameters alongside psychiatric diagnoses, and also conducted a separate validation of the the Smartphone Addiction Inventory (Lin et al. 2014), the existence of smartphone addiction is not included in manuals such as the DSM-V, and so it was not included in the main review (Sect. 3).

attempted control (e.g. nonlinear control problem of a trailer reversing), especially in cases of positive feedback loops. As such, there are a number of potential dangers from the misuse of the aforementioned technologies, if designed to (probabilistically) control a user's mental state on the basis of inaccurate information or controversial theoretical assumptions, such as a potentially restrictive taxonomy (e.g. restrictive taxonomy of distinct emotional states).

These consequences require careful discussion of the ethical, legal, and social issues that emerge from use of machines that can read our minds (Burr et al. 2018). We turn to discuss some specific cases now.

4.2 Consent and Trust

As we act we constantly leak information about our goals, beliefs, orientations, mental states, and psychological traits. An analysis of our behaviour, if combined with sufficient data from a normative group, may allow learning algorithms to infer this information. It seems that several independent research communities have followed a similar trend in exploring this possibility. The result is that this technology is emerging without coordinated oversight.

In our review, we did not make a distinction between cases where the subject is willing or cooperating and the cases where the subject is unaware or opposed to the assessment. In principle, many of the methods could be performed on unknowing or unwilling subjects, for whom the relevant samples of behaviour have been gathered.[29] The issue of consent has already been extensively discussed and debated (Boyd & Crawford 2012; Ioannidis 2013), and has influenced new forms of regulation, such as the European Union's General Data Protection Regulation (GDPR), which seeks to restrict the collection and use of data (e.g. requirement of explicit consent).[30]

However, in relation to the ethical implications that arise from inferring a user's mental state or psychological traits on the basis of some digital sample of behaviour, the issue of consent should not be discussed as a general principle, because specific uses of inferred knowledge will likely lead to differing ethical concerns. For example, individuals may not view a lack of consent as particularly concerning in cases where the inferred information is simply used for choosing which advertisement to display (e.g. persuasion). However, if the information is used in an attempt

[29] Specific instances of data collection without consent have been reported. For example, Purnell (2018) reports that a London-based security firm uncovered a smartphone app that is pre-installed on devices in Myanmar, Cambodia, Brazil, India, and China, which automatically collects and transmits personal information (e.g. device information, location information) without the user's knowledge to a mobile-advertising firm.

[30] For example, recital 32 of the GDPR, which clarifies the definition of 'consent', states, "Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement." (European Commission 2016).

to (probabilistically) control the user's mental state, individuals may likely view the lack of consent as deeply problematic due to overlooking or not respecting their autonomy.

Furthermore, it is not always clear how much understanding a user may have about (a) the information being collected about their online activities, and (b) the types of uses (i.e. diagnosis, prediction, persuasion, or control) the data is collected for. The urgency of this issue has been re-emphasised recently, following the publication of a report from a research team at Vanderbilt University (Schmidt 2018). The report details a number of experiments in which a new Android smartphone was monitored to determine the scope and type of data that is sent to Google's servers. Importantly, the study found that two-thirds of the data collected is by *passive means* (i.e. without user input), and thus possibly without the user's knowledge or explicit consent. In one experiment, the study found that an Android device left idle with no user interaction sent ~900 data samples were sent to a variety of Google's servers over 340 instances and across a 24-hour period. When actively used, this amount of data collection rose to approximately 450 instances (1.4× the passive amount). The type of this data was varied, including personally identifying information (e.g. user name, birthdate, zip code, gender, device identifiers) as well as a range of behavioural information (e.g. websites visited, apps used, purchases made). Perhaps unsurprisingly, *location information* constituted 35% of all the data samples sent to Google, as much of this can be used for advertising purposes. However, it can also be used to determine higher-level behavioral characteristics such as whether a user is walking, cycling, running, etc. Finally, the report states that "Google identified user interests with remarkable accuracy" (ibid., p. 3), and that their findings "indicate that Google has the ability to connect the anonymous data collected through passive means with the personal information of the user." (ibid., p. 4). Although the study's authors used Google's privacy policies as a source of information about the type of data collection that occurs, it was not sufficient on its own to allow them to determine the full extent of the data collection. It should therefore be clear why the type of user consent that can be gathered through privacy policies is not enough.

A related consideration arises for the matter of trust. Psychometrics rests on prior theoretical assumptions about why a particular test measures some postulated construct. Many of the studies in our review demonstrate *surprising* correlations between samples of (public) behaviour and (private) psychological information, which is connected with a key concept in psychological assessment known as *face validity* (i.e. the degree to which a test is subjectively viewed as establishing a sound basis for measuring the postulated construct). Face validity is important in establishing trust between test administrators and participants, and the use of digital footprints for bypassing tests may undermine this trust (e.g. would a participant accept that their gaze data is a strong predictor of personality?). Like consent, this may be problematic to differing degrees in certain domains. For example, an employer may risk upsetting potential candidates by using non-traditional forms of assessment, which despite having high predictive accuracy according to some criterion (e.g. job performance), are not evaluated by the candidate as being valid assessment tools.

These considerations highlight a need for the relevant research communities, and the organisations using the aforementioned techniques, to carefully consider the

specific ethical issues that arise in the inference of particular mental states and psychological traits—it is unlikely that broad, all-encompassing principles will suffice.

4.3 From Galton to Google; from Fechner to Facebook

Our paper is primarily concerned with showing how many of the methods used in psychological assessment can be bypassed, rather than replaced, by utilising ML techniques. Nevertheless, it is worthwhile taking the opportunity to briefly consider some of the consequences that ML may have for the ongoing development and application of psychological assessment.

Firstly, a quick terminological note on psychometrics. The dominant paradigm in psychometrics is item response theory (IRT), a statistical framework that models the relationship between the degree to which an individual's possesses some proposed construct (e.g. a trait, often represented by the greek letter 'θ') and their subsequent performance (response) on a set of items in a given psychometric test (Rust and Golombok 2009).

In IRT, every choice reveals information about a latent variable (θ), under the assumption of conditional independence of choices. Internal calibration (reliability) allows us to know the probability distribution of responses given the latent trait (e.g. the distribution of scores to some item × among extroverts). As already noted, this process is very similar to a class of problems known as "inverse problems" where a hidden (or latent) cause needs to be inferred (or postulated) based on its observable effects. While generally "ill-posed", in practice this class of problems can often be solved, under the appropriate assumptions.

Importantly, an 'item' is defined by the Standards for Educational and Psychological Assessment as "a statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task" (American Educational Research Association et al. 2014, p. 220). This means that in principle, an online behaviour could constitute an item response, under certain assumptions (see LIWC case study, Sect. 2.2).

Furthermore, in IRT the reliability and validity of psychometric assessments can also be evaluated statistically. Validity is "the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test." (American Educational Research Association et al. 2014, p. 225). In short, a valid measure is one that measures what it is intended to measure. Reliability is "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test takers." (American Educational Research Association et al. 2014, p. 223)

The statistical nature of IRT means that ML would be well-posed to automate many (but not all) aspects of the assessment process.[31] Indeed, others have already argued that "principles and techniques from the field of machine learning can help psychology become a more predictive science" (Yarkoni and Westfall 2017). However, as previously noted, the impact of ML on the theoretical development of psychometrics is beyond the scope of this article.

Returning to the consequences of ML techniques for the development and application of psychological assessment, one obvious point is that as new datasets are collected, we may find better signals that predict the various constructs we covered in our review, or maybe enable researchers to predict abilities that we have not considered (e.g. numerical and spatial reasoning). This means that datasets that have already been collected, without appropriate regulatory oversight, could be re-mined and analysed for additional psychological insights.

Another possibility is that existing datasets could be linked together, increasing the reliability and validity of current techniques. As Luhmann (2017, p. 30) states, with regards to a review of big data assessments of subjective well-being: "To date, no single data source seems reliable and valid enough to replace traditional self-report measures of well-being. However, this may change as more data sources are developed, validated, and combined."

By linking datasets together, we may also find further latent traits that are uninterpretable by traditional psychological standards—techniques for unsupervised learning will likely prove to be invaluable in this regard. These undiscovered traits may turn out to be better predictors of behaviour, which would have obvious financial benefit for many organisations that are unconcerned with theoretical constraints such as construct validation, and merely wish to improve their ability to influence user behaviour. Such developments could likely exacerbate ethical concerns as a result of linking certain datasets (e.g. measuring the probability of anxiety among conservatives, and using this information to develop particular campaign strategies; incorrectly detecting positive emotions in individuals suffering from depression and withholding necessary treatment).

As is to be expected, it is clear that there are possible advantages and disadvantages for how this technology could be developed and utilised. Because of the wide-reaching effects of these technologies, it is imperative that ongoing communication between the various communities continues—it is our hope that the current paper demonstrates the importance of ongoing collaboration.

[31] See Alexandrova and Haybron (2016) for a discussion of the epistemological and methodological nature of construct validation, including an argument in favour of the "coherentist logic" that underpins it. This aspect of psychometrics is perhaps the most resistant to automation or replacement by machine learning techniques.

5 Conclusion

Current technologies can already infer probabilistic information about our mental states and psychological traits, and classify us in ways that bypass traditional forms of psychological assessment. Our review identifies just a portion of the many studies in which different types of behavioural samples can be used by an algorithm to read our minds. Many more methods are still being studied and developed across different communities for the same purpose.

As the types and amount of interaction between us and our online devices increases, and as new types of sensors for measuring behavioural signals are developed, there is the expectation that by combining these sources of information a ML algorithm could form a very accurate image about us.

The likely convergence of these technologies and methods raises many ethical issues—beyond the topics of consent and trust that we have explored. Most notably, there are the risks associated with enabling intelligent systems to take actions that aim to control our behaviour, on the basis of inferred psychological information (Burr et al. 2018). These issues will not be solved entirely by legislation, and the individual research communities reviewed should not be expected to develop ethical guidelines on their own. Rather, it is imperative that policymakers and researchers understand the scope of these developments, in order to better facilitate the ongoing discussions about the growing use and convergence of machines that can read our minds and control our behaviours.

We believe that the pace of progress is such that looking at the work of multiple communities within a unified framework can help understand how much progress has been made, and may help us better see what is currently occurring and may continue to emerge in the near future.

Acknowledgements Christopher Burr and Nello Cristianini were supported by European Research Council Advanced Grant ThinkBIG [Advanced Grant (AdG), PE6, ERC-2013-ADG], awarded to Nello Cristianini.

References

Alexandrova, A. (2017). *A Philosophy for the Science of Well-Being*. New York: Oxford University Press.

Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science, 83*(5), 1098–1109.

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., et al. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior, 21,* 361–376.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Publishing.

Annalyn, N., Bos, M. W., Sigal, L., & Li, B. (2018). Predicting Personality from Book Preferences with User-Generated Content Labels. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2018.2808349.

Ariely, D., & Berns, G. S. (2010). Neuromarketing: The hope and hype of neuroimaging in business. *Nature Reviews Neuroscience, 11*(4), 284–292.

Attali, T., & Burstein, J. (2005). Automated Essay Scoring With e-rater® v.2.0. Educational Testing Service [Research Report]. URL: https://www.ets.org/Media/Research/pdf/RR-04-45.pdf.

Baras, K., Soares, L., Paulo, N., & Barros, R. (2016). 'Smartphine': Supporting students' well-being according to their calendar and mood. In *Presented at the 2016 international multidisciplinary conference on computer and energy science (SpliTech) IS-SN-VO–VL* (pp. 1–7).

Beck, A. T., Ward, C. M., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Bone, D., Lee, C. C., & Narayanan, S. S. (2012). A robust unsupervised arousal rating framework using prosody with cross-corpora evaluation. In *Presented at the thirteenth annual conference of the international speech communication association*.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In: *Presented at the proceedings of the 9th international conference on web and social media, ICWSM 2015* (pp. 31–40).

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The selfassessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing, 18*(1), 32–39.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, *28*(4), 735–774.

Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.). (2015). *The Oxford handbook of affective computing*. Oxford: Oxford University Press.

Calvo, R. A., & Peters, D. (2014). *Positive computing: Technology for wellbeing and human potential*. Cambridge: MIT Press.

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2017). The datafication of talent: How technology is advancing the science of human potential at work. *Current Opinion in Behavioral Sciences, 18,* 13–16. https://doi.org/10.1016/j.cobeha.2017.04.007.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology, 9*(3), 621–640. https://doi.org/10.1017/iop.2016.6.

Chen, J., Hsieh, G., Mahmud, J. U., & Nichols, J. (2014). Understanding individuals' personal values from social media word use. In *Presented at the the 17th ACM conference, New York*, New York: ACM Press. doi: http://doi.org/10.1145/2531602.2531608.

Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality, 42,* 96–132.

Cohen, R., & Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy!. In *Presented at the proceedings of the 7th international conference on weblogs and social media, ICWSM 2013*.

Cohn, J., & De La Torre, F. (2015). Automated face analysis for affective computing. In Rafael A. Calvo, Sidney K. D'Mello, Jonathan Gratch, & Arvid Kappas (Eds.), *The oxford handbook of affective computing* (pp. 131–150). Oxford: Oxford University Press.

Collins, R. P., Litman, J. A., & Spielberger, C. D. (2004). The measurement of perceptual curiosity. *Personality and Individual Differences, 36*(5), 1127–1141.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Presented at the proceedings of the 7th international conference on weblogs and social media, ICWSM 2013* (pp. 128–137).

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*(3), 542–575.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the USA, 111*(15), 1–9. https://doi.org/10.1073/pnas.1322355111.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200.

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford: Oxford University Press.

European Commission. (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), Url: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. Accessed 5 July 2018.

Freitas, A., Brito, L., Baras, K., & Silva, J. (2017). Overview of context-sensitive technologies for well-being. In *Presented at the 2017 international conference on internet of things for the global community (IoTGC) IS-SN-VO–VL* (pp. 1–8).

Grover, T., & Mark, G. (2017). Digital footprints. In *Presented at the the 2017 ACM international joint conference on pervasive and ubiquitous computing and the 2017 ACM international symposium on wearable computers, New York* (pp. 41–44), New York: ACM Press. doi: http://doi.org/10.1145/3123024.3123139.

Hao, B., Li, L., Gao, R., Li, A., & Zhu, T. (2014). Sensing subjective well-being from social media. In D. Ślęzak, G. Schaefer, S. T. Vuong, & Y. S. Kim (Eds.), *Active media technology, 10th international conference, AMT 2014* (pp. 324–335). Cham: Springer International Publishing.

Harman, G. (1999). Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society*, *99*(3), 315–331.

Hare, R. D. (2003). *Manual for the hare psychopathy checklist-revised* (2nd ed.). Toronto: Multi- Health Systems.

Hart, S. G., & Staveland, L. E. (1998). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances Psychology, 52,* 139–183.

Hern, A. (2018). *Cambridge Analytica: How did it turn clicks into votes?* The Guardian [Online], URL: https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie. Accessed 5 July 2018.

Hoppe, S., Loetscher, T., Morey, S. A., & Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience, 12,* 81–88. https://doi.org/10.3389/fnhum.2018.00105.

Hu, A., & Flaxman, S. (2018). Multimodal sentiment analysis to explore the structure of emotions. In: *Presented at 2018 ACM SIGKDD international conference on knowledge discovery and data mining*. Preprint: arXiv:1805.10205 [stat.ML].

Huppert, F. A., & So, T. T. C. (2013). Flourishing across Europe: Application of a new conceptual framework for defining well-being. *Social Indicators Research, 110*(3), 837–861. https://doi.org/10.1007/s11205-011-9966-7.

Hyönä, J., & Olson, R. K. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 21*(6), 1430–1440.

Ioannidis, J. P. (2013). Informed consent, big data, and the oxymoron of research that is not research. *The American Journal of Bioethics, 13*(4), 40–42.

Issenberg, S. (2012). How Obama's team used big data to rally voters. *Wired [Online]*, Url: https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/ Accessed 6 July 2018.

Jimison, H., Pavel, M., & Le, T. (2008). Home-based cognitive monitoring using embedded measures of verbal fluency in a computer word game. In *30th annual international IEEE EMBS conference* (pp. 3312–3315).

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology, 120*(2), 263–286.

Karg, M., Kühnlenz, K., & Buss, M. (2010). Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40*(4), 1050–1061. https://doi.org/10.1109/TSMCB.2010.2044040.

Koldijk, S., Neerincx, M. A., & Kraaij, W. (2016). Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on Affective Computing, 9*(2), 227–239. https://doi.org/10.1109/TAFFC.2016.2610975.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70*(6), 543–556. https://doi.org/10.1037/a0039210.

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. https://doi.org/10.1073/pnas.1218772110.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788–8790.

Lambiotte, R., & Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE, 102*(12), 1934–1939. https://doi.org/10.1109/JPROC.2014.2359054.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud faces database. *Cognition and Emotion, 24*(8), 1377–1388.

Lazer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721–723. https://doi.org/10.1126/science.1167742.

Lin, Y. H., Chang, L. R., Lee, Y. H., Tseng, H. W., Kuo, T. B., & Chen, S. H. (2014). Development and validation of the smartphone addiction inventory (SPAI). *PLoS ONE, 9,* e98312. https://doi.org/10.1371/journal.pone.0098312.

Lin, Y. H., Lin, Y. C., Lin, S. H., Lee, Y. H., Lin, P. H., Chiang, C. L., et al. (2017). To use or not to use? Compulsive behavior and its role in smartphone addiction. *Nature Translational Psychiatry, 7,* e1030. https://doi.org/10.1038/tp.2017.1.

Luhmann, M. (2017). Using big data to study subjective well-being. *Current Opinion in Behavioral Sciences, 18,* 28–33. https://doi.org/10.1016/j.cobeha.2017.07.006.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences, 114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114.

Mavani, V., Raman, S., & Miyapuram, K. P. (2017). Facial expression recognition using visual saliency and deep learning. In *Presented at the 2017 IEEE international conference on computer vision workshop (ICCVW), IEEE* (pp. 2783–2788). Doi: http://doi.org/10.1109/ICCVW.2017.327.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90,* 862–877.

Metz, R. (2018). The smartphone app that can tell you're depressed before you know it yourself. MIT Technology Review [Online], Url: https://www.technologyreview.com/s/612266/the-smartphone-app-that-can-tell-youre-depressed-before-you-know-it-yourself/. Accessed 10 Jan 2019.

Mitchell, T. (1997). *Machine learning*. Singapore: McGraw-Hill.

Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N., & Takada, S. (2017). Detecting abnormal word utterances in children with autism spectrum disorders. *Perceptual and Motor Skills, 124*(5), 961–973.

Nettle, D. (2009). *Personality: What makes you the way you are*. Oxford: Oxford University Press.

Nowak, M., & Eckes, D. (2014). United States Patent No. 8825764 - Determining user personality characteristics from social networking system communications and characteristics. Retrieved from http://patft.uspto.gov/netahtml/PTO/index.html. Accessed 6 July 2018.

Pennebaker, J. W. (2011). *The secret life of pronouns: How our words reflect who we are*. London: Bloomsbury.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin: University of Texas at Austin.

Purnell, N. (2018) *App traps: How cheap smartphones help themselves to user data*. Wall Street Journal [Online], Url: https://www.wsj.com/articles/app-traps-how-cheap-smartphones-help-themselves-to-user-data-1530788404. Accessed 6 July 2018.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401.

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Sci., 6*(1), 1–12.

Rello, L., & Ballesteros, M. (2015). Detecting readers with dyslexia using machine learning with eye tracking measures. In *Presented at the the 12th Web for All Conference, New York* (pp. 1–8) New York: ACM Press. Doi: http://doi.org/10.1145/2745555.2746644.

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018). *How Trump consultants exploited the facebook data of millions*. New York Times [Online]. URL: https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html. Accessed 19 Mar 2018.

Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression- vulnerable college students. *Cognition and Emotion, 18,* 1121–1133.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). London: Pearson International.

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London: Routledge.

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69*(4), 719–727.

Schmidt, D. (2018). Google Data Collection. *Digital Content Next [Online]*, Url: https://digitalcontentnext.org/wp-content/uploads/2018/08/DCN-Google-Data-Collection-Paper.pdf. Accessed Aug 21 2018.

Schreiber, D. (2017). Neuropolitics: Twenty years later. *Politics and the Life Sciences, 36*(2), 114–131.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 Countries. *Advances in Experimental Social Psychology, 25,* 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6.

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. In *Questionnaire package of the european social survey* (pp. 259–290).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9), e73791–e73796. https://doi.org/10.1371/journal.pone.0073791.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2017). *Deep image reconstruction from human brain activity*. Preprint: bioRxiv. https://doi.org/10.1101/240317.

Steele, V. R., Rao, V., Calhoun, V. D., & Kiehl, K. A. (2017). Machine learning of structural magnetic resonance imaging predicts psychopathic traits in adolescent offenders. *NeuroImage, 145,* 265–273. https://doi.org/10.1016/j.neuroimage.2015.12.013.

Svoboda, E. (2018). The "neuropolitics" consultants who hack voters' brains. MIT Technology Review [Online], Url: https://www.technologyreview.com/s/611808/the-neuropolitics-consultants-who-hack-voters-brains/. Accessed 22 Aug 2018.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676.

Vizer, L. M., Zhou, L., & Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *Journal of Human Computer Studies, 67*(10), 870–886. https://doi.org/10.1016/j.ijhcs.2009.07.005.

Wachter, S. & Mittelstadt, B. (Forthcoming). A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. In Columbia Business Law Review. Available at SSRN (September 13, 2018), URL: https://ssrn.com/abstract=3248829. Accessed 9 Jan 2019.

Wang, J., Cherkassky, V. L., & Just, M. A. (2017). Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human Brain Mapping, 38*(10), 4865–4881.

Wang, Y., & Kosinski, M. (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. https://doi.org/10.1037/pspa0000098.

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the positive and negative affect schedule—Expanded Form.

Wei, Z., Wu, C., Wang, X., Supratak, A., Wang, P., & Guo, Y. (2018). Using support vector machine on EEG for advertisement impact assessment. *Frontiers in Neuroscience, 12,* 812–821. https://doi.org/10.3389/fnins.2018.00076.

World Health Organisation. (2018). *International statistical classification of diseases and related health problems for mortality and morbidity statistics (ICD-11 MMS)*. Url: https://icd.who.int/browse11/l-m/en. Accessed 17 July 2018.

Yang, C., & Srinivasan, P. (2016). Life satisfaction and the pursuit of happiness on twitter. *PLoS ONE, 11*(3), e0150830–e0150881. https://doi.org/10.1371/journal.pone.0150881.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393.

Zijlstra, F., & van Doorn, L. (1985). The construction of a scale to measure subjective effort. Ph.D. dissertation, Dept. Philosophy Social Sci., Delft Univ. Technol., Delft, CD, The Netherlands.