**RESEARCH ARTICLE**

WILEY

# Personality computing: New frontiers in personality assessment

Le Vy Phan [ORCID]  |  John F. Rauthmann

Bielefeld University, Bielefeld, Germany

**Correspondence**
Le Vy Phan, Universität Bielefeld, Abteilung
Psychologie, Universitätsstraße 25, Bielefeld
33615, Germany.
Email: le.phan@uni-bielefeld.de

## Abstract

Personality Computing (PC) is a burgeoning field at the intersection of personality and computer science that seeks to extract personality-relevant information (e.g., on Big Five trait levels) from sensor-assessed information (e.g., written texts, digital footprints, smartphone usage, non-verbal behavior, speech patterns, game-play, etc.). Such sensor-based personality assessment promises novel and often technologically sophisticated ways to unobtrusively measure individual differences in a highly precise, granular, and faking-resistant manner. We review the different conceptual underpinnings of PC; survey how well different types of sensors can capture different types of personality-relevant information; discuss the evaluation of PC performance and psychometric issues (reliability and validity) of sensor-derived scores as well as ethical, legal, and societal implications; and highlight how modern personality and computer science can be married more effectively to provide practically useful personality assessment. Together, this review aims to introduce readers to the opportunities, challenges, pitfalls, and implications of PC.

## 1 | INTRODUCTION

Personality Computing (PC) is a burgeoning field at the intersection of personality and computer science seeking to extract personality-relevant information (e.g., on Big Five trait levels) from machine-sensed information (e.g., written texts, digital footprints, smartphone usage, non-verbal behavior, speech patterns, game-play, etc.) via

machine learning (ML) approaches. Novel ML approaches offer the tools needed to appropriately analyze vast amounts of sensor-based data and can be used to predict personality scale scores (e.g., Ilmini & Fernando, 2017; Stachl et al., 2020a, 2020b; Vinciarelli & Mohammadi, 2014a). Sensor-based personality assessment is a promising way to unobtrusively measure individual differences in a highly precise, granular, and faking-resistant manner (Harari et al., 2015, 2016). This new area of research may inform personality theory in an unprecedented fashion through the detection of complex patterns that cannot be retrieved by traditional statistical approaches. PC also holds the potential to yield real-world applications that might improve health and well-being through personalized interventions (Alexander et al., 2020). There has been an increase in PC studies in recent years (Mehta et al., 2019), originating not only from personality psychology but also from computer science given the centrality of highly complex data-analytical methods.

The purpose of this article is to give a basic introduction to PC. First, we briefly introduce personality science and discuss the conceptual underpinnings of PC. We then review the field of PC, focusing on the performance of different approaches, limitations regarding performance, and psychometric issues. We conclude with a short overview of some ethical, legal, and societal implications and discuss the role of personality psychologists in PC. Throughout this article, we use the term *Personality Computing*, coined by Vinciarelli and Mohammadi (2014a), to describe ML approaches to personality research and assessment. However, other terms such as computational personality traits assessment (e.g., Ilmini & Fernando, 2017), digital phenotyping (e.g., Onnela & Rauch, 2016), psychoinformatics (e.g., Markowetz et al., 2014), personality sensing (e.g., Harari et al., 2020a, 2020b), automatic personality detection (e.g., Mehta et al., 2019), ML personality assessment (e.g., Bleidorn & Hopwood, 2019), and automatic personality assessment (e.g., Kedar & Bormane, 2015) have been used to describe similar approaches.

## 2 | PERSONALITY SCIENCE AND THE MEASUREMENT OF PERSONALITY

People's personalities consist of sets of individual differences, often referred to as psychological *traits*, that describe and explain their enduring patterns of mental and behavioral processes (Funder, 2001; Wilt & Revelle, 2015), such as: perception, cognition (thoughts, attitudes, etc.), emotion (affect, feelings, etc.), motivation (desires, needs, goals, etc.), and behaviors (e.g., verbal; non-, para-, and extra-verbal; action sequences; etc.). Personality traits cannot be directly observed but are expressed in, and thus can be inferred from, people's recurring (re-)action patterns. They are consequential for people's personal, occupational, and social daily lives (Ozer & Benet-Martínez, 2006; Roberts et al., 2007) and predict many important life outcomes (e.g., mental and physical health, longevity, income, romantic attraction, divorce). The power of personality is far-reaching and pervades every aspect of our lives (Beck & Jackson, 2020; Soto, 2019).

Personality science has produced a cornucopia of measures that allow reliable and valid assessments of individual differences. Many measures assess people's standing on the so-called Big Five traits, the most widely used trait taxonomy. The Big Five are: Openness to new experiences/Open-mindedness (O; e.g., *curious, aesthetically sensitive, imaginative*), Conscientiousness (C; e.g., *organized, productive, responsible*), Extraversion (E; e.g., *sociable, assertive, energetic*), Agreeableness (A; e.g., *compassionate, respectful, trusting*), and Neuroticism/Negative Affectivity (N; e.g., *anxious, depressed, volatile*). These Big Five represent broad domains of individual differences that are to a substantial degree heritable, biologically rooted but also molded by the environment, relatively stable across time (though they can change across the lifespan), and can be identified across many cultures and different trait measures (John & Srivastava, 1999; John et al., 2008).

Most trait measures make use of two fundamental data sources: the *self* and *other people* (e.g., friends, family members). By far the most common way of assessing traits is via self-reports: People describe how they see themselves, for example, by answering items such as "I am someone who is outgoing, sociable" on a Likert-type scale (e.g., from 1 = disagree strongly to 5 = agree strongly). Self-reports tap people's explicit self-concepts about their traits, which are parts of their *identities*. However, the way we see ourselves may not necessarily

correspond to how we behave and how others see us. Other-reports tap how other people see a target-person, which is part of someone's *reputation*. For example, an acquaintance rates a target-person on the item "She/He is someone who is outgoing, sociable." Together, identity and reputation are integral parts of personality (Hogan, 1996; Hogan & Roberts, 2000). Though moderately correlated (Connelly & Ones, 2010), they do not fully overlap and each harbor unique insights and predictive value (Vazire, 2010). Each source has different information available (McAbee & Connelly, 2016) and may be contaminated with different forms of "biases" (the self: e.g., self-serving bias, social desirability, impression management; others: e.g., familiarity, liking, relationship). A multimethod approach considering both self- and other-reports may provide a better, or more accurate, picture of target-persons' "true" personalities (Funder, 1995; Kenny, 1994). For example, *composite trait scores* capturing shared variance between self- and other-reports can be used as proxies for "actual" traits (i.e., what is known to the self *and* others).

Besides self- and other-reports, there are numerous other, more behavior-linked data sources that personality researchers could tap. For example, some bio-physiological data (e.g., pulse, heart rate, skin conductance) may be indicative of people's transient affective states but also their enduring traits (DeYoung, 2015; Eysenck, 1967). These can be understood as involuntary and automatic reactions to circumscribed sets of stimuli. Furthermore, behavioral products that remain in an environment after a behavior has been performed (e.g., purchase decisions, room arrangements, clothing) and digital footprints (e.g., Facebook Likes, Tweets, emails) have also been used to infer people's personality traits with some accuracy (Gosling et al., 2002; Kosinski et al., 2013; Kulkarni et al., 2018; Youyou et al., 2015). Such novel methods have far-reaching consequences because people may be unaware they are giving off personality-relevant information and thus would not be able to fake their scores. However, personality is expressed in any form of behavior (Table 1)—be it real-time physiological reactions, complex movements, or lasting remnants.

## 3 | PERSONALITY COMPUTING

The field of PC is mainly concerned with three broad problems (Pianesi, 2013; Vinciarelli & Mohammadi, 2014a): Automatic Personality Recognition, Automatic Personality Perception, and Automatic Personality Synthesis. Automatic Personality Recognition and Automatic Personality Perception approaches infer trait scores from machine-detectable *personality-expressive signals* (for an overview of behavioral modalities, see Table 1). These signals—pertaining to (re-)action patterns or behavioral processes—can be derived, for example, from information encapsulated in text, audio, video, mobile phone data, digital footprints, wearables, and online games. This information is processed (automatically) by machines without human raters. ML algorithms are then trained to predict self-reported personality trait scores (Automatic Personality Recognition) or other-reported trait scores (Automatic Personality Perception) from machine-detected signals. Targets' self-reports and other-reports serve as benchmarks for *accuracy*. Automatic Personality Synthesis attempts to generate artificial personalities via virtual, or embodied agents (e.g., avatars, robots, smart homes).

Vinciarelli and Mohammadi (2014a) place these three approaches within a Brunswikian lens framework (Brunswik, 1956) based on attribution and externalization processes in interpersonal perception and human-machine interaction (see Figure 1). In this framework, aspects of an individual's personality (e.g., traits) are externalized via machine-detectable behavioral signals (i.e., *distal cues*). Behavioral signals can be actual behavior (e.g., speech, movements, gestures, facial expressions, etc.) or behavioral products. Automatic Personality Recognition is concerned with the automatic inference of a person's self-reported traits based on these signals, or cues. The relationship between (self-reported) traits and distal cues is called *ecological validity* (Brunswik, 1956; not to be confused with external validity; Araújo et al., 2007). Importantly, not all behaviors are trait expressions. All experiences and behavior are always contextualized and determined by both the person and the situation (Rauthmann, 2016). Depending on the situation, the same behavior can vary in how indicative it is for an individual's trait and, thus, differ in its ecological validity.

**TABLE 1** A nonexhaustive overview of behavioral modalities used in PC

| Modality | Sample data sources | Sample features | Sample studies |
|---|---|---|---|
| (Written) Text | Essays, blog posts, conversation transcripts | Word category frequencies, LIWC, MRC PD, N-grams | Mairesse et al. (2007) Oberlander & Nowson (2006) Gill et al. (2009) Majumder et al. (2017) Poria et al. (2013) |
| Audio | Conversations, speech clips, meetings, self-presentations | Prosody, speech activity, voice quality, interaction features, OpenSMILE speech features | Mohammadi & Vinciarelli (2012) Mohammadi et al. (2012) Valente et al. (2012) |
| Visual | Vlogs, videos, images, meetings, movie excerpts, faces, self-presentations | Interpersonal distances, walking velocity, body movements, face/hand/head movements, turn takings, motion activity | Zen et al. (2010) Pianesi et al. (2008) Batrinca et al. (2011) Batrinca et al. (2012) Biel et al. (2011) Biel et al. (2012) Gürpinar et al. (2016) |
| Social media | Social media (Facebook, Twitter, etc.) profiles, posts, blog posts, favorite pictures, likes, tweets | Profile info, egocentric networks, LIWC, MRC PD, number of followers/following, usage statistics, visual patterns, photo content, aesthetic preferences, Facebook likes and activities frequencies | Golbeck et al. (2011a) Golbeck et al. (2011b) Quercia et al. (2011) Nguyen et al. (2011) Cristani et al. (2013) Celli (2012) Kosinski et al. (2013) Youyou et al. (2015) |
| Smartphones | Smartphone/mobile phone usage logs, mobile-sensed data | Locations, proximity of other devices through Bluetooth scanning, event logs of phone calls, text messages, app usage, etc., movement patterns | Chittaranjan et al. (2011, 2013) Staiano et al. (2012) Wang et al. (2018) Harari et al. (2020a) Stachl et al. (2020a) |
| Wearables | Sociometric badge recordings, EAR recordings | Ambient sound, speech/physical activity, interaction, proximity | Olguín-Olguín et al. (2009) |
| Computer games | World of Warcraft profiles | Gaming behavior | Yee et al. (2011) |

*Note:* Modalities can also be combined for *multimodal* PC.

Abbreviations and explanations: EAR = electronically activated recorder; LIWC = linguistic inquiry and word count (text analysis program to categorize percentages of words in pre-defined categories); MRC PD = Medical Research Council Psycholinguistic Database (machine-readable dictionary containing over 150,000 words, including [psycho-]linguistic attributes for each); N-grams = sequence of *N* items (e.g., words, letters) in a text; OpenSMILE = open-source speech and music interpretation by large-space extraction (toolkit for audio feature extraction and music and speech signal classification); prosody = properties of speech and linguistic functions (e.g., intonation, tone, stress).
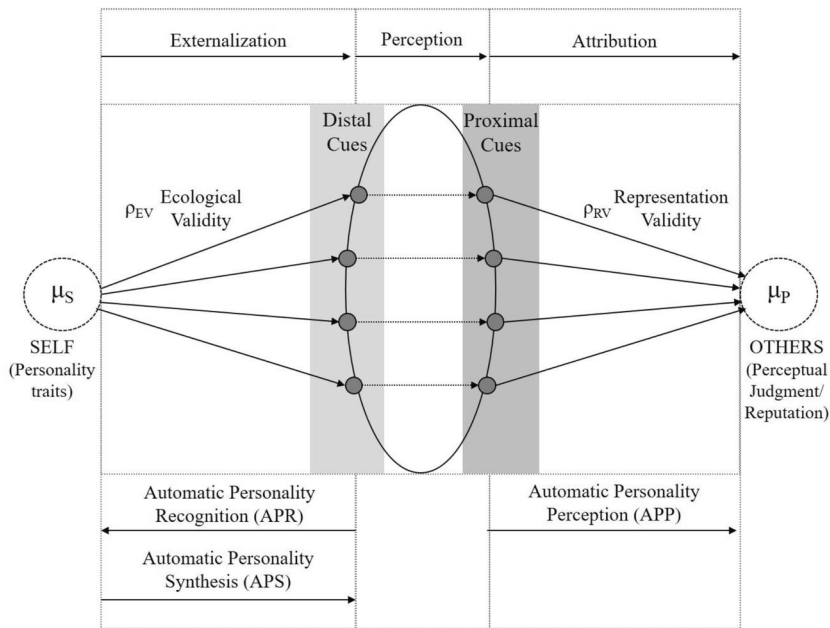
**FIGURE 1** Brunswik lens of processes underlying personality computing. Adapted from Vinciarelli and Mohammadi (2014a). $\mu_S$ = inference of self-reports from distal cues; $\mu_p$ = inference of other-reports from proximal cues; $\rho_{EV}$ = ecological validity = correlation between self-reported personality traits and distal cues; $\rho_{EV}$ = representation validity = correlation between proximal cues and other-reported personality traits

The lens represents sensory-perceptual processes transforming distal cues to *proximal cues*, the cues that an observer actually perceives. Naturally, proximal and distal cues overlap, but some behavioral signals may be omitted in perceptual processes and signals that were not expressed can be perceived erroneously. Proximal cues activate attribution processes and are used to form perceptual judgments about a target's personality (i.e., trait levels attributed by observers). The relationship between a persons' reputations and proximal cues is called *representation validity*.

Automatic Personality Perception theoretically concerns the automatic inference of a person's reputation from proximal cues. In practice, however, it is usually not possible to determine which cues were actually perceived, so most Automatic Personality Perception research uses distal cues to approximate proximal cues. Furthermore, Automatic Personality Perception usually aims at the prediction of average judgments (e.g., of a target's trait) across multiple raters. This means Automatic Personality Perception identifies cues that best predict average ratings, possibly ignoring relevant associations between certain cues and individual perceptual judgments.

Lastly, in Automatic Personality Synthesis, machines (e.g., robots, avatars, virtual agents) express automatically generated distal cues to elicit attributions of certain trait levels in human observers. The main goal is to elicit attributions as planned by machine designers (Vinciarelli & Mohammadi, 2014a). Effective Automatic Personality Synthesis requires further advancements and improvements in Automatic Personality Recognition and Automatic Personality Perception. For example, household robots, smart homes, and assisted programs (e.g., in physiotherapy or psychotherapy) can be built to express certain personalities—either normatively to everyone or tailored idiographically to a user after "reading" the user's trait levels (i.e., Automatic Personality Recognition). Automatic Personality Synthesis informed by Automatic Personality Recognition and Automatic Personality Perception could achieve more trust and compliance as well as more efficiency and effectiveness in human-machine interactions (Salem et al., 2015).

## 3.1 | Performance

How well do Automatic Personality Recognition and Automatic Personality Perception typically perform? This question is difficult to answer because different studies use (a) different signals (e.g., from text, speech, movement, digital footprints) and features (i.e., extracted variables such as mobile phone usage logs, social media profile info, prosody, word category frequencies, etc.); (b) different tasks (e.g., classification or ordinal regression with $n$ classes; linear regression; supervised vs. unsupervised ML); (c) different ways to group into classes (e.g., median-split lower vs. upper half; extreme groups); and (d) different performance measures (for an overview, see Table 2) that cannot easily be converted to a common metric. Furthermore, performance is influenced by the underlying distributions of traits (e.g., normal vs. skewed distribution) and the reliability of a target measure (i.e., performance is limited by a scale's reliability). However, some articles have attempted to summarize Automatic Personality Recognition and Automatic Personality Perception literatures (reviews by Mehta et al., 2019; Vinciarelli & Mohammadi, 2014a; meta-analysis by Azucar et al., 2018).

Vinciarelli and Mohammadi (2014a) provided a comprehensive overview of earlier Automatic Personality Recognition and Automatic Personality Perception works but refrained from meta-analytic integration of results. Here, we will selectively point out a few observations on the performances achieved in some studies included in their review. In binary classification, the most commonly used approach, people are categorized as either high or low on a trait. Binary classification based on communicative signals (i.e., linguistic, para-verbal, nonverbal features) mostly performed at around 60%–75% accuracy (five Automatic Personality Recognition studies, range: ~33%–95%; four Automatic Personality Perception studies, range: ~55%–75%) for all Big Five traits. Binary Automatic Personality Recognition from text (i.e., linguistic features) achieved accuracies mostly around 60% (five studies, range: ~55%–100%). Accuracies decreased by ~10% when using three (low vs. medium vs. high) instead of two classes (three Automatic Personality Recognition studies, range: 17.4% decrease to 2.6% increase). In regression approaches that treat traits as dimensional, performances around $\rho = 0.20$–$0.30$ were achieved (studies: one Automatic Personality Recognition and one Automatic Personality Perception study from social media; one Automatic Personality Recognition study from computer games; range: 0.12–0.55).

In a recent systematic review, Mehta et al. (2019) determined mean best accuracies achieved with deep learning Automatic Personality Recognition and Automatic Personality Perception approaches using popular data sets. For the Big Five, multimodal approaches resulted in higher prediction accuracies (91.7% mean best accuracy) compared to unimodal approaches from just text, audio, or visual signals (best accuracies range: ~58%–91%; however, most hover around ~60%–70%). Among unimodal approaches, visual modalities (e.g., images or videos) yielded the highest accuracies (~91%). The authors further highlighted that for most modalities the best performance had been achieved using deep learning techniques (compared to traditional or "shallow" ML) and that these can be expected to perform even better in the future.

A recent meta-analysis by Azucar et al. (2018) focused solely on Automatic Personality Recognition from digital footprints on social media and reported meta-analytic correlations ranging from $r = 0.29$ for agreeableness to $r = 0.40$ for extraversion. To our knowledge, an extensive and systematic meta-analysis across *all* signal modalities has not yet been attempted.

In conclusion, Automatic Personality Recognition and Automatic Personality Perception increasingly result in above-chance predictions of personality trait scores. However, several factors impede the comparability across studies highlighting the need for a systematic meta-analysis to assess average performance for each trait and across different modalities. Furthermore, there are some problematic common practices in PC that complicate the evaluation of performance. We will elaborate on these issues in the next section.

To evaluate PC performance appropriately, it is also important to clearly distinguish between Automatic Personality Recognition and Automatic Personality Perception given that the self and others have unique insights into a person's personality (Vazire, 2010). For some studies, Automatic Personality Perception showed stronger

**TABLE 2**  Overview of commonly used performance measures

| Performance Measure | Explanation | Task |
|---|---|---|
| Accuracy | Proportion of correct predictions to total number of predictions | C |
| Pearson's $r$ | Linear rank measure of actual and target predicted values | R |
| (Adjusted) $R^2$ | (Adjusted) proportion of variance in the actual target variable that is explained by model-predicted values | R |
| Mean absolute error | Average of all absolute errors (i.e., absolute values of differences between actual target and model-predicted values) | R |
| Mean square error/ Root mean square error | Average of the squares of errors (i.e., difference between actual target and model-predicted values)/ Square root of the average of squared errors | R |
| Area under the receiver operating characteristics | Probability that a classifier ranks a random positive example higher than a random negative example | C |
| F-measure | Harmonic mean of precision (i.e., proportion of true positives to total number of positive predictions) and recall (i.e., proportion of true positives to total number of true positives and false negatives) | C |

Abbreviations: C = classification task; R = regression task.

performances than Automatic Personality Recognition, but this awaits further systematic replication. Furthermore, studies must move beyond binary classifications and acknowledge the continuous, hierarchical, and multifaceted nature of dimensional traits (Wright, 2014). Moreover, multimodal and deep learning approaches are promising avenues to further improve automatic personality prediction. We refer readers interested in the particulars of selected studies to the review articles cited here and to Finnerty et al. (2016).

## 3.2 | Evaluating PC performance

Above, we describe performances typically achieved in PC research without explicitly evaluating whether these can be regarded as good or bad. A *relative* evaluation of performances is difficult for three reasons. First, the evaluation of a model's performance depends on the research question and the purpose of the specific model. For example, if the goal is to replace traditional personality trait measures, performance should at least approach self-other ($rs \approx 0.29$–$0.41$; Connelly & Ones, 2010) or between-inventories correlations ($rs \approx 0.31$–$0.56$; Pace & Brannick, 2010). But if the goal is to complement (rather than replace) traditional assessment, then even with lower performances PC models can yield valuable insights into which behavioral variables are indicative of traits (Stachl et al., 2021). Second, the lack of meta-analytic aggregation prevents conclusive evaluations of PC performances and their heterogeneity. Third, as we detail below, there are common problematic practices in PC that likely distort resulting performances.

It is likely that performance estimates in PC are overly optimistic due to overfitting (Mønsted et al., 2018; Stachl et al., 2020b). A central objective of ML is to find models that generalize from a sample (i.e., training data) to new observations drawn from the same population (i.e., test data). To achieve this, ML models are trained to minimize overfitting. Overfitting means a statistical model erroneously fits sample-specific noise as signal, thus overestimating the performance in a training data set that does not generalize to a test data set. This especially tends to occur with large numbers of predictors and relatively small effect and sample sizes (Yarkoni & Westfall, 2017), which is often the case in PC studies (Mønsted et al., 2018). However, distorted performance estimates can also result from underfitting, which occurs when a statistical model does not adequately represent the true underlying function in the data (e.g., fitting a linear function if the true relationship is exponential). This leads to poor performance both in the training and test data sets (Lantz, 2019). Balancing the error sources associated with overfitting and underfitting is therefore a central goal in ML (Yarkoni & Westfall, 2017). Overfitting can be prevented by using resampling methods like *k*-fold cross-validation (however, this is insufficient for complex modeling settings; Stachl et al., 2020b) and large samples (Yarkoni & Westfall., 2017). However, only few data sets in PC contain large samples (see Mehta et al., 2019; Vinciarelli & Mohammadi, 2014a).

Another issue that may distort performance estimates is the practice of polytomizing or dichotomizing continuous scale scores when training and testing PC models using classification tasks (although some argue that personality scale scores are ordinal in nature and that regression accuracy metrics such as mean square error are not meaningful here; Mønsted et al., 2018). Polytomizing refers to the artificial creation of multiple discrete classes in a continuous variable. Dichotomizing means that continuous scores are divided into high versus low classes. Different approaches are used to assign discrete classes such as mean/median splitting (e.g., Chittanranjan et al., 2013), calculating quantiles (e.g., Mønsted et al., 2018), or clustering (e.g., Wöllmer et al., 2009). Polytomizing and dichotomizing are used to simplify data analysis and are appropriate if the underlying distribution of a variable is truly categorical or if meaningful thresholds exist (DeCoster et al., 2009). However, many personality variables tend to be normally distributed with most of the observations close to the mean/median and few observations at the tails (Schmitt et al., 2007). This leads to poor performance in binary classification tasks (Mariooryad & Busso, 2015). Moreover, using Monte Carlo simulations, DeCoster et al. (2009) showed that using continuous instead of dichotomized variables is generally preferable. Furthermore, dichotomizing can result in the underestimation of effect sizes, reduce statistical power, and increase false-positives (Cohen, 1983; MacCallum et al., 2002). Even in cases where effect sizes increase after dichotomization, this is likely due to sampling error, especially when sample or effect sizes are small (MacCallum et al., 2002). Lastly, dichotomizing impacts the measurement and representation of individual differences and inevitably entails a loss of information, possibly affecting the psychometric properties of a measure (MacCallum et al., 2002). Indeed, true differences between observations that are present in continuous variables might disappear through dichotomization. There are approaches circumventing the problems accompanying polytomization or dichotomization (e.g., regression-based binary classifiers, weighted sum loss functions, Bayesian-optimal classifiers, etc.; see Mariooryad & Busso, 2015), but these are rarely used in PC research.

When evaluating PC performance estimates, researchers should also note that these probably have a ceiling. PC performance can be regarded as the degree of convergent validity between ML measures and questionnaire-based measures of personality. In "traditional" assessment, convergent validities across data sources, measurement methods, or same-source measures with different content are usually moderate due to method artifacts as well as unique information captured by different measures and data sources (Wiernik et al., 2020). Convergent validities are further attenuated by measurement error and poor construct coverage. The same limitations apply to ML measures. Therefore, Wiernik et al. (2020) estimated a ceiling of around $r \approx 0.30–0.50$ for convergent validities between PC and traditional personality measures. They recommended interpreting higher convergent validities with caution because these probably result from overfitting.

Another issue that further challenges PC research is a reproducibility crisis in ML (Hutson, 2018; Raff, 2019). ML code is often not shared openly and has to be recreated from published descriptions. Moreover, modeling decisions affecting performance (e.g., hyperparameter settings) are frequently not reported (Hutson, 2018). Reproducibility is also hampered by inaccessibility of data and hardware (Barber, 2019). However, the culture in ML is slowly beginning to change towards open science. For example, reproducibility checklists (Pineau et al., 2020) as well as frameworks for open and reproducible code (Forde et al., 2018; Van Lissa et al., 2020) are being implemented.

We conclude that despite some promising advances in PC, performance estimates should be interpreted with caution. Surprisingly high estimates might result from overfitting issues and not replicate. It should be noted, however, that PC cannot be evaluated on the basis of performance alone. To truly understand PC and its utility, researchers will also have to evaluate ML measures' psychometric properties. We will elaborate on this issue in the next section.

## 4 | PSYCHOMETRIC ISSUES

Besides performance issues, personality scientists have lamented the atheoretical "black box" approach and lack of psychometric validation in PC (Bleidorn & Hopwood, 2019; Harari et al., 2020b). Algorithm-based personality trait measures rarely undergo the same rigorous psychometric testing as "traditional" assessment tools. Performance measures usually provide some insights into the convergent validity of ML measures. However, few studies go beyond this and focus on (different forms of) reliability or other types of validity (content, discriminant, factorial, criterion, incremental).

Notably, there are a few differences in the psychometric evaluation of ML measures compared to traditional measures. For example, Stachl et al. (2020b) proposed a distinction between *internal* and *external* convergent and discriminant validities. For a supervised ML task, they distinguished between correlations of model-predicted values with the target variable (i.e., internal convergent validity) and correlations with external measures of the same construct (i.e., external convergent validity) or other constructs that were not considered during training (i.e., external discriminant validity). Furthermore, they suggest modifying a model's loss function to minimize associations with theoretically distinct constructs (i.e., internal discriminant validity). High external convergent correlations and low external discriminant correlations would provide more convincing evidence for the validity of a ML measure compared to internal validity correlations. Moreover, as mentioned earlier, researchers should expect a moderate ceiling for convergent validities due to method artifacts and unique construct information in different data sources (Wiernik et al., 2020). Furthermore, ML measures could be particularly prone to discriminant validity issues because the same indicators may be substantially linked to many traits, leading to high intercorrelations between distinct ML measures. Therefore, researchers should assess a broad enough range of indicators so that variables can contribute uniquely to distinct constructs (Wiernik et al., 2020).

That being said, the intended goal of an ML measure needs to be considered to determine whether it should obey traditional psychometric standards (Stachl et al., 2020b). If an ML measure is created to capture a latent variable from classical or probabilistic test theory, psychometric properties (i.e., objectivity, reliability, validity) should be evaluated. Also, its nomological network (Cronbach & Meehl, 1955) should be examined to determine the causes, concomitants, and consequences of the measured variable. If the purpose of an ML measure is to optimize the prediction of a criterion (e.g., job perform5e), psychometric evaluation may be secondary (Stachl et al., 2020b). For a detailed discussion and recommendations on the psychometric evaluation of ML measures, we refer to Bleidorn and Hopwood (2019) and Stachl et al. (2020b).

The reliability and validity of sensor-based features that enter ML models should also be considered. For example, even a simple variable such as "step count"—measured via different pedometers, wearables, and smartphones—varies between devices and deviates from the steps counted by human raters (Case

et al., 2015). Knowledge about the reliability and validity of features is important to inform appropriate data collection and interpretation of the resulting model's content validity (Alexander et al., 2020). Much research is still to be done to ensure that ML personality measures appropriately measure the target constructs. Only then will it be possible to move beyond a black-box approach and substantially inform personality theory.

## 5 | ETHICAL, LEGAL, AND SOCIETAL ISSUES

Despite its potentials and because of its limitations, PC research, and in particular its application, raises some ethical, legal, and societal issues. Especially privacy, autonomy, and fairness are of paramount concern, and we briefly touch upon these issues here.

Privacy has been a longstanding issue in psychological testing, pertaining to an individual's right to decide which, and with whom, personal information is shared and to know one's own test results (Anastasi, 1980). These concerns are amplified in the light of passive data collection and automatized personality prediction. Digital traces may be collected involuntarily and without informed consent to make inferences about people's private traits. Nonconsensual personality profiling does not only undermine people's autonomy and freedom of choice but could also pose a threat to democracy and informed self-determination on a societal level (Boyd et al., 2020; see Cambridge Analytica Scandal: Confessore, 2018).

Another pressing issue is whether PC produces fair results, that is, whether it performs equally well across different groups (Alexander et al., 2020), which may pertain to race, ethnicity, gender, sexuality, disability, confession, class, or culture. Any kind of algorithm-based discrimination might result in amplifications of social inequalities due to systematic discrepancies in access to personalized services or targeted manipulation (Kusner & Loftus, 2020). For example, commercial algorithms used in the US health care system have been shown to discriminate against Black patients assigning them the same risk score as White patients while being considerably sicker (Obermeyer et al., 2019). These risk scores are used to target patients with complex health needs to improve their care by providing additional resources. This racial bias arises because the algorithm predicts health costs rather than health needs. This is problematic because differences in the amount of money spent on care results from unequal access in the first place. Unfortunately, there have been many more reports of instances where discriminatory biases have permeated ML algorithms. For instance, facial recognition algorithms have been shown to perform worse for non-Caucasian faces, voice recognition systems were found to be biased against female voices, and targeted ads for high-paying jobs on Google were displayed far more often to men than women (for an overview, see Howard & Borenstein, 2017). From a psychometric perspective, biases in ML algorithms are reflected in issues with measurement invariance across different groups of people. More research is needed to determine whether algorithms that are trained with samples from certain groups are generalizable to other samples. Another related issue affecting measurement invariance and, thus, fairness is selection bias in the samples studied. There may be class barriers that prevent certain groups from participating in or being recruited for studies such as owning a (certain type of) smartphone. While the issues discussed here relate to ML research in general, the same problems apply also to PC research.

To combat some of the issues mentioned above, data protection laws have been put in place ensuring informational self-determination and regularizing data protection (e.g., EU General Data Protection Regulation; European Union, 2016). However, these laws differ substantially between nations (Guzzo et al., 2015), and PC researchers working with Big Data are prompted to carefully consider the laws and standards of the nation where data is collected, stored, and analyzed. Above and beyond such legal implications, researchers should consider the ethical and societal implications of their work. For example, statements of societal consequences of one's work at conferences can be requested and research on topics such as de-identification or data masking performed (Boyd et al., 2020). Especially for applied settings, such as personnel selection (Tippins et al., 2021), it is important that

researchers and practitioners *together* develop and implement professional guidelines to ensure that legal, ethical, and societal standards are being met.

## 6 | THE ROLE OF PERSONALITY SCIENCE IN PC

There is a discrepancy in how and why PC research is conducted in personality psychology and in computer science. Personality scientists are equipped with theoretical knowledge, rigorous training in measurement and measurement theory, traditional statistical training, and knowledge about study designs. Computational modeling is rarely integrated in psychologists' training. Computer scientists, on the other hand, have expertise in ML modeling, deep mathematical and statistical knowledge, and technical know-how in the collection and processing of behavioral sensor data.

A promising way forward would be to further multi- and interdisciplinary endeavors and specify the tasks and capacities of personality and computer scientists in collaborative research projects. To this end, we conducted a short qualitative survey of PC experts ($N = 8$ experts; four personality scientists; four computer scientists; 50% professors; 75% male), all of whom reported having worked in interdisciplinary PC research projects. Experts commented on the role of personality scientists in PC (in general and in Automatic Personality Recognition, Automatic Personality Perception, and Automatic Personality Synthesis in particular). Personality scientists were further asked for their opinions on how to promote psychometric evaluation in PC. Computer scientists were asked how they thought personality psychologists could improve data collection and labeling to enable powerful and accurate ML models. In the next section, we briefly summarize the results (for a detailed overview of survey results, see https://osf.io/6s8fx/).

The experts agreed that knowledge about personality theory and psychological assessment represents one of the most significant contributions by personality scientists. This involves selecting appropriate personality inventories, conceptual knowledge on personality, and sampling of relevant psychological variables (e.g., behavioral, affective, motivational, cognitive, physiological), situations, groups, and timeframes. Experts also agreed that personality scientists should be involved in designing studies (to ensure that personality-relevant information is captured), executing studies (providing expertise, e.g., on participant recruitment, informed consent, compensation), and selecting features (to ensure a theory-driven approach; not applicable to deep learning models where features are not human-readable).

Another central role of personality scientists is to assist in the interpretation of results. Interpretation (e.g., regarding the relationships between features and personality constructs; embedding results into theory) affords expertise in personality psychology and cannot be provided by computer scientists. Also, personality scientists could improve the communication of results to the scientific community by providing conceptual clarity as well as concise and consistent terminology. Theoretically founded interpretations could improve future ML models, pointing data collection and modeling towards robust effects. Furthermore, personality-psychological expertise is necessary to identify future directions that are relevant to psychological research questions (see also Vinciarelli & Mohammadi, 2014b). For example, personality experts supported the investigation of constructs other than the Big Five (e.g., facets, personality states, values, narratives).

Another important point was that personality scientists should determine what should be used as the "ground truth" in PC. Self- and other-reports are often treated as the "gold-standard"—however, people's actual trait levels might be reflected more accurately by composite scores of shared variance between self- and other-ratings or the latent variance from multiple data sources. Researchers should even consider whether broad traits are appropriate targets in the first place. "Big Few" traits (i.e., Big Five, Five-Factor Model, or HEXACO; Ashton & Lee, 2020; Goldberg, 1990; McCrae & John, 1992) have been instrumental in organizing and integrating personality research. However, if the ultimate goal is to use personality information to predict important outcomes, the question can be raised why multi-indicator information (which are indicators of personality) cannot be used directly instead of intermediately estimating personality traits. In this vein, Mõttus et al. (2020) show that out-of-sample predictive

accuracy increases with higher resolutions of predictors (e.g., a set of individual items predicted outcomes better than their aggregated facets or domains). Furthermore, experts mentioned that the distinction between measures and constructs should be clarified. Other relevant points included contributions of expertise in construct validation and test development; generalizability; promotion of good measurement practices; knowledge about normed distributions of personality constructs; perceptual biases; and choosing adequate methodological and theoretical approaches depending on whether the research goal is explanation, description, or prediction (see also Mõttus et al., 2020).

Lastly, personality scientists are equipped to collect personality-relevant behavioral data in relevant contexts coupled with personality measures that are necessary to build powerful and accurate PC models. These data sets need to be large, clearly labeled, and would ideally include multi-modally sensed behavior (however, computer scientists might be more experienced in collecting sensor data). Optimally, such data sets would be made publicly available (e.g., on a joint platform where data, information on sensing, etc. could be shared) while ensuring privacy and data security. Ultimately, promoting PC might not only require transdisciplinary collaborations between personality and computer scientists but also multi-lab efforts within personality science to provide high-quality data.

## 7 | CONCLUSION

PC opens up new and exciting avenues to gaining insights into complex behavioral aspects of personality, thereby also fostering theory-building and widespread applications. Despite the current limitations in PC and reproducibility issues, much progress should be expected in PC with continuous improvements in technology and computational power, modeling approaches, and an increasing cross-disciplinary integration of computer and personality science. However, personality scientists should be aware that PC research is and will be conducted without their involvement. To ensure conceptual precision and methodological soundness, personality science will have to enhance its influence on PC research. Personality scientists should especially promote psychometric evaluation and nomological validity of ML data. Accordingly, it is important that personality scientists assist in the interpretation of results in PC research to clarify what aspects of personality are being measured (Rauthmann, 2020), but they should also participate in designing studies and data collection. Moreover, state-of-the art knowledge in personality science is essential to move beyond a mere prediction of personality trait scores and to steer PC towards descriptive and explanatory personality research as well as relevant research questions revolving around personality dynamics. To effectively foster personality theory and to develop safe and sound personalized applications, PC research must be a truly interdisciplinary endeavor.

## ORCID
*Le Vy Phan* https://orcid.org/0000-0002-6470-8871

## REFERENCES
AlexanderIII, L., Mulfinger, E., & Oswald, F. L. (2020). Using big data and machine learning in personality measurement: Opportunities and challenges. *European Journal of Personality*. Advance online publication. https://doi.org/10.1002/per.2305n

Anastasi, A. (1980). Psychological testing and privacy. In W. C. Bier (Ed.), *Privacy: A vanishing value* (pp. 348–358). Fordham University Press.

Araújo, D., Davids, K., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on Rogers, Kadar, and Costall (2005). *Ecological Psychology*, *19*(1), 69–78. https://doi.org/10.1080/10407410709336951

Ashton, M. C., & Lee, K. (2020). Objections to the HEXACO model of personality structure—and why those objections fail. *European Journal of Personality*, *34*(4), 492–510. https://doi.org/10.1002/per.2242

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Barber, G. (2019, September 16). Artificial intelligence Confronts a 'reproducibility' crisis. *Wired*. https://www.wired.com/story/artificial-intelligence-confronts-reproducibility-crisis/

Batrinca, L., Lepri, B., Mana, N., & Pianesi, F. (2012). Multimodal recognition of personality traits in human-computer collaborative tasks. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 39–46). https://doi.org/10.1145/2388676.2388687

Batrinca, L. M., Mana, N., Lepri, B., Pianesi, F., & Sebe, N. (2011). Please, tell me about yourself: Automatic personality assessment using short self-presentations. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (pp. 255–262). https://doi.org/10.1145/2070481.2070528

Beck, E. D., & Jackson, J. J. (2020, November 2). A mega-analysis of personality prediction: Robustness and boundary conditions. https://doi.org/10.31234/osf.io/7pg9b

Biel, J. I., Aran, O., & Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in Youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 446–449.

Biel, J. I., Teijeiro-Mosquera, L., & Gatica-Perez, D. (2012). Facetube: Predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 53–56). https://doi.org/10.1145/2388676.2388689

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. https://doi.org/10.1177/1088868318772990

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, *34*, 599–612. https://doi.org/10.1002/per.2254

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Case, M. A., Burwick, H. A., Volpp, K. G., & Patel, M. S. (2015). Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Jama*, *313*(6), 625. https://doi.org/10.1001/jama.2014.17841

Celli, F. (2012). Unsupervised personality recognition for social network sites. In *Proceedings of the Sixth International Conference on Digital Society*.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011). Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *2011 15th Annual international symposium on wearable computers* (pp. 29–36). IEEE. https://doi.org/10.1109/ISWC.2011.29

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433–450. https://doi.org/10.1007/s00779-011-0490-1

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Confessore, N. (2018, April 4). Cambridge Analytica and Facebook: The scandal and the fallout so far. *The New York Times*. https://www.nytimes.com/2017/03/06/us/politics/cambridge-analytica.html

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092-1122. https://doi.org/10.1037/a0021212

Cristani, M., Vinciarelli, A., Segalin, C., & Perina, A. (2013). Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In Proceedings of the 21st ACM international conference on Multimedia (pp. 213–222). https://doi.org/10.1145/2502081.2502280

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

DeYoung, C. G. (2015). Cybernetic Big Five theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, *14*(4), 349–366. https://doi.org/10.1037/a0016956

European Parliament and Council of European Union. (2016). *Regulation (EU) 2016/679*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Eysenck, H. J. (1967). *The biological basis of personality*. Thomas.

Finnerty, A. N., Lepri B., Pianesi F. (2016). Acquisition of personality. In Tkalčič M., De Carolis B., de Gemmis M., Odić A., & Košir A. (Eds.), *Emotions and personality in personalized services. Human–Computer interaction series* (pp. 81–99). Springer. https://doi.org/10.1007/978-3-319-31413-6_5

Forde, J., Head, T., Holdgraf, C., Panda, Y., Nalvarete, G., Ragan-Kelley, B., & Sundell, E. (2018). *Reproducible research environments with repo2docker. ICML 2018 reproducible machine learning.* ICML. Retrieved from https://openreview.net/pdf?id=B1lYOwuoxm

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295x.102.4.652

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*, 197–221.

Gürpinar, F., Kaya, H., & Salah, A. A. (2016). Combining deep facial and ambient features for first impression estimation. In Hua G & Jégou H. (Eds.), *European conference on computer vision* (pp. 372–385). Springer. https://doi.org/10.1007/978-3-319-49409-8_30

Gill, A. J., Nowson, S., & Oberlander, J. (2009). What are they blogging about? Personality, topic and motivation in blogs. In *Third International AAAI Conference on Weblogs and Social Media*.

Golbeck, J., Robles, C., & Turner, K. (2011a). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 253–262). https://doi.org/10.1145/1979742.1979614

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011b). Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 149–156). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*(3), 379–398. https://doi.org/10.1037/0022-3514.82.3.379

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology*, *8*(4), 491–508. https://doi.org/10.1017/iop.2015.40

Harari, G. M., Gosling, S. D., Wang, R., & Campbell, A. T. (2015). Capturing situational information with smartphones and mobile sensing methods. *European Journal of Personality*, *29*(5), 509–511. https://doi.org/10.1002/per.2032

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., Rentfrow, P. J., Campbell, A. T., & Gosling, S. D. (2020a). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, *119*(1), 204–228. https://doi.org/10.1037/pspp0000245

Harari, G. M., Vaid, S. S., Müller, S. R., Stachl, C., Marrero, Z., Schoedel, R., & Gosling, S. D. (2020b). Personality sensing for theory development and assessment in the digital age. *European Journal of Personality*. Advance online publication. https://doi.org/10.1002/per.2273

Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 180–207). Guilford Press.

Hogan, R., & Roberts, B. W. (2000). A socioanalytic perspective on person–environment interaction. In W. B. Walsh & K. H. Craik (Eds.), *Person–Environment psychology: New directions and perspectives* (2nd ed., pp. 1–23). Erlbaum.

Howard, A., & Borenstein, J. (2017). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, *24*(5), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

Ilmini, W. M. K. S., & Fernando, T. G. I. (2017, December). Computational personality traits assessment: A review. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)* (pp. 1–6). IEEE.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). Guilford Press.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guilford Press

Kedar, S. V., & Bormane, D. S. (2015). Automatic personality assessment: A systematic review. In *2015 International Conference on Information Processing (ICIP)* (pp. 326–331). IEEE. https://doi.org/10.1109/infop.2015.7489402

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford Press.

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*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kulkarni, V., Kern, M. L., Stillwell, D., Kosinski, M., Matz, S., Ungar, L., Skiena, S., & Schwartz, H. A. (2018). Latent human traits in the language of social media: An open-vocabulary approach. *PLoS One*, *13*(11), e0201703. https://doi.org/10.1371/journal.pone.0201703

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, *578*(7793), 34–36. https://doi.org/10.1038/d41586-020-00274-3

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing Ltd.

Mønsted, B., Mollgaard, A., & Mathiesen, J. (2018). Phone-based metric as a predictor for basic personality traits. *Journal of Research in Personality*, *74*, 16–22. https://doi.org/10.1016/j.jrp.2017.12.004

Mõttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality*, *34*(6), 1175–1201. https://doi.org/10.31234/osf.io/hvk5p

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989x.7.1.19

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*, 457–500. https://doi.org/10.1613/jair.2349

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, *32*(2), 74–79. https://doi.org/10.1109/MIS.2017.23

Mariooryad, S., & Busso, C. (2015). The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier. *IEEE Transactions on Affective Computing*, *8*(1), 119–130.

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses*, *82*(4), 405–411. https://doi.org/10.1016/j.mehy.2013.11.030

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, *123*(5), 569–591. https://doi.org/10.1037/rev0000035

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2019). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, *53*(4), 2313–2339. https://doi.org/10.1007/s10462-019-09770-z

Mohammadi, G., Origlia, A., Filippone, M., & Vinciarelli, A. (2012). From speech to personality: Mapping voice quality and intonation into personality differences. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 789–792). https://doi.org/10.1145/2393347.2396313

Mohammadi, G., & Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, *3*(3), 273–284. https://doi.org/10.1109/T-AFFC.2012.5

Nguyen, T., Phung, D. Q., Adams, B., & Venkatesh, S. (2011). Towards discovery of influence and personality traits through social link prediction. In *ICWSM* (pp. 566–569).

Oberlander, J., & Nowson, S. (2006). Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 627–634). https://doi.org/10.3115/1273073.1273154

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Olguín- Olguín, D., Gloor, P. A., & Pentland, A. S. (2009). Capturing individual and group behavior with wearable sensors. In *Proceedings of the 2009 AAAI Spring Aymposium on Human Behavior Modeling* (pp.68–74).

Onnela, J., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*(1), 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*, *49*(7), 669–676. https://doi.org/10.1016/j.paid.2010.06.014

Pianesi, F. (2013). Searching for personality [social sciences]. *IEEE Signal Processing Magazine*, *30*(1), 146–158. https://doi.org/10.1109/msp.2012.2219671

Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., & Zancanaro, M. (2008). Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces* (pp. 53–60). https://doi.org/10.1145/1452392.1452404

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., & Larochelle, H. (2020). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *arXiv*. https://arxiv.org/abs/2003.12206

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. (2013). Common sense knowledge based personality recognition from text. In *Mexican International Conference on Artificial Intelligence* (pp. 484–496). Springer. https://doi.org/10.1007/978-3-642-45111-9_42

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 180–185). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

Raff, E. (2019). A step toward quantifying independently reproducible machine learning research. *arXiv*. https://arxiv.org/abs/1909.06674

Rauthmann, J. F. (2016). Motivational factors in the perception of psychological situation characteristics. *Social and Personality Psychology Compass*, 10, 92–108.

Rauthmann, J. F. (2020). A (more) behavioral science of personality in the age of multi-modal sensing, big data, machine learning, and artificial intelligence [special issue editorial]. *European Journal of Personality*. Advance online publication.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1–8). IEEE. https://doi.org/10.1145/2696454.2696497

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits. *Journal of Cross-Cultural Psychology*, 38(2), 173–212. https://doi.org/10.1177/0022022106297299

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, 30(5), 711–727. https://doi.org/10.1177/0956797619831612

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Volkel, S. T., Schuwerk, T., Oldemeier, D., Ullmann, T., Hussmann, H., Bischl, B., & Buhner, M. (2020a). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Stachl, C., Boyd, R. L., Horstmann, K. T., Khambatta, P., Matz, S., & Harari, G. M. (2021, January 20). Computational personality assessment—An overview and perspective. https://doi.org/10.31234/osf.io/ck2bj

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020b). Personality research and assessment in the era of machine learning. *European Journal of Personality*. Advance online publication. https://doi.org/10.1002/per.2257

Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., & Pentland, A. (2012). Friends don't lie: Inferring personality traits from social network structure. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 321–330). https://doi.org/10.1145/2370216.2370266

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021, January 28). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. https://doi.org/10.31234/osf.io/6gczw

Valente, F., Kim, S., & Motlicek, P. (2012). Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *Thirteenth Annual Conference of the International Speech Communication Association* (pp. 1183–1186).

Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A., Peikert, A., Struiksma, M. E., & Vreede, B. (2020, May 31). WORCS: A workflow for open reproducible code in science. https://doi.org/10.3233/DS-210031

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98, 281–300. https://doi.org/10.1037/a0017908

Vinciarelli, A., & Mohammadi, G. (2014a). A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3), 273–291. https://doi.org/10.1109/TAFFC.2014.2330816

Vinciarelli, A., & Mohammadi, G. (2014b). More personality in personality computing. *IEEE Transactions on Affective Computing*, 5(3), 297–300. https://doi.org/10.1109/TAFFC.2014.2341252

Wöllmer, M., Eyben, F., Schuller, B., Douglas-Cowie, E., & Cowie, R. (2009). Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks. In *Proc. Interspeech 2009*, Brighton, UK (pp. 1595–1598).

Wang, W., Harari, G. M., Wang, R., Müller, S. R., Mirjafari, S., Masaba, K., & Campbell, A. T. (2018). Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), 1–21. https://doi.org/10.1145/3264951

Wiernik, B. M., Ones, D. S., Marlin, B. M., Giordano, C., Dilchert, S., Mercado, B. K., Stanek, K. C., Birkland, A., Wang, Y., Ellis, B., Yazar, Y., Kostal, J. W., Kumar, S., Hnat, T., Ertin, E., Sano, A., Ganesan, D. K., Choudhoury, T., & Al'Absi, M. (2020). Using mobile sensors to study personality dynamics. *European Journal of Psychological Assessment*, 36(6), 1–13. https://doi.org/10.1027/1015-5759/a000576

Wilt, J., & Revelle, W. (2015). Affect, behaviour, cognition and desire in the big five: An analysis of item content and structure. *European Journal of Personality*, *29*(4), 478–497. https://doi.org/10.1002/per.2002

Wright, A. G. (2014). Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, *5*(3), 292–296. https://doi.org/10.1109/TAFFC.2014.2332331

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

Yee, N., Ducheneaut, N., Nelson, L., & Likarish, P. (2011). Introverted elves & conscientious gnomes: The expression of personality in world of warcraft. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 753–762). https://doi.org/10.1145/1978942.1979052

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zen, G., Lepri, B., Ricci, E., & Lanz, O. (2010). Space speaks: Towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis* (pp. 37–42). https://doi.org/10.1145/1878039.1878048

## AUTHOR BIOGRAPHIES

**Le Vy Phan** is a PhD student at the Department of Psychology at Bielefeld University, Germany. Her research interests include personality computing, behavioral consistency, behavioral assessment, and theory building in personality science.

**John F. Rauthmann** is the Professor of Personality Psychology and Psychological Assessment at the Department of Psychology at Bielefeld University, Germany. His research interests include dynamic approaches to personality, person-environment relations, and personality measurement via novel technologies.