Computer-based personality judgments are more accurate than those made by humans

Wu Youyou[a,1,2], Michal Kosinski[b,1], and David Stillwell[a]

[a]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and [b]Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

Judging others' personalities is an essential skill in successful social living, as personality is a key driver behind people's interactions, behaviors, and emotions. Although accurate personality judgments stem from social-cognitive skills, developments in machine learning show that computer models can also make valid judgments. This study compares the accuracy of human and computer-based personality judgments, using a sample of 86,220 volunteers who completed a 100-item personality questionnaire. We show that (*i*) computer predictions based on a generic digital footprint (Facebook Likes) are more accurate ($r = 0.56$) than those made by the participants' Facebook friends using a personality questionnaire ($r = 0.49$); (*ii*) computer models show higher interjudge agreement; and (*iii*) computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores. Computers outpacing humans in personality judgment presents significant opportunities and challenges in the areas of psychological assessment, marketing, and privacy.

personality judgment | social media | computational social science | artificial intelligence | big data

Perceiving and judging other people's personality traits is an essential component of social living (1, 2). People use personality judgments to make day-to-day decisions and long-term plans in their personal and professional lives, such as whom to befriend, marry, trust, hire, or elect as president (3). The more accurate the judgment, the better the decision (2, 4, 5). Previous research has shown that people are fairly good at judging each other's personalities (6–8); for example, even complete strangers can make valid personality judgments after watching a short video presenting a sample of behavior (9, 10).

Although it is typically believed that accurate personality perceptions stem from social-cognitive skills of the human brain, recent developments in machine learning and statistics show that computer models are also capable of making valid personality judgments by using digital records of human behavior (11–13). However, the comparative accuracy of computer and human judgments remains unknown; this study addresses this gap.

Personality traits, like many other psychological dimensions, are latent and cannot be measured directly; various perspectives exist regarding the evaluation criteria of judgmental accuracy (3, 5). We adopted the realistic approach, which assumes that personality traits represent real individual characteristics, and the accuracy of personality judgments may be benchmarked using three key criteria: self-other agreement, interjudge agreement, and external validity (1, 5, 7). We apply those benchmarks to a sample of 86,220 volunteers,* who filled in the 100-item International Personality Item Pool (IPIP) Five-Factor Model of personality (14) questionnaire (15), measuring traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Computer-based personality judgments, based on Facebook Likes, were obtained for 70,520 participants. Likes were previously shown to successfully predict personality and other psychological traits (11). We used LASSO (Least Absolute Shrinkage and Selection Operator) linear regressions (16) with 10-fold cross-validations, so that judgments for each participant were made using models developed on a different subsample of participants and their Likes. Likes are used by Facebook users to express positive association with online and offline objects, such as products, activities, sports, musicians, books, restaurants, or websites. Given the variety of objects, subjects, brands, and people that can be liked and the number of Facebook users (>1.3 billion), Likes represent one of the most generic kinds of digital footprint. For instance, liking a brand or a product offers a proxy for consumer preferences and purchasing behavior; music-related Likes reveal music taste; and liked websites allow for approximating web browsing behavior. Consequently, Like-based models offer a good proxy of what could be achieved based on a wide range of other digital footprints such as web browsing logs, web search queries, or purchase records (11).

Human personality judgments were obtained from the participants' Facebook friends, who were asked to describe a given participant using a 10-item version of the IPIP personality measure. To compute self-other agreement and external validity, we used a sample of 17,622 participants judged by one friend; to calculate interjudge agreement, we used a sample of 14,410 participants

Significance

This study compares the accuracy of personality judgment—a ubiquitous and important social-cognitive activity—between computer models and humans. Using several criteria, we show that computers' judgments of people's personalities based on their digital footprints are more accurate and valid than judgments made by their close others or acquaintances (friends, family, spouse, colleagues, etc.). Our findings highlight that people's personalities can be predicted automatically and without involving human social-cognitive skills.

Author contributions: W.Y. and M.K. designed research; W.Y., M.K., and D.S. performed research; W.Y. and M.K. contributed new reagents/analytic tools; W.Y. and M.K. analyzed data; and W.Y., M.K., and D.S. wrote the paper.

Conflict of interest statement: D.S. received revenue as the owner of the myPersonality Facebook application.

This article is a PNAS Direct Submission. D.F. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: The data used in the study are shared with the academic community at mypersonality.org.

[1]W.Y. and M.K. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: yw341@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1418680112/-/DCSupplemental.

*The sample used in this study was obtained from the myPersonality project. myPersonality was a popular Facebook application that offered to its users psychometric tests and feedback on their scores. Since the data are secondary, anonymized, was previously published in the public domain, and was originally gathered with an explicit opt-in consent for reuse for research purposes beyond the original project, no IRB approval was needed. This was additionally confirmed by the Psychology Research Ethics Committee at the University of Cambridge.

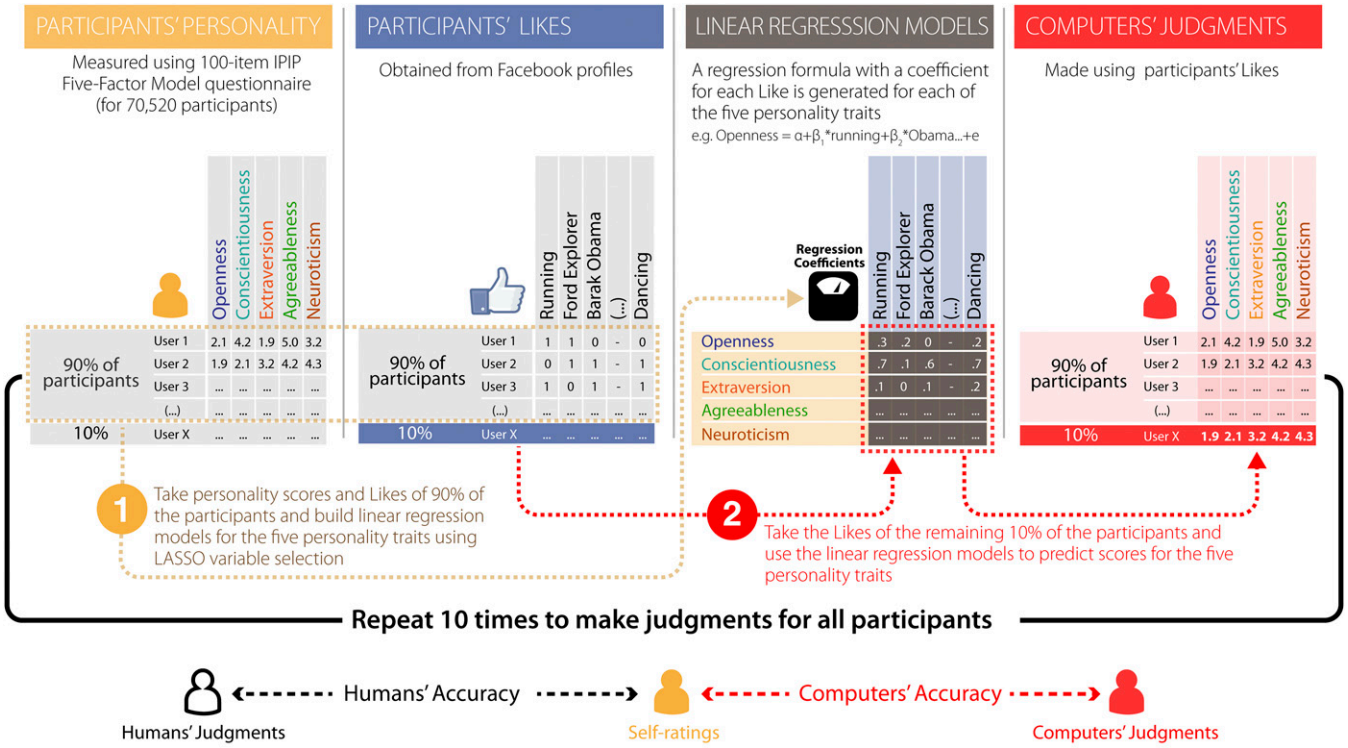

Fig. 1. Methodology used to obtain computer-based judgments and estimate the self-other agreement. Participants and their Likes are represented as a matrix, where entries are set to 1 if there exists an association between a participant and a Like and 0 otherwise (second panel). The matrix is used to fit five LASSO linear regression models (16), one for each self-rated Big Five personality trait (third panel). A 10-fold cross-validation is applied to avoid overfitting: the sample is randomly divided into 10 equal-sized subsets; 9 subsets are used to train the model (step 1), which is then applied to the remaining subset to predict the personality score (step 2). This procedure is repeated 10 times to predict personality for the entire sample. The models are built on participants having at least 20 Likes. To estimate the accuracy achievable with less than 20 Likes, we applied the regression models to random subsets of 1–19 Likes for all participants.

judged by two friends. A diagram illustrating the methods is presented in Fig. 1.

Results

Self-Other Agreement. The primary criterion of judgmental accuracy is self-other agreement: the extent to which an external judgment agrees with the target's self-rating (17), usually operationalized as a Pearson product-moment correlation. Self-other agreement was determined by correlating participants' scores with the judgments made by humans and computer models (Fig. 1). Since self-other agreement varies greatly with the length and context of the relationship (18, 19), we further compared our results with those previously published in a meta-analysis by Connely and Ones (20), including estimates for different categories of human judges: friends, spouses, family members, cohabitants, and work colleagues.

To account for the questionnaires' measurement error, self-other agreement estimates were disattenuated using scales' Cronbach's α reliability coefficients. The measurement error of the computer model was assumed to be 0, resulting in the lower (conservative) estimates of self-other agreement for computer-based judgments. Also, disattenuation allowed for direct comparisons of human self-other agreement with those reported by Connely and Ones (20), which followed the same procedure.

The results presented in Fig. 2 show that computers' average accuracy across the Big Five traits (red line) steadily grows with the number of Likes available on the participant's profile (x axis). Computer models need only 100 Likes to outperform an average human judge in the present sample ($r = 0.49$; blue point).[†] Compared with the accuracy of various human judges reported in the meta-analysis (20), computer models need 10, 70, 150, and 300 Likes, respectively, to outperform an average work colleague, cohabitant or friend, family member, and spouse (gray points). Detailed results for human judges can be found in Table S1.

How accurate is the computer, given an average person? Our recent estimate of an average number of Likes per individual is 227 (95% CI = 224, 230),[‡] and the expected computer accuracy for this number of Likes equals $r = 0.56$. This accuracy is significantly better than that of an average human judge ($z = 3.68$, $P < 0.001$) and comparable with an average spouse, the best of human judges ($r = 0.58$, $z = -1.68$, $P = 0.09$). The peak computer performance observed in this study reached $r = 0.66$ for participants with more than 500 Likes. The approximately log-linear relationship between the number of Likes and computer accuracy, shown in Fig. 2, suggests that increasing the amount of signal beyond what was available in this study could further boost the accuracy, although gains are expected to be diminishing.

Why are Likes diagnostic of personality? Exploring the Likes most predictive of a given trait shows that they represent activities, attitudes, and preferences highly aligned with the Big Five theory. For example, participants with high openness to experience tend to like Salvador Dalí, meditation, or TED talks; participants with high extraversion tend to like partying, Snookie (reality show star), or dancing.

Self-other agreement estimates for individual Big Five traits (Fig. 2) reveal that the Likes-based models are more diagnostic of

[†]This figure is very close to the average human accuracy ($r = 0.48$) found in Connelly and Ones's meta-analysis (20).

[‡]Estimate based on a 2014 sample of $n = 100{,}001$ Facebook users collected for a separate project. Sample used in this study was recorded in the years 2009–2012.

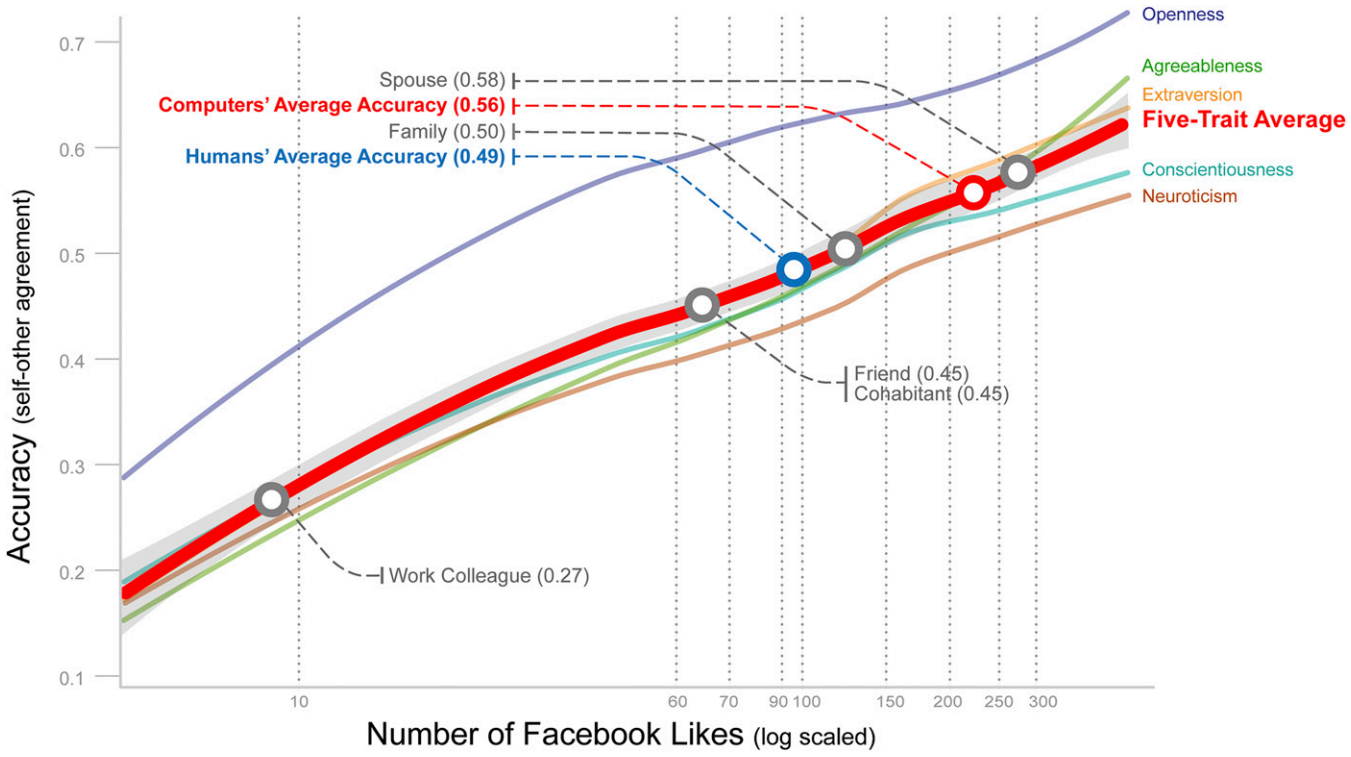

Fig. 2. Computer-based personality judgment accuracy (*y* axis), plotted against the number of Likes available for prediction (*x* axis). The red line represents the average accuracy (correlation) of computers' judgment across the five personality traits. The five-trait average accuracy of human judgments is positioned onto the computer accuracy curve. For example, the accuracy of an average human individual ($r = 0.49$) is matched by that of the computer models based on around 90–100 Likes. The computer accuracy curves are smoothed using a LOWESS approach. The gray ribbon represents the 95% CI. Accuracy was averaged using Fisher's *r*-to-*z* transformation.

some traits than of others. Especially high accuracy was observed for openness—a trait known to be otherwise hard to judge due to low observability (21, 22). This finding is consistent with previous findings showing that strangers' personality judgments, based on digital footprints such as the contents of personal websites (23), are especially accurate in the case of openness. As openness is largely expressed through individuals' interests, preferences, and values, we argue that the digital environment provides a wealth of relevant clues presented in a highly observable way.

Interestingly, it seems that human and computer judgments capture distinct components of personality. Table S2 lists correlations and partial correlations (all disattenuated) between self-ratings, computer judgments, and human judgments, based on a subsample of participants (n = 1,919) for whom both computer and human judgments were available. The average consensus between computer and human judgments ($r = 0.37$) is relatively high, but it is mostly driven by their correlations with self-ratings, as represented by the low partial correlations ($r = 0.07$) between computer and human judgments. Substantial partial correlations between self-ratings and both computer ($r = 0.38$) and human judgments ($r = 0.42$) suggest that computer and human judgments each provide unique information.

Interjudge Agreement. Another indication of the judgment accuracy, interjudge agreement, builds on the notion that two judges that agree with each other are more likely to be accurate than those that do not (3, 24–26).

The interjudge agreement for humans was computed using a subsample of 14,410 participants judged by two friends. As the judgments were aggregated (averaged) on collection (i.e., we did not store judgments separately for the judges), a formula was used to compute their intercorrelation (*SI Text*). Interjudge agreement for computer models was estimated by randomly splitting the Likes into two halves and developing two separate models following the procedure described in the previous section.

The average consensus between computer models, expressed as the Pearson product-moment correlation across the Big Five traits ($r = 0.62$), was much higher than the estimate for human judges observed in this study ($r = 0.38$, $z = 36.8$, $P < 0.001$) or in the meta-analysis (20) ($r = 0.41$, $z = 41.99$, $P < 0.001$). All results were corrected for attenuation.

External Validity. The third measure of judgment accuracy, external validity, focuses on how well a judgment predicts external criteria, such as real-life behavior, behaviorally related traits, and life outcomes (3). Participants' self-rated personality scores, as well as humans' and computers' judgments, were entered into regression models (linear or logistic for continuous and dichotomous variables respectively) to predict 13 life outcomes and traits previously shown to be related to personality: life satisfaction, depression, political orientation, self-monitoring, impulsivity, values, sensational interests, field of study, substance use, physical health, social network characteristics, and Facebook activities (see Table S3 for detailed descriptions). The accuracy of those predictions, or external validity, is expressed as Pearson product-moment correlations for continuous variables, or area under the receiver-operating characteristic curve (AUC) for dichotomous variables.[§]

As shown in Fig. 3, the external validity of the computer judgments was higher than that of human judges in 12 of the 13

[§]AUC is an equivalent of the probability of correctly classifying two randomly selected participants, one from each class, such as liberal vs. conservative political views. Note that for dichotomous variables, the random guessing baseline corresponds to an AUC = 0.50.

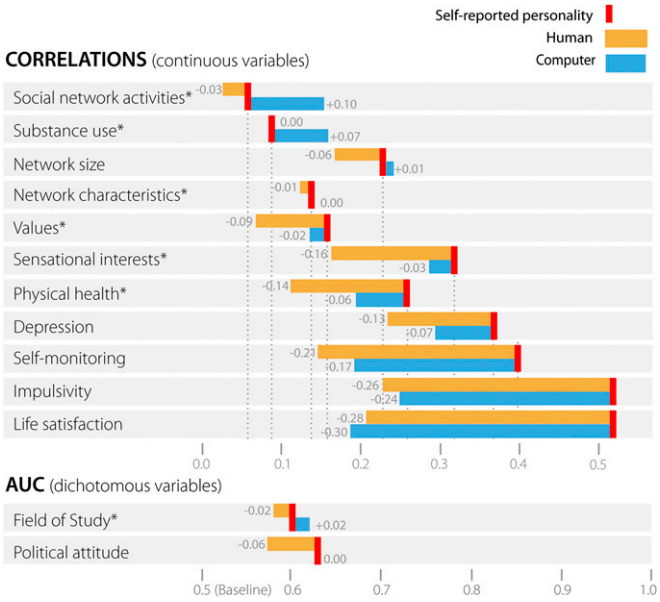

Fig. 3. The external validity of personality judgments and self-ratings across the range of life outcomes, expressed as correlation (continuous variables; *Upper*) or AUC (dichotomous variables; *Lower*). The red, yellow, and blue bars indicate the external validity of self-ratings, human judgments, and computer judgments, respectively. For example, self-rated scores allow predicting network size with accuracy of $r = 0.23$, human judgments achieve $r = 0.17$ accuracy (or 0.06 less than self-ratings), whereas computer-based judgments achieve $r = 0.24$ accuracy (or 0.01 more than self-ratings). Compound variables (i.e., variables representing accuracy averaged across a few subvariables) are marked with an asterisk; see Table S4 for detailed results. Results are ordered by computer accuracy.

criteria (except life satisfaction). Furthermore, computer models' external validity was even better than self-rated personality in 4 of the 13 criteria: Facebook activities, substance use, field of study, and network size; and comparable in predicting political attitudes and social network characteristics. Because most of the outcome variables are self-reports, the high external validity of personality self-ratings is to be expected. It is therefore striking that Likes-based judgments were still better at predicting variables such as field of study or self-rated substance use, despite them sharing more method variance with self-ratings of personality. In addition, the computer-based models were aimed at predicting personality scores and not life outcomes. In fact, Likes-based models, directly aimed at predicting such variables, can achieve even higher accuracy (11).

Discussion

Our results show that computer-based models are significantly more accurate than humans in a core social-cognitive task: personality judgment. Computer-based judgments ($r = 0.56$) correlate more strongly with participants' self-ratings than average human judgments do ($r = 0.49$). Moreover, computer models showed higher interjudge agreement and higher external validity (computer-based personality judgments were better at predicting life outcomes and other behaviorally related traits than human judgments). The potential growth in both the sophistication of the computer models and the amount of the digital footprint might lead to computer models outperforming humans even more decisively.

According to the Realistic Accuracy Model, the accuracy of the personality judgment depends on the availability and the amount of the relevant behavioral information, along with the judges' ability to detect and use it correctly (1, 2, 5). Such conceptualization reveals a couple of major advantages that computers have over humans. First, computers have the capacity to store a tremendous amount of information, which is difficult for humans to retain and access. Second, the way computers use information—through statistical modeling—generates consistent algorithms that optimize the judgmental accuracy, whereas humans are affected by various motivational biases (27). Nevertheless, human perceptions have the advantage of being flexible and able to capture many subconscious cues unavailable to machines. Because the Big Five personality traits only represent some aspects of human personality, human judgments might still be better at describing other traits that require subtle cognition or that are less evident in digital behavior. Our study is limited in that human judges could only describe the participants using a 10-item-long questionnaire on the Big Five traits. In reality, they might have more knowledge than what was assessed in the questionnaire.

Automated, accurate, and cheap personality assessment tools could affect society in many ways: marketing messages could be tailored to users' personalities; recruiters could better match candidates with jobs based on their personality; products and services could adjust their behavior to best match their users' characters and changing moods; and scientists could collect personality data without burdening participants with lengthy questionnaires. Furthermore, in the future, people might abandon their own psychological judgments and rely on computers when making important life decisions, such as choosing activities, career paths, or even romantic partners. It is possible that such data-driven decisions will improve people's lives.

However, knowledge of people's personalities can also be used to manipulate and influence them (28). Understandably, people might distrust or reject digital technologies after realizing that their government, internet provider, web browser, online social network, or search engine can infer their personal characteristics more accurately than their closest family members. We hope that consumers, technology developers, and policymakers will tackle those challenges by supporting privacy-protecting laws and technologies, and giving the users full control over their digital footprints.

Popular culture has depicted robots that surpass humans in making psychological inferences. In the film *Her*, for example, the main character falls in love with his operating system. By curating and analyzing his digital records, his computer can understand and respond to his thoughts and needs much better than other humans, including his long-term girlfriend and closest friends. Our research, along with development in robotics (29, 30), provides empirical evidence that such a scenario is becoming increasingly likely as tools for digital assessment come to maturity. The ability to accurately assess psychological traits and states, using digital footprints of behavior, occupies an important milestone on the path toward more social human-computer interactions.

ACKNOWLEDGMENTS. We thank John Rust, Thore Graepel, Patrick Morse, Vesselin Popov, Winter Mason, Jure Leskovec, Isabelle Abraham, and Jeremy Peang-Meth for their critical reading of the manuscript. W.Y. was supported by the Jardine Foundation; D.S. was supported by a grant from the Richard Benjamin Trust; and M.K. was supported by Microsoft Research, Boeing Corporation, the National Science Foundation, the Defense Advanced Research Projects Agency, and the Center for the Study of Language and Information at Stanford University.

1. Funder DC (2012) Accurate personality judgment. *Curr Dir Psychol Sci* 21(3): 1–18.
2. Letzring TD (2008) The good judge of personality: Characteristics, behaviors, and observer accuracy. *J Res Pers* 42(4):914–932.
3. Funder DC, West SG (1993) Consensus, self-other agreement, and accuracy in personality judgment: an introduction. *J Pers* 61(4):457–476.
4. Letzring TD, Human LJ (2014) An examination of information quality as a moderator of accurate personality judgment. *J Pers* 82(5):440–451.

5. Funder DC (1995) On the accuracy of personality judgment: A realistic approach. *Psychol Rev* 102(4):652–670.
6. Ready RE, Clark LA, Watson D, Westerhouse K (2000) Self- and peer-reported personality: Agreement, trait ratability, and the "self-based heuristic". *J Res Pers* 34(2): 208–224.
7. Funder DC, Kolar DC, Blackman MC (1995) Agreement among judges of personality: interpersonal relations, similarity, and acquaintanceship. *J Pers Soc Psychol* 69(4): 656–672.
8. Macrae C, Quadflieg S (2010) Perceiving people. *Handbook of Social Psychology*, eds Fiske ST, Gillbert DT, Lindzey G (McGraw-Hill, New York), 5th Ed, pp 428–463.
9. Borkenau P, Liebler A (1993) Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *J Pers Soc Psychol* 65(3):546–553.
10. Carney DR, Colvin CR, Hall JA (2007) A thin slice perspective on the accuracy of first impressions. *J Res Pers* 41(5):1054–1072.
11. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci USA* 110(15):5802–5805.
12. Kosinski M, Bachrach Y, Kohli P, Stillwell D, Graepel T (2013) Manifestations of user personality in website choice and behaviour on online social networks. *Mach Learn* 95(3):357–380.
13. Quercia D, Kosinski M, Stillwell D, Crowcroft J (2011) Our Twitter profiles, our selves: Predicting personality with Twitter. *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing* (IEEE, Piscataway, NJ), pp 180–185.
14. Costa PT, McCrae RR (1992) *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Manual* (Psychological Assessment Resources, Odessa, FL).
15. Goldberg LR, et al. (2006) The international personality item pool and the future of public-domain personality measures. *J Res Pers* 40(1):84–96.
16. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58(1):267–288.
17. Taft R (1955) The ability to judge people. *Psychol Bull* 52(1):1–23.
18. Watson D, Hubbard B, Wiese D (2000) Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *J Pers Soc Psychol* 78(3):546–558.
19. Biesanz JC, West SG, Millevoi A (2007) What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *J Pers Soc Psychol* 92(1):119–135.
20. Connelly BS, Ones DS (2010) An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. *Psychol Bull* 136(6): 1092–1122.
21. John OP, Robins RW (1993) Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *J Pers* 61(4):521–551.
22. Vazire S (2010) Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *J Pers Soc Psychol* 98(2):281–300.
23. Vazire S, Gosling SD (2004) e-Perceptions: Personality impressions based on personal websites. *J Pers Soc Psychol* 87(1):123–2.
24. Kenny DA (1991) A general model of consensus and accuracy in interpersonal perception. *Psychol Rev* 98(2):155–163.
25. Funder DC, Dobroth KM (1987) Differences between traits: Properties associated with interjudge agreement. *J Pers Soc Psychol* 52(2):409–418.
26. Funder DC, Colvin CR (1988) Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *J Pers Soc Psychol* 55(1):149–158.
27. Vazire S, Carlson EN (2011) Others sometimes know us better than we know ourselves. *Curr Dir Psychol Sci* 20(2):104–108.
28. Hirsh JB, Kang SK, Bodenhausen GV (2012) Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychol Sci* 23(6):578–581.
29. Kahn PH, et al. (2010) Psychological intimacy with robots? Using interaction patterns to uncover depth of relation. *Proceedings of the Fifth ACM/IEEE International Conference on Human–Robot Interaction* (IEEE, Piscataway, NJ), pp 123–124.
30. Kahn PH, et al. (2012) Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction* (Association for Computing Machinery, New York), pp 33–40.