nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0419-z

A robust data-driven approach identifies four personality types across four large data sets

Martin Gerlach[1], Beatrice Farb[1], William Revelle[2] and Luís A. Nunes Amaral[1,3,4,5*]

Understanding human personality has been a focus for philosophers and scientists for millennia[1]. It is now widely accepted that there are about five major personality domains that describe the personality profile of an individual[2,3]. In contrast to personality traits, the existence of personality types remains extremely controversial[4]. Despite the various purported personality types described in the literature, small sample sizes and the lack of reproducibility across data sets and methods have led to inconclusive results about personality types[5,6]. Here we develop an alternative approach to the identification of personality types, which we apply to four large data sets comprising more than 1.5 million participants. We find robust evidence for at least four distinct personality types, extending and refining previously suggested typologies. We show that these types appear as a small subset of a much more numerous set of spurious solutions in typical clustering approaches, highlighting principal limitations in the blind application of unsupervised machine learning methods to the analysis of big data.

Already in Ancient Greece, philosophers attempted to capture and organize individual differences in behaviour and emotion to understand human personality. However, only in the past few decades has a consensus emerged regarding the basic structure of personality in the form of the Big-5 (ref. [7]), also known as the Five-Factor model (FFM)[8]. The FFM surmises the existence of five traits—neuroticism, extraversion, openness, agreeableness and conscientiousness—that capture the main domains of human personality[3]. These five domains have been reliably identified in numerous empirical studies across different languages and cultures[2] and have been shown to be good predictors of patterns of behaviour, such as well-being and mental health, job performance and marital relations[9]. The FFM also provides a useful framework in a broad set of applications, including clinical assessments of personality disorders[10].

In contrast to the consensus on the existence of personality traits, the existence of personality types remains controversial. The best-supported modern and quantitative description of personality types surmises the existence of three so-called ARC types, named after the authors of the seminal studies by Asendorpf et al.[11], Robins et al.[12] and Caspi et al.[13]. This typology—the claim that people fall into the three distinct categories 'resilient', 'overcontrolled' and 'undercontrolled'—is based in large parts on an extension of the Freudian theory of ego functioning by Block[14]. However, this classification has been challenged extensively on statistical grounds[5,15–17]. Results obtained using different approaches and data sets cannot typically be replicated[4,18] or identify more than three clusters[19]. Even studies confirming the ARC taxonomy show large variation, highlighting the lack of consensus and replicability regarding the three personality types (Fig. 1). The difficulty in obtaining replicable results is exacerbated by the small sample sizes—typically not more than 1,000 individuals—analysed in these studies[6].

Here, we address the controversy related to the existence of personality types by combining an alternative computational approach to clustering with recently available large data sets comprising the responses of hundreds of thousands of users of web-based questionnaires[20–22]. In particular, we use 4 different data sets (each containing 100,000–500,000 respondents) that come from different sources, were collected in different countries, have different demographics with respect to age and gender and use different scales to measure the traits of the FFM. These data sets are among the largest publicly available data sets and allow for insight into whether personality types truly exist. We show that these data can be used to efficiently sample the multidimensional space of personality traits. This richness in data not only allows for a direct visualization of the structure in the space of personality traits but also enables us to formulate robust null models to assess the statistical significance of clustering solutions. Surprisingly, we find that even state-of-the-art clustering techniques[23] yield mostly spurious clusters. However, after developing an alternative clustering approach, we identify four robust clusters that correspond to statistically meaningful personality types. The personality types we uncover provide some support for, but extend and refine, the three ARC types[6].

We first analyse the answers of $N = 145{,}388$ individuals to $L = 300$ items of the IPIP-NEO (International Personality Item Pool[24] implementation of the NEO-PI-R[8]) personality questionnaire (Methods). We use factor analysis, which is a standard approach in psychometrics[25], to extract the five main domains of personality (Methods). Formally, if we denote the answer of respondent j to item i as A_{ij}, factor analysis corresponds to a dimensionality reduction in the form

$$A \approx Q \circ P \tag{1}$$

where $\circ$ denotes matrix multiplication, and Q (the factor loadings) and P (the factor scores) refer to the representations of items and respondents, respectively, in a space of five latent dimensions. Although in many dimensionality reduction applications the latent dimensions are not directly interpretable, here, inspection of Q reveals that they correspond closely to the dimensions from the FFM (Supplementary Fig. 1). In turn, we can identify $\mathbf{P}_j$ as containing the positions of respondent j in a 5D space of personality traits (Methods).

Visual inspection of the 1D and 2D projections of the 5D multivariate distributions does not suggest any obvious cluster structure beyond the trivial peak at the origin (Supplementary Fig. 2).

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [2]Department of Psychology, Northwestern University, Evanston, IL, USA. [3]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. [4]Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA. [5]Department of Medicine, Northwestern University, Evanston, IL, USA. *e-mail: amaral@northwestern.edu

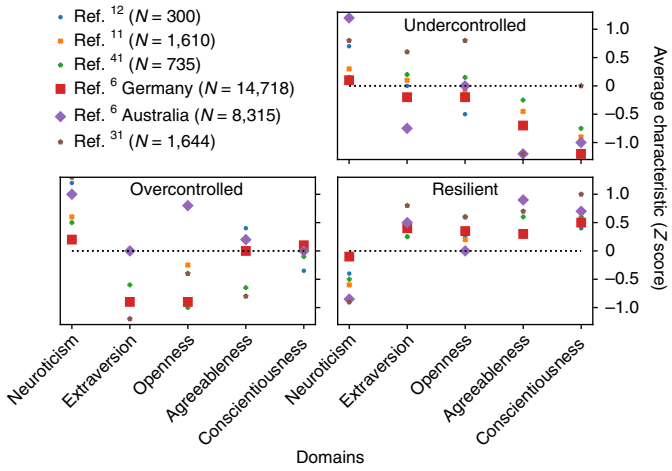

Fig. 1 | Uncertainty in the ARC-type classification. The location in the aggregated trait space of the three ARC personality types as reported in the literature. The values were obtained by visual approximation of the given references. The size of the markers is proportional to (the square root of) the number of respondents in each study. The dotted lines show the average values, $Z = 0$, for the samples and personality traits.

The marginal distributions show an approximately normal distribution of each trait across the population consistent with previous reports[26]. The 2D scatter plots show that, on average, scores between pairs of personality traits are virtually uncorrelated, which is to be expected owing to the assumption of orthogonality of latent dimensions in the factor analysis[25].

We use Gaussian mixture models (GMMs), a standard unsupervised clustering algorithm, to uncover groups of individuals with similar vectors $\mathbf{P}_j$ in the full trait space (Methods). As is recommended[27], we attempt to determine the optimal number of clusters using an information criterion, in this case the Bayesian information criterion (BIC)[28]. We find that the number of clusters (N_c) = 13 provides the 'optimal' fit of the GMM to the data, as any increase in the number of clusters fails to lead to a significant increase in the likelihood (Fig. 2a). This number of clusters is much larger than the ones reported in previous studies[6] (and references therein), but in view of the much larger data set analysed here, it could be rationalized because more data typically allow for the identification of more complex models in any model selection procedure.

Nonetheless, we test whether all identified clusters are truly meaningful. Recent investigations of latent variable models, such as topic models or community detection in networks, have revealed major limitations in the ability to infer the true underlying structure[29,30]. For example, if groups are unequal in size, the models tend to overfit the data to resolve smaller groups, which results in a large number of spurious solutions[30]. Thus, we assess whether inferred clusters correspond to meaningful personality types in the sense that they indicate significant 'peaks' in the distribution of individuals in the space of personality traits. Taking advantage of the large size of our database, our approach consists of, first, directly estimating the density ρ of each cluster and, second, comparing this with the density of a null model obtained from a randomized data set $\widetilde{\rho}$ (Methods). This is analogous to previous approaches in psychometrics that identified the optimal number of factors in factor analysis[31].

Surprisingly, only four of the identified clusters are centred in regions in which we observe a substantially larger fraction of respondents than expected from a random null model (Fig. 2b). In fact, about one-third of the inferred 'clusters' actually occupy regions with lower-than-expected densities, confirming that most correspond to spurious solutions. Although these results suggest that the solution of the GMM with $N_c = 13$ severely overfits the data, a detailed analysis on cluster solutions with different assumed N_c shows a non-trivial dependence of the cluster positions on the number of surmised clusters N_c; that is, fitting with only $N_c = 4$ clusters yields a solution that fails to identify most of the meaningful clusters.

Looking at the different solutions for $N_c = 12,\ldots, 20$, we observe that not only do the number of meaningful clusters stay roughly constant at 4–6 (Fig. 2c) but also that the 'position' of the meaningful clusters remains approximately fixed (Supplementary Fig. 3). This suggests that, to obtain the correct answer, one must consider a model that overfits the data, that is, that searches for a larger number of clusters than one expects to find. This interpretation is supported when analysing synthetic data (Supplementary Fig. 4), in which we know the 'true' number of clusters. Although careful consideration of model selection procedures such as the BIC will prefer a larger number of clusters, it will fail to indicate the large number of 'spurious' clusters required to resolve the structure on a finer scale.

Our analysis strongly supports the hypothesis that the existence of an abnormally high density of individuals around these four cluster centres indicates the existence of robust personality types (Fig. 2d). The least robustly identified cluster (Supplementary Fig. 3), which we denote the 'average' type, is characterized by average scores in all traits and being the only cluster where the univariate Gaussian of each dimension is within one standard deviation from the origin. In addition, the location of several individual traits are characterized by scores both below and above zero across different data sets (see below). The existence of such a type has been reported recently in some studies; however, the empirical evidence reported in the literature is contradictory[6,32]. The remaining three clusters can be roughly organized along the two dimensions of neuroticism and extraversion. One of the most stable clusters (Supplementary Fig. 3), which we denote the 'role model' type because it displays socially desirable traits, is characterized by low scores in neuroticism and high scores in all other traits. It can be unambiguously identified with the resilient type from the ARC taxonomy.

By contrast, the two other clusters are characterized by traits that are less socially desirable when compared to the characteristics of the 'role model' type. One of the clusters is marked by low scores on openness, agreeableness and conscientiousness, whereas the other cluster shows low scores on neuroticism and openness. A comparison with the closest types from the ARC typology (Fig. 1) suggests an identification with the undercontrolled and overcontrolled personality types, respectively. Although it is reassuring that our findings are related to known typological constructs, our analysis goes beyond the replication of such types. Previous studies have characterized the undercontrolled and overcontrolled types with scores varying across the whole spectrum (negative to neutral to positive) in at least two dimensions (extraversion and openness; openness and agreeableness), respectively. Moreover, for some traits, the scores of our types differ substantially from the undercontrolled and overcontrolled types. For example, whereas the overcontrolled type is usually associated with high scores on neuroticism, the closest cluster identified in our analysis displays low scores on neuroticism. This empirical result is also inconsistent with typical theoretical explanations of the differences between the two types in terms of control or, more specifically, internalizing and externalizing problems in the framework of psychopathology[12]. In this view, our analysis provides a different perspective from the classic ARC taxonomy. To highlight this refinement, we denote the two types as 'self-centred' and 'reserved', respectively.

We can obtain a more nuanced view addressing the question of the degree of clustering by visually exploring densities in suitably defined 2D hyperplanes (Supplementary Fig. 5). Although there is

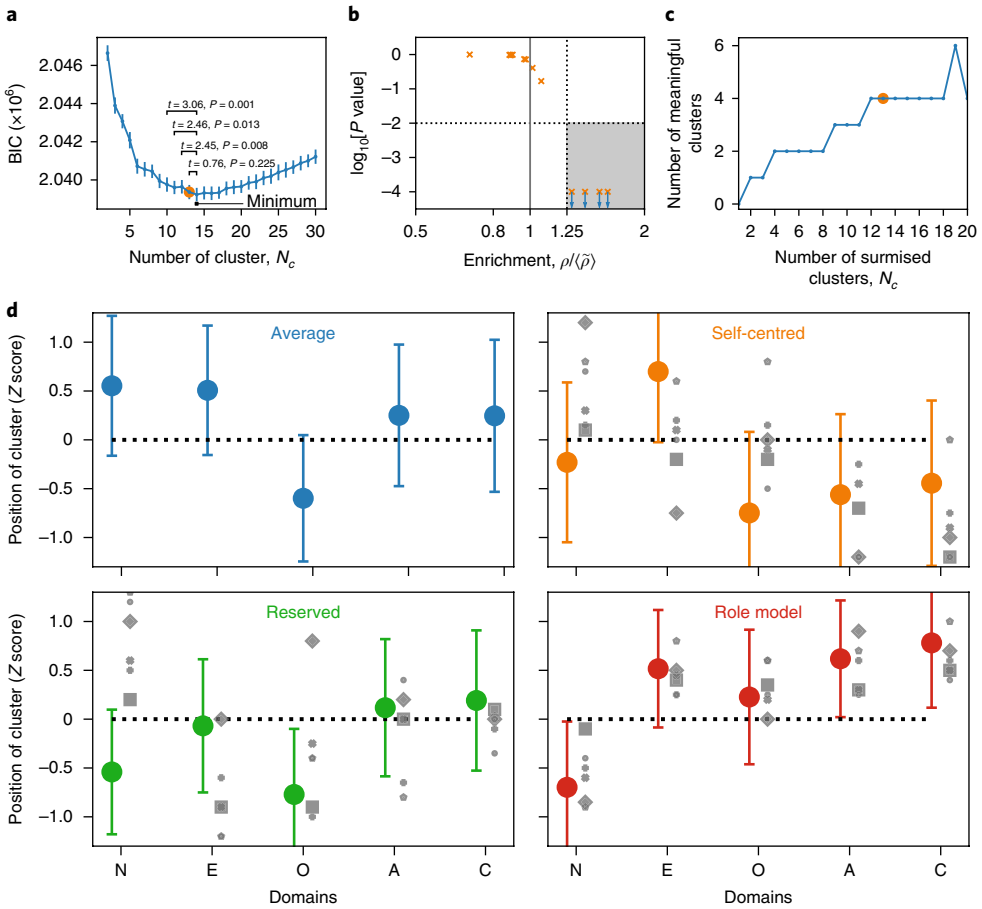

Fig. 2 | Clustering reveals four meaningful personality types. The identification of four meaningful personality types in the Johnson-300 data set ($N = 145{,}388$). **a**, The optimal number of clusters in the GMM according to the BIC as a function of N_c obtained from the best solution of 100 fits with different initial conditions. We obtained error bars by applying the same fitting procedure to 100 bootstrapped samples of the data (±3 standard errors of the mean). We select, as the optimal solution, the smallest value of N_c for which the BIC is statistically indistinguishable from the minimum value of the BIC (two-sided *t*-test for the difference in means from two populations of 100 bootstrapped samples, $P < 0.05$). This identifies $N_c = 13$ (orange dot) as the optimal solution in the Johnson-300 data set as there is no significant improvement in the BIC upon adding further parameters. **b**, The *P* value and enrichment of each cluster (orange x) from the optimal solution identified, suggesting that there are only four meaningful clusters (grey shaded area). The dotted lines (vertical and horizontal) correspond to a threshold for a *P* value of 0.01 and an enrichment of 1.25, respectively. For comparison with a null effect, we show an enrichment of 1.0 (solid grey line). For the randomized data, we obtained an empirical distribution of the estimate of the density from 10,000 different random realizations. The blue arrows indicate $P < 10^{-4}$, which is the maximum resolution given the number of random realizations. **c**, The number of meaningful clusters as a function of the number of surmised clusters, N_c, using the analysis in **b**. **d**, The position of cluster centres (in units of standard deviation in each dimension) in the trait space for the Johnson-300 data set (●). The error bars correspond to the standard deviation in each dimension of each multivariate Gaussian from the GMM (that is, the diagonal entries of the fitted covariance matrix). For comparison, we show the positions of the closest type from the ARC taxonomy reviewed in Fig. 1 (light grey). The dotted lines show the average values, $Z = 0$, for the samples and personality traits. A, agreeableness; C, conscientiousness; E, extraversion; N, neuroticism; O, openness.

a considerable overlap between different clusters, we not only find that individual cluster centres are located in regions of higher-than-expected density but also identify neighbouring significant volumes of 'void' space in which the density of individuals is much lower than expected.

To test the robustness of our finding, we next replicate our analysis on three independent data sets (Johnson-120, myPersonality-100 and BBC-44) from different sources. Although these data sets are of similar magnitude in terms of the number of respondents, they use different scales in their measurement of the traits in the FFM; in particular, they contain only 120, 100 and 44 items, respectively, instead of 300 items (Methods). Applying the same procedure as above, we first perform factor analysis, clearly revealing the Five-Factor structure (Supplementary Fig. 6), and obtain the position of each respondent in the 5D space of personality traits, which yield no obvious clusters in the lower-dimensional projections (Supplementary Figs. 7–9). Next, we perform cluster analysis showing the existence of a similar number of meaningful personality types, confirming our previous finding that the seemingly 'optimal' cluster solution contains a majority of spurious clusters (Supplementary Fig. 10). Comparing the position of the clusters identified in each data set, we find that most of the meaningful and spurious clusters can be matched with a corresponding cluster in the Johnson-300 data set (Fig. 3a), indicating a low occurrence of type I and type II errors when matching clusters across data sets. This is highlighted by the comparison of the location of the matched clusters in the 5D space of personality traits (Fig. 3b and Supplementary Table 1), which shows a high degree of agreement on almost all traits across different data sets. However, although for the Johnson-120 data set we recover all four of the personality types, the myPersonality-100 and the BBC-44 each yield only three of the personality types obtained for the Johnson-300 data set.

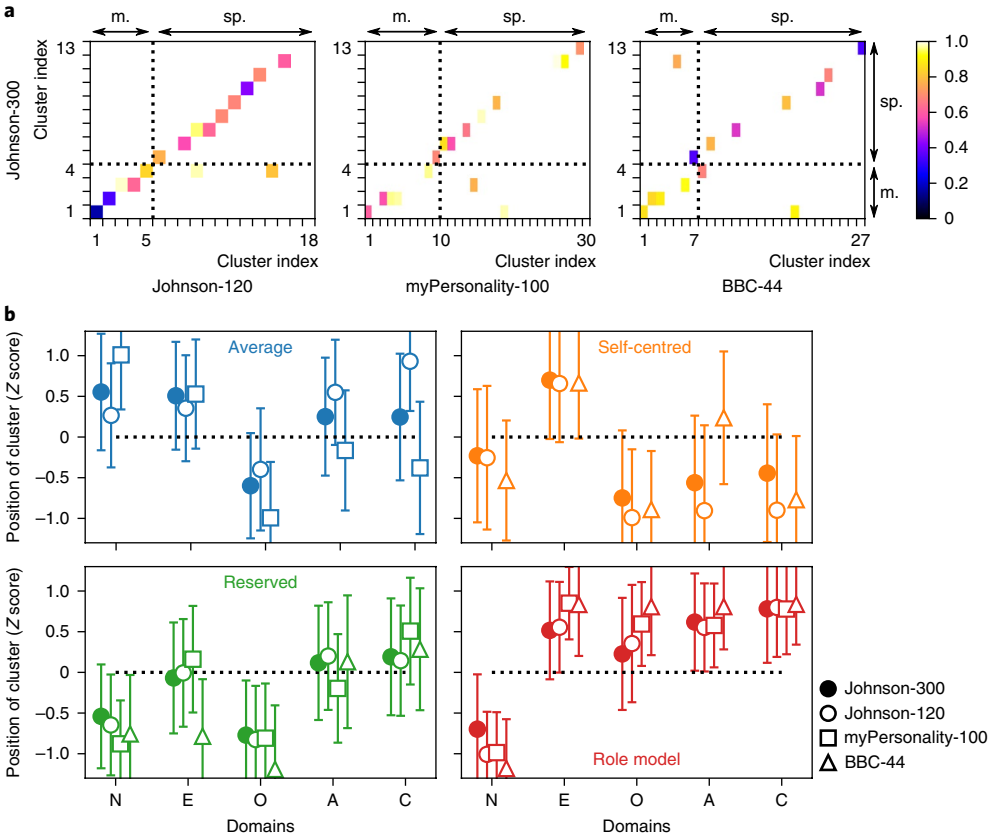

Fig. 3 | Replicability of personality types in three independent data sets. a, The correspondence between clusters found in the Johnson-300 data set ($N = 145,388$) and three independent data sets (Methods): the Johnson-120 ($N = 410,376$), the myPersonality-100 ($N = 575,380$) and the BBC-44 ($N = 386,375$) data sets. We calculate the Euclidean distance between the position of the centres of the clusters identified in the additional data set and the Johnson-300 data set, respectively. For each cluster in the additional data set, we identify the closest cluster in the Johnson-300 data set, that is, with the smallest Euclidean distance, and consider it a match if the distance is smaller than 1. The dotted lines separate meaningful (m.) from spurious (sp.) clusters as identified for the Johnson-300 data set (Fig. 2b) and the three independent data sets (Supplementary Fig. 10b,e,h). **b**, The position of the cluster centres (in units of standard deviation in each dimension) in the trait space for the Johnson-300 data set and the closest matching clusters of the Johnson-120, myPersonality-100 and BBC-44 data sets identified in **a**. The error bars show the values of the diagonal entries of the fitted covariance matrix of each multivariate Gaussian in the GMM. The dotted lines show the average values, $Z = 0$, for the samples and personality traits. A, agreeableness; C, conscientiousness; E, extraversion; N, neuroticism; O, openness.

This suggests that the lower agreement with the cluster solution of the Johnson-300 data set is an artefact of the smaller number of items in the latter data sets. To test this hypothesis, we generate synthetic versions of the Johnson-300 data set in which we only consider a random selection of 120, 100 and 44 items out of the original 300 (Johnson-300*[120], Johnson-300*[100] and Johnson-300*[44]), whereas the responses to the selected items remain the same. Consistent with observations in the additional data sets, we find that the distinction between meaningful and spurious clusters gradually becomes blurred (Supplementary Fig. 11), that is, spurious clusters 'bleed' into the region of significance. More importantly, the degree of overlap between the shorter variations and the original Johnson-300 data set systematically decreases as we keep fewer items (Supplementary Fig. 12). In fact, using 120 or 100 items, we recover the same four personality types, whereas for 44 items, we recover only two of the personality types ('role model' and 'reserved'), respectively, reproducing the very same pattern observed in Fig. 3.

Taken together, these results confirm our initial findings on the existence of at least four robust personality types. Moreover, they demonstrate that reducing the number of items in a questionnaire can lead to a strong decrease in the resolution of the measurement of personality traits of individuals, which is probably an artefact of the increasing discretization of respondents' scores in the trait space due to the smaller number of items (Supplementary Figs. 13–16).

To provide external validity, we next investigate whether the four personality types identified in the Johnson-300 data set are correlated with demographic variables, such as gender and age. For example, from longitudinal studies, we expect an increase of socially desirable traits with maturity[33] or a larger fraction of young males among self-centred individuals[34].

The overall distribution of respondents with respect to gender and age varies considerably (Fig. 4a), both within (females and young individuals are more numerous) and across data sets (the BBC-44 data set is much more skewed with respect to age). Thus, we measure the degree to which a certain combination of age and gender is overrepresented or underrepresented in the vicinity of the location of the respective personality type in comparison to a random sampling of the whole population (Methods). This supervised approach allows us to investigate the composition of all four types, even for data sets in which we were unable to resolve these types in an unsupervised way.

We find strong dependence on age and gender except for the 'reserved' type (Fig. 4b). For the 'role model' type, respondents younger than 21 years of age are slightly underrepresented, whereas respondents older than 40 years of age are strongly overrepresented.

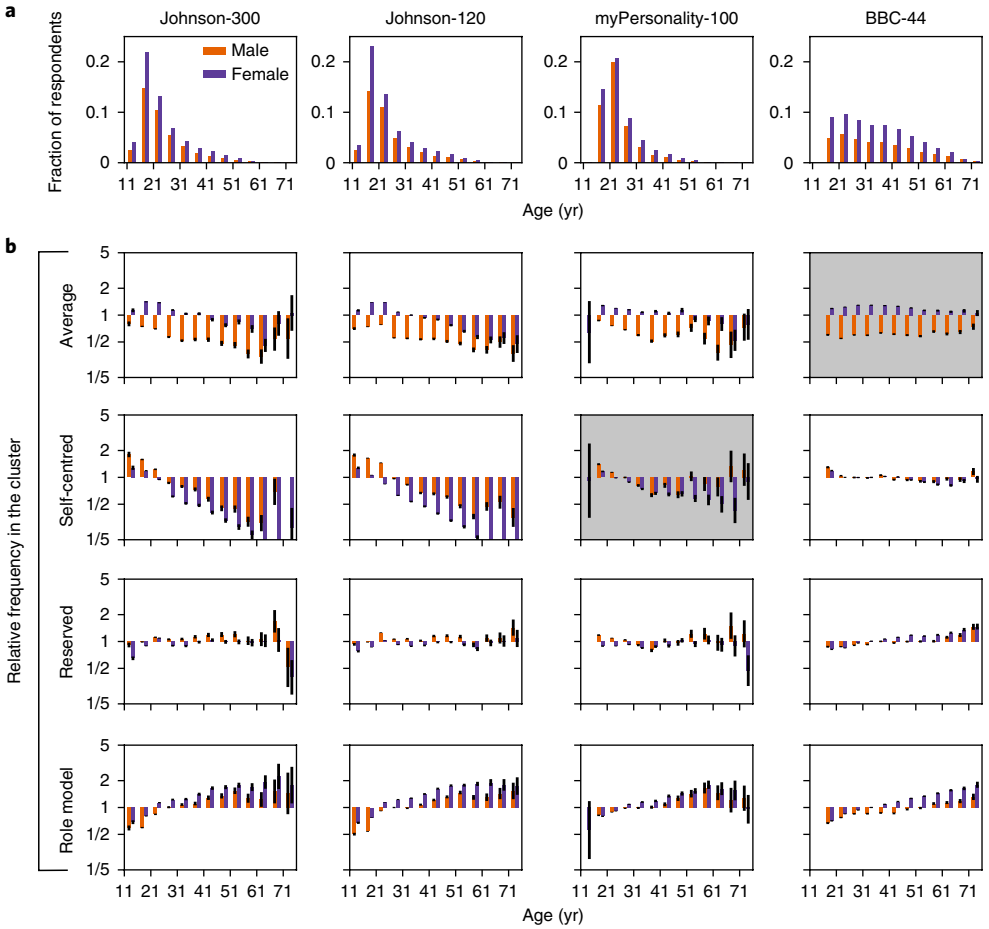

Fig. 4 | The composition of four meaningful clusters is correlated with age and gender and is stable across different data sets. a, The number of respondents according to age and gender for the four different data sets: the Johnson-300 ($N = 145{,}388$), the Johnson-120 ($N = 410{,}376$), the myPersonality-100 ($N = 225{,}117$) and the BBC-44 ($N = 386{,}371$) data sets. **b**, The relative frequency of respondents of a given age and gender at the positions of the four meaningful clusters identified in the Johnson-300 data set (Fig. 2). We measure how much a group, defined by age and gender, is overrepresented (or underrepresented) in each cluster (rows) by calculating the number of respondents of a group within a radius of $r = 1.5$ around each cluster and compare with the expected number of respondents of the same group based on the global distribution of age and gender for each data set. The black bars indicate ±1 standard deviation (Methods). The grey background indicates that the personality type could not be identified in the respective data set.

This dependence is even more pronounced for females than for males. The 'self-centred' type displays almost the exact opposite pattern—the strongest effect can be observed for females older than 60 years of age, showing a more than fivefold decrease in appearing in the vicinity of this cluster. Although young males are overrepresented, females older than 15 years of age are underrepresented. The 'average' type shows an underrepresentation of males. Most importantly, the observed dependence on age and gender for each personality type is consistent across all data sets.

The consistency with measurements of external variables such as gender and age add further support for the robustness of the typology uncovered by our analysis. Furthermore, our analysis highlights how large data sets can offer a complementary approach to small-scale studies of the temporal dynamics of personality structure.

To summarize, our study provides compelling evidence, both quantitatively and qualitatively, for the existence of at least four distinct personality types. Although these types overlap in certain aspects with typologies hypothesized previously—even showing similarities with some of the ancient four temperaments by considering only the two dimensions of neuroticism and extraversion[35]—our data-driven approach minimizes the effect of possible confirmation bias and rationalization of ad hoc typological constructs. The size of our data sets (nearly 1,000-fold larger than typical studies and between 0.1% and 1% of the total population of the United States and the United Kingdom, respectively) makes us confident that the identified typology represents a robust structure.

Several limitations of this study need to be recognized. First, the samples of our study are large and diverse, yet they are not representative of the population, as demonstrated by the distribution of age and gender. Second, different factors induce measurement errors in unknown and potentially biased ways. The small number of items used in some web-based questionnaires results in a decrease of the signal-to-noise ratio. Indeed, our study does not conclusively answer what the minimum number of items needed to reliably assess personality types is. As we have shown, different scales will lead to different factor scores—even when measuring the same five personality domains, two questionnaires might use different items. This issue will be exacerbated when considering alternative representations of the space of personality traits, for example, the 30 facets of the FFM[36], the 6 domains of the HEXACO inventory[18] or the 27-dimensional SAPA Personality Inventory[37]. Third, obtaining data sets of this magnitude is contingent upon the use of self-reports, ignoring unique insights from non-self reports[38]. Although this constitutes an intrinsic limitation due to, for example, differences in response style, poor self-perception or social desirability[39],

we note that self-reports have been repeatedly shown to correlate strongly with peer evaluations[2].

Our study highlights several open challenges for future studies on personality. Despite our results indicating a robust pattern of types across different data sets, there is still no convergence on a unified framework for how many and which types are supported by empirical evidence. Furthermore, although the presented empirical evidence for the identified types is unambiguous, we still lack a theoretical understanding, for example, in analogy to Block's psychodynamic theory for attractor states in the space of personality[14], of why types show particular combination of traits. Finally, the data analysed here does not allow us to address the pertinent question of how much personality types are able to predict life outcomes. Previous research has shown the usefulness of the ARC types in predicting life outcomes[40,41]; however, such data are not available in questionnaires of the type analysed in this study. Fortunately, some large-scale studies, such as the SAPA-Project Database of Individual Differences[22], are now being collected in combination with the assessment of personality.

Our results have important implications. An empirically justified taxonomic system of personality types offers a coarse-grained abstraction on the distribution of personality traits across individuals, in analogy to the distinction between different groups of elementary particles (for example, fermions or bosons) in physics or different species in biology. Such a classification is potentially useful in applied contexts, such as in clinical settings related to psychopathology or vocational settings. Previously found types have been found to correlate with these outcomes, in particular, when the time window for prediction is large[5]. More pragmatically, a type-based approach offers additional possibilities in the design of questionnaires with fewer items, as less information is required for a discrete-type classification than for a continuous trait estimation. Moreover, our analysis establishes a major advance in the interpretation of clustering solutions from a methodological perspective. Indeed, our key technical insight reveals that even state-of-the-art clustering algorithms will only find the correct solution by searching for a larger number of clusters than what could exist in the data and will fail to identify the abundance of mostly spurious clusters even when using a careful consideration of (approximate) model selection techniques, such as the BIC.

Methods

Data. In our analysis, we use four different data sets from web-based questionnaires that measure the personality traits (so-called domains) of the FFM[3] using different scales: the Johnson-300 data set[36] (145,338 respondents), the Johnson-120 data set [36] (410,376 respondents), the myPersonality-100 data set [42] (575,380 respondents) the and BBC-44 data set[43] (386,375 respondents). Participants were asked to state how much they agree with statements such as "I work hard." with possible responses: 1 (very inaccurate), 2 (moderately inaccurate), 3 (neither accurate nor inaccurate), 4 (moderately accurate) or 5 (very accurate). The Johnson-300 data set constitutes by far the most detailed measurement of the FFM owing to the large number of items answered by each individual using the IPIP-NEO. The Johnson-120 data set uses a shorter version of the questionnaire with a subset of 120 items containing independently collected responses. The myPersonality-100 data set uses 100 items from the IPIP representation of the NEO PI-R of which 65 and 37 items are not contained in the Johnson-120 (Johnson-300) data set, respectively. The BBC-44 data set uses 44 items that are not contained in any of the three other data sets. For all data sets, we only consider respondents who gave responses to all of the items. For the myPersonality-100 data set, we considered only one set of responses of each individual (indicated by the variable 'best protocol') in case the same person took the test several times. In addition, we obtained the gender and age of each participant. Although it is clear from these demographics that our data sets are not representative of the general population, it has been well established that data from web-based questionnaires are more diverse and are of at least as good quality as data obtained through more traditional approaches[44]. For the myPersonality-100 and the BBC-44 data sets, gender or age is not available for 350,263 (60.9%) and 4 (0.001%) respondents, respectively. In the analysis involving these demographics, we only considered respondents for which both variables are available.

Factor analysis. We use factor analysis[45], a standard method of dimensionality reduction in the analysis of personality traits[25] similar to, for example, the principal components analysis, to find the underlying (latent) structure of the matrix A_{ij}. In this, we assume the existence of D latent dimensions, such that we can decompose A_{ij} as

$$A_{ij} \approx \sum_{d=1}^{D} Q_{id} P_{dj} \tag{2}$$

where Q_{id} is a $S \times D$ dimensional matrix describing how each latent dimension d is defined as a superposition of items or, put differently, how much the latent factor d influences the answer to item i (the so-called factor loading). $\boldsymbol{P}_{dj}$ is then a $D \times N$ dimensional matrix describing where each respondent is located in the space of the latent dimensions. We use the numerical implementation described in scikit-learn[46].

The matrices Q and P of the factor analysis are not unique in the sense that we can introduce any rotation by an orthogonal matrix R such that

$$A = Q \circ P = Q \circ (R^T \circ R) \circ P = Q' \circ P' \tag{3}$$

with $Q' = Q \circ R^T$ and $P' = R \circ P$ and where $\circ$ denotes matrix multiplication and R^T denotes the transpose of R. We use the most common choice for R, the so-called varimax rotation[47], which maximizes the variance of the squared loadings Q column-wise. This typically leads to the most block-diagonal form of Q, such that one can easily find relations between latent dimensions and items. We use the numerical implementation in the package Factor-rotation[48].

As there is no consensus on the 'correct' rotation and that factor scores from the varimax rotation ignore the known correlations between personality traits, we also consider oblique (non-orthogonal) rotations. In fact, using the so-called quartimin rotation from the oblimin family[49], we reproduce the main findings on personality types reported in the main text (Supplementary Fig. 17).

An alternative to factor analysis are the raw Big-Five scores obtained from direct scoring of the items. However, the small number of items leads to visible discretization artefacts (Supplementary Figs. 13–16), which induce substantial and non-trivial biases when applying clustering approaches for continuous trait variables.

From latent dimensions to personality traits. The vector $\mathbf{P}_j$ corresponds to the position of respondent j in the space of latent dimensions. Upon inspection of the matrix Q, we find a unique mapping between latent dimensions and psychological traits (Supplementary Fig. 1) as

$$j = 1, \ldots, 5 \rightarrow N, E, O, A, C \tag{4}$$

This allows us to interpret the factor scores $\mathbf{P}_j$ as the coordinates of respondent j in the 5D space of personality traits.

GMMs. We cluster the data using a GMM[50]. The GMM is formulated as a generative model in which we assume that the data $\boldsymbol{D} = \{\mathbf{x}_i\}$ with $i = 1, \ldots, N$ observations are generated from $k = 1, \ldots, N_c$ multivariate Gaussian distributions $\boldsymbol{N}(\boldsymbol{\mu}_k, \Sigma_k)$ with the mean $\boldsymbol{\mu}_k$ and the covariance matrix Σ_k. Introducing the weights π_k (obeying $\sum_k \pi_k = 1$), we can express the marginal likelihood as:

$$p(D \mid \{\boldsymbol{\mu}_k, \Sigma_k, \pi_k\}) = \prod_{i=1}^{N} \sum_k \pi_k N(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k) \tag{5}$$

We found the best set of parameters $\{\boldsymbol{\mu}_k, \Sigma_k, \pi_k\}$ by maximizing equation (5) with respect to the parameters. We use the numerical implementation described in scikit-learn[46].

The main advantage of using generative models such as the GMM is that the problem of clustering can be approached in the framework of statistical inference in which the assumptions about the data have to be formulated explicitly. In addition to fitting cluster centres $\boldsymbol{\mu}_k$, we can explicitly take into account unequal sizes of clusters (π_k) as well as the covariance structure in the data (Σ_k). Furthermore, the GMM allows for so-called soft clustering, in which each data point is assigned a probability to belong to any of the clusters. As a result, the GMM is a much more flexible framework for clustering than are more commonly used methods, such as k-means clustering.

Kernel density estimation. Given a set of data $\boldsymbol{D} = \{\mathbf{x}_i\}$ with $i = 1, \ldots, N$ observations where $\mathbf{x}_i$ is a D dimensional vector, we estimate the density $\rho(\mathbf{y})$ for an arbitrary point $\mathbf{y}$ in the trait space using a kernel density estimation[50]:

$$\rho(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} K_h(\mathbf{y}, \mathbf{x}_i) \tag{6}$$

where K_h denotes a kernel with bandwidth h. We use a Gaussian kernel

$$K_h(\mathbf{y}, \mathbf{x}_i) = \frac{1}{(2\pi h^2)^{D/2}} e^{\frac{\sum_{d=1}^{D} (x_{id} - y_d)^2}{2h^2}} \tag{7}$$

with the bandwidth calculated from the average Euclidean distance between the nearest neighbours, yielding $0.20 < h < 0.26$ for the four data sets. We use the numerical implementation described in scikit-learn[46].

We compare the density ρ in the original data set with the density $\tilde{\rho}$ obtained with a random null model. For the latter, we reassign ('shuffle') the values P_{dj} for fixed d across all individuals j. Repeating the shuffling procedure a large number of times allows us to estimate a distribution $p(\tilde{\rho})$. To assess whether the density is larger than expected from chance, we first calculated the P value as $p(\rho < \tilde{\rho})$. This corresponds to a one-sided test of the hypothesis $\rho = \tilde{\rho}$, that is, the density is the same as in the randomized data set. To quantify how much the density exceeds the random expectation in absolute terms, we also calculated the enrichment as $\rho \,/\, \langle\tilde{\rho}\rangle$ where $\langle\tilde{\rho}\rangle$ denotes the mean value of $\tilde{\rho}$.

We use this procedure to define a cluster as meaningful and non-spurious if it meets two criteria: one, the density at the cluster centre is significantly larger than expected from chance ($P < 0.01$); and two, its density exceeds the random density not just marginally but also in absolute terms by at least 25% ($\rho \,/\, \langle\tilde{\rho}\rangle > 1.25$).

Composition of clusters with respect to age and gender. Given the location $\mathbf{P}_k$ of cluster k, we count the number of respondents with a given age a and gender g contained within a sphere of radius $\delta = 1.5$ denoted by $n_k(a,g)$ (with a total of N_k individuals within the sphere). We compare this number with the expected number of individuals with the same age and gender (a,g), $\tilde{n}_k(a,g)$. Assuming a binomial drawing with the fraction of individuals with the same (a,g) in the complete data set, $p(a,g)$, we can calculate the average and the standard deviation as

$$\begin{aligned} \mu(\tilde{n_k}(a,g)) &= N_k p(a,g), \\ \sigma(\tilde{n_k}(a,g)) &= \sqrt{N_k p(a,g)\,(1-p(a,g))} \end{aligned} \tag{8}$$

From this, we can define the relative frequency of age a and gender g in cluster k, which corresponds to the degree to which they are overrepresented or underrepresented, as

$$z_k(a,g) = n_k(a,g) \,/\, \tilde{n_k}(a,g) \tag{9}$$

such that $z > 1$ ($z < 1$) indicates overrepresentation or underrepresentation, respectively.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The code used for data processing and clustering is available in a GitHub repository (https://github.com/amarallab/personality-types).

Data availability

Data are available from https://osf.io/tbmh5/ (Johnson-300 and Johnson-120), http://mypersonality.org (myPersonality-100) and https://doi.org/10.5255/UKDA-SN-7656-1 (BBC-44).

Received: 17 January 2018; Accepted: 25 July 2018;
Published online: 17 September 2018

References

1. Revelle, W., Wilt, J. & Condon, D. M. in *The Wiley-Blackwell Handbook of Individual Differences* (eds Chamorro-Premuzic, T. et al.) 1–38 (Wiley-Blackwell, Oxford, 2013).
2. McCrae, R. R. & Costa, P. T. in *The SAGE Handbook of Personality Theory and Assessment: Volume 1 Personality Theories and Models* (eds Boyle, G. J. et al.) 273–294 (SAGE, London, 2008).
3. Widiger, T. A. *The Oxford Handbook of the Five Factor Model of Personality* (Oxford Univ. Press, Oxford, 2015).
4. McCrae, R. R., Terracciano, A., Costa, P. T. & Ozer, D. J. Person-factors in the California adult Q-set: closing the door on personality trait types? *Eur. J. Pers.* **20**, 29–44 (2006).
5. Donnellan, M. B. & Robins, R. W. Resilient, overcontrolled, and undercontrolled personality types: issues and controversies. *Soc. Pers. Psychol. Compass* **11**, 1070–1083 (2010).
6. Specht, J., Luhmann, M. & Geiser, C. On the consistency of personality types across adulthood: latent profile analyses in two large-scale panel studies. *J. Pers. Soc. Psychol.* **107**, 540–556 (2014).
7. Goldberg, L. R. An alternative "description of personality": the Big-Five factor structure. *J. Pers. Soc. Psychol.* **59**, 1216–1229 (1990).
8. Costa, P. T. & McCrae, R. R. *NEO PI-R Professional Manual* (Psychological Assessment Resources, Odessa, FL, 1992).
9. Ozer, D. J. & Benet-Martínez, V. Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.* **57**, 401–421 (2006).
10. Widiger, T. A. & Costa, P. T. Jr. *Personality Disorders and the Five-Factor Model of Personality* 3rd edn (American Psychological Association, Washington DC, 2013).
11. Asendorpf, J. B., Borkenau, P., Ostendorf, F. & Van Aken, M. A. G. Carving personality description at its joints: confirmation of three replicable personality prototypes for both children and adults. *Eur. J. Pers.* **15**, 169–198 (2001).
12. Robins, R. W., John, O. P., Caspi, A., Moffitt, T. E. & Stouthamer-Loeber, M. Resilient, overcontrolled, and undercontrolled boys: three replicable personality types. *J. Pers. Soc. Psychol.* **70**, 157–171 (1996).
13. Caspi, A. & Silva, P. A. Temperamental qualities at age three predict personality traits in young adulthood: longitudinal evidence from a birth cohort. *Child Dev.* **66**, 486–498 (1995).
14. Block, J. *Lives Through Time* (Bancroft Press, Berkeley, CA, 1971).
15. Costa, P. T., Herbst, J. H., McCrae, R. R., Samuels, J. & Ozer, D. J. The replicability and utility of three personality types. *Eur. J. Pers.* **16**, S73–S87 (2002).
16. Herzberg, P. Y. & Roth, M. Beyond resilients, undercontrollers, and overcontrollers? An extension of personality prototype research. *Eur. J. Pers.* **20**, 5–28 (2006).
17. Altman, N. & Krzywinski, M. Points of significance: clustering. *Nat. Methods* **14**, 545–546 (2017).
18. Ashton, M. C. & Lee, K. An investigation of personality types within the HEXACO personality framework. *J. Individ. Differ.* **30**, 181–187 (2009).
19. Isler, L., Fletcher, G. J. O., Liu, J. H. & Sibley, C. G. Validation of the four-profile configuration of personality types within the Five-Factor model. *Pers. Individ. Dif.* **106**, 257–262 (2017).
20. Rentfrow, P. J. et al. Divided we stand: three psychological regions of the United States and their political, economic, social, and health correlates. *J. Pers. Soc. Psychol.* **105**, 996–1012 (2013).
21. Rentfrow, P. J., Jokela, M. & Lamb, M. E. Regional personality differences in Great Britain. *PLoS ONE* **10**, e0122245 (2015).
22. Revelle, W. et al. in *SAGE Handbook of Online Research Methods* (eds Fielding, N. G. et al.) 578–595 (SAGE, London, 2016).
23. Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.* **31**, 651–666 (2010).
24. Goldberg, L. R. in *Personality Psychology in Europe* Vol. 7 (eds Mervielde, I., Deary, I., De Fruyt, F. & Ostendorf, F.) 7–28 (Tilburg Univ. Press, Tilburg, 1999).
25. Revelle, W. *An Introduction to Psychometric Theory with Applications in R* (Personality Project, 2017); http://www.personality-project.org/r/book/
26. Costa, P. T. & McCrae, R. in *The Oxford Handbook of the Five Factor Model* (ed. Widiger, T. A.) 1–52 (Oxford Univ. Press, Oxford, 2015).
27. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference* 2nd edn (Springer, New York, NY, 2002).
28. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
29. Fortunato, S. & Barthelemy, M. Resolution limit in community detection. *Proc. Natl Acad. Sci. USA* **104**, 36–41 (2007).
30. Lancichinetti, A. et al. A high-reproducibility and high-accuracy method for automated topic classification. *Phys. Rev. X* **5**, 011007 (2015).
31. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).
32. Xie, X., Chen, W., Lei, L., Xing, C. & Zhang, Y. The relationship between personality types and prosocial behavior and aggression in Chinese adolescents. *Pers. Individ. Dif.* **95**, 56–61 (2016).
33. Terracciano, A., McCrae, R. R., Brent, L. J. & Costa, P. T. Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore longitudinal study of aging. *Psychol. Aging* **20**, 493–506 (2005).
34. Meeus, W., Van de Schoot, R., Klimstra, T. & Branje, S. Personality types in adolescence: change and stability and links with adjustment and relationships: a five-wave longitudinal study. *Dev. Psychol.* **47**, 1181–1195 (2011).
35. Eysenck, H. J. & Eysenck, M. W. *Personality and Individual Differences: a Natural Science Approach* (Plenum Press, New York, NY, 1985).
36. Johnson, J. A. Measuring thirty facets of the Five Factor model with a 120-item public domain inventory: development of the IPIP-NEO-120. *J. Res. Pers.* **51**, 78–89 (2014).
37. Condon, D. M. The SAPA personality inventory: an empirically-derived, hierarchically-organized self-report personality assessment model. Preprint at https://psyarxiv.com/sc4p9/ (2018).
38. Vazire, S. & Mehl, M. Knowing me, knowing you: the accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *J. Pers. Soc. Psychol.* **95**, 1202–1216 (2008).
39. Paulhus, D. L. & Vazire, S. in *Handbook of Research Methods in Personality Psychology* (eds Robins, R. W. et al.) 224–239 (Guilford, New York, NY, 2007).
40. Chapman, B. & Goldberg, L. Replicability and 40-year predictive power of childhood ARC types. *J. Pers. Soc. Psychol.* **101**, 593–606 (2011).
41. Steca, P., Alessandri, G. & Caprara, G. V. The utility of a well-known personality typology in studying successful aging: resilients, undercontrollers, and overcontrollers in old age. *Pers. Individ. Dif.* **48**, 442–446 (2010).
42. Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D. Facebook as a social science research tool: opportunities, challenges, ethical considerations and practical guidelines. *Am. Psychol.* **70**, 543–556 (2015).

43. University of Cambridge, Department of Psychology, British Broadcasting Corporation *BBC Big Personality Test, 2009–2011: Dataset for Mapping Personality across Great Britain* [data collection] (UK Data Service, 2015); https://doi.org/10.5255/UKDA-SN-7656-1
44. Gosling, S. D., Vazire, S., Srivastava, S. & John, O. P. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *Am. Psychol.* **59**, 93–104 (2004).
45. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* 2nd edn (Springer, New York, NY, 2009).
46. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
47. Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200 (1958).
48. Factor rotation. Python code for factor rotation (GitHub, 2017); http://github.com/mvds314/factor_rotation
49. Carrol, J. An analytical solution for approximating simple structure in factor analysis. *Psychometrika* **18**, 23–38 (1953).
50. Bishop, C. *Pattern Recognition and Machine Learning* (Springer, New York, NY, 2006).

Acknowledgements

L.A.N.A. thanks the John and Leslie McQuown Gift and support from the Department of Defense Army Research Office under grant number W911NF-14-1-0259. W.R.'s work was partially supported by a grant from the National Science Foundation: SMA-1419324. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank J. Johnson for making the Johnson-300 and the Johnson-120 data sets publicly available; D. Stillwell, M. Kosinski and the myPersonality project for sharing the myPersonality-100 data; and the BBC LabUK for making the BBC-44 data set publicly available.

Author contributions

M.G., B.F., W.R. and L.A.N.A. designed the research. M.G., B.F., W.R. and L.A.N.A. performed the research. M.G. and B.F. analysed the data. M.G., W.R. and L.A.N.A. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0419-z.

Correspondence and requests for materials should be addressed to L.A.N.A

nature research

Corresponding author(s): Amaral, Luis A. Nunes

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used publicly available data on web-based personality questionnaires. Data was downloaded and analyzed using custom Python code.
Data analysis	Custom Python code using standard packages including numpy, scipy, and scikit-learn.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data are available from:
https://osf.io/tbmh5/ (Johnson-300 and Johnson-120),

http://mypersonality.org (myPersonality-100), and
http://doi.org/10.5255/UKDA-SN-7656-1 (BBC-44).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We use 4 different datasets from web-based questionnaires measuring the personality traits (so-called domains) of the Five-Factor model 3 using different scales.
Research sample	Participants are asked to state how much they agree with statements such as "I work hard." with possible responses: 1 (Very Inaccurate), 2 (Moderately Inaccurate), 3 (Neither Accurate nor Inaccurate), 4 (Moderately Accurate), or 5 (Very Accurate). The Johnson-300 dataset with 145,338 respondents constitutes by far the most detailed measurement of the Five-Factor model due to the large number of items answered by each individual using the IPIP-NEO. The Johnson-120 dataset with 410,376 respondents uses a shorter version of the questionnaire with a subset of 120 items containing independently collected responses. The myPersonality-100 dataset with 575,380 respondents uses 100 items from the IPIP representation of the NEO-PI R of which 65 (37) of the items not contained in the Johnson-120 (Johnson-300) dataset. The BBC-44 dataset with 386,375 respondents uses 44 items that are not contained in any of the 3 other datasets.Additionally, we obtain the gender and age of each participant. For the myPersonality-100 and the BBC-44 gender or age is not available for 350,263 (60.9%) and 4 (0.001%) respondents, respectively.
Sampling strategy	We use data collected in other studies. Sampling is described in the original studies.
Data collection	We used publicly available data on web-based personality questionnaires. Data was downloaded and analyzed with custom Python code.
Timing	Data was collected in other studies described in: - Johnson, J. A. Measuring thirty facets of the Five Factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. J. Res. Pers. 51, 78–89 (2014). Data at https://osf.io/wxvth. - University of Cambridge. Department of Psychology, British Broadcasting Corporation. BBC big personality test, 2009-2011: Dataset for mapping personality across Great Britain. [Data collection]. UK Data Service 7656 (2015). Data at http://doi.org/10.5255/UKDA-SN-7656-1. - Kosinski, M., Matz, S., Gosling, S., Popov, V. & Stillwell, D. Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines. Am. Psychol. 70, 543–556 (2015).
Data exclusions	For all datasets, we only consider respondents which gave responses to all of the items. For the myPersonality-100 dataset, we considered only one set of responses of each individual (indicated by the variable "best protocol") in case the same person took the test several times.
Non-participation	n/a
Randomization	n/a

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Data was collected in other studies; details about recruitment are available in the original research papers (see above).