nature human behaviour

Registered Report

https://doi.org/10.1038/s41562-024-02059-4

How trait impressions of faces shape subsequent mental state inferences

Received: 11 August 2020

Accepted: 10 October 2024

Published online: 02 December 2024

Check for updates

Chujun Lin [1] ✉, Umit Keles [2], Mark A. Thornton [3,4] & Ralph Adolphs [2,4]

People form impressions of one another in a split second from faces. However, people also infer others' momentary mental states on the basis of context—for example, one might infer that somebody feels encouraged from the fact that they are receiving constructive feedback. How do trait judgements of faces influence these context-based mental state inferences? In this Registered Report, we asked participants to infer the mental states of unfamiliar people, identified by their neutral faces, under specific contexts. To increase generalizability, we representatively sampled all stimuli from inclusive sets using computational methods. We tested four hypotheses: that trait impressions of faces (1) are correlated with subsequent mental state inferences in a range of contexts, (2) alter the dimensional space that underlies mental state inferences, (3) are associated with specific mental state dimensions in this space and (4) causally influence mental state inferences. We found evidence in support of all hypotheses.

Humans spontaneously form impressions of other people's general characteristics upon seeing their faces[1,2]. For instance, we judge whether people are trustworthy, intelligent or feminine on the basis of how their faces look[3–5]. These trait judgements show high consensus across perceivers in different age groups and from different cultures[6–8]. The accuracy of trait judgements from faces is debated: although some suggest that personality can be accurately judged from static face images[9–11], many argue that trait judgements from faces merely reflect perceivers' biases and stereotypes[12–14]. However valid or invalid they may be, these judgements shape consequential decisions in the real world[15–23]. Such influences are most prominent in situations where understanding a person's general traits plays an important role, such as evaluating which candidate might be a good political leader[15] or which individual on a dating site might be the best long-term partner[21].

Understanding other people's enduring dispositions is only one contributing factor that guides social judgements and decisions. More often, to successfully navigate the complex social world, it is critical also to understand a person's context-dependent mental states in the moment[24]: what is the other person currently thinking about, feeling or intending? For instance, our ability to tell whether a friend is joking or being serious would make all the difference in selecting appropriate behaviour towards them in that particular situation. As with trait inferences from faces, people also make inferences about others' mental states rapidly and automatically[25–30]. This ability develops early on, with evidence suggesting that infants are able to infer goals and intentions from six months of age[31,32]. Inferences of momentary mental states are based on a range of cues, such as facial expressions, body postures and gestures, together with situational information[30,33–36].

Little is known about how judgements of relatively stable traits from faces might bias or influence judgements of more transient mental states. Studies on the recognition of facial expressions show that people perceive faces that are digitally manipulated to look untrustworthy as displaying more negative emotions such as anger[37]. This finding suggests that trait judgements from faces may shape emotion judgements of isolated faces. However, whether this effect generalizes to judging a broader set of mental states (beyond basic emotions) in more realistic settings (for example, with situational information) is unclear. Studies investigating the relation between a wider range of traits and mental states in more naturalistic settings show that trait knowledge does shape mental state inferences[38,39]. However, those studies focus on more reliable trait knowledge, such as trait inferences of participants' friends and family members, and famous people about whom participants already have substantial biographical and contextual information. It remains unknown whether people rely on trait

[1]Department of Psychology, University of California, San Diego, San Diego, CA, USA. [2]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [4]These authors jointly supervised this work: Mark A. Thornton, Ralph Adolphs. ✉e-mail: chujunlin@ucsd.edu

information to the same extent when it is drawn from solely superficial facial judgements.

In the present investigation, we ask: are the diverse mental states that people attribute to different individuals in specific situations biased by the trait judgements of those individuals' neutral faces? Answering this question would advance our understanding of how people make sense of others' momentary and enduring features, two types of information critical for social navigation[40]. If the answer is yes, then the biases and stereotypes in trait impressions from faces[41,42] would probably be carried over to shape inferences of various mental states in a wide range of situations. This may help explain, for instance, why Black males whose faces are stereotypically perceived to be aggressive[43,44] are wrongly attributed the mental state of intending to harm more often in various situations even without any evidence[45]. If trait impressions of faces do influence mental state inferences, then this would also suggest that the impacts of spontaneous trait judgements of faces are much broader than previously thought[46]. They not only would influence decisions in situations where temporally stable trait information is patently important but also could influence moment-to-moment decisions we make in the course of social interactions. Such broader influence could also help explain why trait judgements of faces are sometimes accurate[9–11]. For instance, individuals whose faces look like they are introverted may be attributed the mental state of being unwilling to interact with other people, leading others to reduce interaction with these individuals and in turn exacerbating the social isolation of these individuals.

Our present research tested four main hypotheses. First, when a photograph of a person's neutral face is available, we asked whether inferences of this individual's mental states in specific situations are associated with the trait impressions that are based on the neutral face (Table 1, Q1). As mentioned above, prior research shows that emotion recognition from faces (for example, anger, fear or happiness) is associated with trait judgements of those faces (for example, sociable, dominant or trustworthy)[47,48]. However, in real life, we do not judge others' mental states from their faces in isolation, as participants in most prior studies have done. Instead, we make sense of people's mental states in specific situations. To test this hypothesis, we asked participants to infer how much different specific people, whose neutral faces the participants saw, would feel a certain mental state in a given situation (scenario-state task; Methods). We linked these mental state inferences to the first impressions formed from those specific people's neutral faces on a range of traits (face-trait task; Methods).

To increase the comprehensiveness and generalizability of our study, we representatively sampled mental state terms using deep neural networks and the maximum variation procedure from a comprehensive list of putative mental states (Fig. 1a–d). We verified that our final selected set of 60 mental states were representative of the terms laypeople use in everyday life to describe others' internal, non-pathological, specific mental states (Fig. 2a–c). For each mental state, we selected a situation that people thought co-occurred with the mental state in real life (Methods). We representatively sampled trait terms (Fig. 1e–h) that people spontaneously use to describe faces (Fig. 2d–f), and faces that were diverse with respect to gender, race and age (Fig. 1i–l) and that populate the facial geometry that people see in everyday life (Fig. 2g–i).

Second, we hypothesized that the underlying psychological dimensions people use to represent others' mental states differ when face images are available compared with when they are not (Table 1, Q2). Prior research has shown that people use three dimensions to represent mental states (the 3-D Mind Model dimensions: valence, rationality and social impact)[49,50]. Those studies did not include face information when participants made mental state inferences. However, in real life, we can usually see the individuals to whom we attribute mental states. It remains an open question how adding the information from faces might modify the psychological dimensions that characterize mental state inferences. By analogy, we need three dimensions to represent the location of an object in space. When more information about the object is available, we may be able to use fewer, the same, or more and/or different dimensions to locate the object. For instance, if the new information identifies the object's distance to Earth, then we need only two dimensions; if the new information identifies a moving object, then we need a fourth dimension (time).

Third, we asked whether (and to what extent) the mental state dimensions are associated with trait impressions from faces (Table 1, Q3). For example, recent work has shown that how frequently people judged the targets to experience positive or negative mental states in various situations was most closely related to the trait dimension of warmth[38]. However, that study investigated trait judgements based on more information than just faces. It remains an open question how trait judgements merely from briefly seeing an unfamiliar face might be associated with mental state dimensions. Understanding the association between trait impressions and core mental state dimensions (beyond individual mental state judgements) would allow for further generalizability of our findings. Specifically, it would allow us to predict how inferences of any mental state—beyond the 60 mental states for which we collected data—would be associated with trait impressions of faces.

Finally, we asked whether the associations between trait impressions from faces and mental state inferences might be causal (Table 1, Q4). Prior research has shown that changing the trait impressions of emotional faces shifted people's perception of emotions[37]. Furthermore, changing the trait impressions of target people (beyond faces) has a causal effect on people's mental state inferences of those targets across a range of situations[38]. These results suggest a causal link from trait to mental state inferences. However, it remains unclear whether changing the trait impressions formed merely on the basis of faces would be sufficient to cause people to change their mental state attributions. We tested this hypothesis by digitally manipulating faces so that the same individual generated multiple stimulus images that exhibited different traits. We then measured how participants attributed different mental states to the same individual in the same context as a function of the experimental manipulation of that individual's face. Understanding the causal effects (Q4) beyond correlations (Q1–Q3) is essential to determining the nature of the relation between trait impressions from faces and mental state inferences. Findings of correlations without causation would suggest that the observed correlations were driven by third variables (for example, inferred social roles that shape both inferences of traits and mental states).

Protocol registration

The Stage 1 protocol for this Registered Report was accepted in principle on 30 March 2022. The protocol, as accepted by the journal, can be found at https://doi.org/10.6084/m9.figshare.19664316.v1.

Results

As mentioned in Methods ('Deviations from protocol'), one deviation from the approved registered Stage 1 protocol occurred in the participant recruitment sources for the cross-world-region data. In the approved registered Stage 1 protocol, we had planned to recruit participants in all five world regions (the USA, Africa, Asia, Europe and South America) using the MTurk Toolkit via CloudResearch. We successfully collected all data from the USA (n = 5,260), Asia (n = 961) and Europe (n = 1,153) as planned. Due to the limited number of participants in Africa and South America on MTurk Toolkit, we collected the data in these two regions using an additional recruitment option offered on CloudResearch, Prime Panels, together obtaining the planned amount of data in Africa (n = 1,040) and South America (n = 1,171). We obtained permission from the Editors before carrying out the above data collection. Except for the deviation mentioned above, we adhered precisely to the approved registered experimental procedures, data analysis procedures and result interpretation procedures as detailed in Table 1 and Methods.

https://doi.org/10.1038/s41562-024-02059-4

Table 1 | Design table

Question	Hypothesis	Sampling plan	Analysis plan	Interpretation given to different outcomes
Q1. Are trait impressions from neutral faces associated with mental state inferences about those same people in given scenarios?	H1a. We predicted that mental state inferences of unfamiliar others in given scenarios (scenario-state) would be associated with the trait impressions formed from those individuals' neutral faces when shown in isolation (face-trait). H1b. We predicted that H1a would hold even when we controlled for the mental states that the neutral faces displayed (face-state).	Please refer to 'Sampling plan' in Methods for the details. H1a. We determined the sample size for the scenario-state task to be n=50 participants per mental state. This sample size was estimated empirically on the basis of our pilot data via the jackknife resampling procedure. We determined the sample size for the face-trait task to be n=38 participants per trait. This sample size was estimated empirically on the basis of sequential resampling by prior research[84]. H1b. We determined the sample size for the face-state task to be n=31 participants per mental state. This sample size was estimated empirically on the basis of sequential resampling by prior research[84].	Please refer to 'Analysis plan' in Methods for the details. H1a. We tested this hypothesis using ridge regression with cross-validations for each of the 60 mental states. Each model regressed the average ratings given to the faces for a mental state under the specific scenario (scenario-state) on the ratings given to the faces for 13 traits (face-traits). Multiple comparisons across the 60 mental states were corrected for via maximal statistic permutation tests. H1b. We tested this hypothesis using variance partition analyses. We assessed whether the unique variance in the scenario-state ratings explained by face-traits, when controlling for face-states, is significantly greater than zero across cross-validation resampling.	H1a. If >80% of the scenario-states are significantly predicted by face-traits, the evidence for H1a is very broad; if 60–80% are predicted, the evidence is broad; if 40–60% are predicted, the evidence is moderately broad; if 20–40% are predicted, the evidence is narrow; and if <20% (but at least one mental state) are predicted, the evidence is very narrow. Otherwise, there is no evidence for H1a. H1b. If the unique variance explained by face-traits is greater than zero in >80% of the scenario-states, the evidence for H1b is very broad; if greater than zero in 60–80%, the evidence is broad; if greater than zero in 40–60%, the evidence is moderately broad; if greater than zero in 20–40%, the evidence is narrow; and if greater than zero in <20% (but at least one mental state), the evidence is very narrow. Otherwise, there is no evidence for H1b.
Q2. What are the dimensions that underlie mental state inferences of others when we also see their faces?	H2a. We predicted that mental state inferences (scenario-states) could be represented by a small number of dimensions (<10) even when faces were available. H2b. We predicted that the mental state dimensions in H2a would at least partially overlap with previously found mental state dimensions when no face was available (the 3-D Mind dimensions: rationality, social impact and valence)[49].	We tested H2a and H2b using the same set of data collected for testing H1a.	Please refer to 'Analysis plan' in Methods for the details. H2a. Since no single method is regarded as the best method for determining the optimal number of dimensions, we applied five distinct methods: Horn's parallel analysis, the optimal coordinate index, the empirical Bayesian information criterion, Velicer's minimum average partial test and bi-cross-validation. The optimal number of dimensions was the number that most methods agreed on; or, if all methods disagreed, it was the minimum number that generated the most interpretable dimensions that accounted for ≥75% variance in the data in exploratory factor analysis. H2b. We measured the Spearman correlation between the dimensions in our data and the 3-D Mind dimensions[49] using two different methods: one based on factor loadings and scores, and the other based on participants' ratings of meaning similarity. We deemed an absolute correlation of 0.2–0.39, 0.4–0.59 or ≥0.6 to be an indication of weak, moderate or strong similarity.	H2a. If the optimal number of dimensions was <10, we concluded that mental state inferences are represented by a small number of dimensions even when faces are available. Otherwise, we concluded that mental state inferences are no longer represented by a small number of dimensions when faces are available. H2b. If any dimension in our data was at least moderately correlated (r≥0.4) with any 3-D Mind dimension on the basis of both methods, we concluded that mental state dimensions when faces are available partly overlap with previously found mental state dimensions when no face was available. Otherwise, we concluded that there is no strong evidence for H2b.
Q3. Are the dimensions that underlie mental state inferences of others when faces are available associated with trait impressions formed from those faces?	H3a. We predicted that the dimensions of mental state inferences when faces were available (scenario-state dimensions) would be associated with trait impressions formed from those faces (face-trait). H3b. We predicted that H3a would hold even when we controlled for the mental states that those neutral faces displayed (face-state).	We tested H3a and H3b using the same set of data collected for testing H1a and H1b.	Please refer to 'Analysis plan' in Methods for the details. H3a. We tested this hypothesis using ridge regression with cross-validations as in H1a. The only difference is that each model here corresponded to each core mental state dimension in Q2. Each model regressed the factor scores for a mental state dimension across the faces (factor scores of scenario-states) on the ratings given to the faces for 13 traits (face-traits). H3b. We tested this hypothesis using variance partition analyses as in H1b. The only difference is that the dependent variable here is the factor scores for a mental state dimension across the faces.	H3a. If any mental state dimension was significantly predicted by face-traits, we concluded that the dimension(s) of mental state inferences when faces are available is (are) associated with trait impressions from those faces. Otherwise, we concluded that there is no evidence for H3a. H3b. If for any mental state dimension, the unique explained variance of face-traits was significantly greater than zero, we concluded that there is evidence for H3b. Otherwise, we concluded that there is no evidence for H3b.

Table 1 (continued) | Design table

Question	Hypothesis	Sampling plan	Analysis plan	Interpretation given to different outcomes
Q4. Are mental state inferences of others in a given scenario causally influenced by the trait impressions formed from those individuals' neutral faces?	H4. We predicted that trait impressions formed from neutral faces (face-trait) causally shape mental state inferences of those people in specific scenarios (scenario-state).	Please refer to 'Sampling plan' in Methods for the details. H4. We used a set of n=272 face images to detect the causal effect. This sample size was determined via formal power analysis, with a paired one-sided t-test. See 'Stimuli: Trait-manipulation of face stimuli' in Methods. We checked each trait manipulation. Each subset of facial identities with their manipulated images were rated by n=38 participants per manipulated trait, using a similar procedure as in the face-trait task in H1a. We tested causality using the trait-manipulated faces via a similar procedure as in the scenario-state task. Each subset of facial identities with their manipulated images were rated by n=50 participants per mental state as in H1a.	Please refer to 'Analysis plan' in Methods for the details. H4. We tested causality for the mental states that were strongly correlated with trait impressions in H1 (for example, the top predicted state(s) from each dimension; targeting around six states). For each mental state, we identified a different, strongly correlated trait. For each trait, we digitally manipulated each face to enhance and reduce that trait, generating two versions of face images. We tested the causal effect for each state-trait pair using two methods: one based on aggregate scenario-state ratings, using paired one-sided t-tests between the two versions of faces; and another based on individual scenario-state ratings, using linear mixed modelling to regress the ratings on the face versions while controlling for the random effects of participants and face identities.	H4. For each state–trait pair, if we found a significant effect in the expected direction (as discovered in H1) using both methods, we concluded that there is strong causal evidence for that state–trait pair. If only one method indicated a significant effect, the evidence is weak. If neither method indicated a significant effect, there is no causal evidence for that state–trait pair. Across all tested state–trait pairs, we concluded that the evidence for H4 is very broad if >80% pairs showed strong evidence; broad if 60–80% pairs showed strong evidence; moderately broad if 40–60% pairs showed strong evidence; narrow if 20–40% pairs showed strong evidence; and very narrow if <20% but at least one pair showed strong evidence. Otherwise, there is no strong evidence for H4.

Associations between mental state and trait inferences

We found very broad evidence that mental state inferences of unfamiliar others in given scenarios were associated with trait impressions formed from those individuals' faces (Table 1, H1a). Inferences of every one of the 60 mental states were significantly predicted by inferences of the 13 traits: ridge regression analysis with cross-validations (for increasing generalizability; Methods) showed that the prediction accuracy for the 60 mental state inferences ranged from $r = 0.64$ to $r = 0.97$, with mean $r = 0.88$ (multiple comparisons across 60 mental states were corrected for using maximal statistic permutations; corrected P values ranged from 0.0005 to 0.047; see Supplementary Table 2 for the results with detailed statistics of all 60 mental states). These results suggest that given the same context, how people infer different individuals' mental states is associated with the trait impressions inferred from those individuals' faces.

The strong associations between context-specific mental state inferences and trait impressions from faces remained robust even when controlling for the context-irrelevant affective and cognitive states inferred from the faces (captured when the photos were taken; Table 1, H1b). Using variance partition analysis (three models were fitted per mental state with different predictors; Methods), we found that trait impressions contributed a significant amount of unique explained variance for 47 of the 60 mental state inferences (Fig. 3). Across those 47 mental states, face-trait impressions on average contributed $r^2 = 0.39$ explained variance (the lower bound across 2,000 cross-validation iterations (that is, the 2.5th percentile) ranged from 0.005 to 0.597 across mental states; see Fig. 3 for detailed statistics) beyond the variance that was commonly explained by both face-trait and face-state impressions (see Supplementary Table 2 for the results of all 60 mental states). These results suggest that given the same context, how people infer different individuals' mental states is influenced by first impressions from faces that are specifically about the individual's stable characteristics beyond momentary states such as emotions (see Supplementary Fig. 3 for how each mental state was differently influenced by different trait impressions; we validated this interpretation of the ridge regression coefficients with three additional analysis methods such as ordinary least squares regressions and LASSO regressions; Supplementary Methods and Supplementary Fig. 4).

Trait inferences of faces do not merely influence an isolated mental state judgement when that face is seen in a particular context. Trait inferences also change the psychological space that characterizes the relationships among mental state judgements (Table 1, H2a). We investigated the dimensions underlying these relationships by analysing which mental state inferences were highly correlated with one another across different target faces. As preregistered, we first determined how many dimensions optimally summarized the common variance in the judgements of the 60 mental states. Three of the five planned methods indicated that four dimensions optimally represented mental state judgements (agreed by Horn's parallel analysis, the optimal coordinate index and the empirical Bayesian information criterion; 2 of the 60 mental states were excluded for low factorability: 'bored' and 'indecisive').

The interpretation of these four mental state dimensions is most clearly obtained by examining which types of targets were more (and less) often attributed the mental states associated with each dimension (Figs. 4 and 5). We interpreted the four dimensions as describing sentimental mental states (mental states that are stereotypically associated with exaggerated or self-indulgent emotions), youthful mental states (mental states that are stereotypically associated with youthful people), empathetic mental states (mental states that are stereotypically associated with people who understand others' feelings) and competent mental states (mental states that are stereotypically associated with competent people). These four mental state dimensions together explained 83% of the common variance in the context-specific mental state inferences (each explaining 39%, 19%, 15% and 10%). Since our dimension analysis method (exploratory factor analysis) does not force the dimensions to be orthogonal, it reveals the natural relationships between the dimensions. The sentimental mental state dimension and the empathetic mental state dimension were moderately correlated ($r = 0.47$; $t_{98} = 5.32$; $P = 6.57 \times 10^{-7}$; 95% confidence interval (CI), (0.31, 0.61)); correlations between the other mental state dimensions were weak (all $r \leq 0.25$).

We compared the four mental state dimensions uncovered when targets' faces were available with the three mental state dimensions (valence, rationality and social impact) from prior theory[50] (Table 1,

https://doi.org/10.1038/s41562-024-02059-4

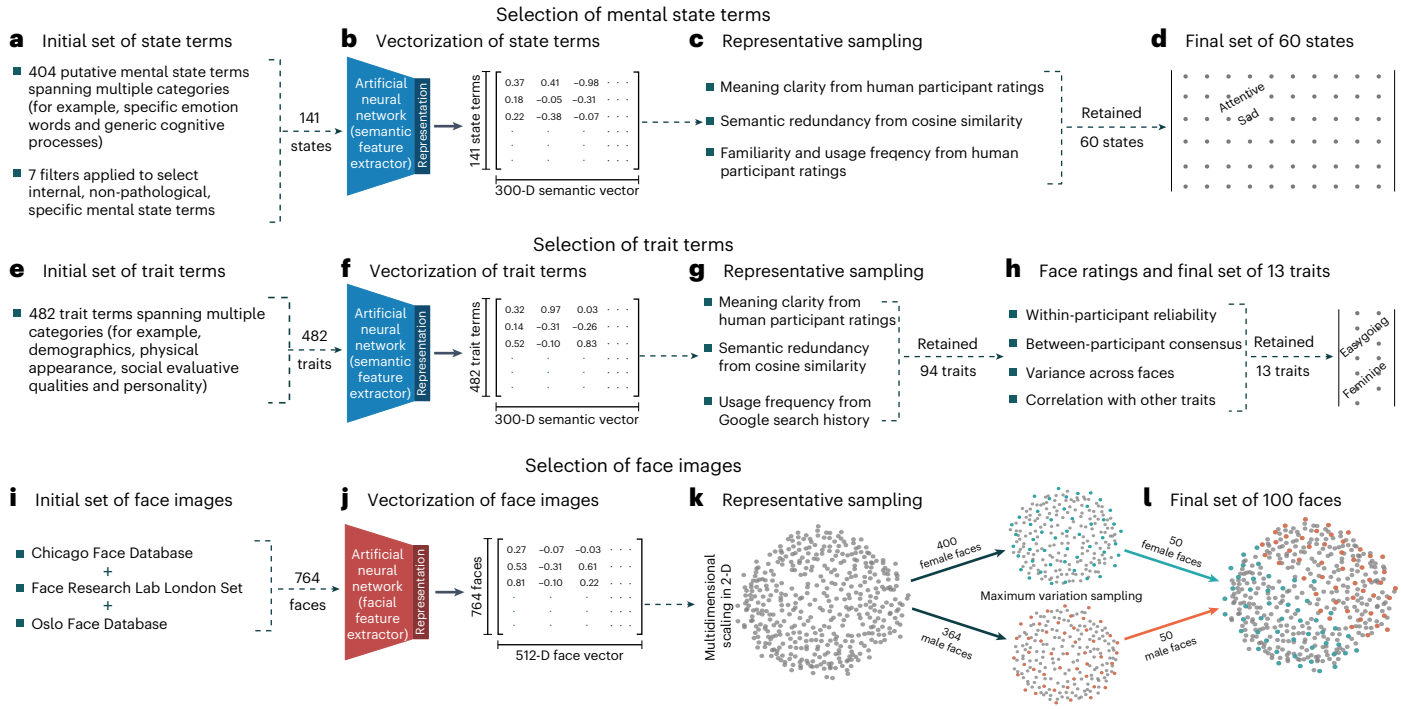

Fig. 1 | Sampling mental states, traits and faces to generate comprehensive sets. a, An extensive list of putative mental state terms was gathered from multiple sources in the literature[50,72–75], to which seven filters were applied to retain internal, non-pathological, specific mental state terms. **b**, Each term was quantified with a vector of 300 semantic features using a natural language model[76]. **c**, Four filters were applied to systematically sample the state terms: (1) terms with unclear meaning were excluded, and (2) for every pair of terms with similar meaning, the one with (3) a lower familiarity or, if the same, (4) a lower usage frequency was excluded from the pair. **d**, The final set of 60 mental states. **e**, An extensive list of trait words was gathered from multiple sources in the literature[4,14,17,19,46,48,85–94]. **f**, Each term was quantified as in **b** with a vector of 300 semantic features. **g**, Three filters were applied to systematically sample the trait terms: (1) terms with unclear meaning were excluded, and (2) for terms with similar meaning, the one with (3) a higher usage frequency was retained. **h**, Four filters were applied to select a smaller subset of traits that have (1) a within-participant test–retest reliability above the mean across all traits, (2) a between-participant consensus above the mean across all traits, (3) a variance across faces above the mean across all traits and (4) a correlation with any other trait below 0.9. **i**, Clear, frontal, relatively neutral faces of all available races (58% white, 26% Black, 16% Asian) were gathered from three popular face databases[70,71,80]; faces with problematic race labels were excluded. **j**, Each face was represented with a vector of 512 facial features using a state-of-the-art neural network[81] that had been pretrained to identify individual identities across millions of faces of all different aspects and races. **k**, Maximum variation stratified sampling (regarding gender (50% male, 50% female) and race (58% white, 26% Black, 16% Asian)) was applied to select faces with maximum variability along the 512 facial features. **l**, The final set of 100 face images (green and orange dots).

H2b). The youthful (reversed) mental state dimension found here was strongly similar to the previously found valence dimension (Spearman $\rho = -0.61$; $P = 1.23 \times 10^{-4}$; 95% CI, (−0.76, −0.38), on the basis of factor scores and loadings; $\rho = -0.67$; $P = 7.53 \times 10^{-9}$; 95% CI, (−0.78, −0.52), on the basis of participant ratings). This result indicates that mental states that were stereotypically associated with youthful people were high in valence. There was no strong evidence for similarity between the other mental state dimensions found here and the other two mental state dimensions previously proposed (rationality and social impact; see Supplementary Table 3 for similarities between all dimensions).

These dimensional analyses together suggest that, compared with prior research where no information on the target individuals was available, with more information that is typically present in real-world interactions (for example, faces) a greater number of dimensions (four instead of three) is needed to represent others' mental states. Importantly, it is not simply that a fourth dimension is added but that the dimensions change on the basis of information: only one of the dimensions found here overlapped with one previously found dimension. These findings suggest that the dimensions people use to represent others' mental states are exceedingly flexible and are strongly influenced by trait information available from faces.

The above associations between mental state dimensions and trait impressions from faces remained robust even when we controlled for the context-irrelevant affective and cognitive states inferred from the targets' faces (Table 1, H3). Trait impressions from faces explained a unique amount of variance beyond that commonly explained together with face-state inferences in every one of the four mental state dimensions (Supplementary Fig. 6). These results, combined with our comprehensive sampling of traits, faces, mental states and scenarios, suggest that the associations between context-specific mental state inferences and judgements of traits from faces are generalizable to a wide range of mental states and situations beyond those measured here.

Causal effects of trait impressions on state inferences

We confirmed that the associations between context-specific mental state inferences and trait impressions from faces are indeed causal; furthermore, we tested generalizability across participants from different regions of the world (Table 1, H4). As preregistered, we tested causation using a subset of the representatively sampled mental states—one from each dimension—and data from participants across five different world regions. The four targeted state–trait pairs were state embarrassed and trait stereotypically strong, state threatened and trait stereotypically white, state jealous and trait stereotypically feminine, and state lonely and trait stereotypically leader-like (Supplementary Methods). We first confirmed that our trait manipulations of faces were successful using data from US participants: the trait-increased versions of the faces received greater trait ratings from participants than the trait-decreased versions ($\Delta = 0.83$ on a seven-point Likert scale; Cohen's $d = 2.59$; $t_{271} = 42.68$; $P = 1.37 \times 10^{-122}$; effect size 95%

https://doi.org/10.1038/s41562-024-02059-4

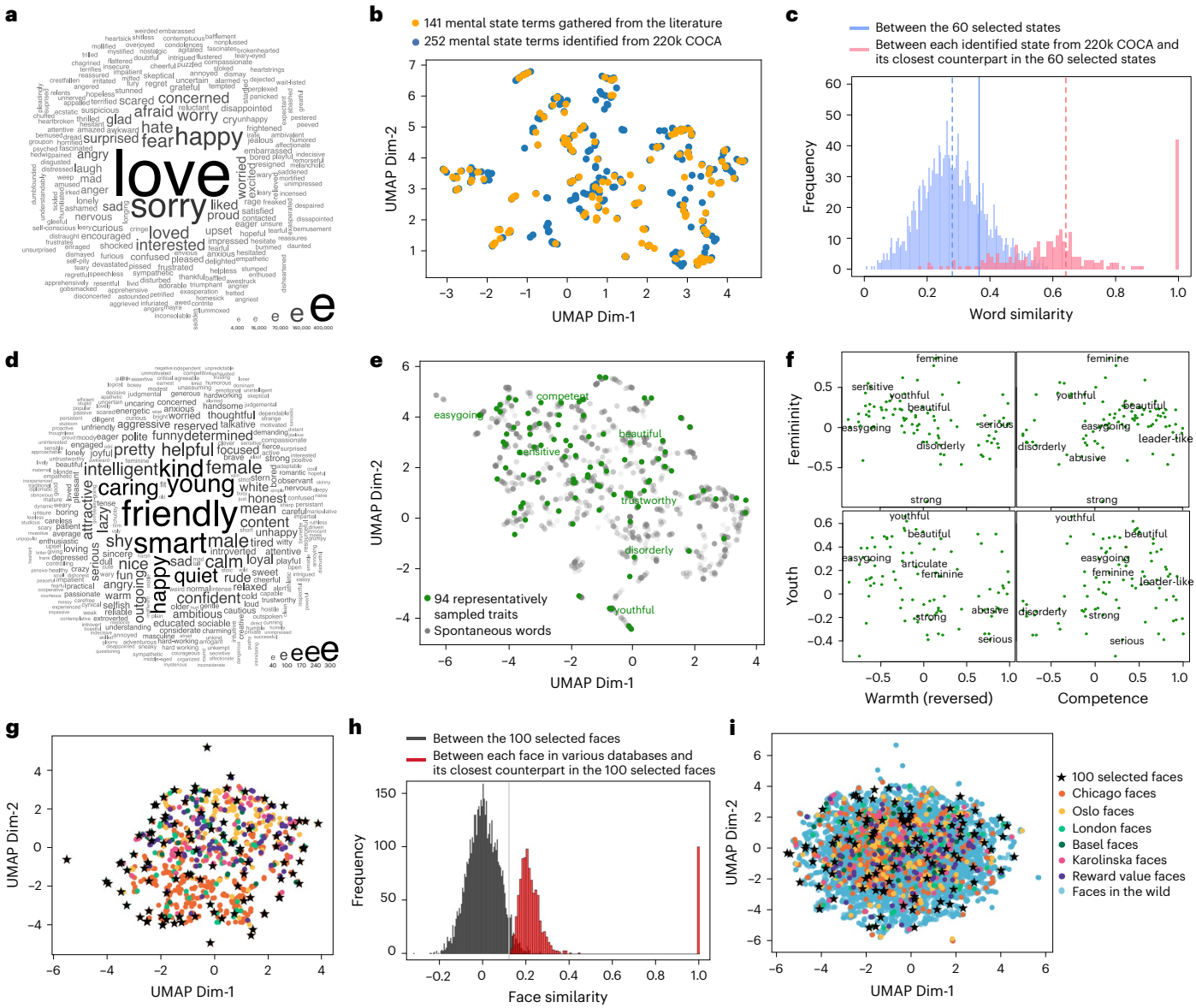

Fig. 2 | Representativeness of the selected mental states, traits and face images. a, Word cloud of internal, non-pathological, specific mental state terms identified by a linear classifier from the most frequent 220,000 words of the one-billion-word Corpus of Contemporary American English (220k COCA) (see 'Design' in Methods). Font size indicates frequency (≥100 are shown). **b**, Uniform Manifold Approximation and Projection (UMAP[78]) of the 141 mental states gathered from the literature (orange) compared with those predicted by a linear classifier from 220k COCA (blue) on the basis of 300 semantic features[76]. **c**, Distributions of word similarities (cosine distance) based on 300 semantic features (the dashed lines indicate the means). All mental states predicted from 220k COCA had a similar counterpart in our 60 selected mental states: their similarity (red) was greater than the 80th percentile of the similarities among the 60 selected states (solid blue line), except for 14 mispredicted terms (indicated by the red bars to the left of the solid blue line). **d**, Word cloud of spontaneous face descriptions. Font size indicates frequency (≥10 are shown). **e**, UMAP of the 94 representatively sampled traits (green; examples are labelled) compared with spontaneous face descriptions (grey) based on 300 semantic features. **f**, Distributions of the 94 traits (green dots) and our selected 13 traits (black text; eight examples are labelled in each panel) per dimension pair. **g**, UMAP of our final selected 100 faces (stars) compared with a broader set of frontal, neutral faces from various databases[95–97] (labels are shown to the right of **i**) (dots, $n = 1{,}009$ faces from individuals of different genders, races and ages) based on 512 facial features. **h**, Distributions of face similarities (cosine distance) based on 512 facial features. All faces from the broader set of databases in **g** had a similar counterpart in our 100 selected faces: their similarity (red) was greater than the 95th percentile of the similarities among the 100 selected faces (solid grey line). **i**, UMAP of our 100 faces (stars) compared with faces from various databases including faces in naturalistic contexts (light blue dots, $n = 12{,}460$) based on 512 facial features.

CI, (2.47, 2.71) for stereotypically strong; $\Delta = 0.13$; $d = 0.49$; $t_{271} = 8.03$; $P = 1.51 \times 10^{-14}$; 95% CI, (0.37, 0.61) for stereotypically white; $\Delta = 0.27$; $d = 1.10$; $t_{271} = 18.13$; $P = 5.88 \times 10^{-49}$; 95% CI, (0.98, 1.22) for stereotypically feminine; $\Delta = 0.06$; $d = 0.27$; $t_{271} = 4.43$; $P = 6.87 \times 10^{-6}$; 95% CI, (0.15, 0.39) for stereotypically leader-like).

Modifying the perceived traits of the faces influenced the judgements of mental states made of the same individual given the same scenario (Fig. 6): making the same individual look stereotypically stronger caused participants to judge that this person would feel less embarrassed when noticing that their shirt is on inside out ($\Delta = -0.40$ on a seven-point Likert scale; $d = -1.73$; $t_{271} = -28.53$; $P = 6.42 \times 10^{-84}$; 95% CI, (−1.85, −1.61) based on aggregate data; $\beta = -0.47$; $t_{6245} = -15.20$; s.e. = 0.03; $P = 3.00 \times 10^{-51}$; 95% CI, (−0.53, −0.41) based on individual-level data in linear mixed model); making the same individual look more stereotypically white caused participants to judge that this person would feel less threatened when reading news about stigma

https://doi.org/10.1038/s41562-024-02059-4

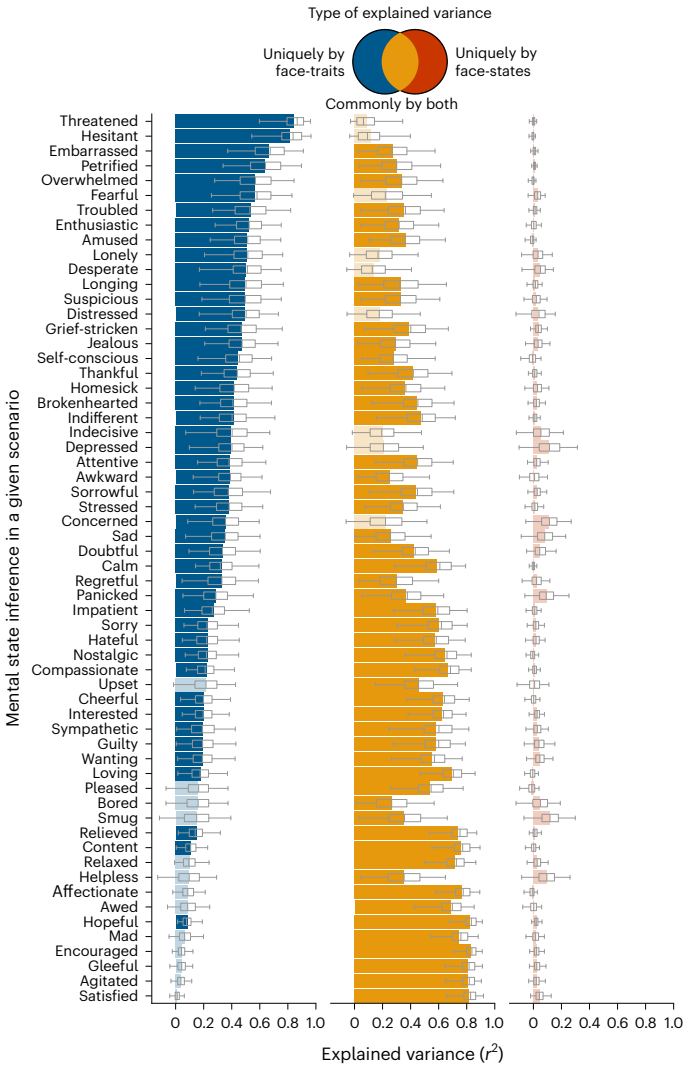

Fig. 3 | Variance partition results of mental state inferences in given scenarios. The *y* axis indicates different dependent variables, one for each mental state. For each mental state, three models were fitted for variance partition: one with face-traits as predictors, the second with face-states as predictors and the third with both as predictors. All models were fitted to the ratings across the $n = 100$ faces. Different types of explained variance for each mental state, indicated along the *x* axes, were computed on the basis of the combination and subtraction of the explained variances returned by the three models. The bar length indicates mean explained variance averaged across 2,000 cross-validation iterations. The error bars indicate the 95% CI (that is, 2.5th and 97.5th percentiles) across these iterations. The boxes span from the 25th to the 75th percentiles, and the mid-lines in the boxes indicate the median values across these iterations. The bar colours indicate different types of explained variance (blue indicates uniquely explained by 13 trait ratings inferred from faces, red indicates uniquely explained by 8 state ratings inferred from faces and orange indicates commonly explained by the 13 trait ratings as well as the 8 state ratings inferred from faces); desaturated colours indicate non-significant results (that is, 2.5th percentile below zero).

against their group ($\Delta = -0.13$; $d = -0.71$; $t_{271} = -11.75$; $P = 2.51 \times 10^{-26}$; 95% CI, (−0.83, −0.59); $\beta = -0.14$; $t_{6114} = -4.60$; s.e. = 0.03; $P = 4.26 \times 10^{-6}$; 95% CI, (−0.21, −0.08)); making the same individual look more stereotypically feminine caused participants to judge that this person would feel more jealous when hearing that their best friend admires another new friend ($\Delta = 0.13$; $d = 0.66$; $t_{271} = 10.82$; $P = 3.32 \times 10^{-23}$; 95% CI, (0.54, 0.78); $\beta = 0.14$; $t_{5572} = 4.03$; s.e. = 0.03; $P = 5.61 \times 10^{-5}$; 95% CI, (0.07, 0.20)); and making the same individual look more stereotypically leader-like caused participants to judge that this person would feel less lonely when being different from everyone else in a group ($\Delta = -0.10$; $d = -0.42$; $t_{271} = -6.96$; $P = 1.26 \times 10^{-11}$; 95% CI, (−0.54, −0.30); $\beta = -0.08$; $t_{6240} = -2.47$; s.e. = 0.03; $P = 0.014$; 95% CI, (−0.14, −0.02)). These results show that, given the same context and the same face identity, how people infer mental states is shaped by their judgements of the individual's traits based on the face—a specific causal effect of traits on mental states that is generalizable to a wide range of mental state inferences and situations.

We replicated the above findings in four other world regions beyond the USA: Africa, Asia, Europe and South America (Fig. 6). As preregistered, we first tested whether the associations between the above four target mental-state–trait pairs remained robust in these different world regions. All mental-state–trait pairs were significantly correlated except for feeling jealous and looking stereotypically feminine in the African sample ($r = 0.013$; $t_{98} = 0.13$; $P = 0.899$; 95% CI, (−0.18, 0.21)) and feeling lonely and looking stereotypically leader-like in the Asian sample ($r = 0.12$; $t_{98} = 1.20$; $P = 0.232$; 95% CI, (−0.08, 0.31)); these causal relations were thus not tested (we preregistered to test causality only if the association was significant). This cross-regional variability in the correlations between mental state inferences and trait impressions from faces suggests that how people understand the conceptual relations between mental states and traits may vary across cultures.

Similar to the US sample, making a face look stereotypically stronger reduced the inference that the individual would feel embarrassed in the given situation by all participant samples (Supplementary Table 4). Making a face look more stereotypically white decreased the inference that the individual would feel threatened in the given situation by all participant samples. Making a face look more stereotypically feminine increased the inference that the individual would feel jealous in the given situation by all participant samples. Making a face look more stereotypically leader-like decreased the inference that the individual would feel lonely in the given situation by most samples except for South America. These results suggest that the causal effects of trait impressions from faces on context-specific mental state inferences are generalizable across a wide range of populations.

Discussion

In the present large-scale, cross-regional investigation (total $n = 9{,}585$), we showed that mental state inferences are sensitive not only to the context but also to judgements of the individual's traits inferred from their face. By systematically sampling stimuli that were more diverse (Figs. 1 and 2), we showed that a wide range of mental state inferences in various situations are associated with trait impressions from faces (Fig. 3 and Table 1, Q1). Moreover, by selectively manipulating the perceived traits of the faces, we demonstrated that the above associations are indeed causal (Fig. 6 and Table 1, Q4): how people attribute mental states to others in given contexts is influenced by trait impressions from faces. Importantly, we showed that this causal effect is generalizable across perceivers in five different world regions (Fig. 6). Together, these findings demonstrate that how people attribute momentary thoughts, intentions and feelings to other people is biased by their facial appearance, and in particular by the durable traits people infer from the face alone.

Understanding others' moment-to-moment thoughts and feelings is key to navigating the social world[51,52]. Prior work has revealed how people use momentary cues such as facial expressions[53,54], behaviour[55–57] and contexts[58–60] to infer each other's mental states. Some recent studies have looked beyond momentary cues and investigated how knowledge or inferences of others' more stable characteristics (for example, whether someone is a good or bad social partner in general) shape mental state inferences, and vice versa[38,39,61]. However, the trait knowledge and inferences used in prior work were predominantly based on more comprehensive information (for example, participants' knowledge of their families and friends' traits or

https://doi.org/10.1038/s41562-024-02059-4

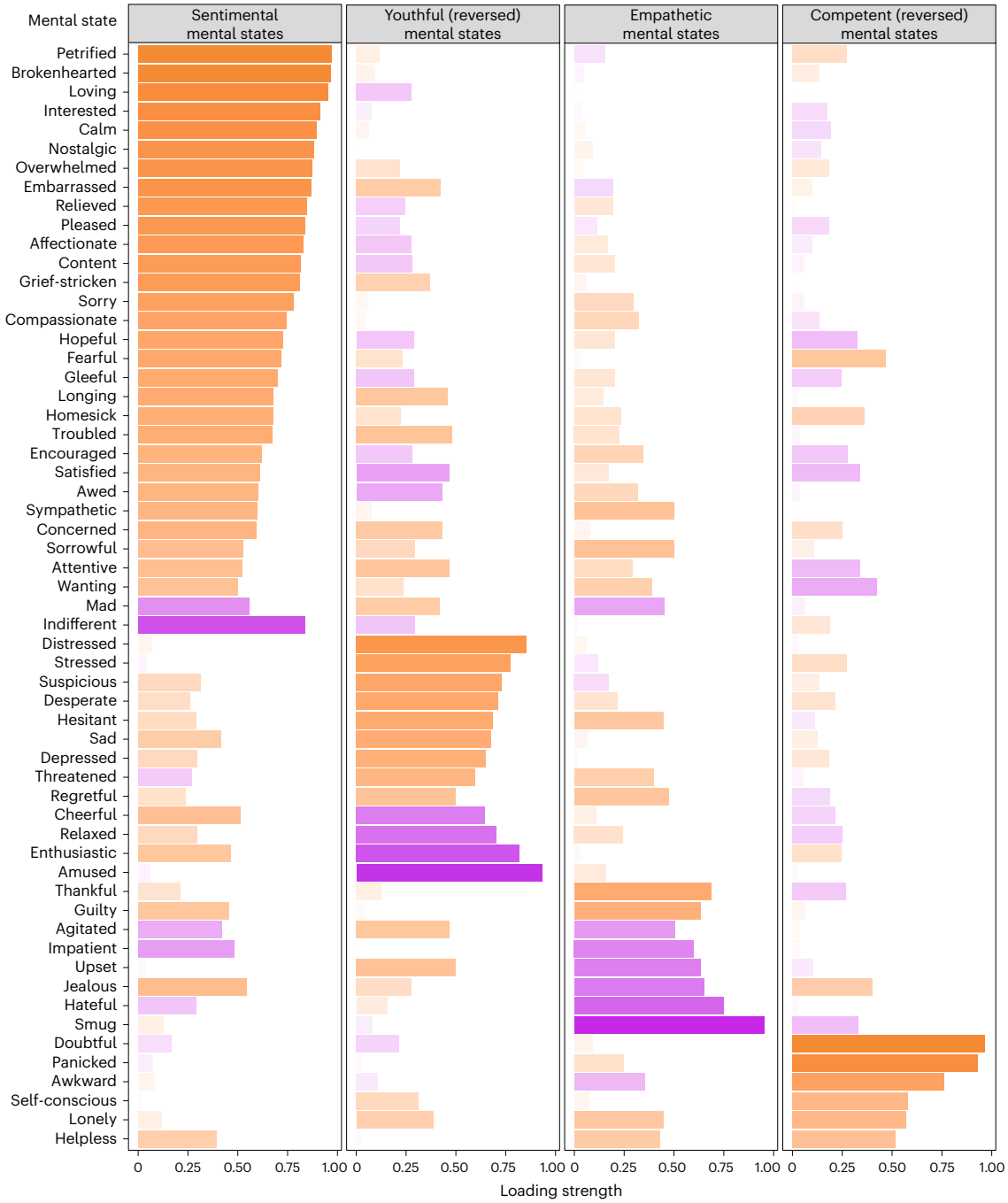

Fig. 4 | Exploratory factor analysis results of mental state inferences in given scenarios. Factor loadings of mental state inferences (rows) on the four mental state dimensions (columns) across the 58 representatively sampled mental states (2 mental states, 'bored' and 'indecisive', were excluded from this analysis as preregistered due to low factorability) using their aggregate ratings across the $n = 100$ representatively sampled faces. The bar lengths indicate absolute loading strength; the bar colours indicate the sign of the association (orange for positive and purple for negative). Top-scored faces received greater ratings on positively loaded mental states (orange); bottom-scored faces received greater ratings on negatively loaded mental states (purple).

famous people's traits, and inferences of targets' traits based on their behaviour). Here we showed that trait inferences based on relatively superficial information—judgements from faces alone—also shape mental state inferences. These findings provide a new framework for understanding how mental state inferences are formed in a more naturalistic context—that is, when both context-specific and person-specific information is available[62] and the processes of mental state inferences and trait inferences are not artificially isolated.

We found substantial variability among the 60 representatively sampled mental states in their associations with face-trait impressions (Table 1, Q1, Fig. 3 and Supplementary Table 2). For example, face-traits explained almost none of people's judgements of whether a target person felt satisfied in a given situation (1%) but most of people's judgements about whether a target person would feel threatened (84%). At least three factors may contribute to this variability. First, the conceptual connection between mental state and trait may differ across mental states. For instance, greater trait-explained variance suggests that people may think that the experiences of those mental states (for example, threat, hesitation or embarrassment) are more heavily driven by one's enduring traits than the experiences of mental states with lower trait-explained variance (for example, satisfaction, agitation or glee)[63]. Second, the salience of context may differ across

https://doi.org/10.1038/s41562-024-02059-4

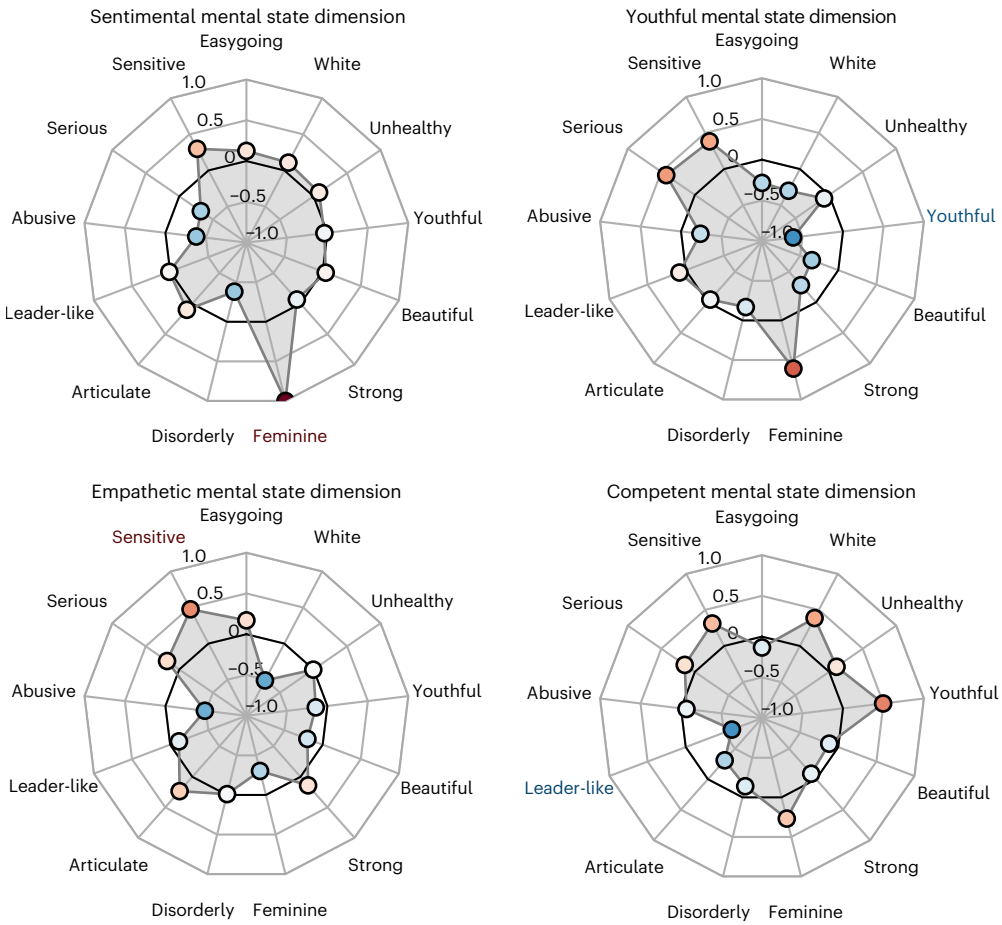

Fig. 5 | Ridge regression coefficients of 13 traits for four mental state dimensions. Each radar plot shows the result for one mental state dimension. For each mental state dimension, we regressed their factor scores across the $n = 100$ faces on the face's 13 traits while controlling for the face's 8 states (that is, the full model in the variance partition analysis). The dots indicate the mean coefficient values averaged across 2,000 cross-validation iterations. The colours the of dots indicate the sign of the coefficient (red for positive and blue for negative; more saturated colours indicate stronger associations; coloured labels indicate the trait with the strongest effect). We have validated these interpretations of the ridge regression coefficients with three additional analysis methods such as ordinary least squares regressions and LASSO regressions (Supplementary Methods and Supplementary Fig. 5).

mental states. For instance, two mental states may be conceptually thought of as equally likely to be driven by one's enduring traits (for example, feeling satisfied and threatened). However, the scenario for one mental state may be more predictive than the other (for example, 'getting back home from a long vacation' may be thought to strongly predict 'feeling satisfied' for everyone in general, whereas 'reading news about stigma against his/her group' may be less predictive of 'feeling threatened'). It is likely that when the scenario is predictive, people rely more on the scenario and less on other cues such as trait impressions from faces for mental state inferences[64]. Finally, the unique explained variance of face-trait impressions is constrained by the signal quality in the scenario-state ratings. A greater consensus in participants' scenario-state ratings for a mental state allows for a greater upper bound of the unique explained variance by face-trait impressions[65]. Regardless of which factors underlie the variability in the trait-explained variance, these results confirm the importance of investigating a diverse set of mental states and scenarios (Figs. 1a–d and 2a–c) for understanding the comprehensive relations between mental state inferences and trait impressions from faces.

We also found a large amount of shared variance in mental state inferences (Fig. 3 and Table 1, Q1) that was commonly explained by both the impression of others' stable traits and their photo-captured affective and cognitive states (mean $r^2 = 0.46$ across 60 mental states; shared variance was significant for 51 of the mental states, mean $r^2 = 0.51$; Supplementary Table 2). These results suggest that context-specific mental state inferences use overlapping cues that are relevant to both trait and state inferences from faces. This finding is consistent with prior research showing bidirectional causal effects between face-trait and face-state inferences (for example, a neutral face's structural resemblance to emotional expressions shapes trait inferences)[37,48]. As expected, the unique explained variance of face-state inferences was minimal across all scenario-state inferences (Fig. 3; mean $r^2 = 0.03$; none was significant). This is probably because all face images used in the present research were neutral. Thus, once the face's structural resemblance to affective- and cognitive-state expressions is taken into account, there is no useful information for mental state inferences beyond those already reflected in trait inferences (for example, facial structure).

Prior work using diverse trait words and white faces showed that the trait impressions people formed from faces are optimally represented by four dimensions (warmth, competence, femininity and youth)[3]. Here we showed that when faces are available, mental states are organized along four dimensions: sentimental mental states, youthful mental states, empathetic mental states and competent mental states (Figs. 4 and 5 and Table 1, Q2). These findings suggest that people may spontaneously use trait information along the four trait dimensions when inferring others' mental states in given scenarios: people may be influenced by the stereotypical femininity of the face when inferring mental states that are stereotypically associated with sentimental people, the youthfulness of the face when inferring mental states that

https://doi.org/10.1038/s41562-024-02059-4

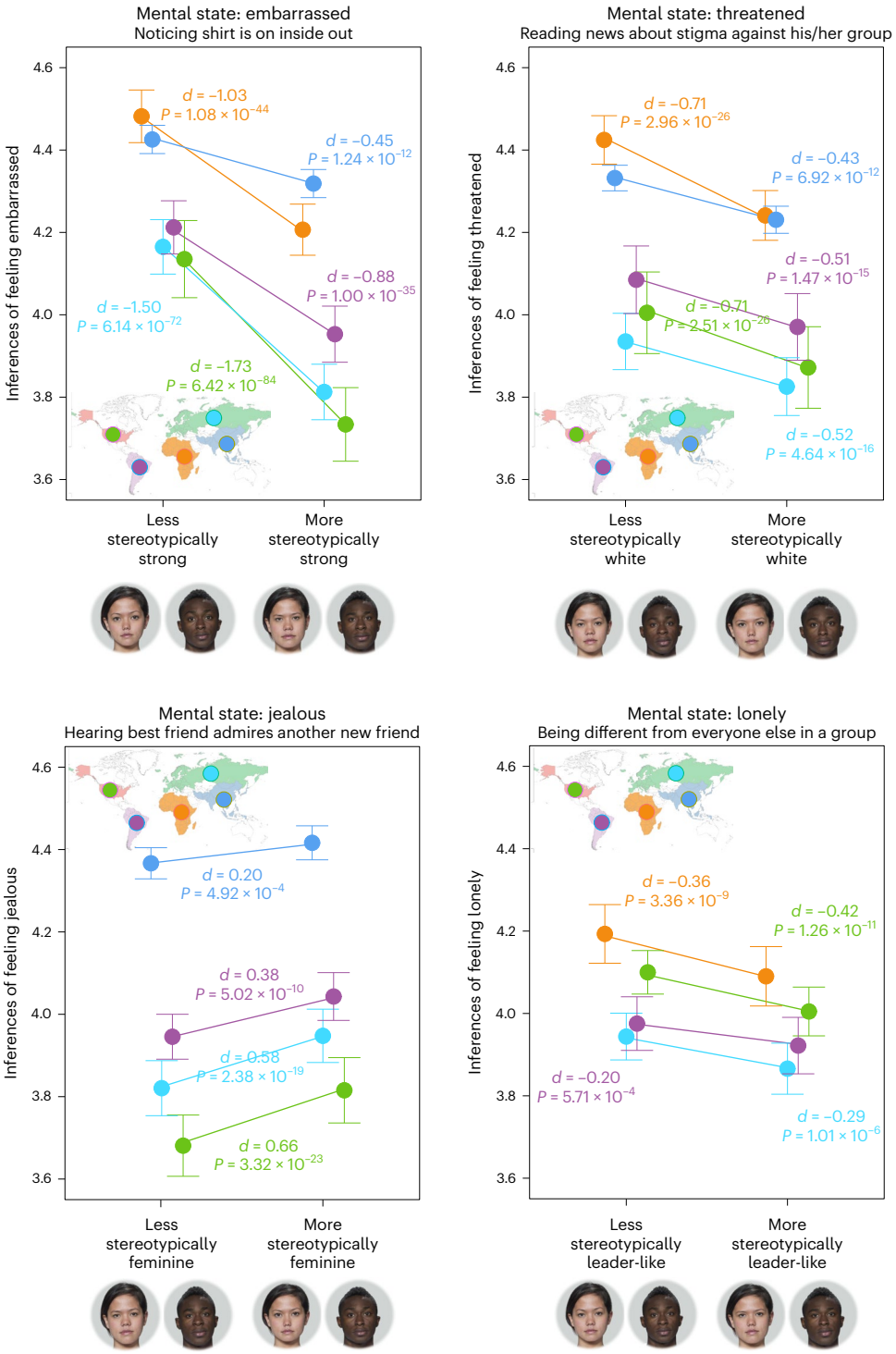

Fig. 6 | Causal effects of trait impressions on mental state inferences in given scenarios. Each panel plots the results for one mental state (specified in the title, with its scenario). As preregistered, causal effects were only tested in regions where the state-trait pairs showed significant correlations (the correlation between state-jealous and trait-stereotypically-feminine was not significant in the African sample; the correlation between state-lonely and trait-stereotypically-leader-like was not significant in the Asian sample). The dots indicate the mean scenario-state ratings (*y* axis) averaged across all participants (*n*s ranging from 31 to 50 participants per region per mental state after preregistered data exclusion) and faces ($n = 272$ images) for the two versions (*x* axis; version examples for two facial identities are shown). The error bars indicate the 95% CIs of the mean ratings for each version. The colours indicate data from different world regions as labelled on the maps. The text indicates the Cohen's *d* effect sizes calculated using aggregate data (that is, ratings per face and version averaged across participants) using one-sided paired *t*-tests as preregistered.

are stereotypically associated with younger people, the warmth of the face when inferring mental states that are stereotypically associated with people who are aware of others' feelings and the competence of the face when inferring mental states that are stereotypically associated with competent people (Fig. 5 and Table 1, Q3). These findings suggest that the mind represents social information in an integrative manner: when the faces of the targets are available, mental state representation integrates both the target information and the mental state

prediction[66]. This also suggests that, to understand social cognition in more naturalistic contexts, investigating each domain in isolation (for example, mental states alone or faces alone) may not provide a complete or ecologically valid picture; instead, more integrative, naturalistic designs may be necessary[64,67,68].

By putting trait impressions in a broader context (that is, for inferring mental states in given situations), our findings provide insights into how judgements from faces may influence real-world outcomes. Prior work investigating the real-world impacts of face impressions has predominately focused on contexts where understanding the target's traits is particularly informative, such as predicting whether someone will be a good leader, partner or employer in the long run[18,19,21,22]. Here we focused on contexts where the target's traits should be less informative than other information such as the scenario or situation—that is, inferring others' mental states in the moment. Our findings suggest that the real-world impacts of trait impressions from faces may be broader than previously believed[46]. They probably shape people's real-time social interactions with each other through influencing momentary judgements of each other's thoughts and feelings. This broader impact helps explain why certain trait judgements of faces have been found to be accurate[9–11]. For instance, perceivers who judge a target to be introverted on the basis of the face may attribute the mental state of not willing to interact to the target and thus avoid approaching the target in social events. This in turn reduces the likelihood of the target interacting with others in social events and reinforces the perceivers' initial trait impression that the target is introverted (that is, a self-fulfilling prophecy).

We found both variabilities and similarities between world regions (Fig. 6 and Table 1, Q4). While perceived stereotypical femininity was associated with and causally shaped the mental state attribution of jealousy in the USA, Asia, Europe and South America (Supplementary Table 4), this association was not significant in the Africa sample ($r = 0.01$, $P = 0.899$). This null result in the Africa sample suggests a lack of evidence for the stereotypical conception of feminine people being prone to feeling jealous in Africa[69]. However, we note that the face femininity ratings by participants in Africa were extremely similar to those made by participants in other regions (r ranged from 0.95 to 0.99). While perceived stereotypical leader-likeness was associated with and causally shaped the mental state attribution of loneliness in the USA, Africa, Europe and South America (Supplementary Table 4), this association was not significant in the Asian sample ($r = 0.12$, $P = 0.232$). This may be a result of the conceptual difference in how the leader-like trait is linked to various mental states in Asia compared with other regions. This may also be a result of the perceptual difference in how people judge leader-likeness from faces in Asia compared with other regions, given that the correlations between leader-like ratings in the Asian sample and those in other samples were only moderate (mean $r = 0.38$).

Several limitations constrain the conclusions of this research. First, although we increased the generalizability of our research by representatively sampling stimuli (faces, mental state terms and trait terms), we used only one scenario for each mental state. As discussed above, the predictiveness of the scenario may influence how much perceivers rely on the context versus the face-trait information when making mental state inferences. We systematically selected the scenario for each mental state to be relatively predictive (that is, above average and below ceiling; Methods). However, using a greater variety of scenarios per mental state may provide a more complete picture of the relationship between mental state inferences and face-trait impressions. Second, although we increased the ecological validity of our research by integrating information streams that tend to be naturally present in everyday life mentalizing (that is, information on the target and the context), our designs were still artificial. For instance, we presented the face of the target statically in isolation from other contextual information (such as clothing), and we presented the context using text. This may limit our conclusions from generalizing to real-world situations where target people are observed dynamically and naturally integrated with the complex context, which requires more than verbal processing to comprehend. Relatedly, all participants in the present research were recruited over the Internet. This approach allowed for participants with more diverse demographic backgrounds (see the laboratory log at https://osf.io/8ynzj/?view_only=f29e741904354cd9919da9b43a2609b7 for demographics) and large-scale data collection in different world regions. However, it compromises the naturalistic aspect of the research, which could be addressed by in-person research that uses an interaction paradigm rather than mere observation.

In conclusion, this investigation provides insights into how people make mental state inferences in more naturalistic contexts and the broader real-world impacts of first impressions from faces. People rely on trait impressions of the targets inferred from their faces for context-specific mental state inferences even though this information is probably invalid and biased[14]. By bridging four different types of social information (faces, contexts, trait inferences and mental state inferences) and using more generalizable designs (representative stimuli and diverse participants), our research provides a more comprehensive understanding of how the mind flexibly integrates different social information to navigate the social world.

Methods

Ethics information

The present research complied with all relevant ethical regulations and has been approved by the Committee for the Protection of Human Subjects of Dartmouth College (00032195) and the Institutional Review Board of the California Institute of Technology (21-1141). Informed consent was obtained from all human participants. The participants were compensated at an hourly rate of no less than US$9.50. (Editor's comment: this paragraph was added to the Methods section at Stage 2 at our request.) The face stimuli used in our experiments were drawn from three publicly available databases: Chicago Face Database, London Face Database and Oslo Face Database. All three datasets were originally obtained with participant consent as follows. Chicago Face Database: "Upon arrival, participants were asked to carefully read a consent/release form, allowing us to use their photos for research purposes"[70]. London Face Database: "All individuals gave signed consent for their images to be 'used in lab-based and web-based studies in their original or altered forms and to illustrate research (e.g., in scientific journals, news media or presentations).' Images were taken in London, UK, in April 2012"[71]. Oslo Face Database: "Participants gave verbal consent to be photographed for a database of face images for use in research".

Deviations from protocol

We adhered precisely to the approved registered experimental procedures, data analysis procedures and result interpretation procedures detailed in Table 1 and Methods. The only deviation from the approved registered Stage 1 protocol was the participant recruitment sources for the cross-world-region data. In the approved registered Stage 1 protocol, we had planned to recruit participants in all five world regions (the USA, Africa, Asia, Europe and South America) using the MTurk Toolkit via CloudResearch. We successfully collected all data from the USA ($n = 5{,}260$), Asia ($n = 961$) and Europe ($n = 1{,}153$) as planned. Due to the limited number of participants in Africa and South America on MTurk Toolkit (through which we collected data from $n = 707$ participants in South America and $n = 21$ participants in Africa), we collected the data in these two regions using an additional recruitment option offered on CloudResearch, Prime Panels, together obtaining the planned amount of data in Africa ($n = 1{,}040$) and South America ($n = 1{,}171$). This deviation was carried out with permission from the Editors.

All data collection details and summaries of participant demographics have been documented in the laboratory log available via the Open Science Framework at https://osf.io/8ynzj/?view_only=f29e741904354cd9919da9b43a2609b7. There is no unregistered post hoc analysis.

Pilot data

To demonstrate the feasibility of our experimental design and analysis method for our main hypothesis, H1a, we collected pilot data for five randomly selected mental states (attentive, gleeful, mad, panicked and sorrow) from 300 participants via Amazon Mechanical Turk (MTurk; 60 participants per mental state). These pilot data were collected following the same experimental procedures that were applied in our planned scenario-state task (see 'Design'). Each participant completed one experiment module. Each experiment module corresponded to one mental state and one scenario. The participants viewed 100 faces (from a previously published study[3]) one by one in randomized order and imagined each of those individual people in the given scenario (for example, listening to safety instructions). The participants rated how much each person would experience the given mental state (for example, attentive) in that given scenario, using a seven-point Likert scale anchored at 1 = not at all and 7 = extremely (Supplementary Fig. 1).

Participant-wise and trial-wise exclusion criteria were applied to the pilot data (see 'Sampling plan'; for example, excluding participants who failed more than one attention check), resulting in a remaining sample size ranging from 45 to 57 participants across the five pilot mental states (for details of the pilot data, see Supplementary Methods). Given that the corresponding scenario for each mental state was selected to elicit that mental state (see 'Design'), as expected, participants' ratings for each of the five mental states across the 100 faces were heavily skewed to high ratings, with a long tail (all means were >4, skewness ranged from −0.28 to −0.14 and kurtosis ranged from −0.71 to −0.40 across the five mental states). These pilot data also suggested minimal floor or ceiling effects (ratings of 1 occurred no more than 2% and ratings of 7 no more than 7% of the time across all five mental states).

To test whether these mental state inferences would be associated with the trait impressions formed from the individuals' faces (Table 1, H1a), we performed ridge regression with cross-validation analyses as planned. For each mental state, we regressed the aggregated mental state ratings (averaged across participants per face) from our pilot data on the aggregated ratings of those faces on 100 different traits from the previously published study[3]. The results showed that inferences of all five mental states in the corresponding scenarios were significantly associated with the trait impressions of the faces: model prediction accuracy was $r = 0.77$ (95% CI, (0.56, 0.90); $P = 1.67 \times 10^{-4}$) for attentive, $r = 0.91$ (95% CI, (0.83, 0.96); $P = 1.67 \times 10^{-4}$) for gleeful, $r = 0.89$ (95% CI, (0.77, 0.96); $P = 1.67 \times 10^{-4}$) for mad, $r = 0.58$ (95% CI, (0.30, 0.79); $P = 0.004$) for panicked and $r = 0.60$ (95% CI, (0.27, 0.83); $P = 1.63 \times 10^{-3}$) for sorrow. The CIs were the 2.5th and 97.5th percentiles of the empirical distribution of the prediction accuracies generated over 2,000 train/test splits in cross-validations. The P values were computed and corrected for multiple comparisons via maximal statistic permutation tests. The pilot data and analysis code are available via the Open Science Framework (see 'Data availability' and 'Code availability').

Stimuli

Selection of mental state terms. To increase generalizability, we representatively sampled a set of mental state terms from a large, inclusive initial set. We defined a mental state as a person's internal, mental, temporally changeable characteristics. Our goal was to sample a comprehensive set of mental state terms that describe people's non-pathological and specific mental states. We first gathered an extensive list of putative mental state terms from the literature[50,72–75] ($n = 404$). We then applied seven criteria to exclude terms that (1) describe other people's actions or attitudes towards a person, which were instead substituted with terms that describe the person's own mental state (for example, 'social exclusion' and 'ostracized' were substituted by 'lonely'); (2) describe pathological states (for example, 'insane' or 'hallucinating'); (3) describe non-specific mental states that were instead substituted with more specific terms (for example, 'unhappy' was substituted by 'sad' or 'angry' or 'afraid'); (4) describe generic psychological processes (for example, 'thinking' or 'feeling'); (5) have unclear meanings (for example, 'interconnected' or 'objective'); (6) describe physical states or behaviours (for example, 'hungry' or 'subordinate'); and (7) could describe both mental states and traits but were derived from traits (for example, 'neurotic' or 'serious'). On the basis of these seven exclusion criteria, two authors independently judged whether each mental state term should be excluded or included or was debatable. A mental state term was excluded if both authors labelled it as to be excluded or debatable. After exclusion, 141 terms remained (Fig. 1a).

We verified that the 141 mental state terms selected from the literature were representative of the mental state terms laypeople use in everyday life to describe others' internal, non-pathological, specific mental states. To this end, we quantified each of the 141 mental state terms with a vector of 300 semantic features using a state-of-the-art neural network that had been pretrained to assign words to their contexts across 600 billion words[76]. We then trained a linear classifier[76] using our manual labels (141 'included' and 263 'excluded' from our putative list of 404) and applied it to 220k COCA[77]. COCA is "the only corpus of English that is large, up-to-date (1990–2020), and balanced between many genres"[77], including spoken, fiction, popular magazines, newspapers, academic texts, TV and movie subtitles, and blogs. The linear classifier performed well (precision = 1.00, recall = 0.89, $F_1 = 0.95$ for 'included' states; precision = 0.95, recall = 1.00, $F_1 = 0.97$ for 'excluded' states). This linear classifier identified 252 words from 220k COCA as internal, non-pathological, specific mental state terms (Fig. 2a). UMAP[78] of the 141 literature-selected and the 252 corpus-identified mental state terms onto the same two-dimensional space showed that the former were a fairly representative sample of the latter (Fig. 2b).

After verifying that the list of mental state terms we selected from the literature were representative and valid, we next aimed to reduce the ambiguity and redundancy of the list. To this end, we applied four filters (Fig. 1b,c): (1) we excluded terms with unclear meaning ($n = 19$) according to an independent sample of MTurk participants ($n = 32$); (2) we computed the cosine similarity between each pair of terms on the basis of their vectors in the 300-dimensional semantic space[76]; for each pair with a similarity greater than two standard deviations from the mean, we (3) excluded the term with a lower rated familiarity (that is, people are less likely to feel that mental state in everyday life) according to an independent sample of MTurk participants ($n = 34$) or (iv) if they had the same familiarity, we excluded the term with a lower rated usage frequency (that is, people are less often to use that word in everyday life) according to an independent sample of MTurk participants ($n = 31$). These four filters eliminated 81 terms, resulting in our final set of 60 mental state terms that are clear and minimally redundant (Supplementary Table 1). These 60 mental state terms were used in the scenario-state task (see 'Procedures').

Selection of mental state scenarios. To understand how people infer another individual's mental states in specific scenarios, we chose one scenario for each of the 60 selected mental states. We aimed to choose one short text scenario (with three to eight words) per mental state that would (1) prompt the intended mental state in an average person, (2) do so without introducing ceiling effects (that is, preventing all faces receiving high ratings because the scenario is too stereotypical or strong) and (3) elicit sufficient variability in terms of the intensity with which different people might feel that mental state. To this end, we generated four putative scenarios for each of the 60 selected mental states. All scenarios were then rated (on a five-point Likert scale) by an independent sample of MTurk participants ($n = 60$) on how much each scenario would elicit the intended mental state in an average person.

Data from participants who passed all attention checks, were native English speakers and had at least high school education ($n = 47$)

were used for scenario selection. As expected, most scenarios (97%) received a high mean rating (>3, which was the midpoint of the rating scale). That is, participants agreed that the scenarios would elicit the intended mental state in an average person. Aiming for a scenario that would probably elicit ratings with a sufficient variance across faces for the target mental state, we selected the scenario with the largest rating variance across participants and an average rating >3. We further confirmed that all selected scenarios were not specific to salient social categories such as any specific gender or race. See 'Data availability' for all mental state terms, putative scenarios and selected scenarios. These selected scenarios, one for each of the 60 mental states, were used in our scenario-state task (see 'Procedures').

Selection of face-trait terms. To increase generalizability, we systematically selected a representative set of trait terms on which participants rated the faces. We defined a trait broadly as a person's temporally stable characteristic, which could include age, gender, race, socio-economic status, social evaluative qualities and personality. Our goal was to sample a small set of trait terms that (1) can be reliability inferred from faces (in terms of within-participant test–retest reliability and between-participant consensus), (2) can elicit a range of ratings across our face stimuli and (3) are representative of the comprehensive space of trait impressions from faces[3,4,79].

We began with a comprehensive and representative list of 94 traits from a previously published study[3]. That study selected these traits from an inclusive set of traits of all important categories for person perception (demographics, physical appearance, social evaluative qualities, personality and emotional traits) from the literature (n = 482). That study systemically sampled these 94 traits according to meaning clarity, semantic similarity and usage frequency (Fig. 1e–g). That study also provides the ratings on the 94 traits for a representative set of faces. On the basis of those ratings, we selected the traits that have (1) a good within-participant test–retest reliability (within-participant Spearman's correlation averaged across all participants was greater than the mean across all traits (0.42)), (2) a good between-participant consensus (between-participant Spearman's correlation averaged across all participant pairs was greater than the mean across all traits (0.14)) and (3) a sufficient variance across faces (the variance of the aggregated ratings across faces was greater than the mean across all traits (0.49)). These three filters selected a subset of 22 traits. Some of these 22 traits were extremely highly correlated in their face ratings. Therefore, for any pair of extremely highly correlated traits (with a Pearson correlation >0.9), we retained only the trait with a higher within-participant test–retest reliability. These procedures selected a subset of 13 traits. We made a slight modification to one of the selected traits for balancing the valence across the 13 traits (substituted 'healthy' with 'unhealthy'; Supplementary Table 1). These 13 traits were used in our face-trait task (see 'Procedures').

Selection of face-state terms. We measured the faces' current mental states (when the photos were taken) to control for their effects on the association between trait impressions from faces and subsequent mental state inferences in specific scenarios. We aimed to representatively sample a wide range of mental states that the face owners might have felt at the moment the photos were taken. We therefore included both affective states such as emotions and cognitive states such as contemplation. We based our selection on a previously published study[50], which investigated a comprehensive set of 60 affective and cognitive mental states. That study found that these mental states could be represented by three dimensions. We selected eight mental states from the 60 terms in that study along those three dimensions. These eight mental states were at the extremes (that is, the corners) as well as in the middle of this comprehensive three-dimensional mental state space (Supplementary Table 1). These eight mental states were used in our face-state task (see 'Procedures').

Selection of face stimuli. We focused on faces from high-quality studio photographs (clear, frontal, direct gaze, plain background), of relatively neutral expression (not depicting any blatant emotions). Factors such as angle, lighting, gaze direction, background and facial expression have been shown to shape mental state and trait inferences from faces; however, the interactions of these factors with the relationship between mental state and trait inferences were beyond the scope of our present study, and so we selected stimuli that do not exhibit noticeable variance in these factors.

We began with three large and popular publicly available high-quality face databases[70,71,80], which included face images of individuals from both genders, different races (white, Black and Asian) and different age groups. We included the variance from these three factors because they represent the important dimensions of diversity in the individuals that people see in everyday life. We applied three filters to exclude faces that (1) were not frontal, (2) were not neutral and (3) had objects obscuring the face, resulting in 764 faces. We aimed to sample a subset of faces that maximally vary in facial geometry. We therefore quantified each face with a vector of 512 facial features using a state-of-the-art neural network that had been pretrained to identify individual identities across millions of faces[81]. These features identify individuals by their facial structures and are not sensitive to factors such as angle, lighting, background, facial expression, race and makeup. We then computed the cosine distance between each pair of faces using their vectors in the 512-dimensional face space. Stratified maximum variation sampling was then applied to sample 50 female faces (58% white, 26% Black, 16% Asian) and 50 male faces (58% white, 26% Black, 16% Asian) that were most distinct from each other within each gender and race (Fig. 1i–l).

We determined the sampling method and race ratio on the basis of two considerations. First, we used maximum variation sampling to select faces that were most different in terms of facial geometry to increase the generalizability of our results. Specifically, the more distinct the faces we sampled, the larger the portion of the face space our samples covered. This allowed for greater generalizability to all diverse faces within the larger sampled space via interpolation. Otherwise, if we only sampled faces that covered a small subspace of the face space (for example, faces that are close to the average faces that people would see in everyday life), then to generalize results to more diverse faces, one would need to rely on extrapolation to faces that are outside of the small sampled space. Such extrapolation is known to be less accurate than interpolation[82,83]. Second, we determined the race ratios on the basis of the proportion of each race in the 764 faces. Basing the race ratios on the sampling database instead of other references (for example, equal ratios or the US population) ensures that the maximum variation sampling procedures do not bias any one of the races. Specifically, the number of faces of different races in the sampling database differs largely (444 white, 198 Black, 122 Asian). Selecting faces proportional to the base rate in the database instead of, for instance, selecting an equal number of faces per race ensures that the selected white faces are no more heterogeneous with respect to each other than the selected Black or Asian faces.

These 100 selected faces were used as stimuli in our scenario-state task, face-trait task and face-state task (see 'Procedures'). They were used as reference images for digitally manipulating the trait impressions of new faces (see 'Stimuli: Trait manipulation of face stimuli').

Trait manipulation of face stimuli. To test the causal effect of face-trait impressions on mental state inferences (H4), we manipulated the trait impressions of new faces. We focused on a subset of the 60 mental states that were strongly associated with trait impressions and that spanned multiple mental state dimensions, ideally around six mental states (10%). Specifically, for each of the mental state dimensions in H2a, we identified the mental states that were (1) most strongly associated with this dimension, and among these, those that were (2) most

strongly and significantly associated with trait impressions in H1b. The selected mental states were the targets for causality testing. For each target mental state, we identified a target trait that was strongly associated with this mental state in H1b (greatest absolute model weight in the full model). In the case where the same target trait coincides with multiple target mental states, we diversified the target traits by selecting the state–trait pairs that have the greatest average association.

For each target trait, we selected two sets of base faces (B+ and B−) from our 100 faces, which were used to digitally increase and decrease the impressions of the target trait in any new face image. The two sets of base faces met the following three criteria: (1) the B+ set had a mean rating on the target trait that was maximally greater than the mean rating of the B− set; (2) the B+ set had a mean rating on face-states that was similar to that of the B− set, to control for the effect of face-states; and (3) the B+ and B− sets included a similar number of faces, ideally around ten. We digitally averaged the faces within the B+ set and the B− set. The B+ average and B− average faces were used to manipulate the impressions of the target trait in new faces.

We selected 272 new faces to which we applied trait manipulation. This sample size of faces was determined by formal power analysis. Our aim was to test whether digitally manipulating a face to look more versus less prominent in a trait would cause participants to shift mental state inferences in the direction that is consistent with our correlational findings in H1. We therefore estimated the number of face images needed on the basis of paired one-sided *t*-tests. The results indicated that to detect a small effect ($d = 0.2$) with 95% power and a significance level of 0.05, we needed to use 272 face images. These 272 new faces were sampled from the three databases[70,71,80] we used before (Fig. 1i), excluding the 100 already selected faces (Fig. 1l). We used the same stratified maximum variation sampling method as before to sample these new faces of different gender, race and age. For each of the 272 new faces and each target trait, we created two versions of the face image: one with increased target trait impression, by digitally adding a proportion of the difference between the B+ average and B− average faces (in shape and texture) to the original face image; and the other with decreased target trait impression, by subtracting the same proportion of the difference from the original face image. The maximum proportion without distorting the faces was used, ideally above 50%. The resulting 544 face images were used in our scenario-state task for understanding the causal relation between the target trait and target mental state (see 'Procedures: Scenario-state task with trait manipulation').

Procedures

Scenario-state task. In the scenario-state task, participants imagined different individuals in specific scenarios and inferred the individuals' mental states in those scenarios. Each participant was randomly assigned to one experiment module. Each experiment module corresponded to one of the 60 mental states and its scenario (see 'Stimuli'). In each module, the participants saw the scenario (for example, 'listening to safety instructions') and rated how much (1) an average person, (2) 100 specific people (our face stimuli) and (3) the participant themselves would feel the corresponding mental state (for example, 'attentive') in that scenario (Supplementary Fig. 1a).

For each of the 100 specific people (shown in random order), the participants first saw their face. Subsequently, the participants saw the text scenario and imagined the person whose face they were looking at in that scenario (for example, 'imagine this person is listening to safety instructions'). In particular, the participants were instructed that when those people took the photos, they were told to keep a neutral face and not to show any smile or other facial expressions; those people's faces might look different when they are in the given scenario. The participants then saw a question asking how much the person would experience the given mental state in that scenario (for example, 'how attentive would this person feel?'). The participants entered their answers by pressing number keys 1 to 7 on their computer keyboard, with 1 = not at all and 7 = extremely (Supplementary Fig. 1a). Participants received an alert message if they mistakenly pressed other keys than the number keys 1 to 7. They saw an alert message if their response time (from when the question appeared to when a response was entered) was too short (under 200 milliseconds) or too long (over 10,000 milliseconds). The participants had the option to take a short break after every 20 trials. There were also seven attention checks randomly scattered between the 100 trials, in which the participants were asked to press a certain number key on their keyboard.

After the experiment, the participants filled out a questionnaire about their understanding of the mental state term and the scenario, demographics, traits (self-report on the 13 traits used in the face-trait task) and mental states (self-report on eight mental states used in the face-state task). The participants provided ratings for the 13 traits and the eight mental states using a seven-point Likert scale. The experiment code is available via the Open Science Framework (see 'Code availability').

Face-trait task. In the face-trait task, participants viewed different individuals' face images and judged those individuals' traits solely on the basis of their face images. Each participant was randomly assigned to one experiment module. Each module corresponded to one of the 13 traits (see 'Stimuli'). In each module, the participants viewed the 100 faces one by one in random order and rated how much each face fitted the description of the given trait. For each face, the participants first saw a fixation cross. Subsequently, they saw the question, the face and the rating scale. The participants entered their answers using a seven-point Likert scale (anchored at 1 = not at all and 7 = extremely) by pressing number keys 1 to 7 on their computer keyboard (Supplementary Fig. 1b). They received an alert message if their response time was too short or too long. The participants had the option to take a short break after every 20 trials.

Face-state task. In the face-state task, participants viewed different individuals' face images and inferred the mental states that the individuals were currently displaying solely on the basis of their face images. Each participant was randomly assigned to one experiment module. Each module corresponded to one of the eight mental states (see 'Stimuli'). In each module, the participants viewed the 100 faces one by one in random order and rated how much each face looked like it was currently displaying the given mental state. For each face, the participants first saw a fixation cross. Subsequently, they saw the question, the face and the rating scale. The participants entered their answer using a seven-point Likert scale (anchored at 1 = not at all and 7 = extremely) by pressing number keys 1 to 7 on their computer keyboard (Supplementary Fig. 1c). They received an alert message if their response time was too short or too long. The participants had the option to take a short break after every 20 trials.

Scenario-state task with trait manipulation. In the scenario-state task with trait manipulation, participants imagined different individuals in specific scenarios and inferred the individuals' mental states in those scenarios. This task aimed to test the causal effect of trait impressions on mental state inferences (H4). Since we used 272 face identities to test this causal effect, each with two manipulated versions (see 'Stimuli'), it would take too long for each participant to rate all 544 face images. We therefore shuffled the 272 face identities and randomly assigned them to four different subsets ($n = 68$ face identities, with 136 face images per subset). Each participant was randomly assigned to one experiment module. Each experiment module corresponded to rating a subset of 136 face images for one mental state in the corresponding scenario. In each module, the participants viewed the 136 faces one by one in random order, with the constraint that the images of the same identity appeared at least one trial apart. For each face, the

participants imagined the face owner in the given scenario and rated how much the person would feel the corresponding mental state, as in the scenario-state task (Supplementary Fig. 1a).

Sampling plan

Recruitment plan. We recruited participants who satisfied our inclusion criteria listed below via MTurk through the CloudResearch platform (formerly known as TurkPrime). All participants completed our study online using desktop or laptop computers. In most studies, we recruited participants who were located in the USA.

To promote participant diversity and increase the generalizability of our study, we also included participants from different world regions. Given our current budget, we collected data from different world regions only for testing hypothesis H4. Specifically, besides the participants located in the USA, we recruited participants from four different continents (South America, Europe, Africa and Asia) via MTurk and Prime Panels through CloudResearch. From each of these international samples, we collected data for testing the correlations between the target state–trait pairs identified from the US samples (see 'Stimuli: Trait manipulation of face stimuli'). That is, we repeated data collection for H1 but only for the target mental states and traits in these international samples. For each significantly correlated state–trait pair in each international sample, we further collected data for testing the causal effect of this state–trait pair.

Inclusion criteria. For the US samples, we included participants who satisfied the following criteria: (1) were age 18 or older, (2) had normal or corrected-to-normal vision, (3) had a good performance history in terms of approval rate (greater than or equal to 99%) and the number of previous submissions (more than 100 tasks), (4) were native English speakers, (5) were located in the USA and (6) had at least high school education.

For each of the international samples across the four continents, we included participants who satisfied the following criteria: (1) were age 18 or older, (2) had normal or corrected-to-normal vision, (3) had a good performance history in terms of approval rate (greater than or equal to 99%) and the number of previous submissions (more than 100 tasks) for participants on MTurk and all active participants on Prime Panels, (4) were fluent in English, (5) were located in countries that are within the targeted continent and (6) had at least high school education.

We maintained a good balance between participants who self-identified as women and men for each sample, and we included participants from a wide range of age groups.

In the case where there were not enough participants who satisfied the above inclusion criteria in an MTurk sample, we relaxed the criterion of performance history to include participants with an approval rate of greater than or equal to 90% and a number of previous submissions more than or equal to 10 tasks.

Sample size. We determined the sample size for the scenario-state task based on our main hypothesis H1a via jackknife resampling of empirical data. Using the pilot data we collected on five randomly selected mental states, we performed the ridge regression with cross-validations as planned (see 'Analysis plan'). The results showed that the sample size we had in the pilot data was sufficient for detecting an accurate prediction in all five mental states. To empirically examine how the predication accuracy for each mental state would change as a function of the sample size of participants, we reduced the sample size one by one using the jackknife resampling procedure. In each iteration, we removed one randomly selected participant and computed the new aggregate ratings on the mental state for all faces using data from the remaining participants; these new aggregate ratings were then used to train and test the model. We repeated these steps 40 times at each sample size (that is, a different participant was randomly selected to be removed each time). The results showed that the minimum sample size for detecting a reliably above-chance prediction accuracy (that is, the 95% CI of the prediction accuracies was above the chance level) in any randomly selected pilot mental state was 37 participants (Supplementary Fig. 2). On the basis of these results and assuming a participant exclusion rate of 25%, we determined the sample size to be 50 participants for each mental state. In the case where the actual participant exclusion rate turned out to be greater than what we expected, we recruited participants until the sample size after data exclusion reached 37 participants per mental state, the minimum sample size indicated by our empirical power analysis.

We determined the sample size for the face-trait task and the face-state task on the basis of the point of stability for aggregate data via sequential resampling of empirical data. Given that we only used the aggregate data of face-trait ratings and face-state ratings for our planned analyses, we targeted a sample size that would generate a stable average—that is, additional samples do not meaningfully change the average. Prior data[84] on face judgements on a variety of traits and mental states (emotions) collected on a seven-point Likert scale showed that the mean sample size needed for generating a stable average with a corridor of stability of ±0.5 and 95% confidence was 28 participants across traits, and 23 participants across mental states. On the basis of these results and assuming a participant exclusion rate of 25%, we determined the sample size to be 38 participants for each trait and 31 participants for each mental state. In the case where the actual participant exclusion rate turned out to be greater than what we expected, we recruited participants until the sample size after data exclusion reached 28 participants per trait and 23 participants per mental state.

Exclusion criteria. For the scenario-state task (with the original faces as well as the trait-manipulated faces), participant-wise exclusion was done if a participant (1) reported that the meaning of the mental state was not clear, (2) reported that the meaning of the scenario was not clear, (3) gave a rating of 1 to any of the two questions about an average person (indicating that our scenario did not work for the participant or was misunderstood by the participant), (4) failed more than one attention check, (5) gave more than 90% of the faces the same rating (indicating inattention to the traits of the faces) or (6) had more than 10% of the trials excluded per trial-wise exclusion criteria. Trial-wise exclusion was done if a trial (1) had a rating of 1 while the average rating across participants for the same face and mental state was above 6 (suggesting the participant flipped the rating scale on that trial) or (2) had a response time shorter than 200 milliseconds or longer than 10,000 milliseconds.

For the face-trait task and the face-state task, participant-wise exclusion was done if a participant (1) gave more than 90% of the faces the same rating or (2) had more than 10% of the trials excluded per the trial-wise exclusion criterion. Trial-wise exclusion was done if a trial had a response time shorter than 400 milliseconds or longer than 10,000 milliseconds.

Analysis plan

Contingency of analysis. Analysis decisions of hypotheses H1–H3 were independent of each other. Given any findings about the associations between scenario-states and face-traits in H1, we carried out the analyses for H2 and H3 about the dimensionality of scenario-states as planned. For dimensionality analysis, we excluded mental states that had low factorability (see below). On the basis of our pilot data of scenario-state ratings for five randomly selected mental states, only one of them was excluded from the dimensionality analysis. We therefore expected to have a non-empty set of factorable mental states. Given any non-empty set of factorable mental states, there is a non-empty set of factors—therefore, we carried out the analyses for H2 and H3 as planned.

Analysis decisions of all sub-hypotheses within each hypothesis were independent of each other. Given any findings about the associations between scenario-states and face-traits in H1a, we carried out the analysis of H1b to examine the unique variance explained by face-traits as planned. Similarly, given any findings about the associations between scenario-state dimensions and face-traits in H3a, we carried out the analysis of H3b to examine the unique variance explained by face-traits as planned. Given any dimensions found in H2a, we carried out the analysis of H2b to compare these dimensions to previously discovered mental state dimensions as planned.

Analysis decisions of hypothesis H4 were contingent on the findings from H1 and H2 in two aspects. First, we tested causal effects (H4) only if there were correlational effects (H1). Second, the state–trait pairs for which we tested causal effects were selected on the basis of the findings from H1 and H2: we focused on a subset of mental states that were strongly associated with trait impressions (H1b) and that spanned multiple mental state dimensions (H2a). On the basis of our pilot data, all of the randomly selected mental states were found to be significantly correlated with face-traits. We therefore expected to find significant correlational effects for a large number of different mental states—thus, we expected to carry out the analyses for H4 as planned.

Statistical analysis. For all analyses, we first processed the data according to our planned exclusion criteria. For testing hypotheses H1–H3, we used the aggregate ratings averaged across participants for each face on each mental state, face-trait and face-state. For testing hypothesis H4, we used both aggregate ratings and individual-level ratings.

To test hypothesis H1a, we performed ridge regression with cross-validations for each of the 60 mental states. Ridge regression was chosen for its ability to handle multicollinearity and low-importance predictors. Each model regressed the scenario-state ratings (dependent variable) on the 13 face-trait ratings (independent variables) across the faces. In each cross-validation iteration, we trained the model on a randomly selected set of 80% of the data and tested the model on the other 20% of the data; model hyperparameter λ (the weighting of the penalty to the loss function) was tuned using nested cross-validation. Model prediction accuracy in each cross-validation iteration was assessed with Pearson correlation (r) in the test data. We assessed the significance of the prediction accuracy and corrected for multiple comparisons (over the 60 mental states) using maximal statistic permutation tests.

To test hypothesis H1b, we constructed three ridge regression models for each of the 60 mental states. The three models regressed the scenario-state ratings (dependent variable) on (1) the 13 face-trait ratings, (2) the 8 face-state ratings and (3) both the 13 face-trait ratings and the 8 face-state ratings. The explained variance by each set of independent variables equals the squared prediction accuracy of the corresponding model as assessed with Pearson correlation. The unique variance explained by face-traits equals the explained variance of model 3 minus the explained variance of model 2. In each cross-validation iteration, the three models were trained on a randomly selected set of 80% of the data and tested on the other 20% of the data, with the hyperparameter λ being tuned using nested cross-validation. Over the 2,000 cross-validation iterations, we obtained an empirical distribution of the unique variance explained by face-traits. We deemed the explained variance to be significantly greater than zero if the 2.5th percentile of the empirical distribution was greater than zero.

To test hypothesis H2a, we first computed the Pearson correlation between each pair of mental states using their aggregate-level scenario-state ratings across the 100 faces. We then excluded any mental state that had a low factorability (that is, an average absolute correlation with all other mental states below 0.20). For the remaining mental state data, we employed five different methods to determine the number of underlying dimensions: Horn's parallel analysis, the optimal coordinate index, the empirical Bayesian information criterion, Velicer's minimum average partial test and bi-cross-validation. If multiple methods agreed on the number of dimensions, we deemed the number that most methods agreed on to be the optimal number of dimensions that describe our data. If none of the five methods agreed on the optimal number of dimensions, we performed exploratory factor analysis to extract different numbers of dimensions. We deemed the minimum number of dimensions for which all dimensions had the clearest interpretation (with oblimin rotation) and that accounted for at least 75% of the common variance in the data to be the optimal number of dimensions.

To test hypothesis H2b, we first extracted the optimal number of dimensions determined in H2a using exploratory factor analysis with oblimin rotation. We then measured the similarity between these dimensions in our data and the 3-D Mind dimensions using two different methods. One method computed the Spearman correlations between the dimensions in our data and the 3-D Mind dimensions using factor loadings and scores. Specifically, a subset of 37 mental state terms that were used in our present study overlapped with those used in the discovery of the 3-D Mind Model[49]. That study collected human participant ratings of various mental states on 16 psychological scales and found that those mental states could be summarized by only three dimensions (valence, rationality and social impact). We used the factor scores on those three dimensions from that study in our present analysis. For each pairwise combination between the 3-D Mind dimensions and our dimensions, we computed the Spearman correlation between the 3-D Mind dimension's factor scores and our dimension's factor loadings across the 37 overlapping mental states. We deemed an absolute correlation of 0.2 to 0.39, 0.4 to 0.59 or 0.6 and above to be an indication of weak, moderate or strong factor similarity, respectively. The second method asked an independent set of MTurk participants to rate how well each of our 60 mental states could be described by the 3-D Mind dimension labels (that is, how positive, rational and socially impactful the mental states are). For each pairwise combination between the 3-D Mind dimensions and our dimensions, we then computed the Spearman correlation between participants' average ratings for the 60 mental states on the 3-D Mind dimension label and the 60 mental states' factor loadings on our dimension. We deemed an absolute correlation of 0.2 to 0.39, 0.4 to 0.59 or 0.6 and above to be an indication of weak, moderate or strong factor similarity, respectively.

To test hypotheses H3a and H3b, we used the same methods as for testing H1a and H1b. The only difference was that the dependent variable per model was no longer the scenario-state ratings across the faces per mental state, but instead the factor scores across the faces per mental state dimension.

To test hypothesis H4, we first performed manipulation checks. For each trait manipulation, we computed the aggregate face-trait ratings averaged across participants per face image. We assessed whether the trait-increased versions of the face images elicited higher face-trait ratings on the manipulated trait than the trait-decreased versions, using paired one-sided t-tests across face identities. If a trait manipulation was successful, we tested its causal effects on the corresponding mental state inferences. We used two different methods: one based on the aggregate-level ratings and the other on the individual-level ratings. For the aggregate data analysis, we computed the aggregate scenario-state ratings averaged across participants per image version for each face identity. To assess whether the scenario-state ratings for the trait-increased version and the trait-decreased version were significantly different in the expected direction according to our correlational findings from H1, we performed paired one-sided t-tests across face identities. For the individual data analysis, we regressed the individual-level scenario-state ratings on the face versions (binary coded) while controlling for the random effects of individual participants and face identities using linear mixed modelling.

Reporting summary

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

Data availability

All experiment materials, the raw and processed data, the data usage guidance, and the laboratory log documenting the details of data collection are available via the Open Science Framework at https://osf.io/8ynzj/?view_only=f29e741904354cd9919da9b43a2609b7. No data for any preregistered study (other than pilot data included at Stage 1) were collected prior to the date of acceptance in principle. All data files were collected after acceptance in principle and appropriately time-stamped according to the approved registered Stage 1 protocol (except that data collection in Africa and South America used a second CloudResearch platform, Prime Panels, beyond the preregistered platform, MTurk Toolkit, with the Editors' permission). Source data are provided with this paper.

Code availability

All experiment code and analysis code are available via the Open Science Framework at https://osf.io/8ynzj/?view_only=f29e741904354cd9919da9b43a2609b7.

References

1. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* **19**, 1508–1519 (2007).
2. Willis, J. & Todorov, A. First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* **17**, 592–598 (2006).
3. Lin, C., Keles, U. & Adolphs, R. Four dimensions characterize attributions from faces using a representative set of English trait words. *Nat. Commun.* **12**, 5168 (2021).
4. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl Acad. Sci. USA* **105**, 11087–11092 (2008).
5. van't Wout, M. & Sanfey, A. G. Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition* **108**, 796–803 (2008).
6. Hester, N., Xie, S. Y. & Hehman, E. Little between-region and between-country variance when people form impressions of others. *Psychol. Sci.* **32**, 1907–1917 (2021).
7. Walker, M., Jiang, F., Vetter, T. & Sczesny, S. Universals and cultural differences in forming personality trait judgments from faces. *Soc. Psychol. Pers. Sci.* **2**, 609–617 (2011).
8. Cogsdill, E. J., Todorov, A. T., Spelke, E. S. & Banaji, M. R. Inferring character from faces: a developmental study. *Psychol. Sci.* **25**, 1132–1139 (2014).
9. Kramer, R. S. S. & Ward, R. Internal facial features are signals of personality and health. *Q. J. Exp. Psychol. (Hove)* **63**, 2273–2287 (2010).
10. Back, M. D. et al. Facebook profiles reflect actual personality, not self-idealization. *Psychol. Sci.* **21**, 372–374 (2010).
11. Penton-Voak, I. S., Pound, N., Little, A. C. & Perrett, D. I. Personality judgments from natural and composite facial images: more evidence for a 'kernel of truth' in social perception. *Soc. Cogn.* **24**, 607–640 (2006).
12. Foo, Y. Z., Sutherland, C. A. M., Burton, N. S., Nakagawa, S. & Rhodes, G. Accuracy in facial trustworthiness impressions: kernel of truth or modern physiognomy? A meta-analysis. *Pers. Soc. Psychol. Bull.* https://doi.org/10.1177/01461672211048110 (2021).
13. Rule, N. O., Krendl, A. C., Ivcevic, Z. & Ambady, N. Accuracy and consensus in judgments of trustworthiness from faces: behavioral and neural correlates. *J. Pers. Soc. Psychol.* **104**, 409–426 (2013).
14. Todorov, A. *Face Value: The Irresistible Influence of First Impressions* (Princeton Univ. Press, 2017).
15. Lenz, G. S. & Lawson, C. Looking the part: television leads less informed citizens to vote based on candidates' appearance. *Am. J. Polit. Sci.* **55**, 574–589 (2011).
16. Ahler, D. J., Citrin, J., Dougal, M. C. & Lenz, G. S. Face value? Experimental evidence that candidate appearance influences electoral choice. *Polit. Behav.* **39**, 77–102 (2017).
17. Todorov, A. Inferences of competence from faces predict election outcomes. *Science* **308**, 1623–1626 (2005).
18. Martin, D. S. Person perception and real-life electoral behaviour. *Aust. J. Psychol.* **30**, 255–262 (1978).
19. Lin, C., Adolphs, R. & Alvarez, R. M. Cultural effects on the association between election outcomes and face-based trait inferences. *PLoS ONE* **12**, e0180837 (2017).
20. Lin, C., Adolphs, R. & Alvarez, R. M. Inferring whether officials are corruptible from looking at their faces. *Psychol. Sci.* **29**, 1807–1823 (2018).
21. Oliviola, C. et al. First impressions and consumer mate preferences in online dating and speed-dating. *ACR N. Am. Adv.* **43**, 51–55 (2015).
22. Hamermesh, D. S. *Beauty Pays: Why Attractive People Are More Successful* (Princeton Univ. Press, 2011).
23. Blair, I. V., Judd, C. M. & Chapleau, K. M. The influence of Afrocentric facial features in criminal sentencing. *Psychol. Sci.* **15**, 674–679 (2004).
24. Caputi, M., Lecce, S., Pagnin, A. & Banerjee, R. Longitudinal effects of theory of mind on later peer relations: the role of prosocial behavior. *Dev. Psychol.* **48**, 257–270 (2012).
25. Naughtin, C. K. et al. Do implicit and explicit belief processing share neural substrates? *Hum. Brain Mapp.* **38**, 4760–4772 (2017).
26. Schuwerk, T., Vuori, M. & Sodian, B. Implicit and explicit theory of mind reasoning in autism spectrum disorders: the impact of experience. *Autism* **19**, 459–468 (2015).
27. Frith, C. D. & Frith, U. Implicit and explicit processes in social cognition. *Neuron* **60**, 503–510 (2008).
28. Roux, P., Smith, P., Passerieux, C. & Ramus, F. Preserved implicit mentalizing in schizophrenia despite poor explicit performance: evidence from eye tracking. *Sci. Rep.* **6**, 34728 (2016).
29. Schneider, D., Bayliss, A., Becker, S. & Dux, P. Eye movements reveal sustained implicit processing of others' mental states. *J. Exp. Psychol. Gen.* **141**, 433–438 (2011).
30. Blakemore, S.-J. & Decety, J. From the perception of action to the understanding of intention. *Nat. Rev. Neurosci.* **2**, 561–567 (2001).
31. Hamlin, J. K., Wynn, K. & Bloom, P. Social evaluation by preverbal infants. *Nature* **450**, 557–559 (2007).
32. Onishi, K. H. Do 15-month-old infants understand false beliefs? *Science* **308**, 255–258 (2005).
33. Chen, Z. & Whitney, D. Tracking the affective state of unseen persons. *Proc. Natl Acad. Sci. USA* **116**, 7559–7564 (2019).
34. Mitchell, P. Mentalizing in autism: interpreting facial expressions, following gaze, reading body language and inferring traits. *J. Educ. Sci. Psychol.* **3**, 111–120 (2013).
35. Barrett, L. F., Mesquita, B. & Gendron, M. Context in emotion perception. *Curr. Dir. Psychol. Sci.* **20**, 286–290 (2011).
36. Masuda, T. et al. Placing the face in context: cultural differences in the perception of facial emotion. *J. Pers. Soc. Psychol.* **94**, 365–381 (2008).
37. Oosterhof, N. N. & Todorov, A. Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion* **9**, 128–133 (2009).
38. Lin, C. & Thornton, M. Evidence for bidirectional causation between trait and mental state inferences. *J. Exp. Soc. Psychol.* **108**, 104495 (2023).
39. Thornton, M. A., Weaverdyck, M. E. & Tamir, D. I. The brain represents people as the mental states they habitually experience. *Nat. Commun.* **10**, 2291 (2019).

40. Uleman, J. S., Adil Saribay, S. & Gonzalez, C. M. Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.* **59**, 329–360 (2008).
41. Kleider-Offutt, H. M., Bond, A. D. & Hegerty, S. E. A. Black stereotypical features: when a face type can get you in trouble. *Curr. Dir. Psychol. Sci.* **26**, 28–33 (2017).
42. Oh, D., Dotsch, R., Porter, J. & Todorov, A. Gender biases in impressions from faces: empirical studies and computational models. *J. Exp. Psychol. Gen.* **149**, 323 (2020).
43. Blair, I., Judd, C. & Fallman, J. The automaticity of race and Afrocentric facial features in social judgments. *J. Pers. Soc. Psychol.* **87**, 763–778 (2005).
44. Blair, I. V., Judd, C. M., Sadler, M. S. & Jenkins, C. The role of Afrocentric features in person perception: judging by features and categories. *J. Pers. Soc. Psychol.* **83**, 5–25 (2002).
45. Gutsche, R. E., Cong, X., Pan, F., Sun, Y. & DeLoach, L. #DiminishingDiscrimination: the symbolic annihilation of race and racism in news hashtags of 'calling 911 on Black people'. *Journalism* **23**, 259–277 (2020).
46. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).
47. Knutson, B. Facial expressions of emotion influence interpersonal trait inferences. *J. Nonverbal Behav.* **20**, 165–182 (1996).
48. Said, C. P., Sebe, N. & Todorov, A. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* **9**, 260–264 (2009).
49. Thornton, M. A. & Tamir, D. I. People represent mental states in terms of rationality, social impact, and valence: validating the 3D Mind Model. *Cortex* **125**, 44–59 (2020).
50. Tamir, D. I., Thornton, M. A., Contreras, J. M. & Mitchell, J. P. Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proc. Natl Acad. Sci. USA* **113**, 194–199 (2016).
51. Wu, Y., Schulz, L. E., Frank, M. C. & Gweon, H. Emotion as information in early social learning. *Curr. Dir. Psychol. Sci.* **30**, 468–475 (2021).
52. Young, L. & Saxe, R. An fMRI investigation of spontaneous mental state inference for moral judgment. *J. Cogn. Neurosci.* **21**, 1396–1405 (2009).
53. Ekman, P. Facial expression and emotion. *Am. Psychol.* **48**, 376–379 (1993).
54. Lee, D. H. & Anderson, A. K. Reading what the mind thinks from how the eye sees. *Psychol. Sci.* **28**, 494–503 (2017).
55. Aviezer, H., Trope, Y. & Todorov, A. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**, 1225–1229 (2012).
56. Cannon, E. N. & Woodward, A. L. Infants generate goal-based action predictions. *Dev. Sci.* **15**, 292–298 (2012).
57. Elsner, B. & Adam, M. Infants' goal prediction for simple action events: the role of experience and agency cues. *Top. Cogn. Sci.* **13**, 45–62 (2021).
58. Ngo, N. & Isaacowitz, D. M. Use of context in emotion perception: the role of top-down control, cue type, and perceiver's age. *Emotion* **15**, 292–302 (2015).
59. Greenaway, K. H., Kalokerinos, E. K. & Williams, L. A. Context is everything (in emotion research). *Soc. Pers. Psychol. Compass* **12**, e12393 (2018).
60. John, O. P., Chaplin, W. E. & Goldberg, L. R. Conceptions of states and traits: dimensional attributes with ideals as prototypes. *J. Pers. Soc. Psychol.* **54**, 541–447 (1988).
61. Woo, B. M., Tan, E., Yuen, F. L. & Hamlin, J. K. Socially evaluative contexts facilitate mentalizing. *Trends Cogn. Sci.* **27**, 17–29 (2023).
62. Lewin, K. in *Field Theory in Social Science: Selected Theoretical Papers* (ed. Cartwright, D.) xx, 346 (Harpers, 1951).
63. Tamir, D. I. & Thornton, M. A. Modeling the predictive social mind. *Trends Cogn. Sci.* **22**, 201–212 (2018).
64. Zaki, J. Cue integration: a common framework for social cognition and physical perception. *Perspect. Psychol. Sci.* **8**, 296–312 (2013).
65. Lage-Castellanos, A., Valente, G., Formisano, E. & Martino, F. D. Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Comput. Biol.* **15**, e1006397 (2019).
66. Lin, C. & Adolphs, R. Trait impressions from faces depend on the goals of the perceiver. *Br. J. Psychol.* **114**, 501–503 (2023).
67. Lin, C., Bulls, L. S., Tepfer, L. J., Vyas, A. D. & Thornton, M. A. Advancing naturalistic affective science with deep learning. *Aff. Sci.* **4**, 550–562 (2023).
68. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).
69. Ntoimo, L. F. C. & Mutanda, N. in *Family Demography and Post-2015 Development Agenda in Africa* (ed. Odimegwu, C. O.) 147–169 (Springer International, 2020); https://doi.org/10.1007/978-3-030-14887-4_8
70. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago Face Database: a free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135 (2015).
71. DeBruine, L. & Jones, B. Face Research Lab London Set. *Figshare* https://doi.org/10.6084/m9.figshare.5047666.v3 (2017).
72. Nummenmaa, L., Hari, R., Hietanen, J. K. & Glerean, E. Maps of subjective feelings. *Proc. Natl Acad. Sci. USA* **115**, 9198–9203 (2018).
73. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl Acad. Sci. USA* **114**, E7900–E7909 (2017).
74. Clore, G. L., Ortony, A. & Foss, M. A. The psychological foundations of the affective lexicon. *J. Pers. Soc. Psychol.* **53**, 751–766 (1987).
75. Hepach, R., Kliemann, D., Grüneisen, S., Heekeren, H. R. & Dziobek, I. Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency—implications for social-cognitive tests and training tools. *Front. Psychol.* **2**, 266 (2011).
76. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. In Lee, L., Johnson, M. & Toutanova, K. (eds) *Transactions of the Association for Computational Linguistics*, 135–146 (MIT Press, 2017).
77. Davies, M. *The Corpus of Contemporary American English (COCA)* (English-Corpora.org, 2008); https://www.english-corpora.org/coca/
78. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2018).
79. Sutherland, C. A. M. et al. Social inferences from faces: ambient images generate a three-dimensional model. *Cognition* **127**, 105–118 (2013).
80. Chelnokova, O. et al. Rewards of beauty: the opioid system mediates social motivation in humans. *Mol. Psychiatry* **19**, 746–747 (2014).
81. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proc. EEE/CVF Conference on Computer Vision and Pattern Recognition* 4690–4699 (2019).
82. Marcus, G. Deep learning: a critical appraisal. Preprint at https://doi.org/10.48550/arXiv.1801.00631 (2018).
83. Marcus, G. F. Rethinking eliminative connectionism. *Cogn. Psychol.* **37**, 243–282 (1998).

84. Hehman, E., Xie, S. Y., Ofosu, E. K. & Nespoli, G. Assessing the point at which averages are stable: a tool illustrated in the context of person perception. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/2n6jq (2018).
85. Sutherland, C. A. M. et al. Facial first impressions across culture: data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Pers. Soc. Psychol. Bull.* **44**, 521–537 (2018).
86. Hehman, E., Sutherland, C. A. M., Flake, J. K. & Slepian, M. L. The unique contributions of perceiver and target characteristics in person perception. *J. Pers. Soc. Psychol.* **113**, 513–529 (2017).
87. Saucier, G. & Goldberg, L. R. Evidence for the big five in analyses of familiar English personality adjectives. *Eur. J. Pers.* **10**, 61–77 (1996).
88. Stolier, R. M., Hehman, E. & Freeman, J. B. Trait knowledge forms a common structure across social cognition. *Nat. Hum. Behav.* **4**, 361–371 (2020).
89. Zebrowitz, L. A. & Montepare, J. M. Social psychological face perception: why appearance matters. *Soc. Pers. Psychol. Compass* **2**, 1497 (2008).
90. Rule, N. O., Ambady, N. & Hallett, K. C. Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *J. Exp. Soc. Psychol.* **45**, 1245–1251 (2009).
91. Olivola, C. Y. & Todorov, A. Elected in 100 milliseconds: appearance-based trait inferences and voting. *J. Nonverbal Behav.* **34**, 83–110 (2010).
92. Todorov, A., Mende-Siedlecki, P. & Dotsch, R. Social judgments from faces. *Curr. Opin. Neurobiol.* **23**, 373–380 (2013).
93. Secord, P. F., Dukes, W. F. & Bevan, W. Personalities in faces: I. An experiment in social perceiving. *Genet. Psychol. Monogr.* **49**, 231–270 (1954).
94. Allport, G. W. & Odbert, H. S. Trait-names: a psycho-lexical study. *Psychol. Monogr.* **47**, i–171 (1936).
95. Walker, M., Schönborn, S., Greifeneder, R. & Vetter, T. The Basel Face Database: a validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLoS ONE* **13**, e0193190 (2018).
96. Lundqvist, D., Flykt, A. & Öhman, A. *The Karolinska Directed Emotional Faces (KDEF)* CD ROM (Karolinska Institute, Department of Clinical Neuroscience, Psychology Section, 1998).
97. Morrison, D., Wang, H., Hahn, A. C., Jones, B. C. & DeBruine, L. M. Predicting the reward value of faces and bodies from social perception. *PLoS ONE* **12**, e0185093 (2017).

Acknowledgements

We thank Y. Xu for helping with face image processing. This work was funded by NIMH (2P50 MH094258) and NSF (BCS-1840756). R.A. was supported in part by the Moonshot R&D JPMJMS2294 (overall programme manager: K. Matsumoto). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

C.L. and R.A. developed the study concept and designed the study. C.L. and U.K. prepared the experimental materials. M.A.T. and R.A. supervised the data collection and analyses. C.L. performed the data collection. C.L. and U.K. performed the data analyses. C.L. drafted the initial manuscript. All authors revised and reviewed the manuscript and approved the final manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-02059-4.

Correspondence and requests for materials should be addressed to Chujun Lin.

Peer review information *Nature Human Behaviour* thanks Eric Hehman, Agata Lapedriza and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.