



ARTICLE

<https://doi.org/10.1038/s41467-020-18566-7>

OPEN

Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings

Lou Safra ^{1,2,3✉}, Coralie Chevallier¹, Julie Grèzes¹ & Nicolas Baumard ^{2✉}

Social trust is linked to a host of positive societal outcomes, including improved economic performance, lower crime rates and more inclusive institutions. Yet, the origins of trust remain elusive, partly because social trust is difficult to document in time. Building on recent advances in social cognition, we design an algorithm to automatically generate trustworthiness evaluations for the facial action units (smile, eye brows, etc.) of European portraits in large historical databases. Our results show that trustworthiness in portraits increased over the period 1500–2000 paralleling the decline of interpersonal violence and the rise of democratic values observed in Western Europe. Further analyses suggest that this rise of trustworthiness displays is associated with increased living standards.

¹Laboratoire de Neurosciences Cognitives, Département d'études cognitives, ENS, PSL, Research University, INSERM, Paris, France. ²Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL Research University, CNRS, Paris, France. ³Sciences Po, CEVIPOF, CNRS, Paris, France.
✉email: lou.safra@sciencespo.fr; nicolas.baumard@ens.fr

A number of historical observations suggest that social trust rose steadily in Europe from the early modern period onwards: religious tolerance increased, witch hunts abated, honor killings and revenge lost their appeal and intellectual freedom became a central value of modern countries^{1,2}. Historians have used a range of cues to document this process: etiquette manuals, registries of friendly societies, or legal changes^{1,3,4}. However, quantitative evidence is scarce and progress in the history of mentalities has been limited by the paucity of tools to capture people's extinct mental life. Quite obviously, we cannot go back in time and ask people to fill out questionnaires or play economic games^{5–7} but we still have access to what their minds produced: books, songs, paintings, sculptures, etc. These cultural artefacts are the remnants of people's past psychologies and can function as cognitive fossils of extinct mentalities and social preferences. Recent work has indeed shown that people's preferences in various areas of social cognition are reflected in cultural artefacts: Costa and Corazza⁸ demonstrated that the people's preference for friendly-looking faces leads painters to exaggerate “neotenic” features in their portraits (big eyes or round faces). Similarly, Morin⁹ has shown that direct-gaze Renaissance portraits are more popular than averted-gaze portraits. Fictions, such as romance novels¹⁰, TV shows¹¹, epic poems¹² or tragedies¹³, are all consistently aligned with humans' universal interest for information related to mating, commitment and status competition for reviews and discussions, see refs. 14,15. These shifts in cultural artefacts reveal global changes in mentalities, reflecting the preference of the sitter, the artist and the audience altogether.

Portraits are particularly promising to document and quantify the level of trust over time. Experimental work have revealed that specific facial features, such as a smiling mouth or wider eyes, are consistently recognized as cues of trustworthiness across individuals and cultures^{16–21}. In this paper, we capitalize on this large empirical literature to build an algorithm that estimates trustworthiness based on a pre-identified set of facial characteristics²². More precisely, we apply recent machine-learning methods to extract quantitative information about the evolution of social cues contained in portraits. The algorithm generates automatic human-like trustworthiness ratings on portraits based on the muscle contractions (facial action units) detected in facial displays using the open software OpenFace²³. This algorithm was trained on avatars controlled for trustworthiness and optimized using a random forest procedure (see Supplementary Methods for more details). To assess the generalizability of our model, we then tested its validity on four databases of natural faces rated by real participants. We first demonstrated that the algorithm produced trustworthiness ratings that were aligned with those produced by human participants in all four controlled databases. Another validation method would have been to also measure the correlation between the estimated trustworthiness of the historical portraits calculated by our algorithm and the evaluations of the actual participants on these paintings. This other method has the major advantage of providing a direct test of the reliability of our algorithm. However, since participant evaluations may be influenced by historical cues not relevant to trustworthiness (such as the sitter's outfit or the painting style) that may bias these evaluations so that older portraits are perceived as less trustworthy, this method of validation is limited. Therefore, we chose to assess the validity and generalizability of our model independently of idiosyncratic biases of participants by relying on well-known effects in the literature, i.e., the effect of emotion, age, gender, and head orientation on facial evaluations.

We thus checked that the algorithm was susceptible to the same biases as humans, i.e., rating younger, feminine, and happy faces as more trustworthy. Third, we checked that the output of the algorithm was robust to variations in head orientation^{21,24}

(see Supplementary Methods for the results). We then replicated all these findings outside well-controlled databases by analyzing all the images (photographs and paintings) obtained from a Google image search for ‘women portraits’ vs ‘male portraits’ ($N = 633$; trustworthiness: $t(632) = 7.89$, $p < 0.001$; dominance: $t(632) = -11.79$, $p < 0.001$). This validation method provides evidence of the ability of our algorithm to produce human-like face evaluations on a large range of images (i.e., controlled photographs, natural photographs and paintings).

Results

Trustworthiness displays in portraits increased throughout history. To assess the evolution of trustworthiness displays in history, we first analyzed the paintings of the National Portrait Gallery (Fig. 1a), the largest online database of historical portraits (analyzed $N = 1962$ English portraits from 1505 to 2016). Because perceived trustworthiness is correlated with perceived dominance²⁴, all the analyses were controlled for dominance. In line with historical work, we found a significant increase of trustworthiness displays with time ($b = 0.14 \pm 0.02$, $z = 7.49$, $p < 0.001$; Table 1; time coded such as one unit corresponds to 100 years, \pm corresponds to standard errors to the mean; Figs. 1b and 2a), suggesting that the value of interpersonal trust increased from the 16th to the 20th century. We then replicated our findings on the Web Gallery of Art, an important fine art repository ($N = 4106$ portraits) spanning 19 Western European countries seven centuries (1360–1918) and found a significant increase in trustworthiness displays with time ($b = 0.07 \pm 0.01$, $z = 5.33$, $p < 0.001$; Table 1; Fig. 2b). Overall, these results are consistent with more qualitative works documenting a so-called ‘Smile Revolution’²⁵ and a rise of prosocial displays in paintings and in novels²⁶. It is worth noting, however, that the historical increase in trustworthiness observed in our datasets parallels the rise of liberal values such as religious tolerance, political freedom and democracy^{2,27,28}.

Whether such increased trustworthiness in portraits parallels an actual shift in social trust remains an open question. To assess the validity of this assumption, we applied our algorithm to selfies posted on Instagram in six cities around the world in 2013 (Bangkok, Berlin, London, Moscow, New York and Sao Paulo; SelfieCity database, pictured analyzed $N = 2277$ ²⁹), we found that people located in places where interpersonal trust and cooperation are higher (as assessed in the European and World Value Surveys^{30,31}) displayed higher levels of trustworthiness in their selfies (cooperation level: $b = 0.13 \pm 0.03$, $z = 3.67$, $p < 0.001$; trust level: $b = 0.81 \pm 0.23$, $z = 3.50$, $p < 0.001$; \pm corresponds to standard errors to the mean; Supplementary Figure 6). Together, this suggests that the display of trustworthiness in portraits can indeed be used as a reliable proxy of the level of social trust in individuals' environment^{32,33}.

Trustworthiness displays in portraits increased with affluence.

Another open question is that of the potential predictors of trustworthiness fluctuations in social displays. We first examined the role of resources. Trust can indeed be construed as an investment in social interactions with potential benefits (in the event of cooperation) and also potential losses (in the event of defection). Because losses have more dramatic effects for poorer individuals, individuals with lower resources are arguably more exposed by exploitation risks and should therefore have lower levels of social trust³⁴. In line with this reasoning, international surveys show a strong association between resources and social trust^{35–38}. Moving beyond correlations, economists have recently demonstrated that childhood resources had a causal impact on adult trust levels using exogenous variations in caloric rationing in post WW2 Germany³⁹.

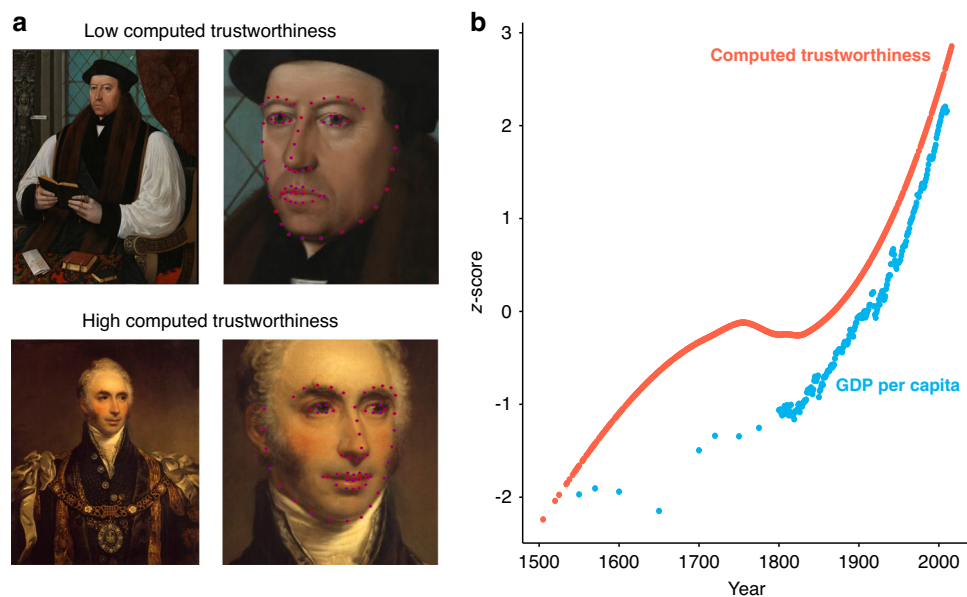


Fig. 1 Evolution of trustworthiness displays in England across time. **a** Example of faces detected in portraits from the National Portrait Gallery and estimated as lowly trustworthy (top; Thomas Cranmer by Gerlach Flicke, 1545-1546, NPG 535 All rights reserved © National Portrait Gallery, London) and highly trustworthy (bottom; Sir Matthew Wood by Arthur William Devis, 1815-1816, NPG 1481 All rights reserved © National Portrait Gallery). **b** Evolution of displays of trustworthiness in the National Portrait Gallery (modeled trustworthiness value adjusted for dominance) and GDP per capita in England. Source data are provided as raw data and scripts on the online depository.

This is particularly relevant in light of the fact that the Middle Ages and the early Modern Period were periods of prolonged economic growth for Europe in general and England in particular^{40,41}. We thus tested whether higher GDP per capita was associated with the rise of trustworthiness in portraits. Our analysis of the National Portraits Gallery database revealed an association between higher levels of affluence and higher levels of trustworthiness displays between the 16th and the 21st centuries ($b = 0.03 \pm 0.01$, $z = 7.13$, $p < 0.001$; Table 1; Fig. 2c), even after adjusting for a monotonous effect of time ($b = 0.02 \pm 0.01$, $z = 3.16$, $p = 0.002$; Table 1). Crucially, GDP per capita accounted for the evolution of trustworthiness displays better than a monotonous effect of time (Bayes Factor: 3.38), which suggests that the observed evolution of trustworthiness displays cannot be reduced to a simple cultural accumulation that would have led to the development of painting techniques making sitters look more trustworthy. We then sought to replicate this result in the Web Gallery of Art database and also found a significant positive association between GDP per capita and trustworthiness displays ($b = 0.09 \pm 0.03$, $z = 3.16$, $p = 0.002$; Table 1; Fig. 2d). This association was robust to adjusting for a monotonous increase of trust displays over time ($b = 0.07 \pm 0.04$, $z = 1.98$, $p = 0.048$; Table 1). Again, the model including GDP per capita provided a better account of the variations of trust displays than time alone (Bayes Factor: 130.16).

Institutional change is another possible predictor of increased trust. The establishment of more democratic, more inclusive and more egalitarian institutions might indeed have created a climate of trust and tolerance^{42,43}. We tested this idea by measuring the association between displays of trustworthiness in paintings and political democratization using the Polity2 index (a composite measure of institutionalized democracy and autocracy available from 1800, see Supplementary Methods). Although a significant association was found between these two variables in the National Portraits Gallery ($b = 0.03 \pm 0.01$, $z = 5.24$, $p < 0.001$), this effect was not robust to the inclusion of time as covariate ($b = -0.01 \pm 0.01$, $z = -0.50$, $p > 0.250$) and the evolution of trustworthiness

displays was better explained by GDP per capita than by changes in the institutions (Bayes Factor: 2.75). Moreover, the positive association between more democratic institutions and higher trustworthiness displays was not replicated in the Web Gallery of Art sample ($b = -0.01 \pm 0.01$, $z = -1.96$, $p = 0.051$; with time as a covariate: $b = -0.01 \pm 0.01$, $z = -0.96$, $p > 0.250$; Bayes Factor of the GDP per capita model compared to the democratic institutions model: 6.16).

Changes in affluence precede changes in trustworthiness displays in portraits. Demonstrating that the association between GDP and the rise of trustworthiness is causal would of course require additional data. Based on our dataset however, we were able to investigate the dynamics of these historical changes by running time-lag analyses on trustworthiness displays and GDP per capita. We found that changes in GDP per capita predicted future changes in trustworthiness displays in the National Portraits Gallery two decades later ($F(40,1) = 12.38$, $p = 0.001$) while changes in political institutions did not ($F(15,1) = 0.11$, $p > 0.250$). The effect of GDP per capita on trustworthiness displays was generalizable to the other European countries (Web Gallery of Art sample, effect of GDP 20 years before on trustworthiness displays: $X(1) = 6.42$, $p = 0.011$; Institutions 20 years before: $X(1) = 0.81$, $p > 0.250$). Importantly, changes in trustworthiness displays did not predict future changes in GDP per capita either in the National Portraits Gallery sample ($F(41,1) = 0.76$, $p > 0.250$) or in the Web Gallery of Art dataset ($X(1) = 2.02$, $p = 0.155$), which suggests that changes in GDP per capita may have preceded changes in trustworthiness displays in this dataset. This conclusion is consistent with other works emphasizing the importance of economic growth and psychological changes in history⁴⁴⁻⁴⁶.

Discussion

To conclude, our analyses—replicated across two independent fine arts databases—reveals that trustworthiness displays increased in early modern period portraits and are suggestive

Table 1 Effect of time, GDP per capita and democratization on the portraits of National Portrait Gallery and the Web Gallery of Art.

Time only	Affluence only		Time + Affluence		Democratization only		Time + Democratization	
	National Portraits Gallery	Web Gallery of Art	National Portraits Gallery	Web Gallery of Art	National Portraits Gallery	Web Gallery of Art	National Portraits Gallery	Web Gallery of Art
Year	0.14 ± 0.02 z = 7.49 p < 0.001	0.07 ± 0.01 z = 5.33 p < 0.001			0.08 ± 0.03 z = 3.17 p = 0.002	0.06 ± 0.02 z = 2.87 p = 0.007	0.32 ± 0.11 z = 2.86 p = 0.004	-0.13 ± 0.14 z = -0.98 p > 0.250
GDP per capita			0.03 ± 0.00 z = 7.13 p < 0.001	0.09 ± 0.03 z = 3.16 p = 0.002	0.02 ± 0.01 z = 3.16 p = 0.002	0.07 ± 0.04 z = 1.98 p = 0.048		
Democracy index							0.03 ± 0.01 z = 5.24 p < 0.001	-0.01 ± 0.01 z = -1.96 p = 0.051
Dominance	-0.79 ± 0.02 z = -40.74 p < 0.001	-0.74 ± 0.01 z = -56.58 p < 0.001	-0.78 ± 0.02 z = -40.10 p < 0.001	-0.75 ± 0.02 z = -46.29 p < 0.001	-0.78 ± 0.02 z = -40.30 p < 0.001	-0.74 ± 0.02 z = -46.05 p < 0.001	-0.77 ± 0.03 z = -30.76 p < 0.001	-0.71 ± 0.04 z = 20.17 p < 0.001
Gender	0.32 ± 0.06 z = 5.64 p < 0.001	0.31 ± 0.03 z = 10.76 p < 0.001	0.29 ± 0.06 z = 5.01 p < 0.001	0.30 ± 0.04 z = 8.31 p < 0.001	0.30 ± 0.06 z = 5.10 p < 0.001	0.29 ± 0.04 z = 7.98 p < 0.001	0.28 ± 0.08 z = 3.61 p < 0.001	0.25 ± 0.07 z = 3.30 p = 0.001
Age	-0.00 ± 0.00 z = -2.03 p = 0.043		-0.00 ± 0.00 z = -1.88 p = 0.060		-0.00 ± 0.00 z = -2.26 p = 0.024		0.00 ± 0.00 z = 0.48 p > 0.250	-0.00 ± 0.00 z = -0.17 p > 0.250
Sample (N)	1962	4106	1943	2706	1943	2706	1115	565

The first line corresponds to the regression coefficient with their associated standard error to the mean (mean ± s.e.m.). Results in bold corresponds to statistically significant effects of the variables of interest. The upper part of the table presents the effects of the variables of interest (time, affluence and democratization), while the lower part presents the effects of the control variables (dominance, gender and age). All the tests are two-sided. Following APA's recommendations, exact p-values are provided for p-values between 0.001 and 0.250. Source data are provided as raw data and scripts on the online depository.

of an actual shift in social trust over the period (although differences across countries might have persisted over the period, see refs. 5–7). This cultural shift is more strongly associated with GDP per capita than institutional change. These findings complement existing qualitative historical accounts and demonstrate how insights from cognitive sciences can enrich our understanding of cultural evolution.

Methods

Construction of an algorithm for modeling trustworthiness and dominance evaluations. We built a model that automatically extracts trustworthiness and dominance evaluations from the all the facial action units detected by the OpenFace algorithm (i.e., both dichotomous and continuous estimations; OpenFace version 1.01 using OpenCV 3.3.0⁴⁷). To do so, we extracted the facial action units of five sets of avatars previously generated with Facegen and controlled for dominance, for trustworthiness or for both (Supplementary Fig. 1)⁴⁸. Each avatar is generated from an initial face and manipulated to either express a specific level of dominance, trustworthiness or both based on the model developed by Oosterhof and Todorov²⁴. These avatar faces have been shown to successfully elicit ratings of dominance and trustworthiness in participants^{48–50}. Thus, compared to participants' ratings on photographs that may be sensitive to the participants characteristics and to experimental protocol factors (such as the type of scale used to give the ratings), using avatars allow us to have well-validated sets of faces to train our model. These sets of avatars correspond to all the existing and available validated avatars controlled for trustworthiness or dominance and generated by Facegen.

3% of the faces were excluded from the modeling process for not having been accurately detected by OpenFace. The total sample of avatar faces were then split in a training sample (80% of the faces) and a test sample (20% of the faces). Importantly, the percentage of avatars coming from each avatar set was equal in the training and test samples for both trustworthiness and dominance (Trustworthiness: $\chi^2(2) = 0.02$, $p > 0.250$; Dominance: $\chi^2(2) = 0.01$, $p > 0.250$).

To determine which type of algorithm (linear model, random forest model from the RandomForest R package⁵¹—Breiman's random forest algorithm⁵²—or support vector model either linear or radial from the kernlab R package⁵³) would provide the most accurate evaluations, we ran a repeated 20-folds cross-validation (five repetitions) on the training test of each of these models separately for dominance and trustworthiness using caret R package⁵⁴. Each model's hyperparameters were optimized using a random search. The hyperparameters optimized for each model are presented in Supplementary Table 1. This analysis revealed significantly better performance for the random forest model than for the linear model and the linear SVM model in terms of mean absolute error, root square mean error and r-squared and was and better than, for the trustworthiness model, and similar to, for the dominance model, the radial SVM model (Supplementary Table 1). For both trustworthiness and dominance, the optimal m_{try} hyperparameter of the random forest models was found to be equal to 9, corresponding to setting the number of variables to consider at each tree to 9. We then tested the predictions of the random forest model with this optimal hyperparameter obtained by cross-validation on our trustworthiness and

dominance test sets. This test revealed a high performance of the model (trustworthiness: $r = 0.85 \pm 0.5$, $t(75) = 14.17$, $p < 0.001$; dominance: $r = 0.86 \pm 0.05$, $t(75) = 14.72$, $p < 0.001$; Supplementary Fig. 2; all the reported statistical tests are two-sided).

Validation of the algorithm for modeling trustworthiness and dominance evaluations. To assess the accuracy our trustworthiness and our dominance generator algorithm, we tested their predictions on four different face databases: the Karolinska database ($N = 70$ distinct faces)⁵⁵, the Oslo Face database ($N = 185$ distinct faces)⁵⁶, the Chicago database ($N = 520$ distinct faces)⁵⁷ and the FEI Face database ($N = 520$ distinct faces)⁵⁸. Given that our model was optimized on avatar faces, comparing our model's prediction to real participants ratings in a second step allows us to assess whether our model would give overall coherent ratings with those of real human beings. Our first analysis confirmed the significant correlation of the modeled trustworthiness and dominance estimates with the actual participants' ratings of trustworthiness and dominance ratings on the faces from these databases (except the FEI Face database which did not provide subjective ratings; Supplementary Figure 3). We found significant correlations for both trustworthiness and dominance estimates (trustworthiness: $r = 0.22$, $p < 0.001$, dominance: $r = 0.16$, $p < 0.001$ — $N = 768$ for each correlation, to not artificially increase the statistical power of this analysis only the neutral and facing version of the faces were used for these correlations), confirming that our model gave trustworthiness and dominance estimates that are coherent with real participants' evaluations on these traits.

Going one step further, we assessed whether our modeled trustworthiness and dominance were able to reproduce classical findings in social cognition on perceived trustworthiness and dominance, namely: gender effect (females appear as less dominant and more trustworthy than males; trustworthiness: real effect: $t(768) = 7.94$, $p < 0.00$; recovered effect: $t(972) = 2.67$, $p = 0.008$; dominance: real effect: $t(769) = -7.80$, $p < 0.001$; recovered effect: $t(972) = -3.63$, $p < 0.001$; Supplementary Fig. 4A, B), emotion effects (angry faces appear as more dominant than neutral faces: $t(167) = 9.42$, $p < 0.001$; happy faces appear as more trustworthy than neutral and angry faces: $t(167) = 10.64$, $p < 0.001$; Supplementary Fig. 4C, D), head orientation effects (trustworthiness and dominance evaluations for a unique identity are correlated across head orientations: trustworthiness: $r = 0.29$, $t(1500) = 11.51$, $p < 0.001$; dominance: $r = 0.34$, $t(1500) = 13.79$, $p < 0.001$; Supplementary Fig. 4E, F) and age effect (older adults appear as more dominant and less trustworthy than younger adults: trustworthiness: real effect: $r = -0.12$, $t(518) = -2.75$, $p = 0.006$; recovered effect: $r = -0.12$, $t(518) = -2.68$, $p = 0.008$; dominance: real effect: $r = 0.32$, $t(518) = 7.72$, $p < 0.001$; recovered effect: $r = 0.16$, $t(518) = 3.70$, $p < 0.001$; Supplementary Fig. 4G, H)^{21,24,59,60}.

All these effects were replicated with the modeled trustworthiness and dominance evaluations. In addition, although dominance and trustworthiness were modeled independently, we also replicated the classical correlation between these two traits, further suggesting the importance of investigating trustworthiness conjointly with dominance (effect on participants' ratings: $r = -0.21$, $t(768) = -5.81$, $p < 0.001$; recovered effect: $r = -0.46$, $t(768) = -14.30$, $p < 0.001$).

Importantly, we further validated our model by replicating the gender effect on all the portraits extracted from a Google image search for 'women portraits' vs 'male portraits' containing both pictures and paintings ($N = 633$; trustworthiness: $t(632) = 7.89$, $p < 0.001$; dominance: $t(632) = -11.79$, $p < 0.001$; Supplementary

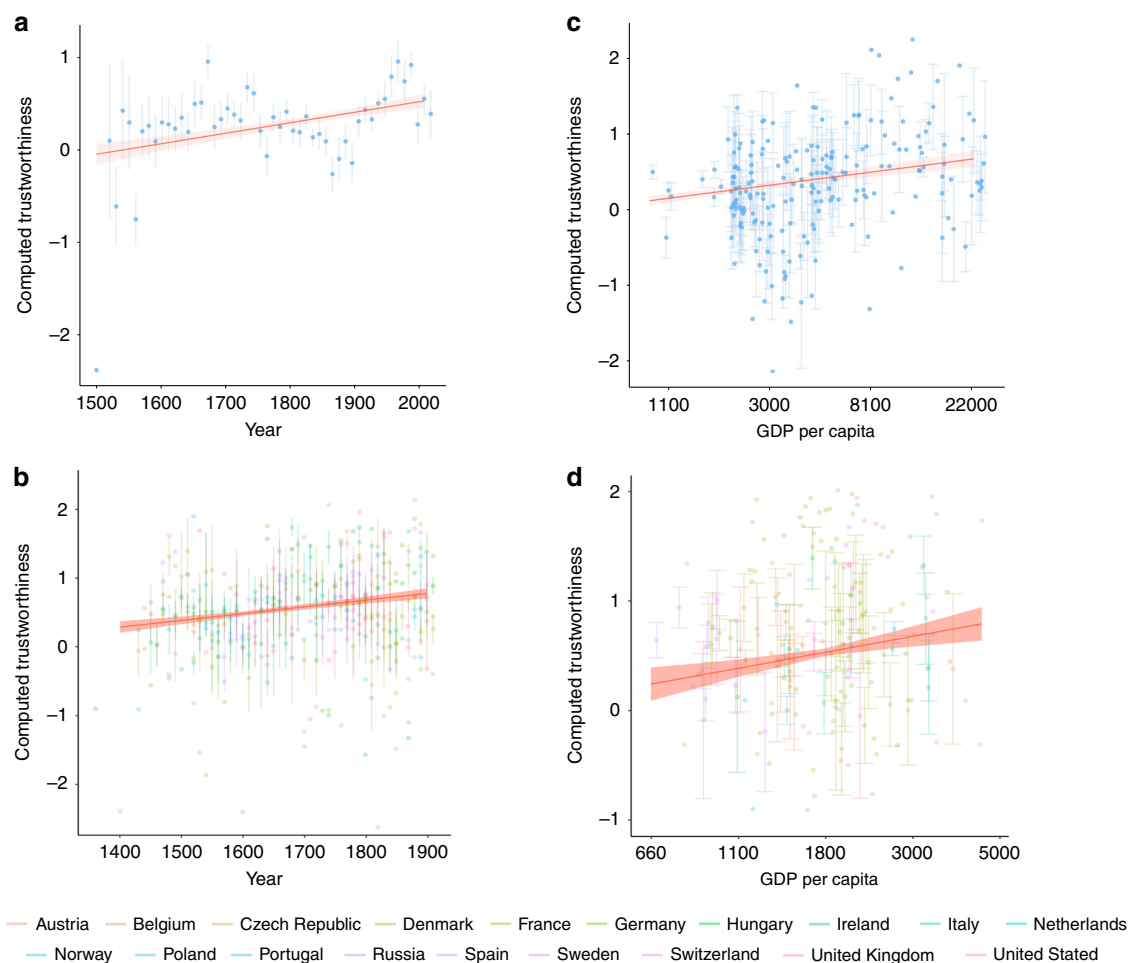


Fig. 2 Effect of time and affluence on trustworthiness displays across time. Time was associated with an increase of trustworthiness displays in both the National Portrait Gallery (**a**) and the Web Gallery of Art (**b**)—data are aggregated by decades). Increased GDP per capita predicted increased trustworthiness displays better than time only-models both in the National Portrait Gallery (**c**) and the Web Gallery of Art (**d**). Data are represented as mean values, error bars represent standard error to the means, the red line corresponds to the estimated effect in the regression adjusting for gender, age (for the National Portrait Gallery only) and dominance, the shaded area represents the standard error to the mean of these effects. Source data are provided as raw data and scripts on the online depository.

Fig. 5A, B). We also replicated the gender effect on the official portrait pictures of US representatives ($N = 419$; gender: trustworthiness: $t(417) = 2.20$, $p = 0.028$, dominance: $t(417) = -4.74$, $p < 0.001$; Supplementary Fig. 5C, D). Importantly, we also replicated the effect found in the literature that conservative representative appear more dominant than democrat representatives ($t(417) = -2.59$, $p = 0.009$; Supplementary Fig. 5E).

Testing the relationship between interpersonal trust and portrait Selfies'

trustworthiness. We tested whether displayed trustworthiness could be used as a proxy for interpersonal trust. To do so, we analyzed the Selficity database²⁹ which includes 3230 selfies posted on Instagram in 2013 from six cities across the world (Bangkok, Berlin, London, Moscow, New York and Sao Paulo; analyzable images: $N = 2277^{29}$).

The identified faces were then individually analyzed by two independent raters who were asked to evaluate, for each picture, the alignment of the OpenFace's face identification points compared to the real face's contours (coded as 0 or 1). The sum of these goodness of fit was then used as weights for the analyses. Therefore, only faces for which the two raters agreed that they were not well detected were removed from the analyses. Faces for which the two raters agreed on their good detection had a weight of 2 in the analyses, and those on which they disagreed had a weight of 1.

Importantly, a preliminary analysis confirmed that the trustworthiness computed with our algorithm recovered the gender effect documented in the literature in this image sample too (trustworthiness: $t(2275) = 13.80$, $p < 0.001$; dominance: $t(2275) = -10.18$, $p < 0.001$; Supplementary Fig. 6A, B). Extracted trustworthiness was analyzed using a linear model taking the sitter's gender, the city longitude and latitude and the sitter's dominance as control variables. The effect of two measures of interpersonal trust were used to assess the link between

displayed trustworthiness and interpersonal trust, extracted from the European and World Value Surveys^{30,31} general social trust question ('most people can be trusted or you cannot be too careful'; Supplementary Fig. 6C) and the sum of five questions bearing on cooperation ('how acceptable is claiming government benefits', 'avoiding a fare on public transport', 'cheating on taxes, keeping money that you have found', 'failing to report damage you've done accidentally to a parked vehicle'; Supplementary Fig. 6D). As the Selficity database is constituted of pictures posted online in 2013, for each country, the most recent vague of the European or World Value Survey was taken (i.e., 2008 for Russia, 2009 for Great Britain, 2011 for the United States, 2013 for Thailand and Germany, and 2014 for Brazil). In line with our hypotheses, people located in places where interpersonal trust and cooperation are higher, displayed higher levels of trustworthiness in their selfies (cooperation level: $b = 0.13 \pm 0.03$, $z = 3.67$, $p < 0.001$; trust level: $b = 0.81 \pm 0.23$, $z = 3.50$, $p < 0.001$; Supplementary Fig. 6C, D).

Analysis of the National portrait gallery. All the paintings of the National Portrait Gallery were downloaded in high resolution from the NPG.uk website. Information about the sitter's age at the date of the portrait were also automatically collected. Portraits' date were automatically coded following the method detailed in the table below (Supplementary Table 2). These values were divided by 100 for the regression analyses such that 1 time unit corresponds 100 years. All the portraits were processed using the OpenFace algorithm. The identified faces were then individually analyzed by three independent raters who were asked to evaluate the model's goodness of fit based on the points' position compared to the real face's contours (coded as 0 or 1). In addition, raters had to note the gender of the sitter. The classification based on the goodness of fit was then used as weights for the analyses. Importantly, in order to ensure that the portraits accurately reflected the level of trust at the time the portrait was painted and to avoid re-interpretation of

past historical figures, only portraits painted during the sitter's lifetime were analyzed (number of analyzed portraits: $N = 1962$). Portraits' dates were automatically coded following the nomenclature reported in Supplementary Table 2.

Level of affluence (countries' GDP per capita) was provided by the Maddison Project⁶¹ and political democratization (Polity 2 index) was provided by the Polity IV project⁶². For the UK, these data exist from 1500 to 2000 for GDP per capita and yearly data from 1800 to 2013 for the democratization index.

In order to keep a maximal temporal resolution, missing values in the GDP per capita and Polity2 indices were completed using the closest previous value, except for the time-lag analyses in which no imputation was made. A total of 1943 data points were included in the analyses looking at the effect of GDP per capita. A total of 1115 data points were included in the analyses looking at the effect of Polity2. Paintings were analyzed using individual linear models (each painting corresponding to one data point), taking the sitter's gender, age and level of dominance as control variables. Bayes factor analyses were conducted using the BIC approximation, which approximates Bayes factors computed under the unit information prior⁶³.

Finally, time-lag analyses were conducted to analyze the temporal dynamics between trustworthiness, GDP per capita and democratization. To do so, data were averaged by decades and analyzed at the aggregated level. The model on trustworthiness at decade d included the simultaneous level of dominance at decade d , the linear effect of the time, the delayed levels of trustworthiness and dominance at $d-2$, and the level of GDP per capita or democratization at $d-2$. On the other hand, models of GDP per capita or democratization included the linear effect of time, the delayed level of GDP at $d-2$ and the delayed levels of trustworthiness and of dominance at $d-2$. For each variable, the model with the delayed variable of interest (GDP per capita or democratization for the trustworthiness models, and trustworthiness for the models on GDP per capita and democratization) were compared with the models in which this variable was removed. Finally, in order to assess the robustness of our effects, we also tested the same models with a delay of one decade instead of two decades (Supplementary Table 3).

Web gallery of art. Data from the Web Gallery of Art (WGA) were analyzed in a similar way as the paintings from the NPG. To better account that the portraits actually reflected the sitter's willingness to display trustworthiness traits, paintings were geocoded using the painter's place of activity at the time of the painting. This geo-coding resulted in 19 countries with paintings ranging from 1360 to 1918. As previously, two independent raters categorized the quality of detection of the faces and these evaluations were used as weights in the linear regression (number of analyzed portraits: $N = 4106$). As for the National Portrait Gallery, the missing levels of affluence and democratization were completed using the previous complete value. The same models as previously were used except that a random effect was included to take the localization of the paintings into account. This resulted, for the analysis of the effect of GDP per capita and democratization in two-level mixed models, taking each painting as an individual data point clustered by the country of production. Correspondingly, for time-lag analyses, we use two-level mixed models but with data aggregated by decades.

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

Data availability

All data analyzed in the main text and in the supplementary materials are accessible online [https://osf.io/j68xu/?view_only=61995a283e9f4c55b43c9f31d6bd1e97] except the World Value Survey [<http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp>] and the European Value Survey [<https://dbk.gesis.org/dbksearch/SDesc2.asp?no=4804&db=E>] which are analyzed in the Selfcity study and are freely downloadable. The source data underlying all the Figures, Tables, Supplementary Figures and Supplementary Tables are provided in the online scripts and data.

A reporting summary for this Article is available as a Supplementary Information file. The images analyzed in this article are available at: Prof. Todorov avatars: <http://tlab.princeton.edu>; Chicago Face database [<https://chicagofaces.org/default/>]; Oslo Face database [<https://sirileknes.com/oslo-face-database/>]; Karolinska Face database [<https://www.kdef.se/index.html>]; FEI Face database [<https://fei.edu.br/~cet/facedatabase.html>]; House of Representative official portraits [<https://www.house.gov/representatives>]; Selfcity [<http://selfcity.net>]; National Portrait Gallery [<https://www.npg.org.uk>]; Web Gallery of Art [<https://www.wga.hu>].

Code availability

All analyses scripts presented in the main text and in the supplementary materials are accessible online [https://osf.io/j68xu/?view_only=61995a283e9f4c55b43c9f31d6bd1e97].

Received: 19 May 2019; Accepted: 10 August 2020;

Published online: 22 September 2020

References

- McCloskey, D. N. *Bourgeois equality: how ideas, not capital or institutions, enriched the world* (University of Chicago Press, 2016).
- Pinker, S. *The better angels of our nature: the decline of violence in history and its causes* (Penguin, UK, 2011).
- Clark, P. *British clubs and societies 1580-1800: the origins of an associational world* (OUP Oxford, 2000).
- Sunderland, D. *Social capital, trust and the industrial revolution: 1780-1880* (Routledge, 2007).
- Putnam, R. D., Leonardi, R. & Nanetti, R. Y. *Making democracy work: civic traditions in modern Italy* (Princeton University Press, 1994).
- Uslaner, E. M. *The moral foundations of trust* (Cambridge University Press, 2002).
- Knack, S. & Keefer, P. Does social capital have an economic payoff? A cross-country investigation. *Q. J. Econ.* **112**, 1251-1288 (1997).
- Costa, M. & Corazza, L. Aesthetic phenomena as supernormal stimuli: the case of eye, lip, and lower-face size and roundness in artistic portraits. *Perception* **35**, 229-246 (2006).
- Morin, O. How portraits turned their eyes upon us: visual preferences and demographic change in cultural evolution. *Evol. Hum. Behav.* **34**, 222-229 (2013).
- Salmon, C. The pop culture of sex: an evolutionary window on the worlds of pornography and romance. *Rev. Gen. Psychol.* **16**, 152-160 (2012).
- Fisher, M. L. Why who shot J. R. Matters: Dallas as the pinnacle of human evolutionary television. *Rev. Gen. Psychol.* **16**, 200-207 (2012).
- Gottschall, J. The rape of troy: evolution, violence, and the World of Homer (Cambridge University Press, 2008).
- Nettle, D. The wheel of fire and the mating game: explaining the origins of tragedy and comedy. *J. Cult. Evol. Psychol.* **3**, 39-56 (2005).
- Gottschall, J., Wilson, E. O., Wilson, D. S., & Crews, F. *The literary animal: Evolution and the nature of narrative*. (Northwestern University Press, 2005).
- Pinker, S. *The stuff of thought: language as a window into human nature* (Penguin, 2007).
- Walker, M., Jiang, F., Vetter, T. & Sczesny, S. Universals and cultural differences in forming personality trait judgments from faces. *Soc. Psychol. Personal Sci.* **2**, 609-617 (2011).
- Xu et al. Similarities and differences in Chinese and Caucasian adults' use of facial cues for trustworthiness judgments. *PLoS ONE* **7**, e34859 (2012).
- Bente et al. Cultures of trust: effects of avatar faces and reputation scores on German and Arab players in an online trust-game. *PLoS ONE* **9**, e98297 (2014).
- 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).
- Birkás, B., Dzhelyova, M., Lábadi, B., Bereczkei, T. & Perrett, D. I. Cross-cultural perception of trustworthiness: the effect of ethnicity features on evaluation of faces' observed trustworthiness across four samples. *Personal Individ. Differ.* **69**, 56-61 (2014).
- 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).
- Sofer et al. For your local eyes only: culture-specific face typicality influences perceptions of trustworthiness. *Perception* **46**, 914-928 (2017).
- Baltrušaitis, T., Robinson, P. & Morency, L. P. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV)*, 2016 IEEE Winter Conference on 1-10 (IEEE, 2016).
- Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl Acad. Sci. USA* **105**, 11087-11092 (2008).
- Jones, C. *The Smile Revolution in Eighteenth-Century Paris*. (Oxford University Press, Oxford, 2014).
- Schama, S. *Citizens: a chronicle of the French Revolution*. (Penguin, UK, 2004).
- McCloskey, D. N. *Bourgeois Equality: How Ideas, Not Capital or Institutions, Enriched the World 3* (University of Chicago Press, Chicago, 2016).
- Mokyr, J. *A culture of growth: the origins of the modern economy* (Princeton University Press, 2016).
- Tifentale, A. & Manovich, L. Selfcity: Exploring Photography and Self-Fashioning in Social Media. in *Postdigital Aesthetics: Art, Computation and Design* (eds. Berry, D. M. & Dieter, M.), 109-122 (Palgrave Macmillan UK, 2015). https://doi.org/10.1057/9781137437204_9.
- EVS (2015): European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008). *GESIS Data Archive, Cologne. ZA4804 Data file Version 3.0.0*.
- Inglehart, R. et al. World Values Survey: Round Six - Country-Pooled Datafile Version: <http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>. (2014).
- Tingley, D. Face-off: facial features and strategic choice. *Polit. Psychol.* **35**, 35-55 (2014).
- Mehu, M., Grammer, K. & Dunbar, R. I. M. Smiles when sharing. *Evol. Hum. Behav.* **28**, 415-422 (2007).

34. McCullough, M. E., Pedersen, E. J., Schroder, J. M., Tabak, B. A. & Carver, C. S. Harsh childhood environmental characteristics predict exploitation and retaliation in humans. *Proc. R. Soc. Lond. B Biol. Sci.* **280**, 20122104 (2013).
35. Trust. Our World in Data <https://ourworldindata.org/trust>.
36. Petersen, M. B. & Aaroe, L. Birth weight and social trust in adulthood: evidence for early calibration of social cognition. *Psychol. Sci.* **26**, 1681–1692 (2015).
37. Haushofer, J. The psychology of poverty: Evidence from 43 countries. Working Paper. <https://www.princeton.edu/haushofer/> (2013).
38. Nettle, D., Colléony, A. & Cockerill, M. Variation in cooperative behaviour within a single city. *PLoS ONE* **6**, e26922 (2011).
39. Kesternich, I., Smith, J. P., Winter, J. K., & Hörl, M. Early-Life circumstances predict measures of trust among adults: evidence from hunger episodes in post-war Germany. *Scand. J. Econ.* **122**, 280–305 (2016).
40. Fouquet, R. & Broadberry, S. Seven centuries of European economic growth and decline. *J. Econ. Perspect.* **29**, 227–244 (2015).
41. Bosker, M., Buringh, E. & van Zanden, J. L. From Baghdad to London: unraveling urban development in Europe, the Middle East, and North Africa, 800–1800. *Rev. Econ. Stat.* **95**, 1418–1437 (2013).
42. North, D. C. & Weingast, B. R. Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. *J. Econ. Hist.* **49**, 803–832 (1989).
43. Acemoglu, D. & Robinson, J. Why nations fail: the origins of power, prosperity, and poverty (Crown Business, 2012).
44. Baumard, N. Psychological origins of the industrial revolution. *Behav. Brain Sci.* **42**, 1–47 (2018).
45. Morris, I. The measure of civilization: how social development decides the fate of nations. (Princeton University Press, 2013).
46. Baumard, N., Hyafil, A., Morris, I. & Boyer, P. Increased affluence explains the emergence of ascetic wisdoms and moralizing religions. *Curr. Biol.* **25**, 10–15 (2015).
47. Baltrušaitis, T., Robinson, P. & Morency, L. OpenFace: An open source facial behavior analysis toolkit. in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) 1–10 (2016). <https://doi.org/10.1109/WACV.2016.7477553>.
48. Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N. & Falvello, V. B. Validation of data-driven computational models of social perception of faces. *Emotion* **13**, 724–738 (2013).
49. Stewart, L. H. et al. Unconscious evaluation of faces on social dimensions. *J. Exp. Psychol. Gen.* **141**, 715–727 (2012).
50. Safra, L., Ioannou, C., Amsellem, F., Delorme, R. & Chevallier, C. Distinct effects of social motivation on face evaluations in adolescents with and without autism. *Sci. Rep.* **8**, 1–8 (2018).
51. Breiman, L. & Cutler, A. Breiman and Cutler's random forests for classification and regression. R package version, **4**, 6–12 (2018).
52. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
53. Karatzoglou, A. et al. kernlab: Kernel-based machine learning lab (2019).
54. Kuhn, M. The caret Package.
55. Lundqvist, D., Flykt, A. & Öhman, A. The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9. (1998).
56. Oslo Face Database. Leknes Affective Brain lab <https://sirileknes.com/oslo-face-database/> (2015).
57. 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).
58. Thomaz, C. E. & Giraldo, G. A. A new ranking method for principal components analysis and its application to face image analysis. *Image Vis. Comput.* **28**, 902–913 (2010).
59. Sutherland, C. A. M., Young, A. W. & Rhodes, G. Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *Br. J. Psychol.* **108**, 397–415 (2017).
60. Rule, N. O., Ambady, N. & Adams, R. B. Personality in perspective: judgmental consistency across orientations of the face. *Perception* **38**, 1688–1699 (2009).
61. Bolt, J. & Zanden, J. L. The Maddison Project: collaborative research on historical national accounts. *Econ. Hist. Rev.* **67**, 627–651 (2014).
62. Marshall, M. G., Jagers, K. & Gurr, T. R. Polity IV project. (Center for International Development and Conflict Management at the ..., 2002).
63. Wagenmakers, E.-J. A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* **14**, 779–804 (2007).

Acknowledgements

We are grateful to the National Portrait Gallery and to the Web Gallery of Art for allowing open access to high-quality paintings online, to Prof. Alexander Todorov for the distribution of the avatars controlled for dominance and trustworthiness as well as to Dr. Lev Manovich and the Selfcity team for allowing the use of their database. We would like to thank Dr. Tadas Baltrušaitis for the creation and free distribution of OpenFace. We would like to thank Anis for his feedback on the construction of our algorithm. We would like to thank Loïa Lamarque, Paul Grignon and Benoît de Courson for their help in coding OpenFace goodness of fit of the portraits. We would like to thank Prof. Alexander Todorov, Dr. Malgorzata Mikucka, Dr. Jeffrey M. Girard and an anonymous reviewer for their insightful comments on our manuscript. This study was supported by the Institut d'Études Cognitives (ANR-17-EURE-0017 FrontCog and ANR-10-IDEX-0001-02 PSL) and by the Fyssen Foundation.

Author contributions

N.B., J.G. and C.C. conceived the project. L.S. designed the study, trained the algorithm and analyzed the data. N.B. and L.S. wrote the first draft of the paper, all authors contributed to the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-18566-7>.

Correspondence and requests for materials should be addressed to L.S. or N.B.

Peer review information *Nature Communications* thanks Alexander Todorov, Malgorzata Mikucka, Christian Bjørnskov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Supplementary Information

Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings.

Correspondence to:

Nicolas Baumard (nicolas.baumard@ens.fr) & Lou Safra (lou.safra@sciencespo.fr)

This PDF file includes:

Supplementary Methods
Supplementary Figures. 1 to 8
Supplementary Tables 1 to 4

Supplementary Methods

In order to quantify trustworthiness displays in historical paintings, we developed an algorithm automatically estimating perceived trustworthiness from faces. Our algorithm also extracted perceived dominance since dominance has been shown to be, together with trustworthiness, one of the main dimensions of social perception¹. Crucially, although dominance displays carry signals of power that are distinct from the cooperation-related signals associated with trustworthiness displays, perceived dominance and perceived trustworthiness are correlated¹. This correlation entails that it is of paramount importance to control for perceived dominance when analyzing perceived trustworthiness. This type of analysis, studying together distinct but related social signals, has already been shown to be particularly promising in the emotion domain by revealing the importance of taking into account the existence of compound emotions².

Construction and validation of an algorithm for modeling trustworthiness and dominance evaluations

We built a model that automatically extracts trustworthiness and dominance evaluations from the all the facial action units detected by the OpenFace algorithm (i.e., both dichotomous and continuous estimations; OpenFace version 1.01 using OpenCV 3.3.0³). To do so, we extracted the facial action units of five sets of avatars previously generated with Facegen and controlled for dominance, for trustworthiness or for both (Supplementary Figure 1)⁴. Each avatar is generated from an initial face and manipulated to either express a specific level of dominance, trustworthiness or both based on the model developed by Oosterhof & Todorov¹. These avatar faces have been shown to successfully elicit ratings of dominance and trustworthiness in participants⁴⁻⁶. Thus, compared to participants' ratings on photographs that may be sensitive to the participants characteristics and to experimental protocol factors (such as the type of scale used to give the ratings), using avatars allow us to have well-validated sets of faces to train our model. These sets of avatars correspond to all the existing and available validated avatars controlled for trustworthiness or dominance and generated by Facegen.

More precisely, one set of avatars was generated from one single face and manipulated for both dominance and trustworthiness ($N = 49$; 7 levels of dominance and 7 levels of trustworthiness, each of the 7 levels corresponds to a standard deviation in Oosterhof and Todorov's¹ model ranging between -3 to +3 SD; set 1). Two other sets of faces correspond to 25 maximally distinct faces manipulated either on trustworthiness only ($N = 175$; 7 different levels of trustworthiness; set 2) or dominance only ($N = 175$; 7 different levels of dominance; set 3). Finally, the two last sets are composed of 25 Caucasian faces manipulated to present the same 7 levels of trustworthiness ($N = 175$; set 4) or of dominance ($N = 175$; set 5). Thus, three sets of avatars were used to build the model automatically extracting trustworthiness levels (sets 1, 2 and 4) and three were used to build the model automatically extracting dominance levels (sets 1, 3 and 5).



Supplementary Figure 1 Sample of the avatar faces used for the algorithm optimization. **Left.** Initial face for the set of avatars controlled for dominance and trustworthiness; **Middle.** Example of the Caucasian faces an initial face for one of the sets of avatars controlled for dominance only and one of the sets controlled for trustworthiness only; **Right.** Example of an initial face of the ‘Maximally distinct faces’ for the other set of avatar controlled for dominance only and for the other set of avatars controlled for trustworthiness only. These three images were created by Prof. Alexander Todorov’s team and is shared under licence CC BY.

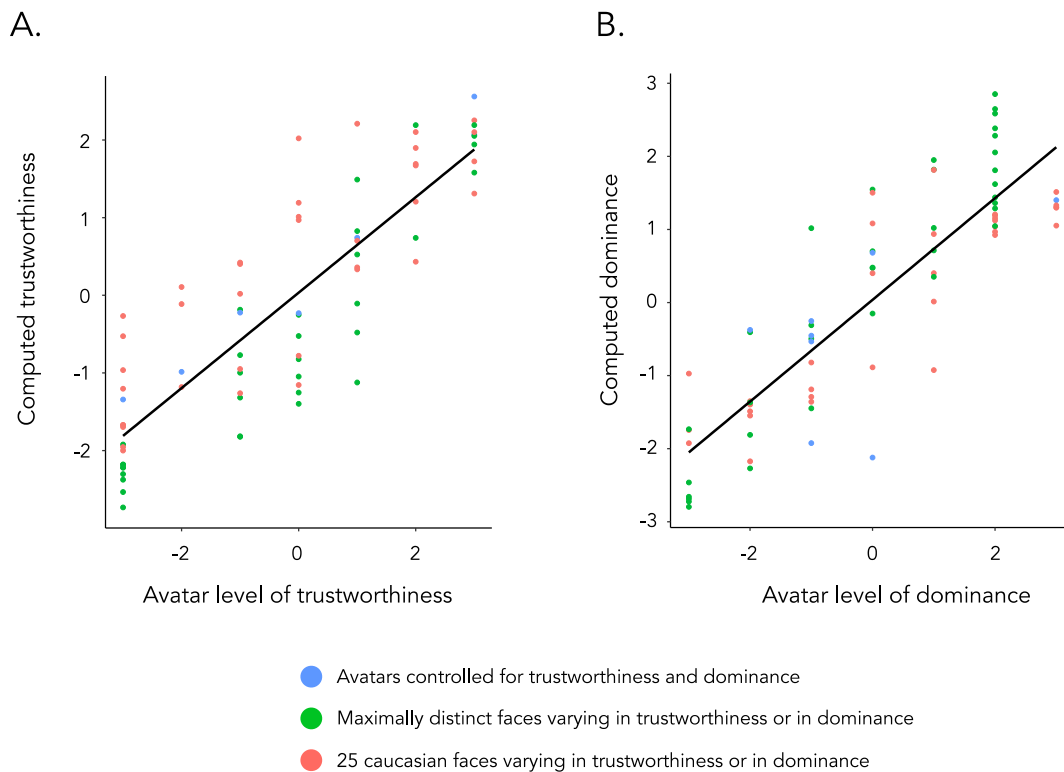
Because all our avatars were generated using the same models for trustworthiness and dominance, actions units with a variance inferior to 0.01 were discarded as not informative enough regarding cues of trustworthiness and dominance. The reason was that they were either too low in frequency or too low in intensity (ten action units discarded over thirty-three in both the trustworthiness and dominance avatar sets).

	SVM linear	SVM radial	Random forest	Linear model
Hyperparameters	<i>Cost (C)</i>	Cost (C) & sigma	<i>mtry</i>	\emptyset
Trustworthiness				
Mean absolute error	0.88 ± 0.02	0.87 ± 0.02	0.82 ± 0.01	0.87 ± 0.01
Root mean squared deviation	1.10 ± 0.02	1.05 ± 0.02	0.99 ± 0.01	1.06 ± 0.02
R squared	0.71 ± 0.01	0.74 ± 0.01	0.78 ± 0.01	0.72 ± 0.01
Dominance				
Mean absolute error	0.92 ± 0.02	0.79 ± 0.02	0.80 ± 0.01	0.90 ± 0.02
Root mean squared deviation	1.14 ± 0.02	0.99 ± 0.02	0.98 ± 0.02	1.11 ± 0.02
R squared	0.68 ± 0.01	0.76 ± 0.01	0.77 ± 0.01	0.70 ± 0.01

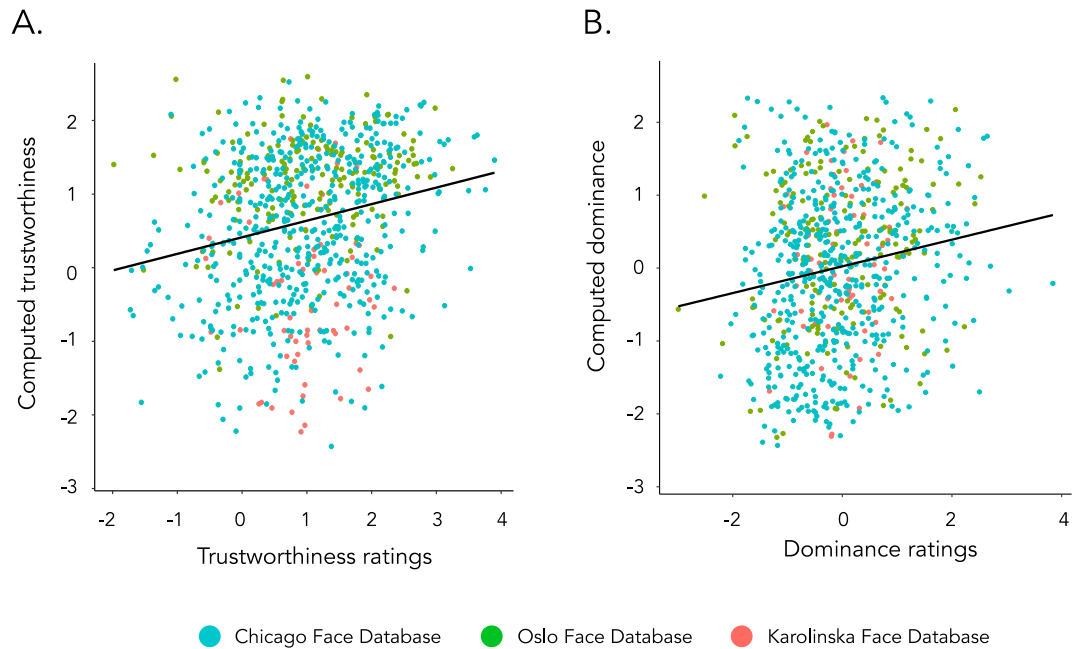
Supplementary Table 1. Model selection for extracting trustworthiness and dominance evaluations. Three indices of fit were computed, two which minimization indicates a better fit (mean absolute error and root mean squared deviation) and one which maximization indicates a better fit (R squared). The random forest was outperforming the linear model and the linear support vector model in the three indices of fit tested: mean absolute error, root mean squared deviation and r-squared. The random forest model was better than the radial support vector model for the trustworthiness model and similar to the radials support vector model for the dominance model. Values are presented as mean ± standard error to the mean. Source data are provided as raw data and scripts on the online depository.

Based on our validation results on the avatar faces, we then trained the trustworthiness and dominance models with the same hyperparameters on the entire avatar dataset in order to

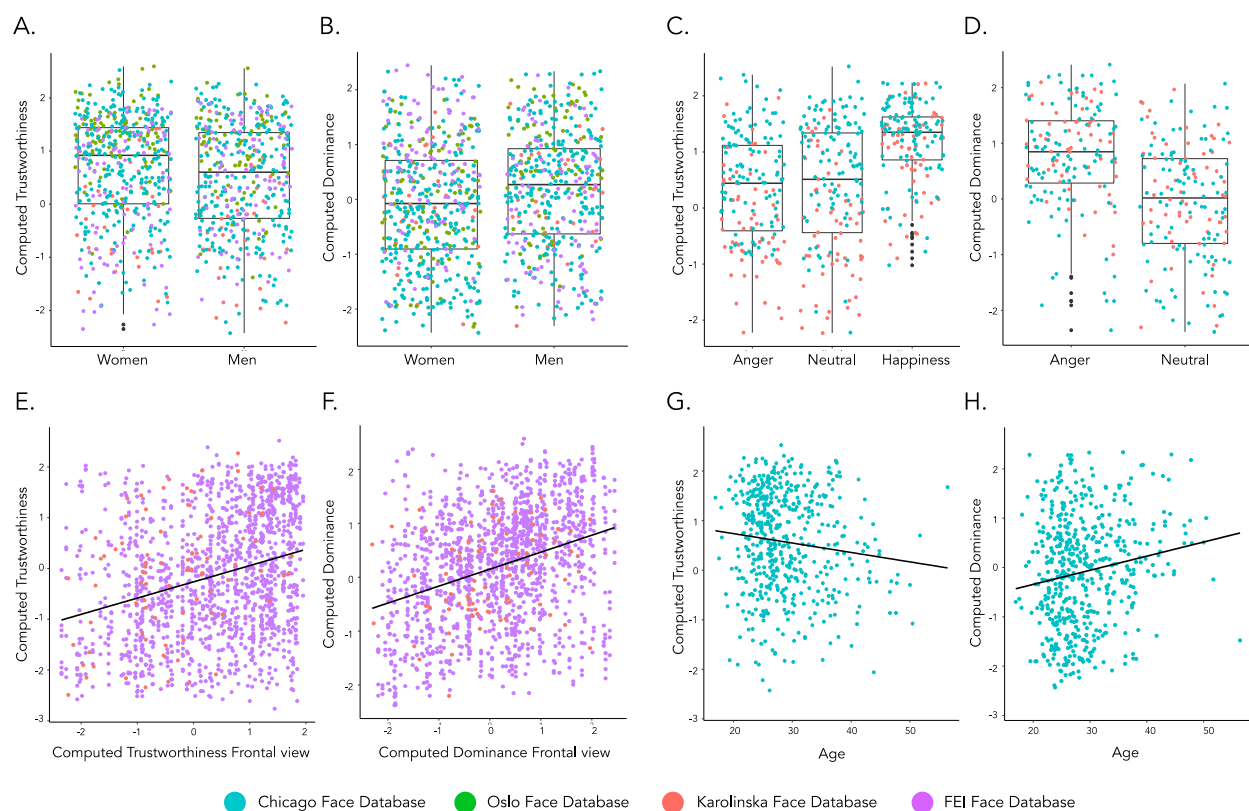
increase the accuracy of our estimates and tested this model on an independent set of photographs. This method differs from the classical train-test split used in machine learning which was not applicable given that each avatar of our dataset presented unique features in terms of luminance, texture and face shape which was important to increase the accuracy of our algorithms. However, our procedure is a highly conservative test of the validity of our models as the test set is completely different and independent of the training set. This conservative method for assessing the validity of the algorithms is particularly critical in the present study as our goal is to generalize the estimated trustworthiness and dominance evaluations to historical portraits, a completely different set of images than those classically used in social cognition research.



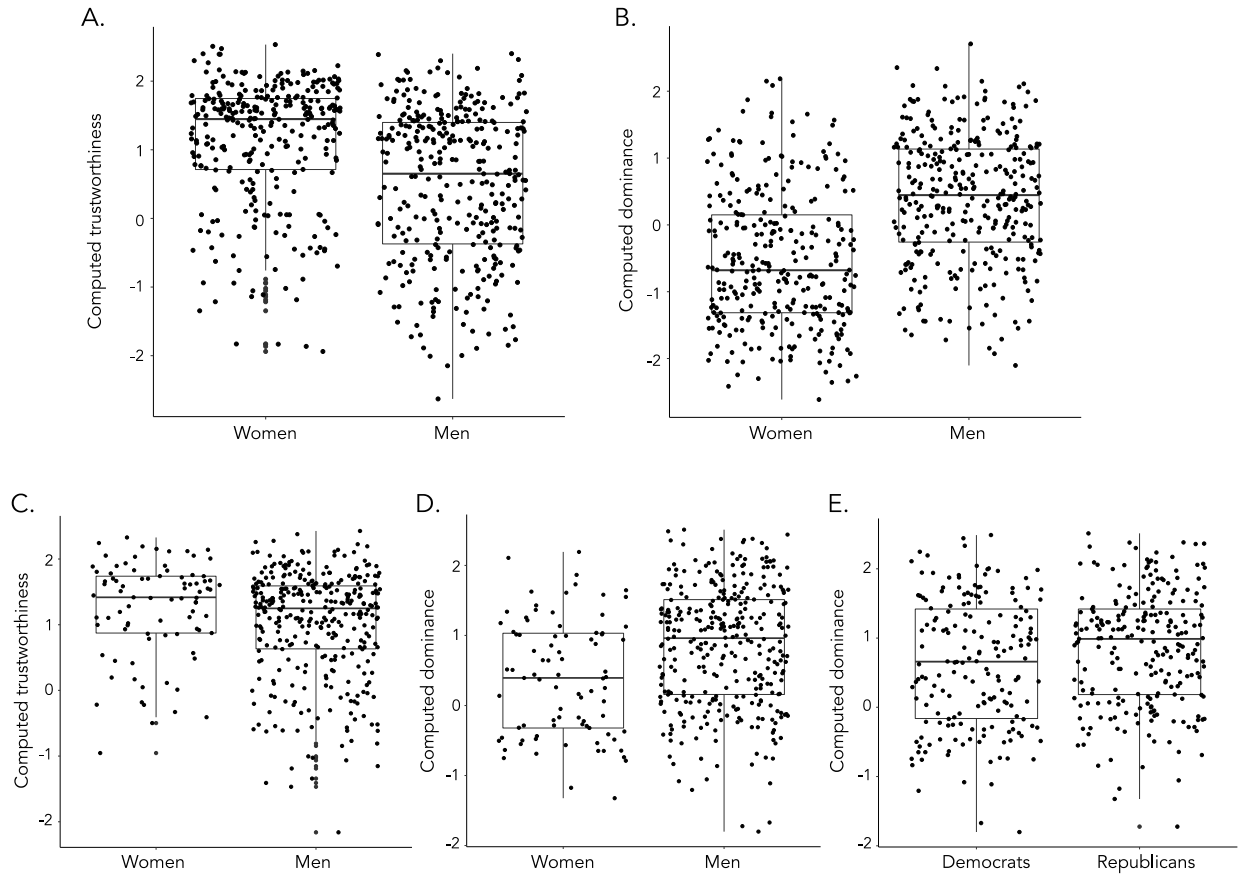
Supplementary Figure 2 Correlation between the avatars' actual level of trustworthiness and dominance in the test set and the computed trustworthiness (A) and dominance (B) based on the model optimized on the training set only. Source data are provided as raw data and scripts on the online depository.



Supplementary Figure 3 Correlation between actual ratings of trustworthiness and dominance in the three databases providing subjective ratings of trustworthiness and dominance (the Chicago Face Database, the Oslo Face Database and the Karolinska Face Database) and the recovered trustworthiness (**A**) and dominance (**B**) levels using the Facial Action Units detected by Open Face and our random-forest model. Source data are provided as raw data and scripts on the online depository.

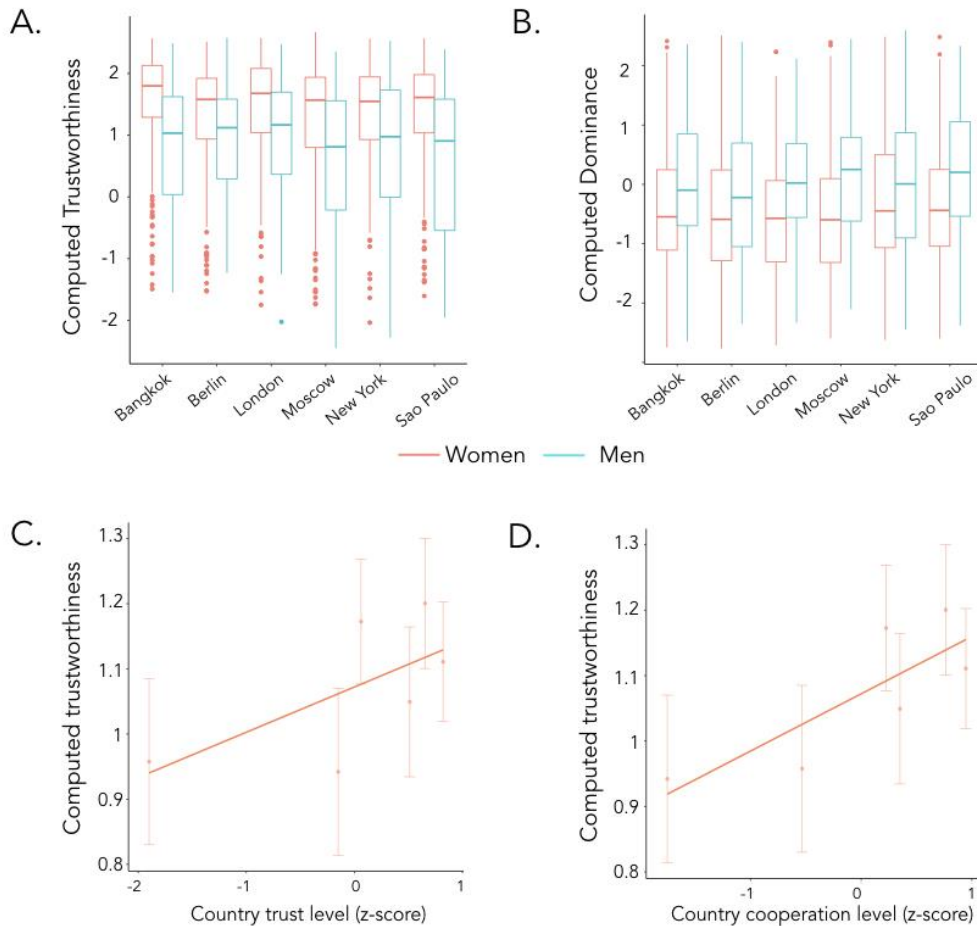


Supplementary Figure 4 Recovery of classical effects of gender (A-B), emotion (C-D), head orientation (E-F) and age (G-H) in the trustworthiness and dominance estimates computed using our random forest algorithm. In the boxplots (A-D), the centre line corresponds to the median, the lower and upper bounds of the box to the 25th and 75th percentiles and the whiskers to the largest and lowest values in a limit of 1.5 times the inter-quartile range from the box bounds. Source data are provided as raw data and scripts on the online depository.



Supplementary Figure 5 Results on natural images

A-B Recovery of the classical effects of gender in Google Image portraits of ‘Women’ ($N = 304$ images) and ‘Men’ ($N = 330$ images); **C-E** Recovery of the classical gender (**C-D**) and party (**E**) effects on the portraits of the House of the Representatives (women : $N = 85$ images ; men : $N = 334$ images ; democrats : $N = 182$ images ; republicans : $N = 237$ images). The centre line corresponds to the median, the lower and upper bounds of the box to the 25th and 75th percentiles and the whiskers to the largest and lowest values in a limit of 1.5 times the inter-quartile range from the box bounds. Source data are provided as raw data and scripts on the online depository.



Supplementary Figure 6 Results on the Selficity Database

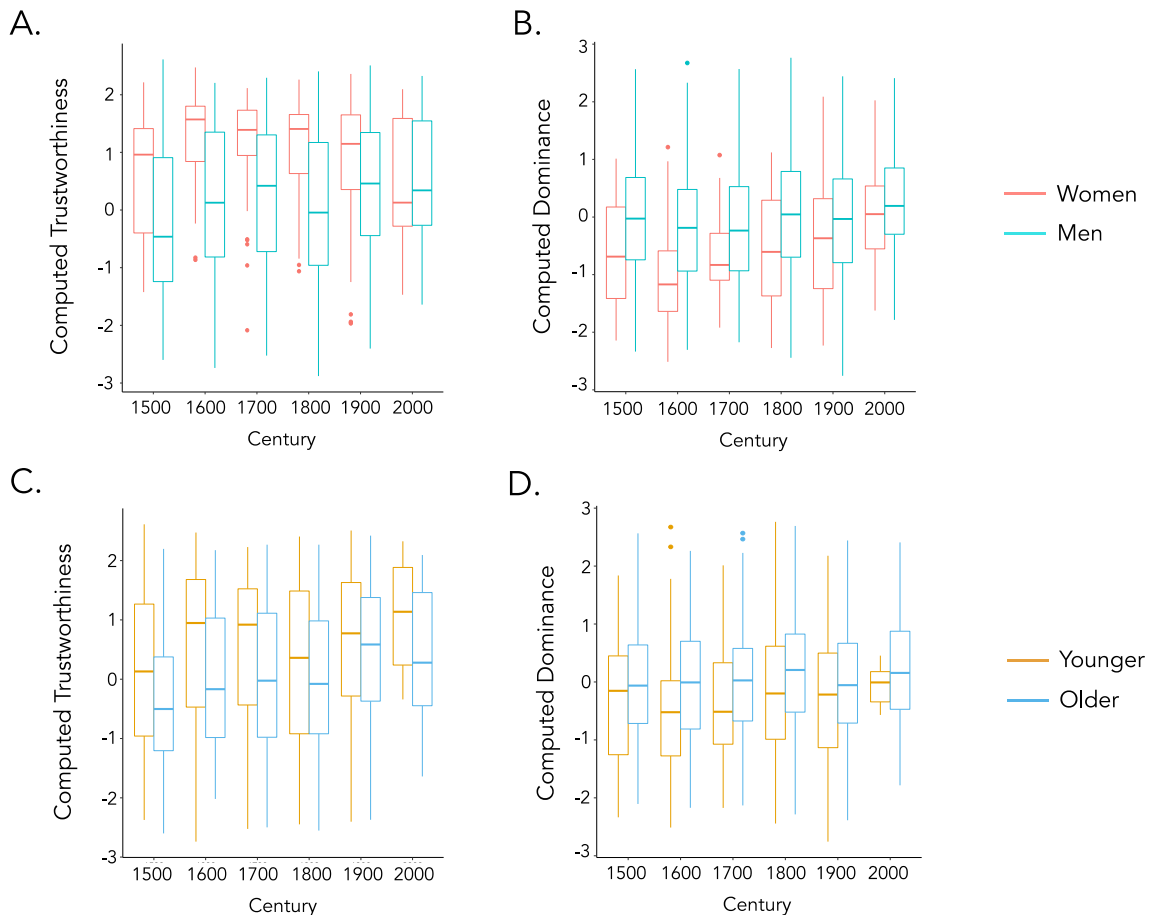
A-B Recovery of the classical effects of gender (Bangkok : $N = 247$ selfies of women, $N = 169$ selfies of men ; Berlin : $N = 239$ selfies of women, $N = 163$ selfies of men ; London : $N = 217$ selfies of women, $N = 134$ selfies of men ; Moscow : $N = 338$ selfies of women, $N = 82$ selfies of men ; New York : $N = 210$ selfies of women, $N = 127$ selfies of men ; Sao Paulo : $N = 231$ selfies of women, $N = 120$ selfies of men). The centre line corresponds to the median, the lower and upper bounds of the box to the 25th and 75th percentiles and the whiskers to the largest and lowest values in a limit of 1.5 times the inter-quartile range from the box bounds.; **C-D** Significant correlation between the country's level of interpersonal trust (**C**) and cooperation (**D**) and the mean trustworthiness estimated on the pictures of the Selficity database averaged between portraits of women and men, the red line corresponds to the effect computed in the regression controlling for the gender of the sitters. Data are represented as mean values and error bars correspond to standard errors to the mean (Bangkok : $N = 416$ selfies ; Berlin : $N = 402$ selfies ; London : $N = 351$ selfies ; Moscow : $N = 420$ selfies ; New York : $N = 337$ selfies ; Sao Paulo : $N = 351$ selfies). Source data are provided as raw data and scripts on the online depository.

Analysis of the National Portrait Gallery and the Web Gallery of Art

Text	Code	Example
Century	Century + 50	16 th century = 1550
Late century	Centruy + 90	Late 16 th century = 1590
Early century	Century + 10	Early 16 th century = 1510
Half of century	Century + 50	Half of 16 th century = 1550
Decade+s	Decade	1650s = 1655
Around/about/perhaps/probably/circa/after + Date	Date	Circa 1655 = 1655
Date 1 – Date 2	Rounded mean of Date 1 and Date 2	1650-1655 = 1652

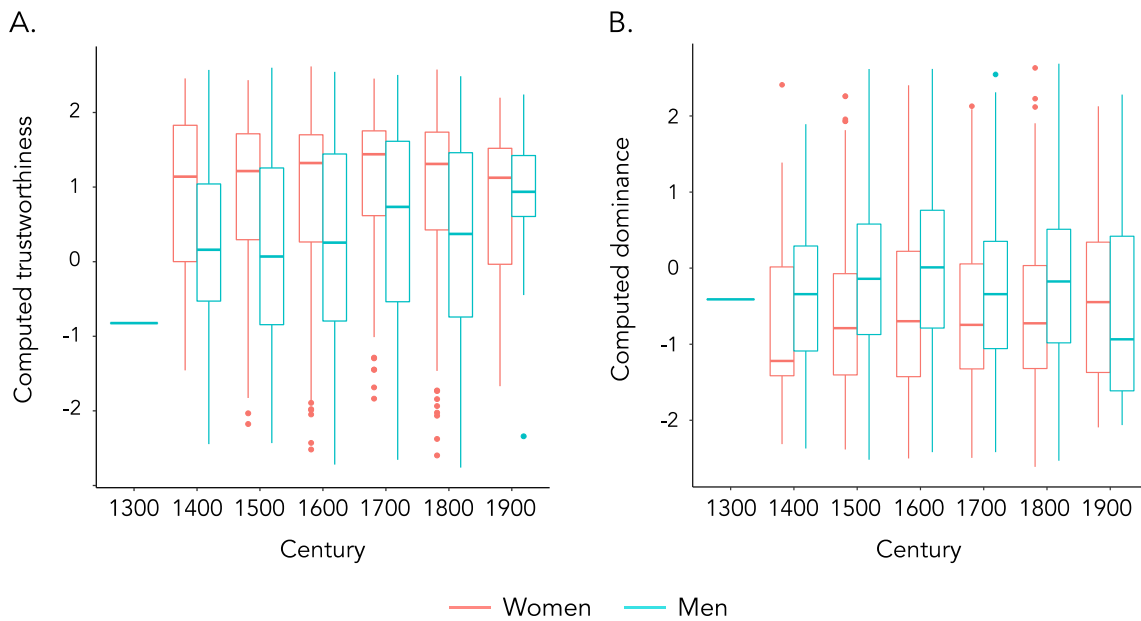
Supplementary Table 2 – Coding of the date of the portraits

The information about the sitters' gender and age allowed us to replicate the classic findings that older sitters appear more dominant and less trustworthy than younger sitters and that female sitters appear more trustworthy and less dominant than male sitters (trustworthiness: gender effect: $t(1960) = 9.69, p < .001$; age effect: $t(1960) = -6.63, p < .001$; dominance: gender effect: $t(1960) = 7.24, p < .001$; age effect: $t(1960) = -9.12, p < .001$; Supplementary Figure 7). As for the NPG, we accurately recovered the gender effect on trustworthiness and dominance on the portraits of the Web Gallery of Art (trustworthiness: $z = 17.70, p < .001$; dominance: $z = -13.35, p < .001$; Supplementary Figure 8).



Supplementary Figure 7 Recovery of the gender (A-B) (1500 : $N = 23$ portraits of women, $N = 68$ portraits of men ; 1600 : $N = 50$ portraits of women, $N = 236$ portraits of men ; 1700 : $N = 53$ portraits of women, $N = 432$ portraits of men ; 1800 : $N = 44$ portraits of women, $N = 609$ portraits of men ; 1900 : $N = 98$ portraits of women, $N = 351$ portraits of men ; 2000 : $N = 19$ portraits of women, $N = 42$ portraits of men) and age (C-D) effects in the National Portrait Gallery database over the centuries (the 'Younger' category is defined as sitters being under 48 year old; 1500 : $N = 61$ portraits of younger sitters, $N = 30$ portraits of older sitters; 1600 : $N = 188$ portraits of younger sitters, $N = 96$ portraits of older sitters ; 1700 : $N = 280$ portraits of younger sitters, $N = 194$ portraits of older sitters; 1800 : $N = 273$ portraits of younger sitters, $N = 345$ portraits of older sitters; 1900 : $N = 187$ portraits of younger sitters, $N = 249$ portraits of older sitters; 2000 : $N = 8$ portraits of younger sitters, $N = 53$ portraits of older sitters). The centre line corresponds to the median, the lower and upper bounds of the box to the 25th and 75th percentiles and the whiskers to the largest and lowest values in a limit of 1.5 times the inter-quartile range from the box bounds. Source data are provided as raw data and scripts on the online depository.

174
175
176
177



178
179
180
181
182
183
184
185
186
187

Supplementary Figure 8 Recovery of the gender effects in the Web Gallery of Art (1300 : $N = 1$ portrait of man ; 1400 : $N = 137$ portraits of men, $N = 41$ portraits of women ; 1500 : $N = 696$ portraits of men, $N = 291$ portraits of women ; 1600 : $N = 963$ portraits of men, $N = 509$ portraits of women ; 1700 : $N = 418$ portraits of men, $N = 350$ portraits of women ; 1800 : $N = 349$ portraits of men, $N = 307$ portraits of women ; 1900 : $N = 22$ portraits of men, $N = 22$ portraits of women) for trustworthiness (A) and dominance (B). The centre line corresponds to the median, the lower and upper bounds of the box to the 25th and 75th percentiles and the whiskers to the largest and lowest values in a limit of 1.5 times the inter-quartile range from the box bounds. Source data are provided as raw data and scripts on the online depository.

Dependent variable	Trustworthiness		GDP per capita	Democratization
Independent variable of interest	GDP per capita	Democratization	Trustworthiness	Trustworthiness
Delay Two decades				
Model comparison	$F(40,1) = 12.38$ $p = .001$	$F(15,1) = 0.11$ $p > .250$	$F(41,1) = 0.76$ $p > .250$	$F(16,1) = 6.54$ $p = .022$
Effect	$b = 0.04 \pm 0.01$ $t(40) = 3.52$ $p = .001$	$b = -0.01 \pm 0.03$ $t(14) = -0.33$ $p > .250$	$b = 0.59 \pm 0.68$ $t(41) = 0.87$ $p > .250$	$b = -5.82 \pm 2.27$ $t(15) = -2.56$ $p = .022$
Delay One decade				
Model comparison	$F(41,1) = 11.40$ $p = .002$	$F(16,1) = 1.11$ $p > .250$	$F(42,1) = 0.01$ $p > .250$	$F(17,1) = 5.26$ $p = .036$
Effect	$b = 0.03 \pm 0.01$ $t(40) = 3.38$ $p = .002$	$b = -0.02 \pm 0.02$ $t(15) = -1.05$ $p > .250$	$b = -0.05 \pm 0.66$ $t(41) = -0.08$ $p > .250$	$b = -4.19 \pm 1.82$ $t(16) = 0.64$ $p > .250$

Supplementary Table 3 Temporal dynamics of trustworthiness, GDP per capita and democratization in the paintings of the National Portrait Gallery. Model comparison corresponds to the comparison of the model that included the delayed variable of interest with the model in which this variable was excluded. Effect corresponds to the estimation of the regression coefficient of the delayed variable of interest. All the tests are two-sided. Following APA's recommendations, exact p-values are provided for p-s between .001 and .250. Source data are provided as raw data and scripts on the online depository.

Dependent variable	Trustworthiness		GDP per capita	Democratization
Independent variable of interest	GDP per capita	Democratization	Trustworthiness	Trustworthiness
Delay One decade				
Model comparison	X(1) = 4.00 p = .046	X(1) = 0.01 p > .250	X(1) = 2.48 p = .115	X(1) = 0.65 p > .250
Effect	b = 0.12 ± 0.05 z = 2.61 p = .009	b = 0.00 ± 0.01 z = -0.11 p > .250	b = -0.03 ± 0.02 z = -1.56 p = .119	b = 0.38 ± 0.49 z = 0.78 p > .250
Delay Two decades				
Model comparison	X(1) = 6.42 p = .011	X(1) = 0.81 P > .250	X(1) = 2.02 p = .155	X(1) = 0.72 p > .250
Effect	b = 0.19 ± 0.06 z = 3.48 p < .001	b = -0.01 ± 0.01 z = -0.84 p > .250	b = -0.05 ± 0.04 z = -1.42 p = .157	b = 0.45 ± 0.55 z = 0.82 p > .250

Supplementary Table 4 Temporal dynamics of trustworthiness, GDP per capita and democratization in the paintings of the Web Gallery of Art. All the tests are two-sided. Following APA's recommendations, exact p-values are provided for p-values between .001 and .250. Source data are provided as raw data and scripts on the online depository.

Copyright of the analysed databases

All the exploited databases (Prof. Todorov's avatar datasets, Karolinska database, Oslo Face database, Chicago Face database, FEI Face database, the National Portrait Gallery database and the Web Gallery of Art database) are free of use for non-commercial research purposes. The use of the Selfiecity database has been authorized by its owner, Dr. Lev Manovuch.

Supplementary References

1. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl. Acad. Sci.* **105**, 11087–11092 (2008).
2. Du, S., Tao, Y. & Martinez, A. M. Compound facial expressions of emotion. *Proc. Natl. Acad. Sci.* **111**, E1454–E1462 (2014).
3. Baltrušaitis, T., Robinson, P. & Morency, L. OpenFace: An open source facial behavior analysis toolkit. in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1–10 (2016). doi:10.1109/WACV.2016.7477553.
4. Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N. & Falvello, V. B. Validation of data-driven computational models of social perception of faces. *Emotion* **13**, 724–738 (2013).
5. Stewart, L. H. *et al.* Unconscious evaluation of faces on social dimensions. *J. Exp. Psychol. Gen.* **141**, 715–727 (2012).
6. Safra, L., Ioannou, C., Amsellem, F., Delorme, R. & Chevallier, C. Distinct effects of

219 social motivation on face evaluations in adolescents with and without autism. *Sci. Rep.* **8**, 1–8
220 (2018).
221
222
223
224
225
226
227
228
229