The power of external influences to modify judgments of facial and moral beauty

Marcus J Glennon & Semir Zeki*

Laboratory of Neurobiology,

Division of Cell & Developmental Biology,

University College London

*Corresponding author

Abstract

Empirical evidence shows that the often-made positive correlation between human physical and moral beauty is tenuous. In this study, we aimed to learn whether facial and moral beauty can be psychophysically separated. Participants (n = 95) provided beauty and goodness (i.e. trustworthiness) ratings for pictures of faces, after which they were presented with a fictitious peer rating for the same face and asked to re-rate the face. We used the difference between the initial and final ratings to quantify the degree of resistance to external influence. We found that judgments of facial beauty were more resistant to external influence than judgments of facial "goodness"; in addition, there was significantly higher agreement within beauty ratings than within goodness ratings. These findings are discussed in light of our Bayesian-Laplacian classification of priors, from which we conclude that moral beauty relies more upon acquired 'artifactual' priors and facial beauty more on inherited biological priors.

Keywords: neuroaesthetics; facial beauty; moral beauty; beauty judgments, judgmental conformity

Introduction

Individuals frequently infer personality traits of strangers based on physical appearance, with many studies indicating links between high facial attractiveness and various positive attributes (Langlois et al., 2000). Chief among these links is the often stated one between beauty and moral "goodness", encapsulated by a single word, *Kalon*, in ancient Greek culture and now more commonly referred to as the "beauty is good stereotype" (Jenkins & Turner, 2009; Dion, Berscheid, & Walster, 1972). Even in spite of its ancient history, how strongly the two are linked is uncertain since the strength of the linkage varies substantially between studies (Eagly, Ashmore, Makhijani, & Longo, 1991). In the current study, we addressed this question in what seemed to us a more manageable way, by restricting ourselves to facial beauty, as representative of physical beauty on the one hand, and to moral beauty (or trustworthiness) on the other.

Past studies have linked facial attractiveness to perceived moral beauty (i.e., trustworthiness; Ma, Correll, & Wittenbrink, 2015, table 3; Shinners, 2009; Xu et al., 2012). This linkage has

implications in many walks of life, particularly the court room, where more attractive individuals benefit from having a higher probability of winning (Zebrowitz & McDonald, 1991). Neuroimaging evidence shows that the medial orbitofrontal cortex (mOFC), which is engaged during the experience of beauty derived from different sources (Ishizu & Zeki, 2011; Kawabata & Zeki, 2004; Zeki et al., 2014) is also engaged when subjects experience facial and moral beauty (Kranz & Ishai, 2006; O'Doherty et al., 2003; Tsukiura & Cabeza, 2011). The issue of a link between facial and moral beauty is possibly confounded further by the common conflation between facial attractiveness and facial beauty, the former often being used as indicative of the latter; these two terms in fact describe separate concepts, which are not always linked to each other, since one can perceive each in a face without necessarily perceiving the other (Yang et al., 2021). In the current study, we are concerned with beauty, not attractiveness.

Our general hypothesis here is that the perception of physical beauty, even if linked to that of moral beauty, can nevertheless be dissociated from it psychophysically and thus shown to be a separate entity, one that is not hostage to external opinion to the same degree as moral beauty. This prediction stems from our Bayesian-Laplacian classification of priors into biological (β) and artifactual (α) categories (Zeki & Chen, 2020) where facial beauty is considered to belong more in the biological category. This proposition is due, *inter alia*, to the finding that infants of varying age, on average, orient preferentially towards face-like stimuli, as well as towards faces judged to be attractive by adult humans, thus indicating the presence of an innate, or rapidly acquired, perceptual mechanism that determines the beauty of a face (Goren, Sarty, & Wu, 1975; Langlois et al., 1991; Slater et al., 1998). On the other hand, history teaches us that moral values are far more flexible and significantly influenced by prevailing opinion and conditions. Indeed, although some believe that many moral values are universally inherited (see Krebs, 2008, for a summary), others suppose that they are established throughout life, with the internalization of a set of standards allowing for self-regulation of behaviour and the development of moral values (Bandura, 1991).

The purpose of the current study was to identify the extent to which judgments of facial and moral beauty can be separated from one another psychophysically. Ma, Xu, and Luo (2016) drew a dissociation between facial attractiveness and moral beauty by finding significantly higher agreement amongst subjects' attractiveness ratings compared to their ratings of the trustworthiness of faces. Concentrating on beauty rather than attractiveness, we wanted to take this a stage further and learn whether two can be dissociated psychophysically, which would be a guide to the tenuous linkage between them. As well as comparing agreement in ratings of facial and moral beauty, we compared the resistance of each to external influence (defined here as peer opinion). Previous studies have addressed conformity solely in the context of physical beauty (e.g., Bignardi, Ishizu, & Zeki, 2020; Klucharev, et al., 2009; Zaki, Schirmer, & Mitchell, 2011). Both Klucharev et al. and Zaki et al. found that, on average, participants change their ratings to conform to those of their peers when assessing the attractiveness of faces while Bignardi et al. (2020) found that ratings of facial beauty also conformed to peer opinion, but to a significantly lesser extent than ratings of the beauty of abstract art.

In the current study, participants were presented with pictures of faces and asked to provide a beauty or goodness (i.e., trustworthiness) rating both before and after exposure to the average rating of peers for the same faces (similar to the design used in Bignardi et al., 2020). We use the terms goodness and trustworthiness interchangeably throughout and participants were so informed. Our hypothesis was that we would be able to separate beauty and goodness ratings psychophysically through two different means: firstly, we predicted that beauty ratings would be more resistant to external information than goodness ratings; and secondly, that there would be significantly greater agreement amongst individuals in their beauty ratings compared to their goodness ratings.

Method

Participants

107 subjects were recruited through an online participant recruitment platform (https://www.callforparticipants.com/) and by word of mouth; of these, 95 (49 female, 5 undisclosed; age range 18 to 61, with a mean of 26.29 (SD = 7.18)) were taken through to analysis (see results section for exclusion justifications). The only inclusion criteria were having fluent English-speaking skills and being between 18 and 65 years of age. Prior to consent, participants were informed that the study involved rating a selection of faces on several different traits. The experiment was approved by the Ethics Committee of University College London.

Participants gave their informed consent and provided, optionally, details of their gender, sexual orientation, age, ethnicity, and primary country of residence (i.e. the country they have lived longest in), all of which could possibly be factors in biasing beauty ratings for faces.

Participants were nationals of 14 different countries, with ethnicity also varying substantially (see supplementary materials 1 for details); they were given the option of declaring their sexual orientation using an adapted Kinsey scale (as used in Zeki & Romaya, 2010) which scales sexuality from 1 (completely heterosexual) to 7 (completely homosexual). Only integer responses were permitted.

Stimuli

The stimuli used in this study were pictures of faces obtained from several face databases (the Face Research Lab London Set, 'FRLLS'; DeBruine & Jones, 2017; Chicago Face Database, 'CFD'; Ma et al., 2015) and online sources (Adobe Stock, Pinterest, and Google images). Based on beauty ratings given in a preliminary study, we selected 32 beautiful faces, 16 average ones, and 16 faces in the 'not beautiful' category; more beautiful faces were selected because the study aimed to learn about the relationship between beauty and morality. Half were female and half male, and they were of varying ethnicity (see supplementary materials 2); each was set to 500x500 pixels and presented against a plain white background.

Procedure

The experimental procedure was designed in PsychoPy (Peirce et al., 2019) and implemented online via Pavlovia (https://pavlovia.org/).

The experiment consisted of 2 blocks of 64 trials each (during which subjects rated the beauty and "goodness" of faces, separately). Both blocks were preceded by 3-trial practice blocks to familiarise participants with the task (see Fig. 1). Each trial had four components: an initial beauty/goodness rating, a certainty rating for the initial rating, presentation of external information (see below for details), and a final beauty/goodness rating in light of the external information (see supplementary materials 3 for instructions provided to participants). Ratings were made on a 7-point scale, with 1 corresponding to not beautiful/good at all, and 7 corresponding to extremely beautiful/good. Originally, participants had a 4s response window for the initial, certainty, and final ratings. This was increased to an 8s window after recruiting the first 32.6% of participants due to a high number of mistrials in which participants ran out of time to respond. Increasing the response window led to fewer mistrials and did not significantly affect average rating change (p = .691 for beauty; and p = .871 for goodness).

The external information consisted of a peer opinion which, participants were informed, was the average rating of other participants in the study. However, we used fictitious, pseudo-randomised peer opinions instead of real ones. Facial beauty is primarily reliant upon stable, biological priors, meaning that ratings of facial beauty are likely to have high agreement among different groups, whereas ratings of facial goodness may be more variable due to their suspected reliance on artifactual priors. Therefore, if we were to use real peer opinions, we may observe an unequal distribution of discrepant peer opinions between beauty and goodness, with peer ratings of beauty being more similar to the ratings of participants than peer ratings of goodness. This would lead to a higher number of opportunities for conformity amongst goodness ratings compared to beauty ratings, thereby making it difficult to compare the data obtained for the two attributes. For this reason, we used fictitious peer opinions, with both beauty and goodness peer ratings having a 50% chance of being different to, and a 50% of being the same as, the initial ratings of participants.

As in Zaki et al. (2011), the fictitious peer opinions were generated using the following procedure: when participants rated a face lower than 3 (i.e. 1 or 2), peer opinion was just as likely to be the same or 2 to 3 points higher than the initial rating (i.e. 3, 4, or 5). Likewise, when participants gave a rating higher than 5 (i.e. 6 or 7), peer opinion was presented as being either the same or 2 to 3 points lower than the initial rating (i.e., 3, 4 or 5), with a 50% chance of being the same and a 50% chance of being different. Finally, for initial ratings between 3 and 5, peer opinion had a 50% chance of being equal to the initial rating, a 25% chance of being 2 to 3 points lower, and a 25% chance of being 2 to 3 points higher.

Figure 1. Practice experimental paradigm. There were 4 steps to each trial. Participants gave an initial beauty or goodness rating on a discrete scale of 1 to 7. Then they were asked to express their certainty regarding their initial rating on a continuous scale of 1 to 7. Following this, they were presented with external information (i.e., the fictitious peer opinion). Finally, once they had clicked to continue, they made a final beauty or goodness rating in light of the peer opinion. A) a practice beauty trial. B) a practice goodness trial.

The order in which trials appeared in each block was randomised, and the blocks were counterbalanced to minimise any order effects.

Analysis

Firstly, we undertook a two-way analysis of covariance (ANCOVA) to determine whether beauty ratings given to faces were affected by the gender and sexual orientation of participants. The gender of the image was a within-subjects factor, the gender of the participant was a between-subjects factor, and sexual orientation (as measured on a 7-point scale) was listed as a covariate.

We next wanted to learn whether beauty and goodness ratings correlate with one another in our sample. We took the average beauty and goodness ratings per face and used a Spearman rank-order correlation coefficient to determine the strength and direction of the association between them. Three faces were removed from this analysis, two of which had exceptionally low average goodness ratings (<2.4) and another of which had an exceptionally high goodness rating (>5.1). Removing these outliers from the analysis did not alter the significance of the results.

To address our hypothesis that subjects would be more resistant to changing their beauty ratings compared to their goodness ratings in light of external opinion, we undertook 5 within-subject comparisons (paired-sample sign tests; 2 for beauty, 2 for goodness, and 1 addressing both). A sign test is a non-parametric test which identifies whether there are

consistent differences between the averages of two dependent samples. We began by determining whether external information that differed from the initial rating given by a participant led to a final rating that was different to the initial rating; next, we investigated whether such external influence was higher when the initial rating of the participant was medium or extreme, that is to say, ratings between 3 and 5 on the one hand and ratings of 1, 2, 6, or 7 on the other; a final sign test was done to determine whether external information influenced beauty ratings and goodness ratings to the same or to different degrees. As there was a total of 5 sign tests included for this analysis, Holm-Bonferroni adjustments were made to correct for the false discovery rate. Rating change (final rating minus initial rating) was the key metric used in the within-subjects comparisons, as it represents the strength of the influence of external information. We calculated absolute values (i.e. removed the negative sign from negative values) to ensure comparability across trials.

After this, we used the certainty ratings given by participants to assess whether they affected susceptibility to external influence and whether participants were any more certain of their beauty ratings compared to their goodness ratings. A Spearman's rank-order correlation was used to determine whether there was a link between the average certainty and rating change exhibited by participants. In addition, another sign test was used to examine whether participants, on average, were more certain of their beauty and goodness ratings.

We then used the Means Minus One test (MM1; Vessel et al., 2018) to calculate the average agreement within beauty ratings and within goodness ratings, separately. The MM1 was calculated by comparing the ratings of each participant to the average ratings of all other participants for the same faces. This process is repeated for each participant, and it results in individual r values that represent the extent to which a participant's ratings correlate with group mean ratings. Subsequently, the r values were converted to z values (using the Fisher r to z transformation), as this has been shown to reduce bias in estimates when determining agreement between participants (Bronstad & Russell, 2007). The average z value was calculated, then converted back to an r value. A paired-sample sign test was used to compare participants' z values for individual beauty and goodness agreement scores.

All analyses were undertaken in either IBM SPSS (Statistics version 27) or R-Studio using R version 4.0.3 (R-Studio team, 2020).

Results

Sexual orientation of participants

Of the 90 out of 95 participants who provided their sexual orientation, 57% identified as either completely homosexual or heterosexual, with the remainder of participants distributed between the two (see supplementary materials 1). This corresponds well with previous data on sexual orientation amongst the younger British population by a 2015 YouGov poll for 18–24-year-olds in the United Kingdom (https://yougov.co.uk/topics/lifestyle/articles-reports/2015/08/16/half-young-not-heterosexual).

Mean beauty and goodness ratings

Twelve participants were excluded from the analysis for several reasons: 8 had too many mistrials in their data (>30% of trials), failing to make analysis feasible; 2 exhibited extremely low average response times to faces in goodness trials (<1.8 seconds) when compared to the average of others (mean = 3.74, SD = .66), implying lack of attention; and, finally, 2 exhibited extremely high average rating changes for all (feedback and control) beauty trials (> 2.5) when compared with the mean of others (mean = 0.24, SD = .20) with no discernible pattern in their responses. Therefore, 95 participants were taken forward to analysis.

As found previously, average beauty ratings for faces from female participants were highly correlated with average beauty ratings from male participants (r = .99, p < .001; Bignardi et al., 2020).

To check whether beauty ratings given to faces were affected by the gender or sexual orientation of participants we undertook a two-way ANCOVA. Data from participants who did not wish to disclose their gender or sexual orientation were excluded from this analysis (therefore n = 85). It revealed that, on average, male participants did not give significantly different ratings for images of female faces, and vice versa (F(1, 82) = .138, p = .711). However, there was a significant effect of the gender of the faces being rated, with female faces averaging a higher beauty rating than male faces (mean = 4.20, SD = .63 and mean = 3.64, SD = .73, respectively; F(1, 82) = 51.021, p < .001; as found by Bignardi et al., 2020). Sexual orientation did not significantly affect beauty ratings (F(1, 82) = .579, p = .449). Effect sizes are given in supplementary materials 4.

The Spearman's rank correlation coefficient revealed that there was a strong, positive correlation between beauty and goodness ratings of the same faces, which was statistically significant (r_s(60) = .786, p < .001; see Fig. 2). However, visual inspection indicates that it is not without considerable variation.

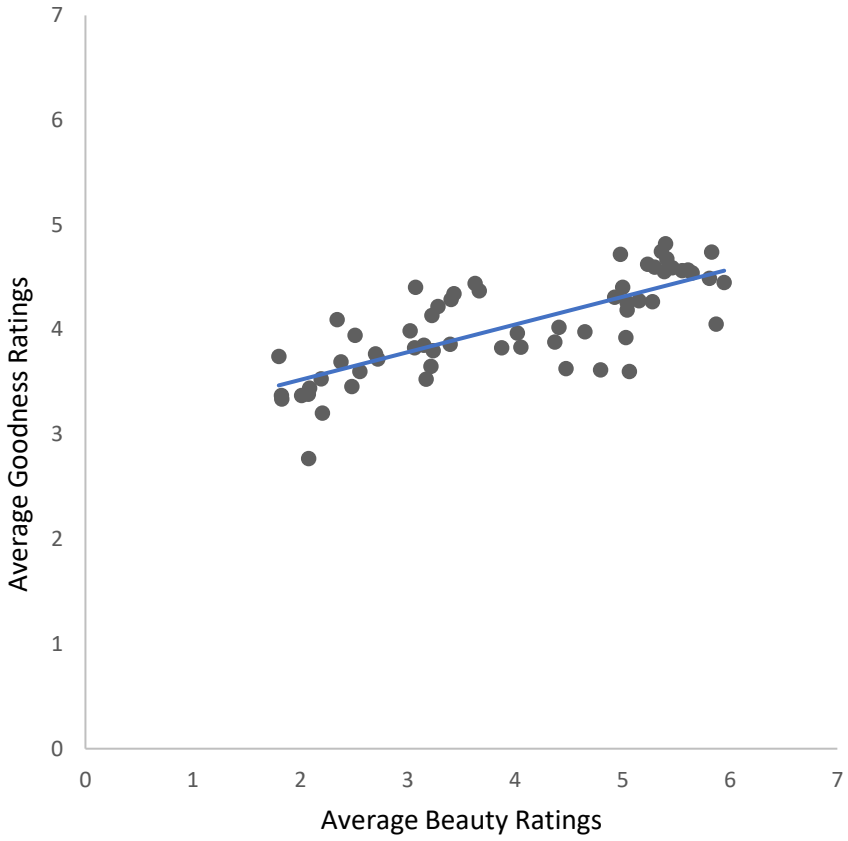

Figure 2. Scatter plot showing the positive correlation between average beauty and goodness ratings per face. Beauty and goodness ratings were taken on a 7-point scale and averaged for each face stimulus in preparation for the Spearman rank-order correlation analysis.

Main analysis: effect of external influence on beauty and goodness ratings

Rating change (RC), quantified as the difference between the initial and final ratings, was the value used to represent the effect of the external influence. The higher the RC, the more susceptible was the initial rating to modification through feedback provided by external influence. Given that the magnitude, as opposed to the direction, of the effect was the focus of this study, negative RCs were converted to positive values. This allowed all values to be compared.

Half of the trials in the study were feedback trials, in which there was a mismatch between the participants' initial rating and the pseudo-randomised peer rating they were presented with. The remaining trials were control ones, in which the initial and peer ratings were equal. To learn whether external information affected rating change, paired-sample sign tests were used; these compared the average rating change produced by feedback trials to that of control trials (see Fig. 3). For beauty trials, there was a significant difference between the average rating change of feedback versus control trials (mean RC = .30, SD = .24 and mean RC = .17, SD = .20, respectively; $p < .001$, one-tailed, Holm-Bonferroni adjusted), with 71 of the 95 participants exhibiting a higher average RC for feedback trials. A significant difference was also found for goodness trials when comparing feedback and control trials (mean RC = .34, SD

= .26 and mean RC = .14, SD = .18, respectively; $p < .001$, one-tailed, Holm-Bonferroni adjusted), with 72 of the 95 participants exhibiting a higher average RC for feedback trials. Although these results show that external information influenced both beauty and goodness ratings, the results below show that there was a difference in the magnitude of the effect.

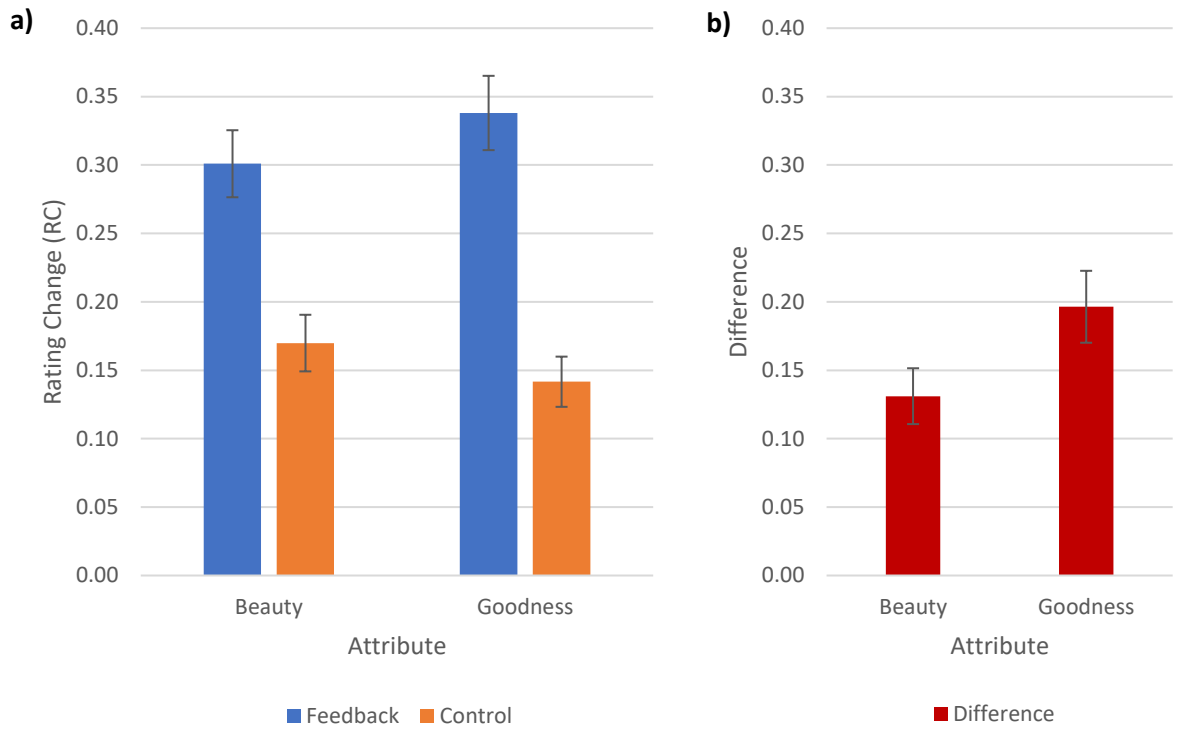

Figure 3. Bar chart showing rating change (RC) in the different study conditions. **a)** Average RC was calculated for feedback and control trials for both beauty and goodness and is plotted on the y-axis. **b)** The bars labelled 'Difference' show the difference in average rating change between feedback and control trials. For both panels, attribute (beauty or goodness) is plotted on the x-axis. Error bars represent standard error.

The next sign tests aimed to establish whether external influence affects medium versus extreme ratings differentially. For beauty trials, no significant difference was detected between extreme and medium ratings (mean RC = .30, SD = .29 and mean RC = .30, SD = .28, respectively; $p = .460$, one-tailed, Holm-Bonferroni adjusted). The same was true for goodness trials (mean RC = .40, SD = .33 and mean RC = .34, SD = .26, respectively; $p = .861$, one-tailed, Holm-Bonferroni adjusted). This shows that external information influences extreme and medium ratings in a similar manner.

To address our hypothesis that beauty ratings would be more resistant to external influence than goodness ratings, we undertook another sign test. To ensure we were using values solely pertaining to the effect of external information, we subtracted participants' average control trial RC from their average feedback trial RC (i.e. the bars labelled 'Difference' in Fig. 3). The values obtained were used for the sign test, which showed a significant difference between the average RC for beauty compared to goodness trials (mean RC = .13, SD = .20 and mean RC = .20, SD = .26, respectively; $p = .045$, one-tailed, Holm-Bonferroni adjusted). Of the 94 out of

95 participants who exhibited a difference, 58 had a higher average RC for goodness trials than beauty trials, and the remaining 36 exhibited the reverse. This shows that, on average, participants were less resistant to external information when re-evaluating goodness ratings than beauty ratings.

Analysis of certainty and consensus

In each trial participants gave certainty ratings regarding their initial judgments of beauty and goodness. Despite a Spearman's correlation analysis, undertaken within each category, showing no significant correlation between the average certainty and RC ($r_S(93) = -.170$, $p = .099$ and $r_S(93) = .135$, $p = .193$, for beauty and goodness judgments respectively), a sign test showed that the average certainty rating for beauty judgments was significantly higher than that for goodness judgments (mean = 4.73, SD = 1.05 and mean = 4.27, SD = .93, respectively; $p = .004$, one-tailed). This is in line with the finding of a higher resistance to external influence for beauty compared to goodness judgments.

We next examined agreement between participants for both beauty and goodness ratings by way of a mean minus one (MM1) analysis (see method section for details). The analysis showed that the average of all the correlations between each participant's ratings and the average ratings of all other participants was higher for beauty ratings than for goodness ratings (MM1 = .81, 95% confidence interval (CI) = [.74, .88] and MM1 = .46, 95% CI = [.41, .51], respectively; see Fig. 4). A paired-samples sign test carried out on the Fisher Z-transformed values revealed that agreement within beauty ratings was significantly greater than agreement within goodness ratings (mean z-score = 1.13, SD = .36 and mean z-score = .50, SD = .25, respectively; $p < .001$, one-tailed). The results of the MM1 analysis were replicated in an intraclass correlation coefficient analysis, a similar method used to examine inter-rater agreement (see supplementary materials 5). Both the certainty and agreement analyses fit with our hypothesis by demonstrating that beauty and goodness judgments can be psychophysically separated.

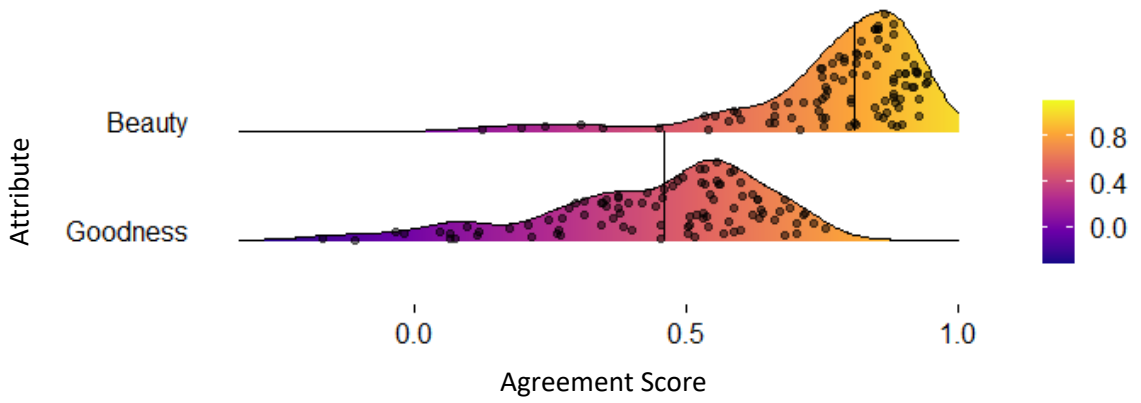

Figure 4. Individual and average agreement scores amongst beauty and goodness ratings. The ratings of each participant were correlated with the average ratings of all the other participants combined. The values obtained are presented in the graph, with each dot representing an individual agreement score. The vertical black lines represent the overall MM1 value for each attribute.

Discussion

We aimed to identify the extent to which judgments of facial and moral beauty can be separated from one another psychophysically. Our results show that, regardless of the correlation between the two, goodness ratings were more hospitable to external influence than were beauty ratings. There was also significantly less agreement among subjects in their ratings for moral beauty, as well as significantly less certainty for them. Each of these findings is in line with our hypothesis, which was that we could separate beauty and goodness ratings using measures of conformity, agreement, and certainty. As well as providing evidence for their differentiation, these findings shed light on how we can classify facial and moral beauty in relation to our theory of the distinction between biological and artifactual categories of priors (Zeki & Chén, 2020).

Biological and artifactual categories refer to the extreme ends of a spectrum of priors (Zeki & Chén, 2020). In colour vision, an inherited concept – that of ratio-taking (Land, 1986) – is applied to incoming chromatic visual signals. This ratio-taking mechanism generates stable biological priors (namely, the constant colour categories and the hues attached to them; Zeki, Javier, & Mylonas, 2020). These priors result in highly consistent colour categorization that is independent of learning, memory, and judgment and remains unchanged when confronted with changes in the wavelength composition of the light reflected from surfaces (Zeki & Chen, 2020).

The recognition of a stimulus as a face is another example of the interfacing of incoming visual signals with an inherited or rapidly acquired brain concept which identifies a certain significant configuration as constituting a face. It is difficult to over-ride this even with prolonged daily exposure (for a month) to mutilated faces; the brain's perceptive system will continue to identify the latter as departures from the norm (Chen and Zeki 2021).

Facial beauty is another example of a process that is dictated by biological priors (see Introduction), but it is not quite as resistant to external influence as the categorization of colour or the determination of a significant configuration as indicating a face (Klucharev, et al., 2009; Zaki et al., 2011); on the other hand, it is more resistant to external influence than artistic judgments (which constitute an example of reliance on artifactual priors; Bignardi et al., 2020).

Where moral beauty fits on this scale of priors is unclear; there has been much discussion on whether a sense of morality is inherited or acquired. Some accounts emphasise the role of evolution in moral acquisition (e.g. Darwin, 1874; Krebs, 2008), whereas others focus on the role of social learning (e.g. Bandura, 1991). There are some moral norms that are apparent across the vast majority of cultures, such as reciprocity (Gouldner, 1960) and obligation to fulfil social duties (Berkowitz, 1972). However, Individuals from different cultures, and even individuals within the same culture, define and prioritise these moral norms differently (Vasquez et al., 2001), as is clear from the history of conflicts in the 20th century, indicating the importance of social context on the development of a sense of morality. We hypothesise that the capacity to develop moral values is inherited, but that the moral values themselves are acquired and subject to influence. The results of the current study, alongside the fact that the acquisition of moral values is heavily dependent upon social experience, supports the proposition that moral beauty is reliant primarily upon artifactual priors as opposed to biological ones, regardless of the evolved mechanisms providing the capacity for moral development.

In conclusion, we have demonstrated that facial and moral beauty can be psychophysically separated. Given this, and previous accounts that have emphasized the importance of social learning and environment in the development of moral values, we conclude that moral beauty relies more upon acquired artifactual priors than biological ones.

Disclosure of Conflict of interest

The authors declare there are no conflicts of interest

Acknowledgements

This work was supported by the Leverhulme Trust, London.

References

Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), Handbook of moral behavior and development (Vol. 1, pp. 45-103). Hillsdale, NJ: Erlbaum.

Berkowitz, L. (1972). Social norms, feelings, and other factors affecting helping and altruism. In *Advances in experimental social psychology* (Vol. 6, pp. 63-108). Academic Press.

Bignardi, G., Ishizu, T., & Zeki, S. (2020). The differential power of extraneous influences to modify aesthetic judgments of biological and artifactual stimuli. *PsyCh Journal*.

Bronstad, P. M. & Russell, R. (2007). Beauty is in the 'we' of the beholder: Greater agreement on facial attractiveness among close relations. *Perception*, *36*(11), 1674-1681.

Chen, C. H. & Zeki, S. (2011). Frontoparietal activation distinguishes face and space from artifact concepts. *Journal of cognitive neuroscience*, *23*(9), 2558-2568.

Darwin, C. (1874). The descent of man and selection in relation to sex. New York: Rand, McNally & Company

DeBruine, L. & Jones, B. (2017). *Face Research Lab London Set*. https://doi.org/10.6084/m9.figshare.5047666.v3

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of personality and social psychology*, *24*(3), 285.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin*, *110*(1), 109.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*(4), 544-549.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American sociological review*, 161-178.

Ishizu, T. & Zeki, S. (2011). Toward a brain-based theory of beauty. *PloS one*, *6*(7), e21852.

Jenkins, I. D. & Turner, V. (2009). *The Greek Body*. Getty Publications.

Kawabata, H. & Zeki, S. (2004). Neural correlates of beauty. *Journal of neurophysiology*, *91*(4), 1699-1705.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140-151.

Kranz, F. & Ishai, A. (2006). Face perception is modulated by sexual preference. *Current biology*, *16*(1), 63-68.

Krebs, D. L. (2008). Morality: An evolutionary account. *Perspectives on psychological science*, *3*(3), 149-172.

Land, E. H. (1986). An alternative technique for the computation of the designator in the retinex theory of color vision. *Proceedings of the national academy of sciences*, *83*(10), 3078-3080.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological bulletin*, *126*(3), 390.

Langlois, J. H., Ritter, J. M., Roggman, L. A., & Vaughn, L. S. (1991). Facial diversity and infant preferences for attractive faces. *Developmental Psychology*, *27*(1), 79.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, *47*(4), 1122-1135.

Ma, F., Xu, F., & Luo, X. (2016). Children's facial trustworthiness judgments: Agreement and relationship with facial attractiveness. *Frontiers in psychology*, *7*, 499.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia*, *41*(2), 147-155.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195-203.

Shinners, E. (2009). Effects of the "what is beautiful is good" stereotype on perceived trustworthiness. *UW-L Journal of Undergraduate Research*, *12*, 1-5.

Slater, A., Von der Schulenburg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S., & Samuels, C. (1998). Newborn infants prefer attractive faces. *Infant Behavior and Development*, *21*(2), 345-354.

Team, R. (2020). RStudio: integrated development for R. RStudio, PBC, Boston.

Tsukiura, T. & Cabeza, R. (2011). Shared brain activity for aesthetic and moral judgments: implications for the Beauty-is-Good stereotype. *Social cognitive and affective neuroscience*, *6*(1), 138-148.

Vasquez, K., Keltner, D., Ebenbach, D. H., & Banaszynski, T. L. (2001). Cultural variation and similarity in moral rhetorics: Voices from the Philippines and the United States. *Journal of Cross-Cultural Psychology*, *32*(1), 93-120.

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, *179*, 121-131.

Xu, F., Wu, D., Toriyama, R., Ma, F., Itakura, S., & Lee, K. (2012). Similarities and differences in Chinese and Caucasian adults' use of facial cues for trustworthiness judgments. *PLoS One*, *7*(4), e34859.

Yang, T., Formuli, A., Paolini, M., & Zeki, S. (2021). The Neural Determinants of Beauty. *bioRxiv*. Manuscript submitted for publication.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological science*, *22*(7), 894-900.

Zebrowitz, L. A. & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and human behavior*, *15*(6), 603-623.

Zeki, S. & Chén, O. Y. (2020). The Bayesian-Laplacian brain. *European Journal of Neuroscience*, *51*(6), 1441-1462.

Zeki, S., Javier, A., & Mylonas, D. (2020). The biological basis of the experience and categorization of colour. *European Journal of Neuroscience*, *51*(2), 670-680.

Zeki, S. & Romaya, J. P. (2010). The brain reaction to viewing faces of opposite-and same-sex romantic partners. *PloS one*, *5*(12), e15802.

Zeki, S., Romaya, J. P., Benincasa, D. M., & Atiyah, M. F. (2014). The experience of mathematical beauty and its neural correlates. *Frontiers in human neuroscience*, *8*, 68.