PNAS
www.pnas.org

Supplementary Information for

The evolution of early symbolic behavior in *Homo sapiens*

Kristian Tylén, Riccardo Fusaroli, Sergio Gonzalez de la Higuera Rojo, Katrin Heimann, Nicolas Fay, Niels N. Johannsen, Felix Riede, Marlize Lombard

Corresponding author: Kristian Tylén (kristian@cc.au.dk)

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1910880117

Supplementary Methods

In the following we present three sets of analyses. First, we describe the stimulus generation process, as well as a number of analyses validating the stimuli images used in the experiments. Second, we present detailed materials and methods for each of the five experiments making up the main investigation in this study. Last, we present an analysis investigating relations between the measurements obtained in the experiments in order to assess the extent to which they regard correlated or orthogonal aspects of human cognition. All data, analysis scripts, and stimulus files are available on the OSF repository DOI 10.17605/OSF.IO/RBTK4 (https://osf.io/rbtk4/).

Table of contents:

1 Stimuli and sampling

Through all the five experiments, we rely on the same stimulus set, consisting of outlines of the Blombos and Diepkloof engravings.

Sampling a corpus of ancient engravings for the purpose of testing experimental predictions is not trivial. Many choices have to be made to ensure representativeness. However, every such choice can potentially bias the collection. We took as a starting point the corpus of Blombos ochre engravings originally reported in Henshilwood et al (1) consisting of 19 engraved ochres, and the corpus of ostrich egg engravings reported in Texier et al (2) consisting of 408 engraved egg fragments out of which 73 are photographically documented in the publication (several of which are assumed to be fragments of the same pattern). From these, we sampled a corpus of twenty-four engravings (12 from Blombos and 12 from Diepkloof) based on criteria of temporal origin and pattern type.

Since we were particularly interested in the development of pattern composition over time, we chose outlines of patterns dated to different time periods. Given the uncertainty in the dating of the engravings (especially in the case of the Diepkloof findings, 3, 4) - the stimulus items were divided within each site in three periods, early, intermediate and late, corresponding to the classification used in Texier et al (2, see table 2). For the ochre engravings, these corresponds to the grouping presented in Henshilwood et al (1, see fig. 21), dating the early period engravings to approx 109 - 100 kya, the intermediate ones to approx 100 - 70 kya, and the late to approx 70 - 52 kya. Importantly, through all analyses, period is treated as an ordinal variable which profiles the order of their appearance over the exact time spans separating individual patterns.

In order to ensure a representative sample, we also strived to include items that belong to different types of patterns. In several studies, the Blombos and Diepkloof patterns have been classified with respect to their compositions and their developments over time (fig. S1):

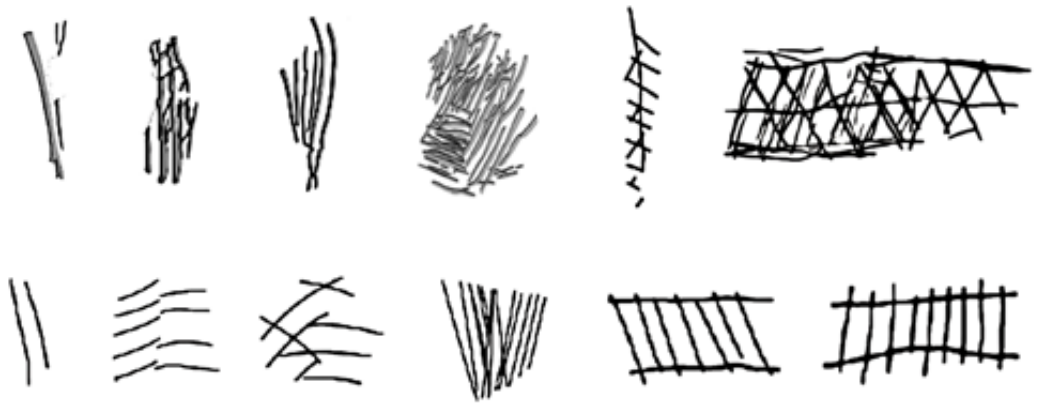

Fig. S1: The original outlines of engraved patterns used in this study and their temporal order from left (early) to right (late). The upper row represent outlines of engraved ochres from Blombos adapted with permissions from ref. 5, and the lower row outlines of engraved egg shells from Diepkloof adapted from ref. 2, with permissions from Elsevier.

The original published outlines of line patterns varied greatly in visual complexity (e.g. number and length of lines) making them unsuitable as stimuli in perception experiments. Moreover, we were mainly interested in development of pattern *composition*, that is, how lines were organized in relation to each other in pattern-like structures and how these patterns change over time. For this purpose, we derived a more controlled stimulus set consisting of stimulus patterns that both closely resembled the originals in terms of compositional traits (e.g. line position and orientations) while maintaining number and length of lines within a pattern constant. The final stimulus set consisted of two tokens of each pattern type for each of the three periods and each of the two archeological sites making up twenty-four patterns in all each consisting of six lines of equal length.

We know little about the originally intended orientation of the patterns. Yet pattern orientations potentially have huge effects on the visual system (6-8). In an attempt to control for such effects, the two variations of each pattern type were rotated relative to each other by 35-180 degrees (for a more systematic experimental manipulation of orientation, see experiment 5, section 2.5). See full stimulus set in fig 2.

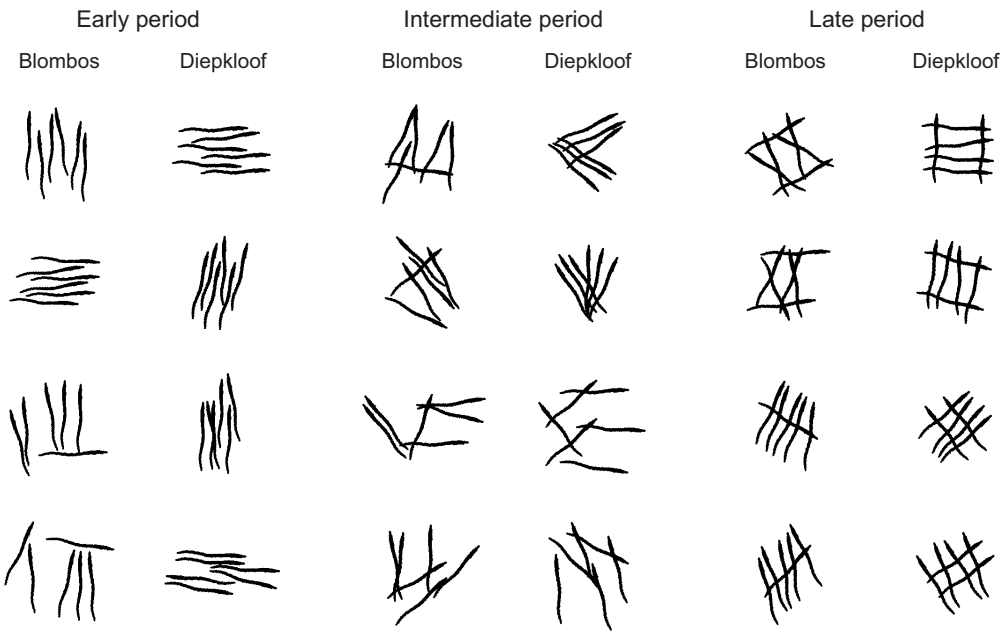

Fig. S2: The full stimulus set used across all experiments. Patterns are organized horizontally to represent the temporal development from early to late periods. Column 1, 3, and 5 are patterns derived from the Blombos outlines, while column 2, 4, and 6 are patterns derived from the Diepkloof outlines.

In order to ensure that the derived, controlled versions of the stimuli closely resembled the originals in terms of perceived and objective qualities, we carried out a number of validation checks reported below.

Furthermore, in addition to the controlled stimuli, in all experiments (except experiment 3) we ran an additional session using the original outlines as stimuli to test if we could replicate results even with the more unorganized original stimuli and thus ensure that our observations were not an artifact of the controlled versions. This was not possible in experiment 3 because the experiment involved placing and rotating six lines of identical length to reproduce a stimulus pattern, and the original patterns vary in number and length of lines.

Finally, to accommodate the concern that the temporal development could be different in the two archeological sites, control analyses are performed allowing temporal changes to vary by site.

1.1 Validation of stimuli

The controlled, derived stimulus patterns were validated in two complementary ways: i) with regard to how well they matched objective structural properties of the originals measured in image entropy and complexity, and ii) with regard to their perceived similarity with the originals as judged by human informants.

The purpose of both approaches was to ensure that the controlled stimulus items share critical structural features with the original outlines and, especially, that they display similar changes over the three periods of time.

1.1.1 Entropy

In order to obtain an objective measure of the order or predictability of elements in an image we can measure its Shannon entropy. The image entropy *S* is calculated based from histogram of a bitmap image file, as the sum of the probability *P* of a value *i* multiplied the binary log of the probability of that value:

$$S_{image} = \sum P_i \times Log_2(P_i)$$

Higher entropy values are associated with more randomness and thus lower predictability. In order to ensure that the controlled stimulus items resembled the original outlines in terms of image entropy, we derived entropy measures from each of the stimuli and each of the originals using the package CulturalAnalytics for *R* (9). In order to compare entropy as a function of time for stimuli and originals, entropy values were z-scored for each of the two sets. Fig. S3, panel a and b suggest that the derived stimuli items largely resemble the originals in terms of entropy as a function of time period[1].

1.1.2 Kolmogorov Complexity

One problem with entropy measures is that they do not take into consideration the spatial arrangement of elements – a critical element in this context. We therefore complement image entropy with a measure of Kolmogorov Complexity (10). Kolmogorov defines the complexity of an object to be the length of the shortest binary computer code that can describe it. Since Kolmogorov complexity cannot be calculated directly (11), we follow the procedure of (12), and use image compression techniques to assess the image file sizes of maximal lossless compression in order to approximate image complexity. In other words, image files of original outlines and derived experimental stimuli were compressed to the JPEG2000 format in Photoshop CC and resulting file sizes in kilobytes were z-scored and analyzed in order to compare image complexity as a function of period between originals and stimuli. Again, fig. S3, panel c and d suggest that the derived stimuli items resemble the originals in terms of Kolmogorov complexity as a function of time period.

[1] Since the number of comparisons is small, statistical treatment is not very meaningful.

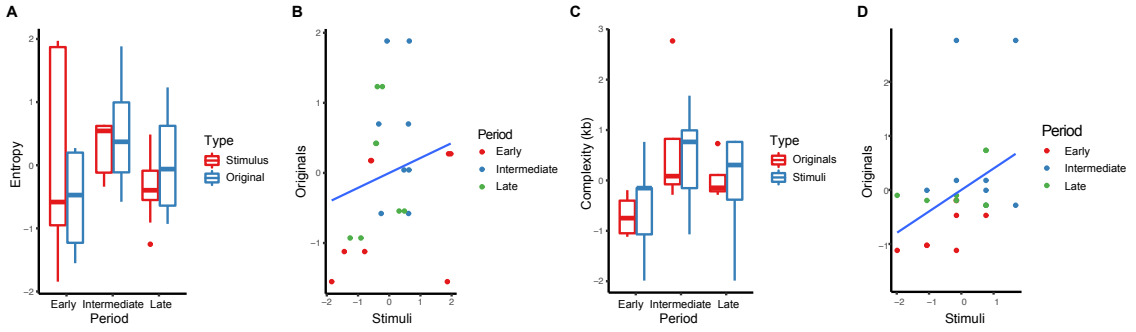

Fig. S3: Graphs displaying the results of stimulus validation procedures (see SI for details). In order to assure that the controlled experimental stimuli shared critical compositional properties with the original outlines, and more importantly, how these properties evolve as a function of period, image entropy and complexity was calculated for both original outlines and controlled stimuli. **a,** Shannon Entropy of original outlines and experimental stimuli as a function of the three periods. **b**, Entropy of originals and stimuli plotted against each other. **c**, Kolmogorov Complexity (measures by JPEG2000 image file size) of original outlines and experimental stimuli as a function of the three periods. **d**, Kolmogorov Complexity of originals and stimuli plotted against each other.

1.1.2 Survey

In order to validate the derived, controlled stimulus items in regard to their perceptual similarity to the originals, we collected similarity judgments from fifty-eight informants (23 m, 33 f, and 2 "other", mean age 33.76, sd 9.73). Fifty-five of the participants reported to have a university degree, two additionally reported having had archeological training. The study was conducted as an online survey using the resource SoSci Survey (13) and made available to the participants via www.soscisurvey.de. The link was distributed using social media.

The survey consisted of 24 trials. In each trial, participants were presented with a 4x3 grid showing the 12 original stimulus outlines on the upper part of the screen, and a single item from the controlled, derived stimulus set on the lower part of the screen. The position of the individual stimulus elements in the grid was counterbalanced between participants. Each of the controlled stimulus items was presented once, and the order was randomized between participants. For each of the 24 controlled stimuli, hence targets, participants were instructed to choose the original outline in the grid that they found most similar to the target pattern in the bottom in terms of line composition.

We analyzed participants' responses in two complementary ways, i) with respect to accuracy, and ii) with respect to intercoder reliability. A correct response was considered a situation in which a target was associated with an original pattern from the same period (early, intermediate, late). To assess accuracy, we first coded the data in terms of the extent to which participants associated a target to one of the 4 original

stimuli of the period that the target was designed to represent (by being derived from one of them).

To assess general classification accuracy, that is, the extent to which stimulus patterns were correctly associated with originals from the same period, we did pairwise comparisons between informants' responses against the correct mapping using Cohen's Kappa (14), which corrects for chance level effects. The mean kappa came out k = 0.5 (95% CI 0.45, 0.54), which is considered in the range "moderate agreement" (15).

An exploratory analysis of accuracy for the individual periods suggest that in general participants had greater difficulty recognizing targets intended to represent the earliest originals. We speculate that this is because the controlled stimuli always would consist of six lines while a characteristic trait of the earliest originals was that they often consisted of only two to four parallel lines. This, in some cases, made participant wrongly associate the "early" targets with intermediate period originals. We do, however, not consider this as a critical problem for the experiments.

To assess participants' agreement in their judgments, we calculated intercoder reliability using Fleiss's kappa (16) relying on the irr package for R (17). The kappa came out k = 0.426, which is again considered in the range "moderate agreement" (15).

The stimulus validations procedures indicate that even if the derived stimulus data set is standardized on some parameters (length and number of lines) it still maintains critical properties associated with the patterns' compositional structure and, importantly, how these behave as a function of period. The derived stimuli thus resemble the original outlines in their image entropy and complexity over the three periods and human informants recognize the stimulus items as belonging to the same periods as the originals they were derived from.

2 Experiments

2.1 Experiment 1 - Time to emerge

In order to investigate temporal developments in the low-level visual salience of the engravings (cf. $H1_{p1}$), experiment 1 used the psychophysics technique *continuous flash suppression* (CFS, 18, 19) A mirror stereoscope is used to present different visual stimuli, in this case suppressor and target stimuli, to each of the two eyes. When the visual stimuli sent to teach of the eyes differ, the more dominant stimulus - in this context of a colorful suppressor stimulus - will initially override the less dominant stimulus target due to the phenomenon of binocular rivalry. However, after a variable time, the stimulus target will gain ocular dominance and enter conscious perception. This *time to emerge* is considered indicative of the relative saliency of a stimulus with shorter emergence times being associated with higher salience.

2.1.1 Participants[2]:

Seventy-one participants (36 f, 33 m, 2 other), mean age 23.58 (SD 3.49) took part in the experiment. One participant was excluded due to too fast reaction times, consistently below 1 second, when the other participants had average reaction times above 2 seconds (analyses with and without the exclusion resulted in almost identical estimates). Participants were recruited through the Cognition and Behavior Lab participant database and were mainly students studying at Aarhus University, screened for normal or corrected-to-normal vision. All participants signed informed written consent in correspondence to the procedures of the local ethical committee and were compensated with DKK 100,- (~ $15) for their participation.

2.1.2 Apparatus and Procedure:

Upon entering the lab, participants were notified about the purpose and procedure of the experiment and then signed informed consent. For stimulus presentation, we used a 19" CRT monitor and a mirror stereoscope mounted on a chin rest at a viewing distance of 50 cm. (fig. S4). Prior to data collection, the apparatus was calibrated by instructing the participant to adjust two frames presented to each of the eyes until they had a perfect overlap using the arrow keys on a standard computer keyboard. For the actual experiment, we used a standard color noise suppressor updated at a constant rate of 100 Hz presented to the participant's dominant eye. Targets were faded in in either the left or right visual field after a randomized delay of 0 - 400 ms and stayed on screen until the participant responded, or a timeout occurred after 15 s. Participants were instructed to respond as soon as they could see the target by pressing the arrow key on the keyboard corresponding to the side of appearance of the target (left/right). Participants first completed a practice round of 20 trials with a replacement target stimulus (a cartoon figure). Then followed the actual experiment consisting of six repetitions of the 24 stimulus patterns organized in three blocks of 48 trials yielding in all 144 trials. The order of stimulus patterns was randomized within each repetition. Participants were invited to take a short break after each block. After the main session with the controlled stimuli followed a shorter session with the original stimulus outlines consisting of two blocks of 24 trials. Stimuli presentation and response recording was controlled using the Python based software PsychoPy2 (20).

Upon completion of the experiment, participants were informally debriefed in order to identify potential problems with perceiving or perceptually merging the stimuli. No such problems were recorded so data from all participants entered statistical analysis.

[2] Concerning sample sizes: Throughout the five experiments, we aimed to get as many participants as possible within the period we had labs and gear available. For each experiment, we reserved time and equipment in the Aarhus University Cognition and Behavior lab for a designated period of time. When this time frame (typically two weeks) expired, we considered the data collection complete for the individual experiment. This results in quite varied numbers of participants in the different experiments.

In the first experiment we investigated whether the period in which the engraving was produced (early, intermediate, late) had an effect on its perceptual saliency as measured in a continuous flash suppression paradigm. For all analyses, reaction time below 200 ms were excluded as implausible and likely due to errors. This excluded 39 trials (over 11570).

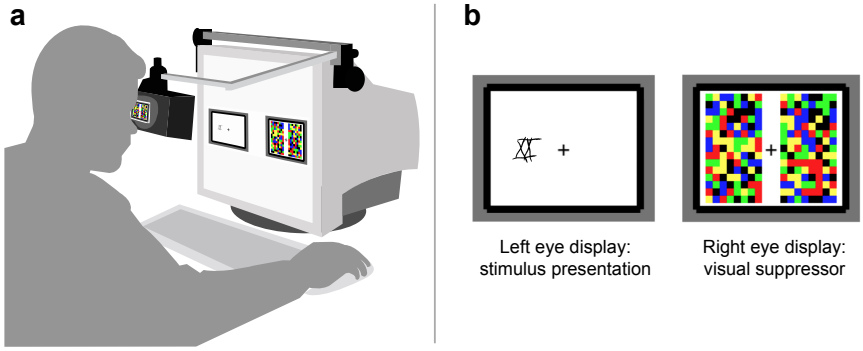

Fig. S4: Experimental setup for experiment 1. **a**, Participants were presented with conflicting stimuli to the two eyes which were fused using a mirror stereoscope. **b**, Example of stimulus presentation. A vivid flickering suppressor stimulus was presented to the dominant eye while the stimuli (engraved patterns) were presented to the non-dominant eye.

2.1.3 Analysis

2.1.3.1 Accuracy: We first assessed whether the participants' ability to identify the stimulus source (right or left eye) was affected by the period using Signal Detection Theory analysis (SDT) in the form of a Bayesian multilevel binomial regression with a probit link function (DeCarlo, 1998). The binomial response (judgment of "right" versus "left") was predicted by intercept (equivalent to criterion, or response bias) and actual source of the stimulus (right eye vs. left eye; equivalent to sensitivity or d prime). Criterion indicates whether the participants had a bias in judging stimuli as more likely to be from the right eye even when the actual source was left. Note that negative bias values indicate a tendency to respond "left." Sensitivity indicates how well the source can be correctly identified, controlling for criterion. We tested whether our independent variable (period of the stimulus) affected criterion and sensitivity by introducing main effects (relations to criterion) and interactions (relations to sensitivity) of period. Period was modeled as "-1" for the earliest artifacts, "0" for the intermediate range ones, and "1" for the most late. To reflect the ordinal nature of the period variable, we built two multilevel probit regressions to be compared: a model assuming a linear effect of period, and a

model including a quadratic term enabling potential slowdowns or speedups of cultural evolution over the three periods.

Linear model: Response $\approx$ $\text{Intercept}_{ps} + ß_{1ps}\,\text{Source} + ß_{2p}\,\text{Period} + ß_{3p}\,\text{Source Period}$

Quadratic model: Response $\approx$ $\text{Intercept}_{ps} + ß_{1ps}\,\text{Source} + ß_{2p}\,\text{Period} + ß_{3p}\,\text{Period}^2 + ß_{4p}\,\text{Source Period} + ß_{5p}\,\text{Source Period}^2$

Note that the subscript "p" indicates that the parameter values are expected to vary from participant to participant (random or varying effect of participant), while the subscript "s" indicates expected variability by stimulus (random or varying effect of stimulus). All effects of participant and stimulus were modeled as centered, that is, as expression of an underlying central distribution (fixed effects). To reduce overfitting and facilitate convergence of the model we used weakly skeptical priors centered at 0. The prior for intercept (expected criterion, or tendency to choose right, when the source is left and the period is intermediate) was modeled as a normal distribution with a standard deviation of 1. Priors for sensitivity and effects of period were modeled as a normal distribution with a standard deviation of 1 enabling a broad range of effects, but with expectations rapidly decreasing for more extreme effects. Priors for varying effects were modeled as normal distributions with a standard deviation of 1. The prior for correlations between random effects was modeled as a LKJ distribution with an eta of 5, with low expectations for extreme correlation values (close to 1 or -1). We performed prior predictive checks to ensure these were weakly regularizing priors for the model, that is, priors that would exclude implausibly high values for the effects of the experimental manipulations while not affecting too much our results (21).

In order to more directly test the support for our hypotheses, we calculated an evidence ratio in the form of the posterior probability of the directed hypothesis against the posterior probability of all the alternative, that is if we expected a positive effect of period, we would test the posterior probability of this against the posterior probability of a null or negative effect. This procedure is equivalent to a Bayes Factor test, but extended to directional hypotheses (e.g. increased perceptual saliency for later engravings) instead of point-wise hypotheses (e.g. a relation of exactly 0, which is highly implausible). We also report the credibility of the estimated parameter distribution, that is, the probability that the true parameter value is above 0 if the mean estimate is positive, or below 0 if it is negative. When our hypotheses were not supported by the data (evidence ratio below 3), we also tested for evidence in favor of the null (equivalent to a Bayes Factor of the null against the alternative hypothesis, or BF01).

The models were fitted using Hamiltonian Monte Carlo samplers (as implemented in Stan, relying on rstan and brms (22, 23)), with 2 parallel chains with 8000 iterations each, an adapt delta of 0.99 and a maximum tree-depth of 20 to ensure no divergence in the estimation process. Estimates from the models are reported as mean and 95% Credibility Intervals (CI) of the posterior estimates.

The quality of the models was assessed by: i) ensuring no divergences in the estimation process; ii) visual inspection of the Markov chains to ensure stationarity and overlapping between chains; iii) ensuring Rhat statistics to be minor than 1.1 and number of effective samples to be above 200; and iv) performing predictive posterior checks to ensure no obvious issue in the model predictions (akin to residuals checks). The relevance of the predictors (quadratic effects of period, control analyses by site and stimulus type) was assessed by stacking weights model comparison relying on estimated out-of-sample error via Leave-One-Out Information Criteria (LOOIC, (24)). LOOIC is an estimation of out-of-sample error, that is, an estimation of the generalizability of the model to new data, accounting for overfitting and uncertainty in the predictions.

2.1.3.2 Reaction Time: We excluded incorrect trials from the current analysis, to focus on the time used to correctly identify stimuli. This excluded 195 trials, bringing the dataset to 11263 trials. We assessed the relation between period and perceptual saliency also as the time the stimuli took to be recognized (reaction time during trials with an accurate response). We modeled reaction time according to a gamma distribution with a logarithmic link, given it presented a long tail (non-null possibility of reaction times up to 15 seconds) and reaction times were expected to be all above 200 ms. Gamma distributions are defined according to two parameters, rate and shape (roughly corresponding to a measure of variance in the outcome). We built models first conditioning rate only on period, stimuli and participants, then also shape on period and participants.
To reduce overfitting and facilitate convergence of the model we used weakly skeptical priors centered at 0. The prior for the intercept was modeled as a normal distribution with a standard deviation of 0.3. Priors for effects of period were modeled as a normal distribution with a standard deviation of 0.1. Priors for varying effects were modeled as normal distributions with a standard deviation of 0.3. The prior for correlations between random effects was modeled as a LKJ distribution with an eta of 5, with low expectations for extreme correlation values (close to 1 or -1). The model building and comparison procedures were analogous to those employed in the analysis of accuracy.

Control analysis by Site: We assessed whether the site of origin of the engravings had an impact on the effects observed by creating two additional models for the SDT model of accuracy:

$$\text{Response} \approx \text{Intercept}_{ps} + ß_{1ps}\,\text{Source} + ß_{2p}\,\text{Period} + ß_{3p}\,\text{Source Period} + ß_{4p}\,\text{Site} + ß_{5p}\,\text{Source Site} + ß_{6p}\,\text{Site Source}$$

$$\text{Response} \approx \text{Intercept}_{ps} + ß_{1ps}\,\text{Source} + ß_{2p}\,\text{Period} + ß_{3p}\,\text{Source Period} + ß_{4p}\,\text{Site} + ß_{5p}\,\text{Source Site} + ß_{6p}\,\text{Site Period} + ß_{7p}\,\text{Source Site Period}$$

We analogously created two additional Gamma regression for reaction time:

$$\text{Reaction Time} \approx \text{Intercept}_{ps} + ß_{1p}\,\text{Period} + ß_{2p}\,\text{Site}$$

$$\text{Reaction Time} \approx \text{Intercept}_{ps} + ß_{1p}\,\text{Period} + ß_{2p}\,\text{Site} + ß_{3p}\,\text{Site Period}$$

The additional parameters related to site had weakly informative priors modeled as normal distributions centered at 0, with a standard deviation of 1 for the SDT model and 0.3 for the Gamma model. We then assessed whether the second model credibly improved our ability to explain the data via LOOIC model comparison, and if so we explored the interaction effects, that is, the difference in effects of period between the two sites.

Control analysis of the original engravings: We assessed whether the original engravings showed analogous effects as the controlled stimuli by adding them to the dataset and creating two additional models (for both accuracy and reaction time):

$$\text{Outcome} \approx \text{Intercept}_{ps} + ß_{1p}\,\text{Period} + ß_{2p}\,\text{Original}$$

$$\text{Outcome} \approx \text{Intercept}_{ps} + ß_{1p}\,\text{Period} + ß_{2p}\,\text{Original} + ß_{3p}\,\text{Period * Original}$$

The parameters related to site had weakly informative priors modeled as normal distributions centered at 0, with a standard deviation of 1 for the SDT model and 0.3 for the Gamma model. We then assessed whether the second model credibly improved our ability to explain the data via LOOIC model comparison, and if so we explored the interaction effects, that is, the difference in effects of period between originals and controlled stimuli.

2.1.4 Results

2.1.4.1 Accuracy: The model with a linear effect of period was better than that with a quadratic effect: LOOIC 1823.94 (SE: 101.28) vs. 1827.35 (SE: 101.54), with a credible difference of -3.41 (SE = 1.27). We therefore opted to focus on the simpler model. The

model provides some evidence that sensitivity increases from earlier to later stimuli (see table S1).

2.1.4.2 Reaction times: The model conditioning rate on linear effects of period was credibly better than that with a quadratic effect: LOOIC 25678.88 (SE: 275.05) vs. 25681.51 (SE: 275.16), with a credible difference of 2.63 (SE = 0.92). However, the model also conditioning the shape parameter on linear effects of period credibly outperformed the previous one: LOOIC 24035.02 (SE: 288.51) with a credible difference of 1643.86 (SE: 110.01). Additionally, a posterior predictive check clearly revealed a superior fit for this latter model. We therefore opted to focus on the gamma model conditioning both rate and shape on linear effects of period, stimuli and participants. Note that the linear effects of period on rate were identical (to the second decimal) in both models. The model indicates that later stimuli are related to faster reaction times in correctly identifying them. On average, earlier stimuli were identified in 2.27 seconds, intermediate stimuli in 2.03 seconds and late ones in 1.82 seconds (table S2 and fig. 3a in the main article).

2.1.4.3 Controlling for site: Signal Detection Theory model of accuracy. Controlling for site did not reveal any likely interaction between site and period. The original model had a LOOIC of 1712.28 (SE: 101.62); adding a main effect of site yielded a LOOIC of 1717.49 (SE: 101.87) and did not credibly improve the model (LOOIC difference of -5.21, SE: 3.58). Adding the interaction of site and period yielded 1728.91 (SE: 102.67) and credibly decreased the performance of the model (LOOIC difference of -16.63, SE: 4.28).

Gamma model of Reaction Time. Controlling for site did not reveal any likely interaction between site and period: adding a main effect of site yielded a LOOIC of 24004.67 (SE: 282.81) and adding the interaction of site and period yielded 24012.79 (SE: 282.44), with no credible difference between the two: -8.12 (SE: 14.47).

2.1.4.4 Controlling for type: Signal Detection Theory model of accuracy. Controlling for different effects in the originals did not reveal any likely interaction between type and period: adding a main effect of site yielded a LOOIC of 1839.31 (SE: 106.38) and adding the interaction of type and period yielded 1840.29 (SE: 106.47), with no credible difference between the two: -0.98 (SE: 3.71).

Gamma model of Reaction Time. Controlling for type revealed a likely interaction between type and period: adding a main effect of type yielded a LOOIC of 26706.41 (SE: 302.90) and adding the interaction of type and period yielded 26685.36 (SE: 300.84), with a credible difference between the two: 21.05 (SE: 10.48). However, assessing the

latter model revealed that while the stimuli took less time to be perceived than the originals (rate: 0.04, 95% CI: -0.04 0.12; shape: -0.29, 95 % CI: -0.48-0.10), there was no credible interaction of type with period (rate: 0.00, 95% CI: -0.07 0.08; shape: -0.01, 95% CI: -0.12 0.10) and the estimates of effects of period stayed analogous (rate: -0.11, 95% CI: -0.18 -0.05; shape: 0.20, 95% CI: 0.10 0.29).

Table S1: Experiment 1 – Criterion and Sensitivity

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Criterion	β = -2.23, 95% CI = -2.41 -2.09	4732, 1.00	
Sensitivity	β = 4.63, 95% CI = 4.40 4.88	4655, 1.00	
Period	β = -0.05, 95% CI = -0.17 0.07	8000, 1.00	
Sensitivity : Period	β = 0.12, 95% CI = -0.09 0.32	5975, 1.00	
Random Effects	*SD*		
Stimulus criterion	0.08, 95% CI = 0.00 0.19	3036, 1.00	
Stimulus sensitivity	0.17, 95% CI = 0.02 0.34	2438, 1.00	
Participant criterion	0.44, 95% CI = 0.32 0.59	2726, 1.00	
Participant sensitivity	0.58, 95% CI = 0.40 0.80	2238, 1.00	
Participant period criterion	0.08, 95% CI = 0.00 0.22	3177, 1.00	
Participant period sensitivity	0.13, 95% CI = 0.00 0.33	2519, 1.00	

Table S2: Experiment 1 – Time to emerge

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Intercept (rate)	β = 0.72, 95% CI = 0.65 0.79	479, 1.00	
Period (rate)	β = -0.11, 95% CI = -0.15 -0.07	2042, 1.00	>1000 / 100%
Intercept (shape)	β = 1.73, 95% CI = 1.59 1.86	755, 1.00	
Period (shape)	β = 0.19, 95% CI = 0.13 0.25	3588, 1.00	
Random Effects	*SD*		
Stimulus intercept (rate)	0.07, 95% CI = 0.05 0.10	2285, 1.00	
Participant intercept (rate)	0.27, 95% CI = 0.23 0.32	919, 1.00	
Participant period (rate)	0.04, 95% CI = 0.01 0.03	4196, 1.00	
Participant intercept (shape)	0.57, 95% CI = 0.48 0.67	1532, 1.00	
Participant period (shape)	0.21, 95% CI = 0.17 0.27	4826, 1.00	

Response times (time to emerge) of correct trials as a function of period for the shape and rate parameters of the gamma model (see SI for detailed analysis). The *evidence ratio* is the posterior probability of the directed hypothesis against the posterior probability of all alternative hypotheses. The *credibility* is the probability that the true parameter value of the estimated parameter distribution is above 0 if the mean estimate is positive, or below 0 if it is negative, given the assumptions of the model.

2.2 Experiment 2 - Intentionality

In order to test hypothesis 2, that is, whether engraved patterns from the three periods become increasingly recognized as purposefully produced by humans, we conducted a two-item-forced-choice ranking study. This enabled us to calculate how the perceived intentionality of individual patterns vary as a function of the period of their origin.

2.2.1 Participants

Fifty-one participant (27 f, 21 m, 3 other) mean age 23.2 (SD 3.2). All participants of experiment 2 also took part in experiment one, that is, recruitment, consent procedures and compensation are identical to experiment 1. Participants always completed experiment 1 before experiment 2.

2.2.2 Apparatus and Procedure:

Participants were seated at a standard windows computer with a 22" LCD screen. After receiving instructions, through 276 trials, they were presented with all pairwise combinations of stimulus patterns in randomized order. For each trial, the participants task was to indicate which of two competitor stimuli presented next to each other on the screen they found was more likely to have been intentionally produced by a human. Participants responded by pressing the arrow key on the computer keyboard that corresponded to the side of presentation of the stimulus (left/right). The task was self-paced with no time-out. After completion of the controlled stimuli followed a shorter session of 66 trials with the original outlines following identical procedures.

2.2.3 Analysis

In order to assess whether more stimuli were perceived as more intentional we modeled the data according to an outcome contest model. Outcome contest models have been developed to model sport tournaments, where at each game (in our case trial) individual variability has to be modelled for two teams (in our case stimuli) separately (25).

$$Judgment_1 \approx int_{p1} - int_{p2}$$

Where the likelihood function is a Bernoulli distribution, $Judgment_1$ indicates the log odds of choosing stimulus 1 as intentional when compared to stimulus 2 and int_{pn} is the estimated intentionality score a participant j perceives in a given stimulus. Given two stimuli are simultaneously presented, the probability of choosing of one over the other depends on their relative scores. int_{pn} is further defined, if linear effects of period are assumed, as:

$$int_{pn} \approx a_{sn} + ß_p * Period$$

Otherwise as

$int_{pn} \approx a_{sn} + ß_{1p} * Period + ß_{2p} * Period^2$

Note that the intercept is conditioned on stimulus (random or varying intercept), while the effects of period are condition on participant (random or varying slope of period on participant). Intercept by participants were unnecessary, since each trial is a comparison between two stimuli and the intercept would cancel out. Priors for the intercept, for the period parameter, as well as for effects of site and stimulus type were weakly informative: normal distributions centered at 0, with a standard deviation of 1. Individual variability for effects of period, type and site by participant and for stimuli were analogously modeled as normal distributions centered at 0, with a standard deviation of 0.5. Correlations between random effects had a LKJ prior with an eta of 5. Model building and model comparison procedures were analogous to those in previous models.

2.2.4 Results

The model conditioning rate on linear effects of period was credibly better than that with a quadratic effect: LOOIC 15419.3 (SE: 108.2) vs. 16855.8 (SE: 95.6), with a credible difference of 1465.3 (SE: 140.6). We therefore focused on the linear model. The model indicated that the average early stimulus had a 18% and a 33% chance of being indicated as intentional respectively against a late and an intermediate stimulus. An average intermediate stimulus had a 33% and a 67% chance of being indicated as intentional respectively against a late and an early stimulus (see table S3 and figure 3b of the main article).

2.2.4.1 Controlling for site: Controlling for site did not reveal any likely interaction between site and period. The original model had a LOOIC of 15419.3 (SE: 108.2); adding a main effect of site yielded a LOOIC of 16575.1 (SE: 100.6), while adding the interaction of site and period yielded 16005.7 (SE: 106.3), neither of which credibly improved the performance of the model.

2.2.4.1 Controlling for type: Controlling for different effects in the originals did not reveal any likely interaction between type and period. A main effect of type would cancel out, since stimuli are only compared with stimuli and originals with originals. Adding an interaction between type and period yielded a LOOIC of 15662.1 (SE: 96.9). The estimated interaction between type and period was -0.01 (95% CI: -0.54, 0.50), indicating that a difference in effects for the originals compared to the stimuli is not credibly supported by the evidence.

Table S3: Experiment 2 - Intentionality judgements

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Period	β = 0.73, 95% CI = 0.4 1.05	786, 1.00	**587 / 99.9%**
Random Effects	SD		
Stimulus intercept	0.54, 95% CI = 0.32 0.99	1525, 1.00	
Participant period	0.72, 95% CI = 0.5 1.12	1023, 1.00	

Participants' ranking of intentionality of the stimulus patterns (i.e. how likely the stimulus pattern is to be purposefully produced by a human) as a function of period.

2.3 Experiment 3 - Memorability

Experiment 3 tests the hypothesis that, as a function of time, the engraved patterns become easier to encode and reproduce from memory. It relies on a delayed reproduction task where participants are presented with a stimulus pattern and has to recreate it after a short delay. We measure the reproduction fidelity as the mean squared error of pixels displaced between stimulus and reproduction and predict the error to decrease over the three periods.

2.3.1 Participants:

Seventy-five participants (45 f, 25 m, 5 other) mean age 24.6 (SD 4.4), different from the ones in experiment 1 and 2, took part in the experiment. Participants were recruited through the Cognition and Behavior Lab participant database and were mainly students studying at Aarhus University. All participants signed informed written consent in correspondence to the procedures of the local ethical committee and were compensated with DKK 100,- (~ $15) for their participation.

2.3.2 Apparatus and Procedure:

Participants were seated at a standard windows computer with a 22" LCD screen running the Python based experimental software PsychoPy2 (20). In each trial, participants were presented with a stimulus pattern for 3 s. Then the stimulus would disappear for 2 s after which the participant was instructed to reproduce the pattern as accurately as possible. This was done in the following way: a single pattern line would appear on the screen replacing the mouse cursor. The participant could then move it to a position and rotate the line using the mouse scroll wheel. Once the participant found it to be in the right position, she could click the left mouse button which would leave the line at the designated position, and the next line would appear. The same procedure repeated until all six lines were placed in positions to reproduce the stimulus pattern from the participant's memory.

Participants would first go through a practice trial with a non-target stimulus. Then followed twenty-four trials with patterns from the three periods presented in randomized order.
By the end of each trial the experimental software would save a bitmap screenshot of the resulting reconstructed pattern.

2.3.3 Analysis

The reproduction accuracy was calculated as the *mean squared error* (mse) between the bitmap image of the stimulus pattern and the corresponding response bitmap screenshot in pixels:

$$mse = \sum(image_{stim} - image_{copy})^2$$

A lower number is thus indicative of lower reproduction error, that is, higher reproduction fidelity. This measure was then transformed on a 0 to 1 scale, by dividing all values by the maximum error to facilitate estimation of the model. The models were linear regressions following the same procedure as in previous experiments. As weakly informative priors for all parameters we chose normal distributions centered at 0. The intercept had a standard deviation of .3, the beta coefficients of .1, and the individual variability of the engravings and participants random effects a standard deviation of 0.5. Correlations between random effects had a LKJ prior with an eta of 5. A prior predictive check showed that the priors covered a much broader range than the actual data.
Note that given the structure of the task - reproducing previously seen stimuli using six lines - it was not possible to run the original engravings as stimuli in this experiment.

2.3.4 Results

The quadratic component of time did not credibly improve the model, with a LOOIC of -2475.75 (SE: 64.95) against -2474.68 (SE: 64.86) and a difference of -1.07 (SE: 1.11). We therefore chose to focus on the linear model. The model indicates an average reduction in error of 0.05 per period, with early stimuli producing average errors of 0.71, intermediate ones of 0.64 and late ones of 0.59 (see table S4 and figure 3c of the main article).

Controlling for site did not improve the model: from a LOOIC of -2475.75 (SE: 64.95) adding a main effect of site yielded -2474.44 (SE: 65.19) and adding the interaction of site and period yielded -2472.76 (SE: 65.12). Neither had credible differences from the base model (1.31, SE: 2.45; and 1.68, SE: 3.13).

Table S4: Experiment 3 - Memory and reproduction fidelity

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Intercept	β = 0.64, 95% CI = 0.61 0.67	1903, 1.00	
Period	β = - 0.05, 95% CI = -0.08 -0.02	2000, 1.00	**887.89**
Sigma	0.11, 95% CI = 0.10 0.11	8000, 1.00	
Random Effects	SD		
Drawing intercept	0.06, 95% CI = 0.05 0.09	2712, 1.00	
Participant Intercept	0.05, 95% CI = 0.04 0.06	2724, 1.00	
Period over Participant	0.01, 95% CI = 0.00 0.02	3107, 1.00	
Correlation Participant / Period	-0.14, 95% CI = -0.63 0.42	8000, 1.00	

Pattern reproduction error, measured in mean squared error of displaced pixels, as a function of period.

2.4 Experiment 4 - Cultural traditions

In order to investigate if engraved patterns evolved elements of cultural traditions or *style* over the three periods making it increasingly easier to recognize patterns as originating from one or the other site, we conducted two complementary experiments. Experiment 4b was part of the initial submission, but based on suggestions from a reviewer it was subsequently replaced by experiment 4a. While only experiment 4a is reported in the main article, here we include both experiments.

2.4.1 Experiment 4a

Experiment 4a used a two-item-forced choice paradigm. Participants were instructed to indicate whether a target pattern was likely to originate from the same site as one or another competitor pattern coming from each of the two sites. We hypothesized that participants would more successfully recognize patterns from the later periods as belonging to the same site suggesting that the patterns evolved elements of style over time.

2.4.2 Participants

XX participants (X f, X m, X other), mean age X (SD X), different from the ones in the previous experiments, took part in this experiment. Participants were recruited through the Cognition and Behavior Lab participant database and were mainly students studying at Aarhus University. All participants signed informed written consent in correspondence to the procedures of the local ethical committee and were compensated with DKK 50,- (~ $7.50) for their participation.

2.4.3 Apparatus and Procedure

Participants were seated at a standard windows computer with a 22” LCD screen running the Python based experimental software PsychoPy3 (Peirce, 2007). After a practice round, the main experiment proceeded through 228 trials. In each trial, a target stimulus pattern was presented in the center of the lower part of the screen, while two competitor stimuli would be presented side-by-side on the upper part of the screen. The target would either come from Blombos or Diepkloof, while one competitor would be from Blombos

and the other from Diepkloof. The order of items, periods, and position of competitors was randomized within and between participants. The task of the participant was to indicate if they thought the target originated from the same site as the competitor to the left or right by pressing the corresponding arrow key on the keyboard. After the main experiment followed a shorter session with the original outlines following identical procedures.

2.4.4 Analysis

We first assessed whether the participants' ability to match target stimuli to stimuli from the same site was affected by the period using Signal Detection Theory analysis (SDT). Then we modelled reaction times in correct trials as a gamma Regression model, as in experiment 1. In order to account for the variation in similarity between pairs of stimuli, we added a random or varying effect by Stimulus Pair. In the gamma regression model, conditioning shape on our predictors caused a large number of divergences. The model therefore only conditions the rate parameter on the predictors. All other statistical procedures were identical to those in the previous experiments.

2.4.5 Results

2.4.5.1 SDT: The model with a quadratic effect of time was not credibly better than that with a linear effect: LOOIC 9414.24 (SE: 96.54) vs. 9414.79 (SE: 96.41), with a difference of 0.55 (SE = 1.08). We therefore opted to focus on the simpler linear model.

The model indicates credible changes in criterion and sensitivity over time. Criterion credibly decreases over time, while sensitivity increases (see table S5 and figure 3d in article). In other words, stimuli from more recent times are easier to correctly match to stimuli from the same site, with false positives decreasing and true positives increasing.

2.4.5.2 RT: The model with a linear effect of period was credibly better than that with a quadratic effect: LOOIC 18522.06 (SE: 233.84) vs. 18949.48 (SE: 228.84), with a credible difference of 427.42 (SE = 53.44). We therefore focus on the linear model. The model indicates a credible effect of period on reaction time, with speed of accurate responses being faster for more recent stimuli than for earlier ones. Participants take on average 2.27 seconds to correctly match stimuli from the earlier period, 2.12 seconds for stimuli from the intermediate period and 1.97 seconds for stimuli from the most recent period (see table S6).

Table S5: Experiment 4a – Criterion and Sensitivity

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Criterion	β = -0.01, 95% CI = -0.31 0.30	4404, 1.00	
Sensitivity	β = 0.30, 95% CI = -0.19 0.79	4743, 1.00	
Period	β = -0.21, 95% CI = -0.36 -0.05	4287, 1.00	
Sensitivity : Period	β = 0.31, 95% CI = 0.05 0.57	4013, 1.00	**101.56 / 0.99**
Random Effects	SD		
Stimulus criterion	0.52, 95% CI = 0.43 0.63	3333, 1.00	
Stimulus sensitivity	0.99, 95% CI = 0.82 1.19	3248, 1.00	
Stimulus pair criterion	0.28, 95% CI = 0.22 0.35	4393, 1.00	
Stimulus pair sensitivity	0.52, 95% CI = 0.42 0.64	4366, 1.00	
Participant criterion	0.06, 95% CI = 0.15 1.00	3183, 1.00	
Participant sensitivity	0.09, 95% CI = 0.00 0.20	3446, 1.00	
Participant period criterion	0.04, 95% CI = 0.00 0.07	2276, 1.00	
Participant period sensitivity	0.05, 95% CI = 0.01 0.10	1688, 1.00	

Signal Detection Theory analysis of accuracy of pattern matching as a function of period.

Table S6: Experiment 4a - Reaction time

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Intercept (rate)	β = 0.89, 95% CI = 0.71 1.08	2916, 1.00	
Period (rate)	β = -0.07, 95% CI = -0.15 -0.00	2862, 1.00	**48.69 / 0.98**
Intercept (shape)	β = 1.08, 95% CI = 0.95 1.22	2114, 1.01	
Random Effects	SD		
Stimulus intercept (rate)	0.09, 95% CI = 0.06 0.13	4377, 1.00	
Stimulus pair intercept (rate)	0.14, 95% CI = 0.11 0.17	3514, 1.00	
Participant intercept (rate)	0.37, 95% CI = 0.29 0.46	3848, 1.00	
Participant period (rate)	0.08, 95% CI = 0.06 0.11	3892, 1.00	
Participant intercept (shape)	0.41, 95% CI = 0.32 0.53	3544, 1.00	

Reaction time of pattern discriminability as a function of period.

2.4.5.3 Control analysis against originals.

SDT: The model including an interaction between Type and Period was not a credibly better model than that without: LOOIC of 10313.50 (SE: 94.61) vs. 10312.75 (SE: 94.47), with a difference of -0.74 (SE: 2.09). This indicates that the controlled stimuli generate similar patterns to the original ones.

RT: The model including an interaction between Type and Period was not a credibly better model: LOOIC of 19852.39 (SE: 236.06) vs. 19852.61 (SE: 235.18), with a difference of 0.22 (SE: 3.55). This indicates that the controlled stimuli generate similar patterns to the original ones.

2.4.6 Experiment 4b

Experiment 4b used a sorting task. Participants were asked to sort patterns from each of the periods in two equally sized groups based on their intuitions about which of them belong together in terms of line composition. We hypothesized that participants would group stimuli from the later periods in ways corresponding more accurately to the actual origin of the patterns from the Blombos and Diepkloof sites indicating that the patterns evolve elements of explicit style.

2.4.7 Participants

Fifty-four participants (27 f, 26 m, 1 other), mean age 24.13 (SD 3.57), different from the ones in the previous experiments, took part in this experiment. Participants were recruited through the Cognition and Behavior Lab participant database and were mainly students studying at Aarhus University. All participants signed informed written consent in correspondence to the procedures of the local ethical committee and were compensated with DKK 50,- (~ $7.50) for their participation.

2.4.8 Apparatus and Procedure

Participants were seated at a standard windows computer with a 22" LCD screen running the Python based experimental software PsychoPy2 (Peirce, 2007). The instruction was to sort stimulus patterns from each of the three periods into equally sized groups. Each trial would present the participant with eight patterns from the same period (early, intermediate, or late), on a horizontal line in the lower part of the screen. The position of patterns in the line was randomized. The task of the participant was then to use the computer mouse to drag and drop individual patterns to one of two areas in each side of the upper part of the screen thus dividing the patterns in two equally sized groups based on the participants intuitions about which of the patterns would originate from the same or different archeological sites (fig. S5a). Once all patterns were placed in one of the target fields, the participant would press the enter key on the keyboard in order to proceed to the next trial presenting patterns from another period. The order of periods was counterbalanced between participants. Once participants had completed the sorting task for the controlled stimulus patterns followed a session with the same task applied to the original outlines.

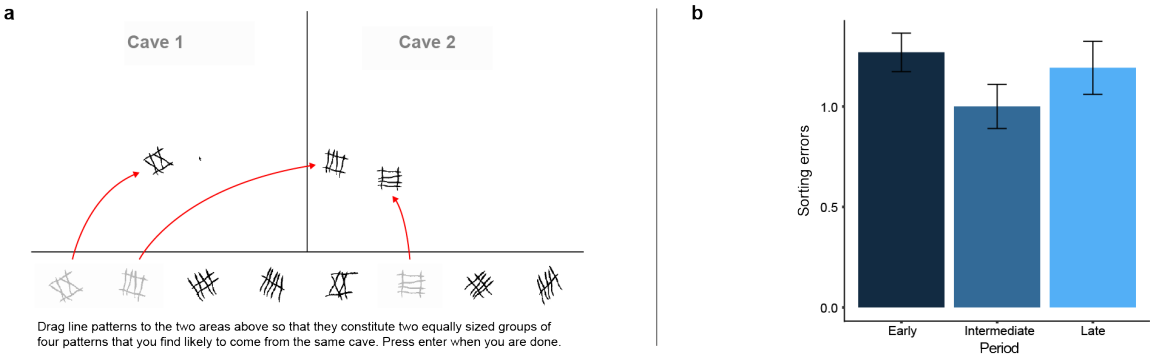

Figure S5: Experiment 4 – experimental setup and results. **a**, Depiction of the experimental setup for the sorting task. Participants were presented with the patterns from one of the three periods at a time in the lower part of the screen, and were instructed to drag them to one of the designated regions on the upper part of the screen based of their judgments on which patterns belonged together and thus were likely to originate from the same archeological site. **b**, Results from the sorting task, measured as mean sorting error as a function of period. The x-axis represents period, the y-axis the average number of miscategorized stimuli.

2.4.9 Analysis

Any resulting group of stimuli could thus have 0 to 4 stimuli from the same site. Four stimuli from the same site counted as perfect performance (2 points), 3 as 1 error (thus giving 1 point), 2 as 2 errors (thus giving 0 points), 1 stimulus would count as 1 error (as the group would be identified as pertaining to the other site) and 0 as 0 errors. We thus modeled the data as a multilevel aggregated binomial regression model, with points (out of a maximum of 2 possible points) as the outcome, and a logistic link function. The remainder of the statistical procedure was analogous to the other experiments.
We chose weakly skeptical priors: normal distributions centered at 0. The intercept and beta coefficients had a standard deviation of 1, the standard deviation of the random effects a standard deviation of 0.5. The correlations between random effects had a LKJ prior with an eta of 5.

2.4.10 Results

Adding a quadratic component to the model did not credibly improve it: LOOIC of 328.38 (SE: 10.38) against 328.15 (SE: 10.37) with a difference of 0.23 (SE: 1.18). No credible effects of time could be inferred, BF01 = 4.38, credibility = 81% (see Table S5 and Figure S5b).

2.4.10.1 Control analysis against originals: The model including an interaction between Type and Period was a credibly better model: LOOIC of 627.38 (SE: 13.38) vs. 653.69 (SE: 13.13), with a difference of 26.30 (SE: 9.17). However, further explorations of the model indicate that where the stimuli have a very unreliable trend of later stimuli being more difficult to categorize, original engravings have an equally unreliable trend of later engravings being easier to categorize. We speculate that in regard to the original outlines, participants might base their judgments on properties of the patterns that we control in the experimental stimuli (number, thickness, and length of lines), rather than the compositional traits (relative position and orientation of lines), which are stable across the two sets (the stimuli and the original outlines). Participants' responses could thus be contingent on (and possibly confounded by) differences in the way that the original outlines were produced and presented in their respective articles: these are more detailed for the Blombos materials (e.g. in terms of varied thickness of lines) and more stylized for the Diepkloof materials (same line thickness).

Table S7: Experiment 4 – Sorting task

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Intercept	β = -0.37, 95% CI = -0.70 -0.02	5813, 1.00	
Period	β = 0.07, 95% CI = -0.40 0.53	5361, 1.00	**1.56 / 61%** **BF01: 4.38, Credibility: 81%**
Random Effects	*SD*		
Participant Intercept	0.77, 95% CI = 0.40 1.15	3785, 1.00	
Period over Participant	1.24, 95% CI = 0.77 1.73	4337, 1.00	
Correlation Participant / Period	0.73, 95% CI = 0.31 0.96	2473, 1.00	

Mean error of grouping of stimulus patterns as a function of period. Notice that since the hypotheses were not supported by the data (evidence ratio below 3), we also report the Bayes Factor (BF01), which provide an index of the evidence in favor of the null against the alternative hypothesis.

2.5 Experiment 5 - Discriminability

In order to investigate whether the engraved patterns evolved to become increasingly easy to discriminate from each other over time, we conducted a discrimination task, where participants had to indicate as fast as possible whether a target stimulus was identical to one or another competitor stimulus. Faster reaction times would thus indicate better discrimination. We used the same experiment to investigate two hypotheses. First, we contrasted discrimination between patterns coming from the same site and patterns coming from different sites. If patterns served as markers of cultural identity, we should predict that they would evolve to become increasingly similar within site (and thus harder to discriminate) while they would become increasingly dissimilar between the two sites (making them faster to discriminate). Second, if patterns served as denotational symbols pointing to each their referent meaning, we would predict discriminability of patterns to increase also within each of the two sites, as it becomes important not to confuse one symbol from another.

2.5.1 Participants

Fifty-seven participants (33 f, 24 m) mean age 23.91 (SD 4.75), different from the ones in the previous experiments, took part in this experiment. Participants were recruited through the Cognition and Behavior Lab participant database and were mainly students studying at Aarhus University. All participants signed informed written consent in correspondence to the procedures of the local ethical committee and were compensated with DKK 50,- (~ $7.50) for their participation.

2.5.2 Procedure and Apparatus

Participants were seated at a standard windows computer with a 22” LCD screen running the Python based experimental software PsychoPy2 (20). After a practice round, the main experiment proceeded through 396 trials. In each trial, a target stimulus pattern was presented in the center of the lower part of the screen, while two competitor stimuli

would be presented side-by-side on the upper part of the screen. The target would be identical to one of the competitors, and the task of the participant was to indicate as fast as possible whether it was the competitor to the left or right by pressing the left or right arrow key on the keyboard using index and middle finger. In order to control for robustness of the discrimination, the target would either be presented at the same orientation or rotated 45 or 135 degrees relative to the matching competitor pattern. This is motivated by the observation that discrimination is generally subject of the *oblique effect*, that is, more difficult when line patterns are not presented in horizontal or vertical orientations (6, 26). By presenting stimulus patterns in different orientations we also wanted to ensure that discrimination was based on perception of the pattern as a whole (i.e. the relative position and intersection of multiple lines) and not a simple heuristic of attending to the absolute orientation of the pattern which would otherwise in many cases be a reliable cue to discrimination. After the main experiment followed a shorter session with the original outlines following identical procedures.

2.5.3 Analysis

We first assessed whether the participants' ability to correctly match the target to one of the competitors was affected by the period using Signal Detection Theory analysis (SDT), then their reaction times in correct trials as a gamma Regression model. In order to account for the variation in similarity between pairs of stimuli we added a random or varying effect by Stimulus Pair. In the gamma regression model, conditioning shape on our predictors caused a large number of divergences, the model therefore only conditions the rate parameter on the predictors. All other statistical procedures were identical to those in the previous experiments.

2.5.4 Results

2.5.4.1 SDT: The model with a quadratic effect was credibly better than that with a linear effect: LOOIC 8220.57 (SE: 88.96) vs. 20049.28 (SE: 98.45), with a credible difference of 11828.71 (SE = 102.56). We therefore opted to focus on the quadratic model. The model indicates a change in criterion: for early stimuli the criterion is -0.13, for intermediate ones is 0.37, for late ones is -0.53. However, no credible changes are highlighted for sensitivity for neither between nor within site discrimination (see Table S6).

Adding a parameter to control for the rotation of the stimuli credibly worsened the performance of the model in terms of LOOIC (21032.71, SE = 159.22), credibly higher than the baseline model (difference: 983.8, SE = 100.49). Controlling for potential differences in the effects of period between those stimulus pairs belonging to the same site and those who did not credibly worsened the performance of the model in terms of LOOIC (20051.67, SE = 98.43), credibly higher than the baseline model (difference:

2.76, SE = 0.73). This suggests that rotation did not play a credible role in affecting accuracy of discrimination, and that participants performed equally in discriminating between stimuli from the same site as between stimuli from different sites.

2.5.4.1 RT: The model with a linear effect of period was credibly better than that with a quadratic effect: LOOIC -8590.95 (SE: 332.31) vs. -6712.77 (SE: 394.32), with a credible difference of 1878.18 (SE = 170.17). We therefore focus on the simpler model. The model does not indicate any credible effect of period on reaction time for neither between nor within site discrimination, BF01 = 9.42, credibility = 90% (See Table S7 and Fig. S6).

Adding a parameter to control for the rotation of the stimuli yielded a credibly higher estimated out-of-sample error (LOOIC = -7120.03, SE = 395.19), with a difference from baseline of 1470.92 (SE = 175.27). Controlling for potential differences in the effects of time between those stimulus pairs belonging to the same site and those who did not credibly worsened the performance of the model (in terms of estimated out of sample error, or LOOIC): -6661.06, SE = 394.99, with a credible difference of 1929.89, SE = 169.89.

Table S8: Experiment 5 – Criterion and Sensitivity

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Criterion	β = 0.37, 95% CI = 0.18 0.56	1577, 1.00	
Sensitivity	β = -0.01, 95% CI = -0.31 0.30	1397, 1.00	
Period	β = -0.20, 95% CI = -0.35 -0.06	2822, 1.00	
Period^2	β = -0.70, 95% CI = -0.94 -0.47	2054, 1.00	
Sensitivity : Period	β = -0.04, 95% CI = -0.30 0.22	1406, 1.00	
Sensitivity : Period^2	β = 0.09, 95% CI = -0.34 0.52	1278, 1.00	
Random Effects	*SD*		
Stimulus criterion	0.03, 95% CI = 0.00 0.13	1675, 1.00	
Stimulus sensitivity	0.04, 95% CI = 0.00 0.13	2322, 1.00	
Stimulus pair criterion	0.04, 95% CI = 0.00 0.11	1061, 1.00	
Stimulus pair sensitivity	0.03, 95% CI = 0.00 0.08	1640, 1.00	
Participant criterion	0.18, 95% CI = 0.02 0.39	798, 1.00	
Participant sensitivity	1.73, 95% CI = 1.54 1.92	892, 1.00	
Participant period criterion	0.11, 95% CI = 0.00 0.28	1251, 1.00	
Participant period^2 criterion	0.24, 95% CI = 0.03 0.45	559, 1.01	
Participant period sensitivity	1.08, 95% CI = 0.931.25	1821, 1.00	
Participant period^2 sensitivity	2.82, 95% CI = 2.57 3.10	1207, 1.00	

Signal Detection Theory analysis of accuracy of pattern discrimination as a function of period. Note that since the estimates for between sites and within site discrimination were almost completely overlapping and estimated differences of parameters between the models were tightly centered at 0, we do not report separate results for within and between sites.

Table S9: Experiment 5 - Reaction time

Predictor	*Estimates*	*Effective Samples and Rhat*	*Evidence Ratio and credibility*
Intercept (rate)	-0.31, 95% CI = -0.38 0.24	261, 1.00	
Period (rate)	0.02, 95% CI = -0.02 0.06	1927, 1.00	**0.18, 15%** **BF01: 9.42** **Credibility: 90%**
Intercept (shape)	2.73, 95% CI = 2.61 2.86	337. 1.01	
Random Effects	*SD*		
Stimulus intercept (rate)	0.05, 95% CI = 0.03 0.09	2011, 1.00	
Stimulus Pair intercept (rate)	0.05, 95% CI = 0.04 0.07	1302, 1.00	
Participant intercept (rate)	0.24, 95% CI = 0.20 0.29	667, 1.00	
Participant period (rate)	0.02, 95% CI = 0.01 0.02	2231, 1.00	
Participant intercept (shape)	0.48, 95% CI = 0.40 0.59	805, 1.00	

Reaction time of pattern discriminability as a function of period. Note that since the estimates for between sites and within site discrimination were almost completely overlapping and estimated differences of parameters between the models were tightly centered at 0, we do not report separate results for within and between sites.

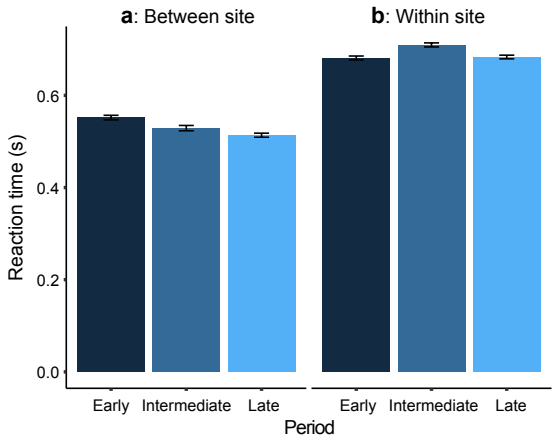

Figure S6: Experiment 5 – reaction time results. Reaction times of correct discrimination trials as a function of period. **a**, Discrimination of stimulus patterns between the two archeological sites. Discrimination differences (decrease in reaction time by period) would support the Cultural Connotation Hypothesis as they could indicate the emergence of style elements. The effects are, however, weak. **b**, Discrimination of stimulus patterns within the two sites. Discrimination differences would support the Symbolic Denotation Hypothesis as they could indicate adaptive pressures for the differentiation of forms within a system of signs.

3 Relations between cognitive affordances of engraved patterns

Our suite of experiments implicitly assume that each experiment is capturing independent cognitive properties of the engraved patterns (the saliency, memorability, intentionality etc. respectively). However, are our experimental measures in fact partially orthogonal cognitive properties of the individual patterns, or are they expressions of the same basic cognitive affordance? In other words, would a more salient stimulus (as measured by experiment 1) be also judged more intentional (as measured in experiment 2), and if so, how much of the variance across stimuli is shared across the domains?

In order to address these questions, we extracted the individual variability of the stimuli from the models in the first 3 experiments, the latter two not having direct estimates of varying effects by stimulus (experiment 4 being analyzed in terms of overall errors in grouping, experiment 5 being about relative similarity between pairs of stimuli). To increase interpretability of the regression coefficients we standardized the posterior estimates of varying effect, dividing each posterior sample by the standard deviation across the whole population. This makes the regression coefficient an equivalent of a Pearson correlation coefficient. To preserve the uncertainty estimates in the varying effects, we summarized the mean and the standard deviation of the posterior estimate for each stimulus in each experiment. We then created regression models assessing the relations between the possible pairs of experiments: Intentionality (experiment 2) as predicted by saliency (experiment 1), errors in memorability (experiment 3) as predicted by intentionality, and errors in memorability as predicted by saliency. We included the standard deviation of each estimate as measurement error in both outcome and predictor. In order to assess how much we should trust these more exploratory analyses we compared the models against the relevant null model containing the same outcome in term of difference in LOOIC (Δ LOOIC, a negative difference in LOOIC indicating the pairwise model being worse than the null model) and stacking weights (indicating the probability of our models to be better than the null)(27). We then report correlation coefficients, a Bayesian measure of R^2 (variance of the predicted values divided by the variance of the predicted values plus the variance of the errors)(28) and the evidence ratio and credibility of the correlation coefficients. Note that these latter should only be considered reliable if the model is better than the null model.

The analyses (reported in Table S8) indicate that there is evidence only for a relation between the cognitive affordances investigated by experiments 2 and 3: stimuli more likely to be judged as intentional are also easier to remember and reproduce (lower errors in reproduction). However, the shared variance in the effects is small, with only 7% of

the variance shared between the two measures. The results suggest that while the cognitive affordances might be related in interesting ways, their variance by stimulus is only minimally overlapping beyond the effect of period.

Table S10: Relations between experimental measurements

Difference from null model			*Effect*		*Evidence*	
	Δ LOOIC, SE	Stacking weight	Estimate, 95% CI	R^2, 95% CI	Evidence ratio	Credibility
Exp 1 ~ Exp 2 (Intentionality vs. Saliency)	-1.21 (1.60)	0%	-0.12 (-0.52 0.31)	0.04 (0 0.17)	2.75	73%
Exp 3 ~ Exp 2 (Memorability vs. Intentionality)	1.05 (2.05)	99.7%	-0.24 (-0.63 0.14)	0.07 (0 0.26)	7.47	88%
Exp 3 ~ Exp 1 (Memorability vs. Saliency)	-1.11 (0.98)	0%	0.07 (-0.31 0.45)	0.03 (0 0.14)	1.82	65%

Pairwise regression models assessing relations between varying effects by stimuli across the first three experiments.

Supplementary References

1. C. S. Henshilwood, F. d'Errico, I. Watts, Engraved ochres from the middle stone age levels at Blombos Cave, South Africa. *Journal of human evolution* **57**, 27-47 (2009).
2. P.-J. Texier *et al.*, The context, form and significance of the MSA engraved ostrich eggshell collection from Diepkloof Rock Shelter, Western Cape, South Africa. *Journal of Archaeological Science* **40**, 3412-3431 (2013).
3. Z. Jacobs, R. G. Roberts, An improved single grain OSL chronology for the sedimentary deposits from Diepkloof Rockshelter, Western Cape, South Africa. *Journal of Archaeological Science* **63**, 175-192 (2015).
4. Z. Jacobs, E. H. Hayes, R. G. Roberts, R. F. Galbraith, C. S. Henshilwood, An improved OSL chronology for the Still Bay layers at Blombos Cave, South Africa: further tests of single-grain dating procedures and a re-evaluation of the timing of the Still Bay industry across southern Africa. *Journal of Archaeological Science* **40**, 579-594 (2013).
5. D. Hodgson, Decoding the Blombos engravings, shell beads and Diepkloof ostrich eggshell patterns. *Cambridge Archaeological Journal* **24**, 57-69 (2014).
6. S. Appelle, Perception and discrimination as a function of stimulus orientation: the" oblique effect" in man and animals. *Psychological bulletin* **78**, 266 (1972).
7. P. Jolicoeur, M. J. Landau, Effects of orientation on the identification of simple visual patterns. *Canadian Journal of Psychology/Revue canadienne de psychologie* **38**, 80 (1984).
8. A. A. Schoups, G. A. Orban, Interocular transfer in perceptual learning of a pop-out discrimination task. *Proceedings of the National Academy of Sciences* **93**, 7358-7362 (1996).
9. R. Myers (2012) CulturalAnalytics: Functions for statistical analysis and plotting of image properties. (https://R-Forge.R-project.org/projects/rca/).
10. A. N. Kolmogorov, On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A* **25**, 369-376 (1963).
11. R. Cilibrasi, P. M. Vitányi, Clustering by compression. *IEEE Transactions on Information theory* **51**, 1523-1545 (2005).
12. H. Yu, S. Winkler (2013) Image complexity and spatial information. in *Fifth International Workshop on Quality of Multimedia Experience* (IEEE, Austria).
13. D. J. Leiner (2014) SoSci survey.
14. J. Cohen, A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* **20**, 37-46 (1960).
15. J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174 (1977).
16. J. L. Fleiss, Measuring nominal scale agreement among many raters. *Psychological bulletin* **76**, 378 (1971).
17. M. Gamer, J. Lemon, I. Singh (2012) irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.

18. L. Wang, X. Weng, S. He, Perceptual grouping without awareness: superiority of Kanizsa triangle in breaking interocular suppression. *PLoS One* **7** (2012).
19. M. Wilke, N. K. Logothetis, D. A. Leopold, Generalized flash suppression of salient visual targets. *Neuron* **39**, 1043-1052 (2003).
20. J. W. Peirce, PsychoPy--Psychophysics software in Python. *J Neurosci Methods* **162**, 8-13 (2007).
21. J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, A. Gelman, Visualization in Bayesian workflow. *arXiv preprint arXiv:1709.01449* (2017).
22. B. Carpenter *et al.*, Stan: A probabilistic programming language. *Journal of statistical software* **76** (2017).
23. P.-C. Bürkner, brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1-28 (2017).
24. A. Vehtari, A. Gelman, J. Gabry, Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. *arXiv preprint arXiv:1507.04544* (2015).
25. H. Manner, Modeling and forecasting the outcomes of NBA basketball games. *Journal of Quantitative Analysis in Sports* **12**, 31-41 (2016).
26. A. A. Schoups, R. Vogels, G. A. Orban, Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity. *The Journal of physiology* **483**, 797-810 (1995).
27. Y. Yao, A. Vehtari, D. Simpson, A. Gelman, Using stacking to average Bayesian predictive distributions. *Bayesian Analysis* **13**, 917-1007 (2018).
28. A. Gelman, B. Goodrich, J. Gabry, I. Ali, R-squared for Bayesian regression models. *Unpublished via http://www. stat. columbia. edu/~ gelman/research/unpublished* (2017).