Cognition 237 (2023) 105452

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Iterated learning reveals stereotypes of facial trustworthiness that propagate in the absence of evidence

Stefan Uddenberg [a,*], Bill D. Thompson [b], Madalina Vlasceanu [c], Thomas L. Griffiths [d], Alexander Todorov [a]

[a] University of Chicago Booth School of Business, United States of America
[b] University of California, Berkeley, United States of America
[c] New York University, United States of America
[d] Princeton University, United States of America

ARTICLE INFO

Keywords:
Face perception
Social cognition
Iterated learning

ABSTRACT

When we look at someone's face, we rapidly and automatically form robust impressions of how trustworthy they appear. Yet while people's impressions of trustworthiness show a high degree of reliability and agreement with one another, evidence for the accuracy of these impressions is weak. How do such appearance-based biases survive in the face of weak evidence? We explored this question using an iterated learning paradigm, in which memories relating (perceived) facial and behavioral trustworthiness were passed through many generations of participants. Stimuli consisted of pairs of computer-generated people's faces and exact dollar amounts that those fictional people shared with partners in a trust game. Importantly, the faces were designed to vary considerably along a dimension of perceived facial trustworthiness. Each participant learned (and then reproduced from memory) some mapping between the faces and the dollar amounts shared (i.e., between perceived facial and behavioral trustworthiness). Much like in the game of 'telephone', their reproductions then became the training stimuli initially presented to the next participant, and so on for each transmission chain. Critically, the first participant in each chain observed some mapping between perceived facial and behavioral trustworthiness, including positive linear, negative linear, nonlinear, and completely random relationships. Strikingly, participants' reproductions of these relationships showed a pattern of convergence in which more trustworthy looks were associated with more trustworthy behavior — even when there was no relationship between looks and behavior at the start of the chain. These results demonstrate the power of facial stereotypes, and the ease with which they can be propagated to others, even in the absence of any reliable origin of these stereotypes.

1. Introduction

Faces are perhaps the most information-dense stimuli we encounter in our daily lives, and certainly the most socially relevant. On viewing a face we automatically extract and/or impute a host of different properties of both the face and the mind behind it (for a review, see Todorov, 2017). Among such properties are those we "read out", including demographic characteristics such as age (Henss, 1991; Montepare & Zebrowitz, 1998). However, there are also properties we "read into", as when we succumb to vivid impressions of what we think that person is like *as a person* — forming, for example, beliefs about how dominant or competent someone might be solely on the basis of their face (Bar, Neta, & Linz, 2006; Oosterhof & Todorov, 2008). Such property extraction/imputation can occur in the blink of an eye; a face need only be visible for as little as 33-50 ms before observers have access to all the visual information they will use to make stable, self-consistent judgments regarding both kinds of properties discussed, as with demographics (Colombatto, Uddenberg, & Scholl, 2021) or more ineffable and complex impressions such as "trustworthiness" (Todorov, Pakrashi, & Oosterhof, 2009). Such character-based facial impressions emerge early in human development, appearing at ages as young as 3 years old (Charlesworth, Hudson, Cogsdill, Spelke, & Banaji, 2019; Cogsdill & Banaji, 2015; Cogsdill, Todorov, Spelke, & Banaji, 2014).

Facial impressions matter a great deal across many domains of human experience and endeavor. For example, greater impressions of competence can lead to greater CEO compensation (Graham, Harvey, &

* Corresponding author at: 5807 S. Woodlawn Ave, Chicago, IL 60637, United States of America.
E-mail address: stefan.uddenberg@chicagobooth.edu (S. Uddenberg).

https://doi.org/10.1016/j.cognition.2023.105452
Received 26 August 2022; Received in revised form 23 March 2023; Accepted 29 March 2023
Available online 11 April 2023

Puri, 2017) and predict real-life electoral success well above chance (Todorov, Mandisodza, Goren, & Hall, 2005), even when children are the ones making the competence judgments (Antonakis & Dalgas, 2009). In contrast, looking less trustworthy can lead to more severe criminal sentencing decisions (at least among mock juries; Porter, ten Brinke, & Gustaw, 2010; for real-life sentencing decisions see Wilson & Rule, 2015 and Kramer & Gardner, 2020 for a recent failed replication), while appearing more trustworthy can pay dividends, as in the contexts of online credit applications (Duarte, Siegel, & Young, 2012) or lab-based economic games such as the trust game (Chang, Doll, van't Wout, Frank, & Sanfey, 2010; Rezlescu, Duchaine, Olivola, & Chater, 2012; van't Wout & Sanfey, 2008).

Although we readily judge people based on their facial appearance, these judgments can be misleading or inaccurate. For example, although we have already noted that more competent-looking CEOs attract higher salaries, there is little evidence to suggest that their performance is any better than that of their less leaderly-looking counterparts (Graham et al., 2017). The evidence for the accuracy of trustworthiness judgments is also mixed, as when participants make face-based inferences as to the behavior of their partners in an economic trust game. Some such studies have shown slightly above-chance performance for predicting whether or not one's partner in the game will behave in a trustworthy manner based on a single face image, but only under very specific conditions (e.g., when the image has been heavily cropped and rendered in grayscale; Bonnefon, Hopfensitz, & De Neys, 2013; see Todorov, Funk, & Olivola, 2015 for a response). However, other studies find no reliable effect on participants' ability to detect cheating behavior (Jaeger et al., 2022) — indeed, in some cases participants would be better off ignoring the face photographs altogether, since relying solely on past reputational information (or even a simple "trust all other players" heuristic) would earn them more money by the end of the game (Efferson & Vogt, 2013; Todorov et al., 2015).

If personality-based facial impressions are inaccurate, how do they come to be in the first place? And how might they persist in the population, despite being inaccurate? The current research aims to answer this question using a novel face-based iterated learning paradigm (Kirby, 2001).

1.1. Iterated learning: A method for exploring our priors

Much of what we learn in our lives is taught to us by others, whether directly or indirectly. A simplified model of this is provided by the "iterated learning" framework (Kirby, 2001), in which data are passed from one generation of participants to the next, much as simple messages are passed from one player to the next in the children's game of "Broken Telephone".

This type of paradigm has a long history within cognitive psychology, dating back to Bartlett's (1932) seminal explorations of how memories for stories and drawings break down as they are transmitted from one person to another via what he called "serial reproduction". For example, in his "Portrait d'homme" series, Bartlett showed that a transmission chain of drawings that began with a vaguely face-like illustration soon failed to resemble the initial drawing at all, converging on more schematic drawings of faces within only a few participants' reproductions. While this particular study may not have survived psychology's ongoing replication crisis (Carbon & Albrecht, 2012), it remains illustrative of a deep truth about memory: it is a reconstructive process, such that the errors in people's reproductions are not entirely random. Instead, errors accrue systematically in the direction of participants' "inductive biases," or what they believe the originally observed input was *most likely* to be. This means that if the participants in a given transmission chain tend to share the same inductive biases, and if the chain is long enough (i.e., involves enough error-prone transmissions from one person to another) then the final output will converge on representations of those inductive biases. Crucially, this should occur regardless of both (a) the original input given at the start of the chain and (b) however long the chain should continue past the point of convergence (Griffiths, Christian, & Kalish, 2006; Kalish, Griffiths, & Lewandowsky, 2007; Xu & Griffiths, 2010).

By exploiting the fact that memory is both reconstructive and error-prone, iterated learning (and related paradigms, such as serial reproduction) can therefore help us characterize our inductive biases without asking about them directly. As such, iterated learning has recently found broad applicability across myriad domains. These include investigations of: our priors for abstract category learning (Canini, Griffiths, Vanpaemel, & Kalish, 2014; Griffiths et al., 2006); the emergence of color terms in language (Xu, Dowman, & Griffiths, 2013); spatial biases in visual working memory (Langlois, Jacoby, Suchow, & Griffiths, 2021); and even racial biases in reproductions of faces (Uddenberg & Scholl, 2018); among others (Jacoby & McDermott, 2017; Kirby, Tamariz, Cornish, & Smith, 2015; Verhoef & Ravignani, 2021). For example, in one study, participants were implicitly taught a simple function mapping the magnitudes of two arbitrary variables: the width and height of two different colored bars (Kalish et al., 2007). Over the course of many trials, participants at the start of each transmission chain were taught either a positive linear relationship, a negative linear relationship, a quadratic relationship, or no relationship at all between them (i.e., random magnitudes of width and height). After implicitly learning the initial function, the first participant then had to reproduce that relationship on both old and new (unseen) magnitudes to the best of their ability. Unbeknownst to the participant, their data at test time then became the initial training data for the next participant in the transmission chain, and so on. This procedure yielded a striking pattern: reproductions converged on simple (mostly positive) linear relationships, regardless of the initial input, and even when all the first participant saw amounted to random noise. This work demonstrated that participants held a strong bias toward positive linear relationships — at the very least when dealing with two arbitrary variables (Kalish et al., 2007).

1.2. The current experiments: "iterated trustworthiness"

Although iterated learning has been used to successfully explore inductive biases across many different domains, including face memory (Uddenberg & Scholl, 2018), to our knowledge it has never been used to answer questions about the nature of our biases for facial impressions. Across two pre-registered experiments, we test whether and how participants' memories may be biased toward associating trustworthy-looking faces with more trustworthy behavior in the context of a simple economic game. As in the iterated learning study described above (Kalish et al., 2007), participants observed one of four different starting relationships between trustworthy looks and economic game behavior: positive linear, negative linear, quadratic, or no relationship at all (i.e., a random mapping between looks and behavior). Insofar as participants demonstrate a shared set of inductive biases relating superficial looks to behavior, we can expect the transmitted relationships to converge onto some pattern. To our minds, the most plausible pattern would be positive linear — demonstrating a belief that more trustworthy-looking people behave in more trustworthy ways. However, there are many other possible outcomes, such as a negative linear (i.e., opposite) relationship, or perhaps some nonlinear mapping, such that faces toward the extremes of the perceived trustworthiness appearance spectrum seem the most insincere or untrustworthy, while the faces toward the middle of the spectrum seem like they are most likely to behave well in the context of the game. Of course, participants could have no strong inductive biases at all, in which case the data transmitted from person to person would devolve into random noise, or fluctuate between different modes.

2. Experiment 1: Iterated learning of facial and behavioral trustworthiness

We first explored participants' inductive biases relating perceived

trustworthiness and economic game behavior across four initial function conditions: positive linear, negative linear, nonlinear, and random noise.

2.1. Method

2.1.1. Participants

We decided before data collection began to test 10 reproduction chains of 10 participants (or generations) for each of our four conditions, for a total of 400 participants. These pre-registered values (see https://osf.io/acspu/?view_only=ccbb97a7e3954cc8ae3404a02778fcc0) were chosen arbitrarily to be roughly in line with past iterated learning studies (e.g., Kalish et al., 2007; Suchow, Pacer, & Griffiths, 2016). Our final analyzed sample therefore included 400 U.S.-based participants (215 females; mean age = 38.07, SD = 13.23; 294 self-identified as White; 23 East Asian; 22 Latinx/a/o or Hispanic; 21 Black/African American; 13 South Asian; 3 Southeast Asian; 2 Native American/American Indian; 1 Middle Eastern; 17 identified as two or more races; 2 preferred not to report their race; and 2 reported that their race/ethnicity was not listed) using the Amazon Mechanical Turk online labor market (MTurk). (For discussion of this pool's nature and reliability, see Crump, McDonnell, & Gureckis, 2013; Germine et al., 2012). A total of 238 were excluded for either failing an attention check or data quality checks, as described in the "Data quality checks" section below. Exclusion criteria were performed automatically at the conclusion of each participant's experiment session, requiring no input from the experimenters. Excluded counts were similar across conditions (positive linear: 53; negative linear: 73; nonlinear: 53; random: 59). An additional 185 participants returned the assignment before completion, while 42 participants started, but did not complete, the experiment.

2.1.2. Apparatus

The experiment was conducted using custom software written with a combination of Python, JavaScript, CSS and HTML. Key libraries used included the Dallinger online experiment platform (https://dallinger.readthedocs.io) and jsPsych (de Leeuw, 2015). All analyses were conducted in Python. Participants completed the experiment via a custom web page which could be loaded in any modern web browser on their own laptop or desktop computers; mobile devices such as phones and tablet computers were explicitly disallowed, and attempts to access the experiment from such a device led to its immediate termination along with an error message.

2.1.3. Stimuli

Stimuli consisted of 100 computer-generated faces created with the FaceGen software development kit (Singular Inversions, https://facegen.com), which allows for the creation of arbitrary 3D faces based on a statistical model derived from laser scans of 271 real human faces (for details, see Blanz & Vetter, 1999). In FaceGen, faces are represented as points in 100-dimensional face space (50 shape and 50 reflectance dimensions). Moving a point (i.e., a face) along a single dimension changes the shape or reflectance map of a face in specific ways. Meaningful social dimensions, such as perceived trustworthiness or dominance, can be modeled as linear combinations of these basic FaceGen dimensions based on subjective trait judgments of random points in the space (for a detailed description of this procedure, see Oosterhof & Todorov, 2008). Some methodological details here are reproduced from Todorov, Dotsch, Porter, Oosterhof, and Falvello (2013), as this work employs a procedure derived from and inspired by that work.

Because facial diversity was crucial to the experiment (for reasons that will become more clear in the procedure below) we created a sample of maximally distinctive identities, following a standard procedure (Oh, Buck, & Todorov, 2019; Oh, Dotsch, & Todorov, 2019; Todorov et al., 2013). To create such distinctive stimuli, we first generated a random sample of 5000 faces within the face space, and chose the 100 faces that differed maximally from each other based on the average Euclidean distance to all other faces. This resulted in a sample of maximally distinctive faces, but also in faces that looked atypical. To reduce this atypicality, we scaled the face coordinates with a factor of 0.5, essentially bringing them closer to the average face. This procedure preserves the ratio of differences so that the faces still differ maximally from each other, yet look more typical. We then transformed each face along the dimension of perceived trustworthiness (along shape dimensions only[1]) between -6SD to +6SD in 100 discrete levels (yielding a total of 10,000 faces — 100 identities × 100 levels). We then arbitrarily assigned each face identity to a single one of those 100 trustworthiness levels (e.g., random face identity #1 being assigned to trustworthiness level #1), yielding a total of 100 distinct face images used throughout the experiment. Example stimuli from the continuum can be seen in Fig. 1a below.)

2.1.4. Procedure

Participants first gave their informed consent and agreed to answer open-ended questions, using a procedure designed to reduce participant attrition (Zhou & Fishbach, 2016). They then read detailed written instructions for the task, as described below, and then answered simple comprehension questions about the instructions before being allowed to proceed to the main experiment (getting any of the comprehension questions wrong twice in a row terminated the experiment immediately). The instructions began with a cover story that all faces shown in the experiment were computerized and photographically altered to protect the identities of the depicted individuals. Participants were told that these people took part in a type of economic game (in practice, our task description was equivalent to a dictator game; Guala & Mittone, 2010), and their job would be to learn and predict the decisions these people made in the game. Participants then read how each person depicted was assigned to the role of Player A in the (dictator) game, were given a $1.00 endowment by the experimenter, and then had to decide what percentage of their endowment they would share with their partner (which was always an amount between 1% and 100%). The experiment was split into two phases: the learning phase, and the testing phase.

2.1.4.1. Learning phase. On each of the 30 trials of the learning phase, participants were shown a single face (300px × 300px), along with the question, "What percentage of their endowment did this person share with their partner?" and a slider (slider area: a 700px × 15px rounded light grey rectangle; slider handle: a 12.5px radius navy blue circle) whose value was allowed to range between 1% and 100% (stimuli were presented toward the center of the browser window). The slider handle was not visible at the start of each trial, but became immediately visible at the point along the slider at which the participant clicked to record their first guess, along with the slider's current value (presented in navy blue below the slider) and feedback of the correct percentage shared (presented in orange below the slider value). If the participant's initial click was close to the correct amount (in practice, within 12.5% of the correct answer) they were given a bonus of $0.01 (with such a notice presented in green below the slider), to incentivize them to pay attention and learn the relationship between facial appearance and behavioral trustworthiness quickly. Participants could move on to the next trial by moving the slider so that its value matched that of the correct amount feedback in orange (in practice, within 1% of the correct amount) and then pressing the "Next" button. An example of what one such learning

[1] It is worth noting that the "shape" dimensions of perceived trustworthiness (along which the faces were transformed) comprise both structural and expression cues to trustworthiness to some degree, especially at the extremes of the continuum (Eggleston, Tsantani, Over, & Cook, 2022). As with past work relying on such models, our conclusions do not rely on the faces being morphed solely along structural dimensions, as we focus on the more general case of high-level social impressions.

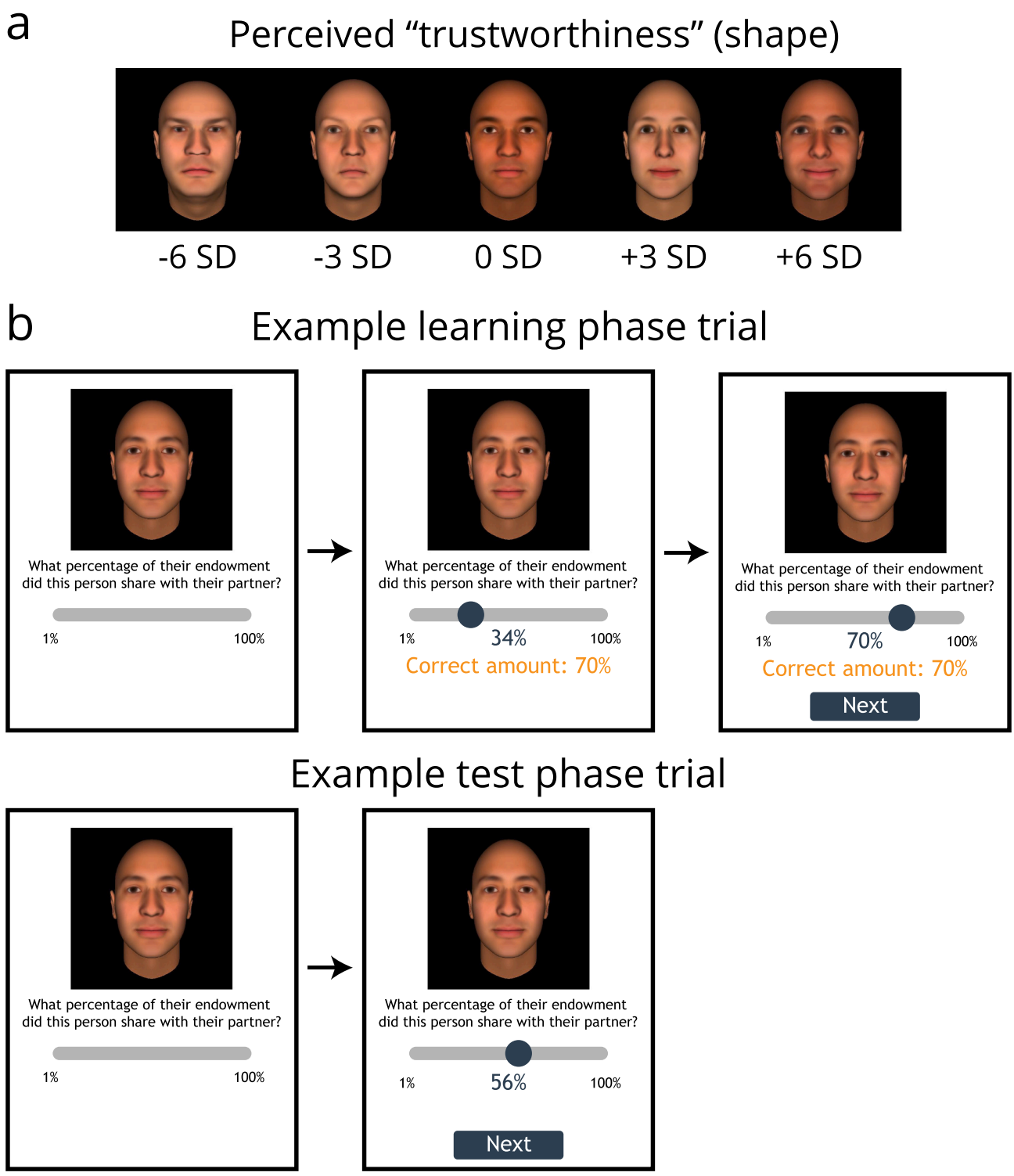

Fig. 1. Examples of (a) face stimuli used throughout the experiments and (b) experimental trials in the learning and test phases. (a) Stimuli were drawn from a continuum of perceived facial trustworthiness measured in standard deviations (SD) from the mean (ranging between -6SD to +6SD, with 100 faces in total). Each face represents a distinct identity with a distinct perceived trustworthiness level. The depicted faces are evenly spaced along the continuum (e.g., the face in the middle is at 0SD or mean perceived trustworthiness). (b) An overview of the iterated trustworthiness task as illustrated via two example trials, in which participants had to learn and then reproduce some relationship between perceived facial trustworthiness and behavioral trustworthiness.

phase trial looked like is presented in Fig. 1b above.

2.1.4.2. Testing phase. The 30 trials in the testing phase were identical to those of the learning phase, except that participants were not provided feedback of any kind; they simply recorded their response by clicking/dragging the slider and then pressing the "Next" button to move on. This phase was designed to test how well they had learnt the initial relationship presented in the learning phase. Half of the trials (15) chosen at random tested their knowledge on previously seen faces, while the other half of the trials (15) selected from the remaining pool of 70 unseen faces, as a test of generalization. Unbeknownst to the participants, the answers provided during the testing phase would become the

face-amount pairings shown to the next participant assigned to the chain during their learning phase. As such, the data presented to each participant during the learning phase was determined by the data provided by the immediately previous participant, with one notable exception: the very first participant in each chain.

2.1.4.3. Starting conditions. The 30 face-amount pairings shown to the very first participant in each chain during the learning phase was determined according to random assignment to one of four functions, corresponding to the four experimental conditions: positive linear ($y = x$; where y represents the behavioral trustworthiness or percentage endowment shared between 1 and 100, and x represents the level of perceived facial trustworthiness between 1 and 100), negative linear ($y = -x$), nonlinear ($50.5 + 49.5sin[\pi/2 + x/(5\pi)]$), and random (with random one-to-one pairings of x- and y-coordinates in which both x, $y \in \{1, \ldots, 100\}$. These starting functions were chosen to test how iterated learning of facial/behavioral trustworthiness unfolds over a variety of initial conditions, as well as to align with past iterated learning research (Kalish et al., 2007).

2.1.4.4. Data quality checks. Participants were excluded automatically and in real-time if they failed to meet a number of quality checks. Firstly, as mentioned previously, participants were not allowed to complete the study if they failed to answer any of the instruction comprehension questions twice in a row (after having been explicitly given the correct answers the first time they got the answer wrong). Additionally, due to the necessarily serial nature of data collection, participants were excluded automatically based on the variance and quality of their data in the test phase, decided in advance of data collection. Specifically, we excluded (1) those whose answers all fell within a highly restricted range (i.e., response variance <200), indicating either inattention or unwillingness to use much of the response scale, and (2) those who failed to adequately learn the relationship in the learning phase — failing to even adequately represent the values previously observed — or who produced answers in the test phase that were too inconsistent with the original endowment-face associations shown in the learning phase, as evidenced by their having increased or decreased the maximal information coefficient (MIC) of the learning phase data by over 0.39 in either direction.[2] MIC is a non-parametric method for quantifying the strength of a wide range of relationships between two variables (Reshef et al., 2011). MIC ranges from 0 to 1, so this relatively liberal criterion allowed the inclusion of participants who demonstrated either imperfect learning (e.g., in cases where there was a relationship to be learned in the first place) or who moderately increased or decreased the amount of structure in the data during the testing phase (e.g., in the random starting condition there was no relationship to begin with). We chose to use the MIC as opposed to Pearson correlation (our primary metric of interest) for this data quality check in order to limit the chances of preselecting only participants who produced data in line with our hypotheses, and to allow for a wide range of possible valid relationships to emerge from the data (Reshef et al., 2011).

2.2. Results and discussion

Fig. 2 shows the median correlation of behavioral trustworthiness and (perceived) facial trustworthiness reproduced in the testing phase for each of the 10 participants at each generation (or 'step') in the chains, while the raw data are depicted in Fig. 3. These data suggest a clear pattern: participants produced progressively more positive linear relationships from generation to generation, regardless of initial starting condition. Indeed, 3 of the 4 starting conditions converged on highly positive correlations by the final generation(s); only 1 starting condition (i.e., negative linear) remained somewhat negative, but it nonetheless ended up much more positive than it started out. These impressions were verified via the statistical analyses reported below.

As we hypothesized that participants' reproductions would quickly converge toward their priors relating facial and behavioral trustworthiness, and in order to increase the robustness of the analyses, we compared the final three generations of each starting condition to one another (as opposed to merely comparing the absolute final generations, which of course only involved 10 participants per condition). A series of Bonferroni-corrected t-tests ($\alpha_{corrected} = 0.0083$) on the Fisher z-transformed correlations revealed that the final three generations of the positive linear condition ($M = 1.10$, $SD = 0.30$) significantly differed from those of the negative linear condition ($M = -0.14$, $SD = 0.56$; $t(58) = 10.64$, $p < .001$, $d = 2.75$) and the nonlinear condition ($M = 0.37$, $SD = 1.03$; $t(58) = 3.73$, $p < .001$, $d = 0.96$), but not the random condition ($M = 0.80$, $SD = 0.69$; $t(58) = 2.22$, $p = .030$, $d = 0.57$). The random condition's final three steps also differed significantly from those of the negative linear condition ($t(58) = 5.74$, $p < .001$, $d = 1.48$) but not from the nonlinear condition's ($t(58) = 1.87$, $p = .066$, $d = 0.48$). Lastly, the first three steps of the nonlinear (M = 0.17, SD = 0.50), random (M = 0.28, SD = 0.47), and negative linear (M = -0.92, SD = 0.63) conditions differed significantly from their final three steps ($\alpha_{corrected} = 0.0125$, all $t(58)$s > 2.60, all ps < 0.012, all ds > 0.673), but this was not the case for the positive linear condition ($t(58) = 0.07$, $p = .943$, $d = 0.019$). However, the final three steps of the nonlinear condition did not differ from the negative linear conditions' ($t(58) = 2.39$, $p = .020$, $d = 0.62$).

These results suggest that participants possess a prior that favors a positive linear relationship between how trustworthy a person's face looks and how trustworthily that person will behave. This is evidenced by the steady increase in positive correlations observed as the learned function passed from mind to mind in the experiment, regardless of the initial starting condition or starting data. It is worth pointing out that although the negative linear condition did not become positively correlated by the 10th generation of reproduction, we would expect it to eventually converge on positive correlations as the other conditions did, based both on past empirical data and theoretical work in Bayesian modeling (e.g., Kalish et al., 2007; Xu et al., 2013; Xu & Griffiths, 2010). This can be appreciated intuitively by looking closely at Fig. 2: notice how the negative linear condition (denoted by the red line) ended with a value similar to early generations of the nonlinear condition (the green line), which nonetheless ended up becoming very positively correlated.

3. Experiment 2: A matter of wording?

Our results from Experiment 1 demonstrated that participants held a prior that strongly favored a positive linear relationship between (perceived) facial and behavioral trustworthiness. However, other iterated learning studies with functions also show evidence for a positive linear bias (e.g., Kalish et al., 2007; Suchow et al., 2016). According to one account, this paradigm may simply yield positive linear functions because it is the simplest response strategy participants could possibly have — "As one quantity varies, I will vary the other to suit." To confirm that our observed results could not be due merely to the paradigm employed, we modified our procedure so that half of the participants performed the same task as before — predicting how much money each person shared with their partner (positive wording condition) — but the other half were asked to predict the *opposite* quantity — how much each person *kept for themselves*, and did not share with their partner (negative wording condition). If it is the case that the paradigm produces positive linear relationships regardless of the question asked, we should see convergence toward positive linear relationships for both conditions. However, if we should replicate our earlier results in the positive wording condition, but also see a negative linear relationship for the negative wording condition, then we could conclude that our observed results are not due to some quirk of the testing paradigm.

[2] These threshold values were chosen via pilot testing to exclude observed outliers (i.e., those whose absolute z-scores were >3), and are kept constant for all experiments.

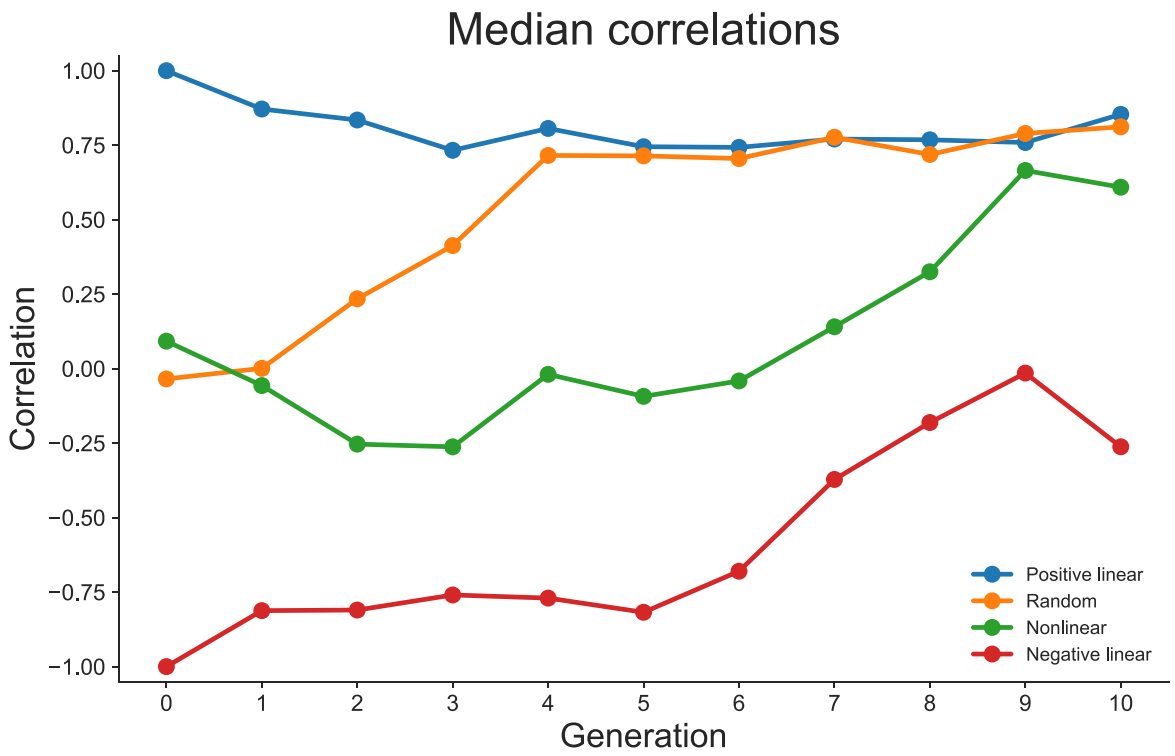

Fig. 2. The median correlations for each starting condition at each generation (or step) of the chains. Regardless of the initial starting condition and its data-generating function, the data produced in the testing phase became more positively correlated as it passed from participant to participant.

3.1. Method

All methodological details are identical to those of Experiment 1 except where listed below.

3.1.1. Participants

We decided before data collection began to test 30 reproduction chains of 10 participants (or generations) for each of the two conditions, for a total of 600 participants. These preregistered values (see https://osf.io/acspu/?view_only=ccbb97a7e3954cc8ae3404a02778fcc0) were chosen arbitrarily to be roughly in line with past iterated learning studies (e.g., Kalish et al., 2007; Suchow et al., 2016). Our final analyzed sample therefore included 600 U.S.-based participants (310 females; mean age = 39.32, SD = 12.24; 456 self-identified as White; 28 East Asian; 18 Latinx/a/o or Hispanic; 40 Black/African American; 7 South Asian; 8 Southeast Asian; 3 Native American/American Indian; 1 Middle Eastern; 38 identified as two or more races; and 1 reported that their race/ethnicity was not listed) using the Amazon Mechanical Turk online labor market (MTurk). (For discussion of this pool's nature and reliability, see Crump et al., 2013; Germine et al., 2012). 426 were excluded for either failing an attention check or data quality check, as described in Experiment 1. Excluded counts were similar across conditions (positive wording: 218; negative wording: 208). An additional 342 participants returned the assignment before completion, while 89 participants started, but did not complete, the experiment.

3.1.2. Procedure

The procedure was identical to that of Experiment 1, except for the following changes. Every chain's starting data was random (as in the random starting condition). Every participant was randomly assigned to one of two wording conditions: Positive and Negative. In the Positive condition, the procedure was identical to Experiment 1, in that participants were asked how much each person depicted shared with their partner in an economic game. In practice, this condition directly replicated the random condition from Experiment 1. However, in the Negative condition, participants were instead instructed to answer how much each person kept for themselves, and therefore did not share with their partner.

3.2. Results and discussion

The median reproduction chains are depicted in Fig. 4 below (see Fig. S1 in the supplement for raw data). Inspection of these aggregate reproduction chains demonstrates that participants did indeed produce more positive linear relationships in the positive wording condition, but more negative linear relationships in the negative wording condition. These impressions were verified via the statistical analyses reported below.

We first compared the final three generations of each wording condition via a series of Bonferroni-corrected *t*-tests ($\alpha_{corrected} = 0.0083$) on the Fisher z-transformed correlations. This analysis revealed that the final three generations of the positive wording condition ($M = 0.41$, $SD = 0.96$) significantly differed from those of the negative wording condition ($M = -0.46$, $SD = 0.77$; $t(178) = 6.69$, $p < .001$, $d = 1.00$). In addition, the first three steps for both the positive (M = -0.01, SD = 0.64), and negative wording conditions (M = -0.01, SD = 0.55) differed significantly from their final three steps (all $t(178)$s > 3.48, all $ps < 0.001$, all $ds > 0.52$).

These results show that participants did indeed link behavioral and perceived facial trustworthiness in a stereotypic fashion, and that our earlier results were not due to any infelicities in the iterated learning paradigm itself.

4. Discussion

Across two experiments and 1000 participants, we found strong evidence for a positive linear prior linking perceived facial and behavioral trustworthiness. This prior was strong enough to not only emerge from random noise, but overturned even the exact opposite linear relationship, as when the two variables were initially (but not ultimately)

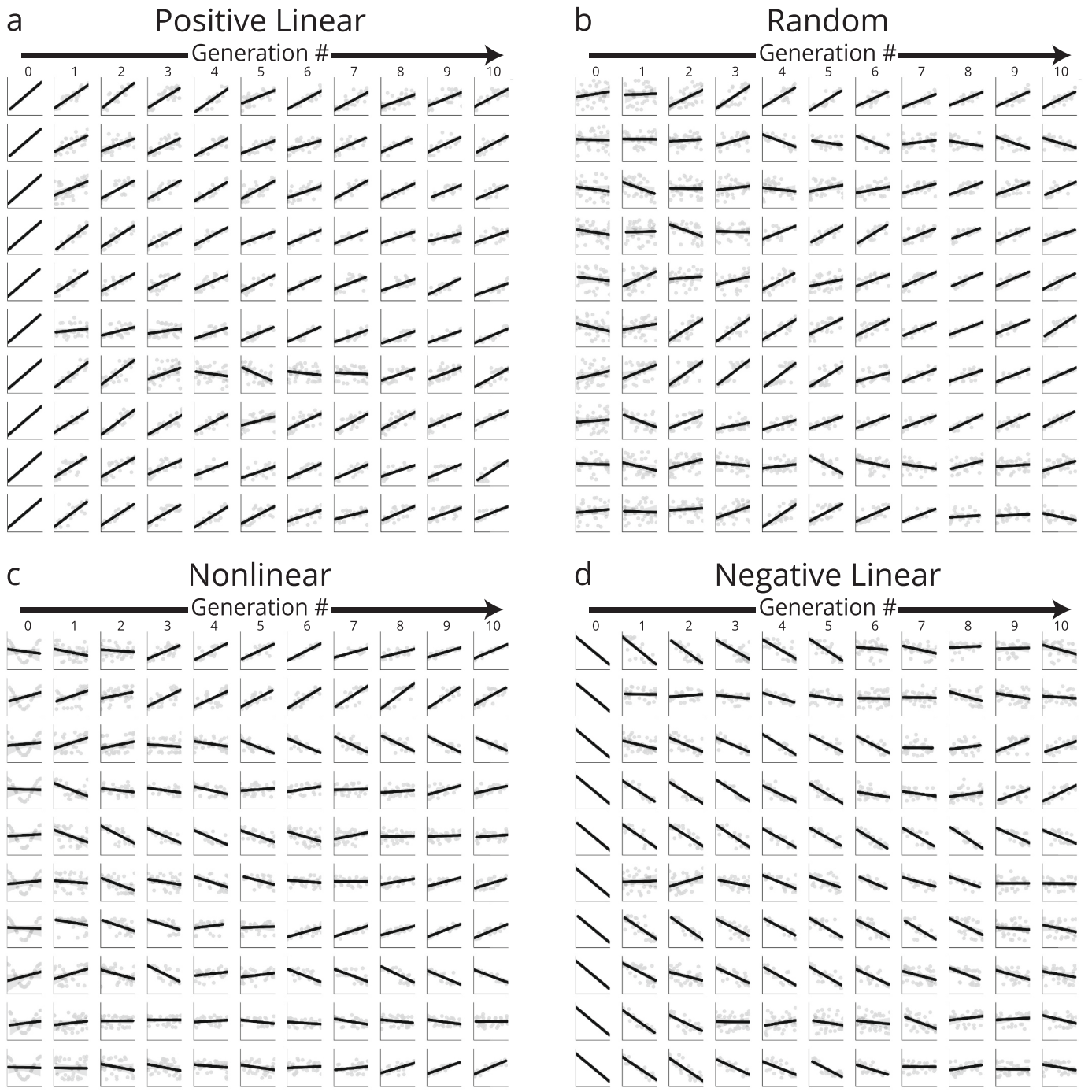

Fig. 3. Raw data (shown in grey) with lines of best fit (shown in black) for Experiment 1 for (a) positive linear, (b) nonlinear, (c) random, and (d) negative linear conditions. Each row represents the data of a single chain of participants who took part in the iterated trustworthiness task, with participant generation (or "step" in the chain) increasing from left to right. The very first (or leftmost) graph of data in each row represents the starting data shown to participants.

related via a negative linear function in Experiment 1. Indeed, Experiment 2 showed that this prior was about trustworthy behavior per se, as people linked more trustworthy-looking faces to more trustworthy behavior, rather than simply associating such faces with larger response magnitudes in general.

There are many plausible explanations for how such prior beliefs or stereotypes could emerge in the first place. According to overgeneralization accounts, people overweight faces' structural resemblance to otherwise informative cues in other contexts (Zebrowitz, 2017). For example, baby-faced individuals (i.e., those possessed of more neotenous features such as large eyes or a small chin) are stereotyped as being more child-like and less competent (Berry & McArthur, 1986; Montepare & Zebrowitz, 1998; Zebrowitz, Fellous, Mignault, & Andreoletti, 2003). Some such stereotypes may also have arisen from — or are at least reinforced via — lessons from our culture (Cook, Eggleston, & Over, 2022), as when popular media regularly depicts heroes, villains, and other character tropes in highly stereotypic ways; no one has ever been confused about who the good guys are in a Disney movie. Many such stereotypes reinforced via cultural products and media have historically contained overt racist and/or sexist content (Wilson II, Gutierrez, & Chao, 2012). Such impressions can be formed de novo relatively quickly, as shown in single-session laboratory studies

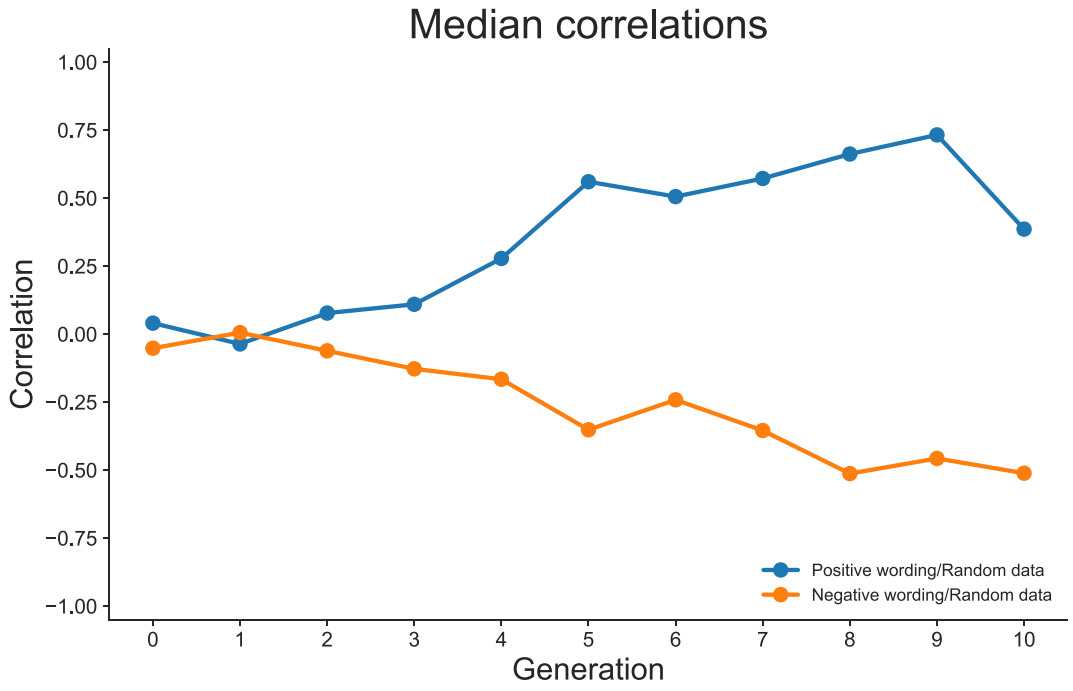

Fig. 4. The median correlations for each question wording condition at each generation (or step) of the chains. Participants answered very differently depending on what was asked of them.

with adults intuiting face-trait mappings (Dotsch, Hassin, & Todorov, 2016) or even abstract stimulus-trait mappings (as with "Greebles"; Lee, Flavell, Tipper, Cook, & Over, 2021). One recent study demonstrated how such stereotypic face-trait mappings may be learned quickly and early in development, as when 5- to 7-year-old children form first impressions via the non-verbal cues given by adults around them (i.e., via social referencing) (Eggleston, Flavell, Tipper, Cook, & Over, 2021). And these culturally transmitted associations may be further exacerbated by how children are treated by the adults within their lives, thus leading to self-fulfilling prophecies. For example, parents may mete out harsher punishments to more mature-looking children (Zebrowitz, Kendall-Tackett, & Fafel, 1991) or give more trustworthy-looking children the benefit of the doubt when confronted with ambiguous evidence of their misbehavior (Thierry & Mondloch, 2021). In light of these findings and other evidence, it stands to reason that children would not merely be passive recipients of such beliefs, but would have some role to play in circulating them among similarly aged peers. To what degree are children susceptible to novel stereotypes (relative to adults)? Can stereotype formation be disrupted or corrected early? Future work may fruitfully explore such questions using a modified iterated learning and/or serial reproduction paradigms.

It is worth noting that the impact of such visual stereotype content could not be explored in the current studies due to the fact that the underlying face models used to generate the stimuli were disproportionately trained on white faces (Blanz & Vetter, 1999). Additionally, the model of perceived trustworthiness used here was trained entirely on white faces (Oosterhof & Todorov, 2008). The combination of these two biases in our stimulus generation procedure — one common throughout the field — resulted in the use of disproportionately white-appearing face stimuli in the current experiments. Even so, there were several faces that would likely be racialized by typical observers as people of color, including Black-appearing faces, as shown in Fig. 1. The current experiments were not designed to test the contributions of racial cognition to the propagation of impression-based stereotypes, in part because the underlying face space and models were ill-suited to answering such questions. However, future work may overcome these limitations by making use of models trained on more comprehensive face spaces with highly diverse stimulus samples, such as those recently developed by our group (Peterson, Uddenberg, Griffiths, Todorov, & Suchow, 2022). Indeed, this was a key motivation behind these newer models' development, as the literature on face perception (including social cognition work on facial impressions) is arguably overly reliant on the use of white faces as the default stimulus class (Cook & Over, 2021).

A high-impact real-world manifestation of the chain of social influence captured in our paradigm is the algorithmic propagation of preexisting stereotypes or sentiments (Vlasceanu & Amodio, 2022). Artificial intelligence (AI) algorithms are trained on historic data that often embeds preexisting societal biases (Baker & Hawn, 2022; Suresh & Guttag, 2021). This feature of AI acts like the training phase of our iterated learning paradigm. Then, when these algorithms are used by new people to inform their cognitive concepts and decisions, they can propagate the stereotypes they had been trained on (Broussard, 2018; Dastin, 2018; Kadiresan, Baweja, & Ogbanufe, 2022; O'Neil, 2016). This more recently uncovered effect of AI on society mirrors the testing phase of our paradigm. Such instances of algorithmic propagation of preexisting stereotypes have been found in all aspects of society ranging from university admissions (Santelices & Wilson, 2010), and hiring decisions (Dastin, 2018), to criminal sentencing decisions (Angwin, Larson, Mattu, & Kirchner, 2016) and healthcare allocation (Obermeyer, Powers, Vogeli, & Mullainathan, 2019).

Social media is another ecosystem rife with chained social influence. Social media algorithms are optimized to maximize engagement (Fisher, 2022), and people tend to mostly engage with attitude-consistent information (Evans, 1989; Meppelink, Smit, Fransen, & Diviani, 2019; Weeks, Lane, Kim, Lee, & Kwak, 2017). Therefore, social media prioritizes content that reinforces preexisting stereotypes and beliefs, maintaining and amplifying them even when false (Farkas, Schou, & Neumayer, 2018; Williams, McMurray, Kurz, & Hugo Lambert, 2015).

Beyond the digital world, stereotype transmission through face-to-face interactions in social networks has been documented for decades, repeatedly finding that stereotype-consistent information is transmitted with greater fidelity than stereotype-inconsistent information (Kashima, 2000; Lyons, Clark, Kashima, & Kurz, 2008; Lyons & Kashima, 2001). Stereotypes have even been found to propagate over time within societies despite evidence of their inaccuracy (Kunda & Oleson, 1995), and this may be particularly germane given that social information is

propagated more easily than similar non-social information (as via gossip; Mesoudi, Whiten, & Dunbar, 2006). The present study adds to this rich body of literature by providing a compelling mechanism by which this counterintuitive phenomenon may have persisted in human communities. Taken together, our results demonstrate that such stereotype-consistent priors, strong as they are, may help explain how stereotypes persist in the population — even when untrue.

CRediT authorship contribution statement

Stefan Uddenberg: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing - original draft. **Bill D. Thompson:** Conceptualization, Methodology, Software, Investigation, Writing - review & editing. **Madalina Vlasceanu:** Conceptualization, Writing - review & editing. **Thomas L. Griffiths:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing. **Alexander Todorov:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was made possible by a grant from Princeton University's DataX Fund, supported by the Schmidt Futures Foundation, and the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105452.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.
Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science, 323*(5918), 1183.
Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092.
Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*(2), 269–278.
Bartlett, F. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.
Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin, 100*(1), 3–18.
Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques - SIGGRAPH '99* (pp. 187–194). ACM Press.
Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology. General, 142*(1), 143–150.
Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. Cambridge, MA: MIT Press.
Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review, 21*(3), 785–793.
Carbon, C.-C., & Albrecht, S. (2012). Bartlett's schema theory: the unreplicated "portrait d'homme" series from 1932. *Quarterly Journal of Experimental Psychology, 65*(11), 2258–2270.
Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87–105.
Charlesworth, T. E. S., Hudson, S.-K. T. J., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2019). Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology, 55*(7), 1400–1413.
Cogsdill, E. J., & Banaji, M. R. (2015). Face-trait inferences show robust child–adult agreement: Evidence from three types of faces. *Journal of Experimental Social Psychology, 60*, 150–156.
Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science, 25*(5), 1132–1139.
Colombatto, C., Uddenberg, S., & Scholl, B. J. (2021). The efficiency of demography in face perception. *Attention, Perception, & Psychophysics, 83*(8), 3104–3117.
Cook, R., Eggleston, A., & Over, H. (2022). The cultural learning account of first impressions. *Trends in Cognitive Sciences, 26*(8), 656–668.
Cook, R., & Over, H. (2021). Why is the literature on first impressions so focused on White faces? *Royal Society Open Science, 8*(9), Article 211146.
Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), Article e57410.
Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting tool that showed bias against women*. Reuters.
Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour, 1*(1), 0001.
Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies, 25*(8), 2455–2484.
Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports, 3*(1), 1047.
Eggleston, A., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Culturally learned first impressions occur rapidly and automatically and emerge early in development. *Developmental Science, 24*(2), Article e13021.
Eggleston, A., Tsantani, M., Over, H., & Cook, R. (2022). Preferential looking studies of trustworthiness detection confound structural and expressive cues to facial trustworthiness. *Scientific Reports, 12*(1), 17709.
Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.
Farkas, J., Schou, J., & Neumayer, C. (2018). Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. *New Media & Society, 20*(5), 1850–1867.
Fisher, M. (2022). *The Chaos machine: The inside story of how social media rewired our minds and our world*. New York, NY, USA: Hachette Book Group.
Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847–857.
Graham, J. R., Harvey, C. R., & Puri, M. (2017). A corporate beauty contest. *Management Science, 63*(9), 3044–3056.
Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2006). Revealing priors on category structures through iterated learning. In *, Vol. 199. Proceedings of the 28th annual conference of the cognitive science society* (pp. 1394–1399).
Guala, F., & Mittone, L. (2010). Paradigmatic experiments: The Dictator Game. *The Journal of Socio-Economics, 39*(5), 578–584.
Henss, R. (1991). Perceiving age and attractiveness in facial photographs. *Journal of Applied Social Psychology, 21*(11), 933–946.
Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology, 27*(3), 359–370.
Jaeger, B., Oud, B., Williams, T., Krumhuber, E. G., Fehr, E., & Engelmann, J. B. (2022). Can people detect the trustworthiness of strangers based on their facial appearance? *Evolution and Human Behavior, 43*(4), 296–303.
Kadiresan, A., Baweja, Y., & Ogbanufe, O. (2022). Bias in AI-based decision-making. In M. V. Albert, L. Lin, M. J. Spector, & L. S. Dunn (Eds.), *Bridging human intelligence and artificial intelligence* (pp. 275–285). Cham: Springer International Publishing.
Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review, 14*(2), 288–294.
Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin, 26*(5), 594–604.
Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5*(2), 102–110.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87–102.
Kramer, R. S. S., & Gardner, E. M. (2020). Facial trustworthiness and criminal sentencing: A comment on Wilson and Rule (2015). *Psychological Reports, 123*(5), 1854–1868.
Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology, 68*, 565–579.
Langlois, T. A., Jacoby, N., Suchow, J. W., & Griffiths, T. L. (2021). Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences, 118*(13), Article e2012938118.
Lee, R., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Spontaneous first impressions emerge from brief training. *Scientific Reports, 11*(1), 15024.
de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12.
Lyons, A., Clark, A., Kashima, Y., & Kurz, T. (2008). Cultural dynamics of stereotypes: Social network processes and the perpetuation of stereotypes. In *Stereotype dynamics: Language-based approaches to the formation, maintenance, and transformation of stereotypes* (pp. 59–92). New York, NY, USA: Taylor & Francis.
Lyons, A., & Kashima, Y. (2001). The reproduction of culture: Communication processes tend to maintain cultural stereotypes. *Social Cognition, 19*(3), 372–394.
Meppelink, C. S., Smit, E. G., Fransen, M. L., & Diviani, N. (2019). "I was right about vaccination": Confirmation bias and health literacy in online health information seeking. *Journal of Health Communication, 24*(2), 129–140.
Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology, 97*(3), 405–423.
Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. In M. P. Zanna (Ed.), *Vol. 30. Advances in experimental social psychology* (pp. 93–161). Academic Press.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces. *Psychological Science, 30*(1), 65–79.
Oh, D., Dotsch, R., & Todorov, A. (2019). Contributions of shape and reflectance information to social judgments from faces. *Vision Research, 165*, 131–142.
O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092.
Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences, 119*(17), Article e2115228119.
Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law, 16*(6), 477–491.
Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting novel associations in large datasets. *Science, 334*(6062), 1518–1524.
Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS One, 7*(3), Article e34293.
Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review, 80*(1), 106–134.
Suchow, J. W., Pacer, M. D., & Griffiths, T. L. (2016). Design from zeroth principles. In *Proceedings of the 38th annual conference of the cognitive science society* (p. 6).
Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization* (pp. 1–9). Association for Computing Machinery: New York, NY, USA.
van't Wout, M., & Sanfey, A. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796–803.
Thierry, S. M., & Mondloch, C. J. (2021). First impressions of child faces: Facial trustworthiness influences adults' interpretations of children's behavior in ambiguous situations. *Journal of Experimental Child Psychology, 208*, Article 105153.
Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.
Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion, 13*(4), 724–738.
Todorov, A., Funk, F., & Olivola, C. Y. (2015). Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences, 19*(8), 422–423.
Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626.
Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833.
Uddenberg, S., & Scholl, B. J. (2018). Teleface: Serial reproduction of faces reveals a whiteward bias in race memory. *Journal of Experimental Psychology: General, 147*(10), 1466–1487.
Verhoef, T., & Ravignani, A. (2021). Melodic universals emerge or are sustained through cultural evolution. *Frontiers in Psychology, 12*, Article 668300.
Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences, 119*(29), Article e2204529119.
Weeks, B. E., Lane, D. S., Kim, D. H., Lee, S. S., & Kwak, N. (2017). Incidental exposure, selective exposure, and political information sharing: Integrating online exposure patterns and expression on social media. *Journal of Computer-Mediated Communication, 22*(6), 363–379.
Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change, 32*, 126–138.
Wilson, C. C., II, Gutierrez, F., & Chao, L. (2012). *Racism, sexism, and the media: Multicultural issues into the new communications age*. SAGE Publications.
Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325–1331.
Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758), 20123073.
Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology, 60*(2), 107–126.
Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science, 26*(3), 237–242.
Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review, 7*(3), 194–215.
Zebrowitz, L. A., Kendall-Tackett, K., & Fafel, J. (1991). The influence of children's facial maturity on parental expectations and punishments. *Journal of Experimental Child Psychology, 52*(2), 221–238.
Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493–504.