Cognition 217 (2021) 104916

Contents lists available at ScienceDirect

Cognition

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

Short Communication

Revising mental representations of faces based on new diagnostic information

Samuel A.W. Klein *, Ryan J. Hutchings, Andrew R. Todd *

University of California, Davis, 1 Shields Avenue, Davis, CA 95616, United States of America

ARTICLE INFO

Keywords:
Face impressions
Mental representations
Reverse correlation
Social cognition
Updating

ABSTRACT

Extending evidence for the rapid revision of mental representations of what other people are like, we explored whether people also rapidly revise their representations of what others *look* like. After learning to ascribe positive or negative behavioral information to a target person and generating a visualization of their face in a reverse-correlation task, participants learned new information that was (a) counter-attitudinal and diagnostic about the person's character or (b) neutral and non-diagnostic, and then they generated a second visualization. Ratings of these visualizations in separate samples of participants consistently revealed revision effects: Time 2 visualizations assimilated to the counter-attitudinal information. Weaker revision effects also emerged after learning neutral information, suggesting that the evaluative extremity of visualizations may dilute when encountering any additional information. These findings indicate that representations of others' appearance may change upon learning more about them, particularly when this new information is counter-attitudinal and diagnostic.

1. Introduction

First impressions are lasting impressions (Asch, 1946)—or so it has been assumed, particularly for implicit (i.e., unintentional) impressions. Indeed, some dual-process theorizing has maintained that, although people readily revise their explicit impressions of others when encountering countervailing target information, implicit impression revision occurs more slowly, if at all (e.g., Rydell & McConnell, 2006). This claim was initially supported by research that failed to find implicit impression revision based on countervailing target information (e.g., Gregg, Seibt, & Banaji, 2006). Although the malleability of implicit impressions has long been recognized (Gawronski & Bodenhausen, 2006), an assumption guiding this literature has been that exposure to abundant countervailing information is required for implicit impression revision.

Counter to this assumption, accumulating evidence now indicates that people can rapidly revise their implicit impressions when encountering even a single piece of diagnostic information that contradicts their initial impression (Ferguson, Mann, Cone, & Shen, 2019). For example, participants who first learned positive behavioral information about a person fully reversed their initially favorable impression after learning about his child molestation conviction (Cone & Ferguson, 2015). Such updating generalizes beyond the context in which the impression was formed (Brannon & Gawronski, 2017) and is evident days later (Cone, Flaharty, & Ferguson, 2021), suggesting genuine revision.

Notably, all documented instances of rapid impression revision have emerged in mental representations of the target person's *character*, commonly assessed with sequential-priming tasks (e.g., affect misattribution procedure [AMP]; Payne, Cheng, Govorun, & Stewart, 2005) and self-report measures. Although tracking impression revision with such measures provides insight into evaluative assumptions about the person's character, it is silent on how the person's physical *appearance* is initially represented or potentially revised.

Here, we explored the revision of facial appearance representations using reverse correlation (Mangini & Biederman, 2004), a data-driven approach for visualizing the features underlying face classifications. This technique imposes no pre-existing assumptions about these features, thereby affording an unconstrained assessment of what another person *looks* like. The measurement outcomes of reverse-correlation tasks, unlike those of the AMP and other indirect measures of what a person *is* like, include conditionally variable features of physical appearance (Brinkman, Todorov, & Dotsch, 2017). Exactly how these features relate to measures that serve as proxies of character representations remains an open question (Dotsch, Wigboldus, & Van Knippenberg, 2013).

Our investigation comprised an image-generation experiment and three image-assessment experiments. In the image-generation

* Corresponding authors.
E-mail addresses: sawklein@ucdavis.edu (S.A.W. Klein), atodd@ucdavis.edu (A.R. Todd).

https://doi.org/10.1016/j.cognition.2021.104916
Received 3 March 2021; Received in revised form 21 September 2021; Accepted 22 September 2021
Available online 28 September 2021

experiment, participants visualized a target person twice: first after learning to ascribe positive or negative behaviors to him, and again after receiving new information that was (a) diagnostic, extreme, and contradictory to the initial information or (b) neutral and non-diagnostic. This procedure produced classification images for each of 8 experimental conditions. In three image-assessment experiments, with three distinct image-processing procedures, separate samples of participants rated these images on traits that are detectable in faces. Data for all experiments are available here: https://osf.io/7u6cd/

2. Image-generation experiment

2.1. Method

2.1.1. Participants

Prior research on implicit impression revision has revealed large effects (η_p^2s > 0.12; Cone & Ferguson, 2015); however, whether appearance representations shift comparably remains unknown. Rounding up to the nearest number divisible by 50, we thus set our target sample size near our smallest effect of interest ($\eta_p^2 = 0.05$; 90% power) in our 2 × 2 × 2 mixed design.[1,2] After surpassing our target sample (250 participants), we continued data collection until the week's end. In total, 338 undergraduates participated for course credit. We excluded data from 53 participants who pressed the same key for ≥95% of image-generation trials at Time 1 or Time 2. The final sample comprised 285 participants (see Table 1 for participant demographics in all experiments). Across experiments, participants provided informed consent prior to participating.

2.1.2. Procedure

Participants first learned to ascribe positive or negative behaviors to a target person, Robert (Cone & Ferguson, 2015). They read (in randomized order) 64 behaviors (32 positive, 32 negative; Rydell & McConnell, 2006) and indicated whether each was characteristic or uncharacteristic of him, after which condition-specific feedback appeared for 2.5 s. In the *positive-induction* condition, a blue *correct* message appeared after classifying a positive (negative) behavior as characteristic (uncharacteristic), and a red *incorrect* message appeared after classifying a negative (positive) behavior as characteristic (uncharacteristic). Accompanying each message was a summary statement (e.g., "Giving flowers to his mother is characteristic of Robert"). In the *negative-induction* condition, these feedback contingencies were reversed. Participants then reported their impressions of Robert on 7 traits that are important for person impressions (Oosterhof & Todorov, 2008): trustworthy, attractive, dominant, caring, intelligent, aggressive, and mean (1 = *not at all*, 7 = *extremely*).

Next, participants completed a reverse-correlation task (Brinkman et al., 2017). On each of 350 trials, they selected which of two side-by-side degraded face images looked more like Robert. Each pair of images comprised a random noise pattern[3] and its inverse superimposed onto a base face image.[4] This technique maximizes between-image contrast (Dotsch & Todorov, 2012). Responses <100 ms or >4000 ms after target onset prompted a message to respond more slowly or more quickly, respectively.

Participants then received one new piece of information about Robert. In the *counter-attitudinal* condition, the information was diagnostic about his character and contradicted the valence of their Time 1 induction ("Robert was recently convicted of child molestation" after a positive induction; "Robert donated one of his kidneys to a child in need he had never met before" after a negative induction). These behaviors were rated similarly in diagnosticity and valence extremity (see Cone & Ferguson, 2015). In the *neutral* condition, the information was neutral in valence ("Robert recently bought a soda"). Finally, participants again reported their impressions of Robert and completed the reverse-correlation task in a newly randomized order.

Using the *rcicr* package (Dotsch, 2014), we created group classification images by superimposing onto the base face the average noise patterns of the selected images across all participants in each condition. Group images reflect the average features visualized of Robert within that condition (Fig. 1).[5]

2.2. Results

All analyses were conducted via linear mixed-effects models (LMEMs), with each model containing fixed effects for Time, Time 1 induction, Time 2 information, and all possible interactions. For each model, we began with its maximal random-effects structure (i.e., random intercepts and all appropriate random slopes for each source of variance; Barr, Levy, Scheepers, & Tily, 2013) and downsized to solve problems of non-convergence and singularity. The sources of variance were participants and traits in the image-generation experiment and image-assessment Experiment 1, and participants and stimuli in image-assessment Experiments 2A and 2B.[6]

We reverse-scored responses for aggressive, dominant, and mean, ensuring that all traits were directionally consistent in valence, and considered the 7 traits as having been sampled from the population of positive traits on which impressions could be formed. Because the trait ratings were highly correlated (see the Supplementary Materials) and fitting separate models for each trait can inflate Type-I error (Herzog, Francis, & Clarke, 2019), we included random effects for traits.[7]

This analysis revealed a significant three-way interaction, $b = -0.37$, $SE = 0.03$, $F(1, 280.97) = 206.33$, $p < .001$ (Fig. 2). To explicate this interaction, we examined contrasts of the model's Time × Time 2 information interactions separately in the positive-induction and negative-induction conditions. This interaction was significant in both the positive-induction condition, $b = 2.27$, $SE = 0.15$, $t(281) = 15.68$, p

[1] To our knowledge, no formal power analysis procedures exist for the image-generation phase in reverse-correlation paradigms (see also Brown-Iannuzzi, Cooley, Marshburn, McKee, & Lei, 2021).

[2] Our planned analyses were conducted at the level of participants; however, based on editorial feedback, we shifted to analyses that account for other sources of variance (i.e., traits or stimuli, depending on the experiment). Thus, the reported a priori power analyses, conducted with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), only considered the number of participants, but not the number of traits or stimuli.

[3] The noise patterns comprised 4092 superimposed truncated sinusoid patches in all possible combinations of 2 cycles in 6 orientations (0°, 30°, 60°, 90°, 120°, 150°) × 5 spatial frequencies (1, 2, 4, 8, 16 patches per image) × 2 phases (0, π/2), with random contrasts.

[4] The base face, which Krosch and Amodio (2014) created by morphing 100 White and 100 Black male faces, has been used in prior reverse-correlation research (e.g., Lei & Bodenhausen, 2017).

[5] We report in the Supplementary Materials additional measures collected in all experiments.

[6] See the Supplementary Materials for a detailed description of the random-effects structures for each mixed-effects model reported in the main text, and a discussion of how problems of non-convergence and singularity in the maximal models led to those reported in the main text.

[7] An alternative analytic approach entails using data-reduction techniques (e.g., exploratory factor analysis) to fit these models to latent factor(s). Because our focus was on testing for revision effects in representations of appearance, regardless of the trait, we do not report this approach in the main text (but see the Supplementary Materials for results using this alternative approach).

Table 1
Participant demographics in each experiment.

Experiment	Age		Gender (%)			Race/Ethnicity (%)				
	M	*SD*	Male	Female	Nonbinary	W	B	A	L	M
IG	19.9	2.2	20.7	76.8	1.4	18.9	2.5	46.0	20.4	12.3
IA 1	35.3	10.6	63.2	36.1	0.0	38.7	36.8	6.5	7.7	10.3
IA 2A	38.5	12.6	48.2	50.0	0.0	76.3	6.1	6.1	1.8	9.6
IA 2B	20.0	2.7	18.2	79.3	0.0	15.7	0.4	50.8	20.2	12.8

Note. IG = image generation, IA = image assessment. Some participants did not report their gender or race/ethnicity. For race/ethnicity, W = White or European American, B = Black or African American, A = Asian American or Pacific Islander, L = Latinx or Hispanic, and M = reported other or more than one race/ethnicity.

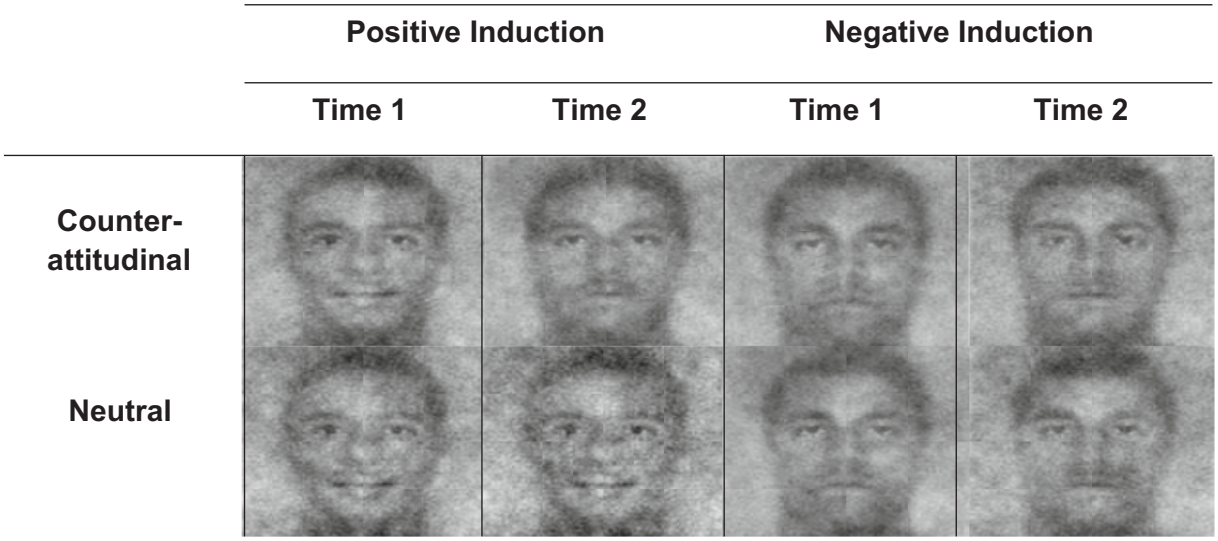

Fig. 1. Group classification images by Time, Time 1 induction, and Time 2 information.

$< .001$, and the negative-induction condition, $b = -0.68$, $SE = 0.15$, $t(281) = -4.66$, $p < .001$, with significantly greater positive-to-negative than negative-to-positive revision, $b = -1.59$, $SE = 0.21$, $t(281) = -7.78$, $p < .001$.[8]

Next, we conducted pairwise comparisons in each induction condition. In the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 2.42$, $SE = 0.13$, $t(30.00) = 19.40$, $p < .001$, but learning neutral information did not, $b = 0.15$, $SE = 0.13$, $t(31.60) = 1.21$, $p = .235$. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.91$, $SE = 0.13$, $t(32.20) = -7.17$, $p < .001$, but learning neutral information did not, $b = -0.24$, $SE = 0.13$, $t(30.00) = -1.88$, $p = .070$.[9]

These results replicate prior findings indicating that learning diagnostic counter-attitudinal target information prompts character representation revision (Cone & Ferguson, 2015). As in this prior work, we also observed an asymmetry whereby positive-to-negative revision was stronger than negative-to-positive revision.

3. Image-assessment experiments

To examine whether learning countervailing diagnostic information prompts appearance representation revision, we conducted several image-assessment experiments. In Experiment 1, a new sample of participants rated the 8 group images (Fig. 1) on the same traits from before. Experiments 2A and 2B used two alternative image-processing procedures (detailed below) and two new samples of raters to assess apparent trustworthiness.

3.1. Experiment 1

3.1.1. Method

3.1.1.1. Participants. We again considered the large revision effects (η_p^2s > 0.12) in Cone and Ferguson (2015) but allowed for weaker effects. To detect a medium-sized three-way interaction ($\eta_p^2 = 0.06$) with 80% power in a $2 \times 2 \times 2$ within-participants design, we set our target sample size at 126. Amazon's Mechanical Turk (MTurk) workers ($N = 155$) participated for pay. No data were excluded; thus, the final sample comprised 155 participants.

3.1.1.2. Procedure. Participants rated the 8 group images on the same 7 traits from the image-generation experiment.

3.1.2. Results

A LMEM revealed a significant three-way interaction, $b = -0.11$, $SE = 0.01$, $F(1, 8358) = 52.32$, $p < .001$ (Fig. 3). We again examined contrasts of the model's Time × Time 2 information interactions separately in the two induction conditions. This interaction was significant in both the positive-induction condition, $b = 0.65$, $SE = 0.08$, $t(8358) =$

[8] This additional post-hoc test assessed the magnitude of revision, as reflected in the difference between the Time × Time 2 information contrast in the positive-induction condition and this same contrast in the negative-induction condition. Due to the opposing numerical directions of the contrasts, we first multiplied all ratings in the negative-induction condition by a constant of −1, ensuring that the difference between the two contrasts reflects the *magnitude* of difference. We used this same approach in all three image-assessment experiments.

[9] For detailed information on the descriptive statistics for these and all other experiments, see the Supplementary Materials.

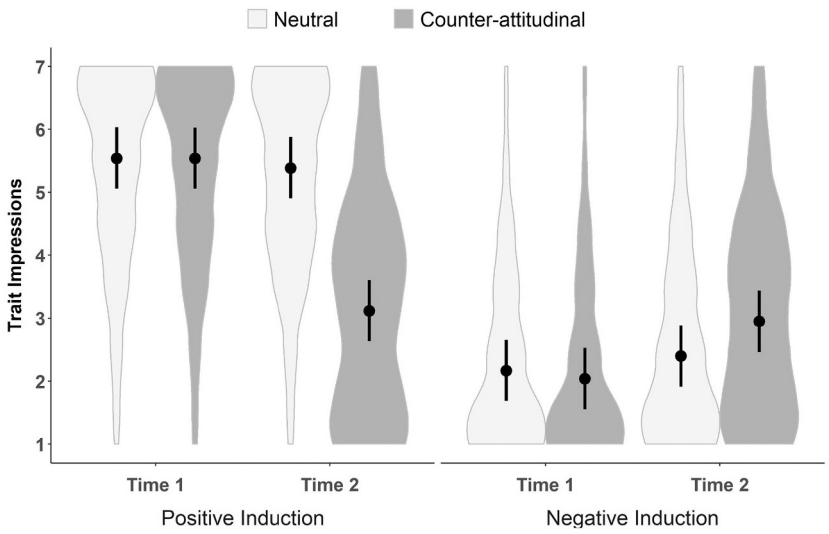

Fig. 2. Estimated marginal means of trait impressions of Robert by Time, Time 1 induction, and Time 2 information in the image-generation experiment. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image generators' responses after a smoothing function was applied.

7.69, $p < .001$, and the negative-induction condition, $b = -0.21$, $SE = 0.08$, $t(8358) = -2.54$, $p = .021$, with significantly greater positive-to-negative than negative-to-positive revision, $b = -0.43$, $SE = 0.13$, $t(8358) = -3.31$, $p = .001$.

Pairwise comparisons in each induction condition revealed that, in the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 0.66$, $SE = 0.06$, $t(8358) = 11.16$, $p < .001$, but learning neutral information did not, $b = 0.02$, $SE = 0.06$, $t(8358) = 2.66$, $p = .790$. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.37$, $SE = 0.06$, $t(8358) = -6.63$, $p < .001$, but so did learning neutral information, $b = -0.16$, $SE = 0.06$, $t(8358) = -2.69$, $p = .007$, albeit to a lesser extent.[10]

Visualizations of Robert's appearance grew less favorable upon learning negative counter-attitudinal information and more favorable upon learning positive counter-attitudinal information. Notably, positive-to-negative revision was stronger than negative-to-positive, replicating findings from the image-generation experiment and elsewhere (Cone & Ferguson, 2015). When learning neutral information, participants did not visualize Robert's appearance differently if their initial visualization was positive; however, they did visualize him more favorably if their initial visualization was negative. Because the neutral information was non-diagnostic, revision here may reflect a dilution effect (Nisbett, Zukier, & Lemley, 1981), whereby highly negative initial visualizations become less extreme upon learning any additional information.

3.2. Experiments 2A and 2B

Image-assessment Experiment 1 relied exclusively on group images that were created by aggregating across the responses of all image generators per condition. This practice, though normative in reverse-correlation research (Brinkman et al., 2017), can artificially augment between-condition differences, thereby inflating Type-I error (Cone, Brown-Iannuzzi, Lei, & Dotsch, 2021). Image-assessment Experiments 2A and 2B used alternative image-processing procedures—subgroup and individual classification images—that avoid this limitation (Cone, Brown-Iannuzzi, et al., 2021; Hutchings, Simpson, Sherman, & Todd, 2021). Both experiments assessed apparent trustworthiness, given its centrality in face impressions (Oosterhof & Todorov, 2008).

3.2.1. Method

3.2.1.1. Participants. We considered the possibility that subgroup and (perhaps especially) individual images are noisier than group images, potentially producing smaller effects. To detect medium-sized three-way interactions (Experiment 2A: $\eta_p^2 = 0.06$; Experiment 2B: $\eta_p^2 = 0.03$) with 80% power, we set target sample sizes of 126 (Experiment 2A) and 257 (Experiment 2B). In total, 125 MTurkers (Experiment 2A) and 259 undergraduates (Experiment 2B) participated for pay and course credit, respectively. We excluded data from participants who gave the same response on ≥95% of ratings (Experiment 2A: $n = 6$; Experiment 2B: $n = 5$) or who did not finish the entire experiment (Experiment 2A: $n = 5$; Experiment 2B: $n = 12$). The final samples comprised 114 participants in Experiment 2A and 242 participants in Experiment 2B.

3.2.1.2. Procedure. We used the *rcicr* package (Dotsch, 2014) to create subgroup and individual images. In Experiment 2A, we created subgroup images by aggregating the noise patterns selected by 12 random subsets of image generators in each condition and superimposing them onto the base face (Cone, Brown-Iannuzzi, et al., 2021). The total stimulus set included 96 subgroup images, with each image comprising the average selected noise patterns from 5 to 7 image generators. Participants rated all 96 subgroup images (order randomized). In Experiment 2B, we created individual images by aggregating the noise patterns selected by each image generator, separately for each time point, and superimposing them onto the base face. The total stimulus set comprised 570 images. To minimize fatigue, we had participants rate one of three sets of 95 randomized pairs of Time 1 and Time 2 images, totaling 190 images. In both experiments, participants rated how trustworthy the person looked (1 = *extremely untrustworthy*, 7 = *extremely trustworthy*).

3.2.2. Results

3.2.2.1. Experiment 2A (Subgroup Images). A LMEM revealed a significant three-way interaction, $b = -0.11$, $SE = 0.03$, $F(1, 45.27) = 10.67$, $p = .002$ (Fig. 4). Next, we examined contrasts of the model's Time × Time 2 information interactions separately in the two induction conditions. This interaction was significant in the positive-induction condition, $b =$

[10] A difference between counter-attitudinal and neutral information in the negative-induction condition is evidenced by the significant Time × Time 2 contrast in the negative-induction condition reported above.

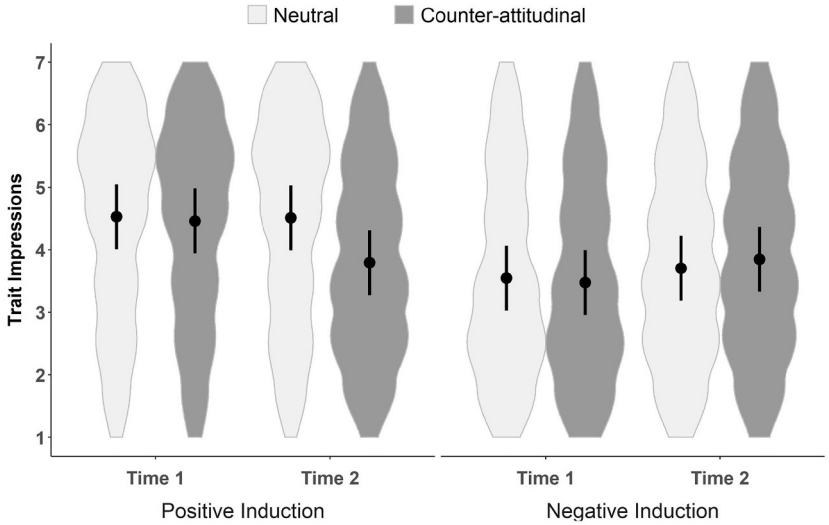

Fig. 3. Estimated marginal means of trustworthiness impressions of group classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 1. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image raters' responses after a smoothing function was applied.

0.72, $SE = 0.19$, $t(44.8) = 3.89$, $p < .001$, but not in the negative-induction condition, $b = -0.14$, $SE = 0.19$, $t(44.8) = -0.75$, $p = .456$, with significantly greater positive-to-negative than negative-to-positive revision, $b = -0.58$, $SE = 0.26$, $t(44.3) = -2.23$, $p = .031$.

Pairwise comparisons in each induction condition revealed that, in the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 0.92$, $SE = 0.13$, $t(48.4) = 6.87$, $p < .001$, but learning neutral information did not, $b = 0.20$, $SE = 0.13$, $t(48.40) = 1.48$, $p = .145$. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.65$, $SE = 0.13$, $t(48.40) = -4.90$, $p < 001$, but so did learning neutral information, $b = -0.52$, $SE = 0.13$, $t(48.40) = -3.86$, $p < 001$.

3.2.2.2. Experiment 2B (Individual Images). Once again, the three-way interaction was significant, $b = -0.03$, $SE = 0.01$, $F(1, 280.36) = 5.90$, $p = .016$ (Fig. 5). As before, we examined contrasts of the model's Time × Time 2 information interactions separately in the two induction conditions. This interaction was significant in the positive-induction condition, $b = 0.20$, $SE = 0.08$, $t(281) = 2.55$, $p < .001$, but not in the negative-induction condition, $b = -0.07$, $SE = 0.08$, $t(282) = -0.89$, $p = .372$, with no significant difference in the magnitude of positive-to-negative versus negative-to-positive revision, $b = -0.13$, $SE = 0.11$, $t(279) = -1.17$, $p = .244$.

Pairwise comparisons in each induction condition revealed that, in the positive-induction condition, learning about Robert's child molestation conviction prompted negative revision, $b = 0.39$, $SE = 0.06$, $t(303) = 6.88$, $p < .001$. Learning neutral information also prompted negative revision, $b = 0.19$, $SE = 0.06$, $t(303) = 3.24$, $p = .001$, albeit to a lesser extent. In the negative-induction condition, learning about Robert's kidney donation prompted positive revision, $b = -0.26$, $SE = 0.06$, $t(293) = -4.43$, $p < 001$, but so did learning neutral information, $b = -0.19$, $SE = 0.13$, $t(294) = -3.29$, $p < 001$.

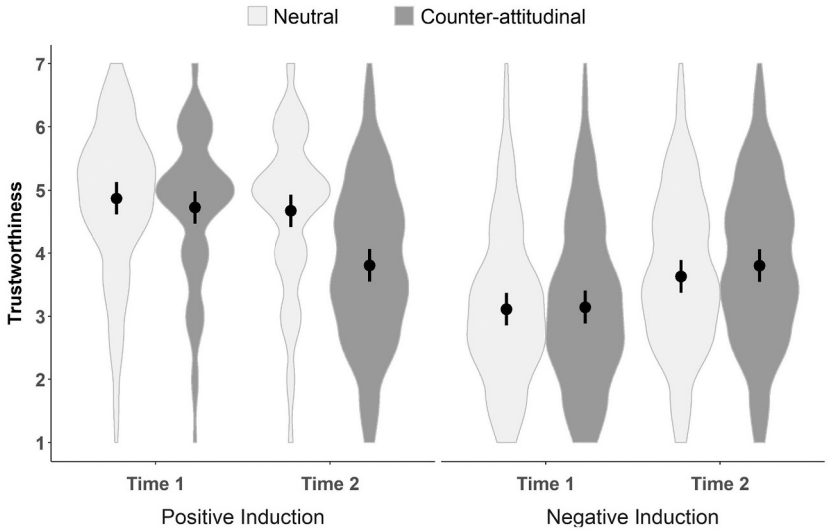

Fig. 4. Estimated marginal means of trustworthiness impressions of subgroup classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 2A. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image raters' responses after a smoothing function was applied.

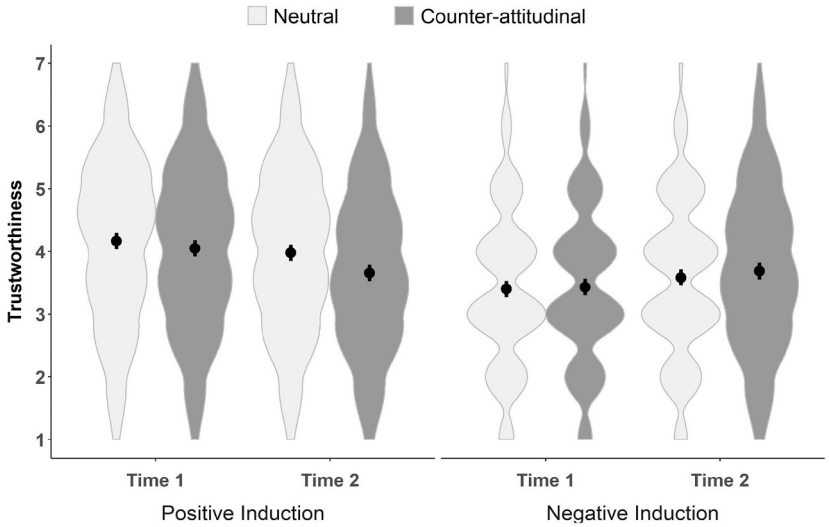

Fig. 5. Estimated marginal means of trustworthiness impressions of individual classification images by Time, Time 1 induction, and Time 2 information in image-assessment Experiment 2B. Error bars represent 95% confidence intervals. The surrounding violin plots illustrate mirrored density distributions of image raters' responses after a smoothing function was applied.

4. Discussion

Using a reverse-correlation paradigm, we found that visualizations of a target person's face consistently assimilated to new information that was extreme, diagnostic, and contradictory to the initial information learned about him. Initially positive visualizations were revised to appear less trustworthy after learning about his child molestation conviction. The opposite pattern emerged when participants with negative initial visualizations learned about his kidney donation, albeit sometimes to no greater extent than the revision prompted by learning neutral information.[11] These results complement other evidence of rapid revision in mental representations of what others are like under similar conditions (Ferguson et al., 2019). We also found a valence asymmetry, whereby greater positive-to-negative (vs. negative-to-positive) revision emerged in most cases (cf. Cone & Ferguson, 2015).

Some evidence of revision also emerged, albeit more weakly, when learning new neutral information about the person. We suspect that revision here might reflect a dilution effect, whereby new non-diagnostic information diluted the extremity of the initial visualizations (Nisbett et al., 1981). If so, one implication of this finding is that extreme appearance representations may dissipate over time upon learning any additional target information.

A strength of this work is its use of group, subgroup, and individual classification images, with the latter two procedures reducing concerns about Type-I error inflation (Cone, Brown-Iannuzzi, et al., 2021). Because subgroup aggregation is a new technique, future work should explore optimal points at which subgroup images minimize noise but preserve image-generator variability. Furthermore, although some results (e.g., revision after neutral information) varied across experiments, the key effect (i.e., stronger revision after counter-attitudinal vs. neutral information) emerged consistently. Such convergence helps bolster our conclusions. Future work should identify boundary conditions of these effects. For example, if the information initially learned about a person (e.g., broke into his neighbor's house) is reinterpreted based on new information (e.g., the house was on fire and children were inside), do we revise our representations of their appearance accordingly (Mann & Ferguson, 2015)?

Notably, we found a sizable correlation between image generators' trustworthiness ratings of Robert in the image-generation experiment and image raters' trustworthiness ratings of individual images of Robert in image-assessment Experiment 2B, $r(568) = 0.51$, $p < .001$, suggesting a correspondence between revision in (explicit) character representations and revision in appearance representations, at least when the new information learned about the person is diagnostic, extreme, and (presumably) believable (see Ferguson et al., 2019). What remains for future research is determining whether similar correspondence emerges if one of these elements is missing or under conditions in which corresponding revisions in implicit and explicit character representations have not materialized in prior work (e.g., Gregg et al., 2006). Future research should also explore whether revisions in character representations precede (and/or cause) changes in appearance representations. Answering these questions promises a richer understanding of how various components of person impressions are integrated.

The current findings indicate that mental representations of others' appearance are far from static. As we learn new information about someone, not only do we revise our representations of what they are like; we also revise our representations of what they *look* like.

Acknowledgements

This research was facilitated by National Science Foundation Grant BCS-1764097 (awarded to ART). We thank Andre Wang for statistical advice, Rebecca Neufeld for assistance with data collection, and the members of the Attitudes and Social Cognition Lab for constructive feedback at various stages of this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104916.

References

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290. https://doi.org/10.1037/h0055756.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context-(in)dependent updating of implicit evaluations. *Social Psychological and Personality Science, 8*, 275–283. https://doi.org/10.1177/1948550616673875.

[11] Less conservative analyses that did not account for random effects of stimuli consistently revealed revision effects in both induction conditions in all experiments. We report these analyses in the Supplementary Materials.

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology, 28*, 333–361. https://doi.org/10.1080/10463283.2017.1381469.

Brown-Iannuzzi, J. L., Cooley, E., Marshburn, C. K., McKee, S. E., & Lei, R. F. (2021). Investigating the interplay between race, work ethic stereotypes, and attitudes toward welfare recipients and policies. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550620983051.

Cone, J., Brown-Iannuzzi, J., Lei, R., & Dotsch, R. (2021). Type I error is inflated in the two-phase reverse correlation procedure. *Social Psychological and Personality Science, 12*, 760–768. https://doi.org/10.1177/1948550620938616.

Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*, 37–57. https://doi.org/10.1037/pspa0000014.

Cone, J., Flaharty, K., & Ferguson, M. J. (2021). The long-term effects of new evidence on implicit impressions of other people. *Psychological Science, 32*, 173–188. https://doi.org/10.1177/0956797620963559.

Dotsch, R. (2014). *rcicr: Reverse correlation image classification toolbox*. R package. Retrieved from http://ron.dotsch.org/rcicr.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*, 562–571. https://doi.org/10.1177/1948550611430272.

Dotsch, R., Wigboldus, D. H. J., & Van Knippenberg, A. D. (2013). Behavioral information biases the expected facial appearance of members of novel groups. *European Journal of Social Psychology, 43*, 116–125. https://doi.org/10.1002/ejsp.1928.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/bf03193146.

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science, 28*, 331–336. https://doi.org/10.1177/0963721419835206.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731. https://doi.org/10.1037/0033-2909.132.5.692.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*, 1–20. https://doi.org/10.1037/0022-3514.90.1.1.

Herzog, M. H., Francis, G., & Clarke, A. (2019). The multiple testing problem. In *Understanding statistics and experimental design (pp. 63–66)*. Cham: Springer.

Hutchings, R. J., Simpson, A. J., Sherman, J. W., & Todd, A. R. (2021). Perspective taking reduces intergroup bias in visual representations of faces. *Cognition, 214*, 104808. https://doi.org/10.1016/j.cognition.2021.104808.

Krosch, A. R., & Amodio, D. M. (2014). Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences, 111*, 9079–9084. https://doi.org/10.1073/pnas.1404448111.

Lei, R. F., & Bodenhausen, G. V. (2017). Racial assumptions color the mental representation of social class. *Frontiers in Psychology, 8*, 519. https://doi.org/10.3389/fpsyg.2017.00519.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*, 209–226. https://doi.org/10.1016/j.cogsci.2003.11.004.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*, 823–849. https://doi.org/10.1037/pspa0000021.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13*, 248–277. https://doi.org/10.1016/0010-0285(81)90010-4.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*, 11087–11092. https://doi.org/10.1073/pnas.0805664105.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293. https://doi.org/10.1037/0022-3514.89.3.277.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*, 995–1008. https://doi.org/10.1037/0022-3514.91.6.995.