A learning model can explain both shared and idiosyncratic first impressions from faces

Richard Cook[a,b,1] and Harriet Over[b]

In their contribution to PNAS, Sutherland et al. (1) document stable individual differences in the character traits spontaneously attributed to strangers based solely on their facial appearance ("first impressions"). Data from the accompanying twin study suggest that these idiosyncratic first impressions are products of individuals' direct social experience. These findings accord well with earlier laboratory-based work indicating that social experience influences observers' subsequent impressions of facial trustworthiness (2, 3).

Documenting stable individual differences in observers' first impressions is valuable, not least because these data potentially inform accounts of developmental origin. It remains unclear, however, why there is also considerable consensus between observers in terms of the character traits inferred spontaneously from the faces of strangers. This is an important question to address as some forms of consensus—negative evaluation of particular faces by many members of a community—can result in systematic discrimination (4).

Many authors, including Sutherland et al., fall back on genetic explanations to explain high levels of interrater agreement, where observed (5, 6). However, the logic of this position is not at all straightforward. Importantly, first impressions—even those where high levels of consensus exist—bear little relation to the ground truth; people who are judged to be untrustworthy are frequently trustworthy, and vice versa (4). If evolution endowed us with a mechanism for inferring the character traits of others, suffice to say it does not do a very good job! It is debatable whether such unreliable first impressions would have conveyed any adaptive advantage. Our ancestors may have been better off assuming nothing about the traits of strangers from their appearance (7).

A learning framework, on the other hand, can be used to understand both consensus, where observed, and stable individual differences in first impressions. The "Trait Inference Mapping" framework assumes that first impressions are the result of learned associations between points in face space and trait space (8). These mappings allow excitation to spread automatically from perceptual descriptions of face shape to representations of particular trait profiles, conceived of as points in a high-dimensional trait space (9).

Idiosyncratic mappings acquired as a result of direct social interactions with others may account for the kinds of individual differences studied by Sutherland et al. (1). Crucially, however, a permissive learning mechanism can also produce consistent first impressions within a culture. Exposure to depictions of "good guys" and "bad guys"—"leaders" and "followers" in illustrated storybooks, film, television, ritual, art, and iconography—may lead different individuals within a society to acquire similar face–trait mappings (8).

An interesting implication of this view is that, earlier in human history, before mass media and the invention of the printing press, there may have been far greater variability in first impressions than we see today. In contemporary Western, educated, industrialized, rich, and democratic (WEIRD) cultures (10), stereotypical depictions of particular character types may attenuate some variability in first impressions. In the absence of systematic cultural influences, however, we hypothesize that in some non-WEIRD cultures, first impressions may be more idiosyncratic.

Acknowledgments

R.C. (ERC-STG-715824) and H.O. (ERC-STG-755719) both receive funding from the European Research Council.

1 C. A. M. Sutherland *et al.*, Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10218–10224 (2020).

2 V. B. Falvello, M. Vinson, C. Ferrari, A. Todorov, The robustness of learning about the trustworthiness of other people. *Soc. Cogn.* **33**, 368–386 (2015).

3 O. FeldmanHall *et al.*, Stimulus generalization as a mechanism for learning to trust. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1690–E1697 (2018).

[a]Department of Psychological Sciences, Birkbeck, University of London, London WC1E 7HX, United Kingdom; and [b]Department of Psychology, University of York, York YO10 5DD, United Kingdom

Author contributions: R.C. and H.O. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: richard.cook@bbk.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.2008816117

4 C. Y. Olivola, F. Funk, A. Todorov, Social attributions from faces bias human choices. *Trends Cogn. Sci. (Regul. Ed.)* **18**, 566–570 (2014).
5 M. Van Vugt, A. E. Grabo, The many faces of leadership: An evolutionary psychology approach. *Curr. Dir. Psychol. Sci.* **24**, 484–489 (2015).
6 L. A. Zebrowitz, The origins of first impressions. *J. Cult. Evol. Psychol*. **2**, 93–108 (2004).
7 A. Todorov, F. Funk, C. Y. Olivola, Response to Bonnefon et al.: Limited "kernels of truth" in facial inferences. *Trends Cogn. Sci. (Regul. Ed.)* **19**, 422–423 (2015).
8 H. Over, R. Cook, Where do spontaneous first impressions of faces come from? *Cognition* **170**, 190–200 (2018).
9 D. Hassabis *et al.*, Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cereb. Cortex* **24**, 1979–1987 (2014).
10 J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world? *Behav. Brain Sci*. **33**, 61–83, discussion 83–135 (2010).