Examining the durability of incidentally learned trust from gaze cues

James W. A. Strachan and Steven P. Tipper

University of York

Corresponding author:

J.W.A. Strachan,

Department of Psychology, University of York,

York, North Yorkshire, YO10 5DD.

Contact: js756@york.ac.uk

Abstract

In everyday interactions we find our attention follows the eye gaze of faces around us. As this cueing is so powerful and difficult to inhibit, gaze can therefore be used to facilitate or disrupt visual processing of the environment, and when we experience this we infer information about the trustworthiness of the cueing face. However, to date no studies have investigated how long these impressions last. To explore this we used a gaze-cueing paradigm where faces consistently demonstrated either valid or invalid cueing behaviours. Previous experiments show that valid faces are subsequently rated as more trustworthy than invalid faces. We replicate this effect (Experiment 1) and then include a brief interference task in Experiment 2 between gaze cueing and trustworthiness rating, which weakens but does not completely eliminate the effect. In Experiment 3, we explore whether greater familiarity with the faces improves the durability of trust learning and find that the effect is more resilient with familiar faces. Finally, in Experiment 4, we push this further and show evidence of trust learning can be seen up to an hour after cueing has ended. Taken together, our results suggest that incidentally learned trust can be durable, especially for faces that deceive.

Keywords: gaze-cueing; memory; trait inference; trustworthiness; familiarity

When we see a person's eyes move, our own attention reorients towards what they are looking at. This results in faster and easier processing of targets in the gazed-at location than of targets in an ignored location (Bayliss, di Pellegrino, & Tipper, 2004; Bayliss & Tipper, 2005; Driver et al., 1999; Friesen & Kingstone, 1998; Frischen & Tipper, 2006; Langton & Bruce, 1999).

Faces also carry with them social information, and this allows viewers to infer the cueing face's internal state from their gaze direction. Bayliss, Paul, Cannon, and Tipper (2006) found that objects that were gazed at were liked more than when gaze was oriented away from objects, and Bayliss, Frischen, Fenske, and Tipper (2007) found that this was modulated by the face's emotional expression, suggesting a sophisticated interpretative mechanism to detect objects that others find desirable. Droulers and Adil (2015) went further to show that looked-at objects were remembered better than were ignored objects.

Gaze cueing can impact more than processing of the objects involved – it can also influence how the cueing faces are processed. Bayliss and Tipper (2006) used a paradigm where faces either always cued the correct location of a subsequent target (provided valid cues) or always cued the incorrect location (invalid cues). Later, when participants were presented with pairs of faces – one valid, one invalid – and asked which they felt was the most trustworthy, they consistently picked the valid face, despite having been told that the faces they saw were irrelevant to the task they were instructed to complete.

This incidental learning of trust from gaze cues has been replicated in several studies since then. Manssuer, Roberts, and Tipper (2015) adapted the two-alternative forced choice (2AFC) measure of trust used in the original study and replaced it with scalar ratings of trustworthiness at the beginning and end of

the experiment – a technique that allows for examination of more subtle variations in trustworthiness, and one that allows for the tracking of trustworthiness over the course of the experiment (see also Manssuer, Pawling, Hayes, & Tipper, 2016; Strachan, Kirkham, Manssuer, & Tipper, 2016 for further replications).

However, despite several replications of this original effect, one question that has never been explored directly concerns the stability of the incidentally learned trustworthiness of a person. That is, no studies to date have examined how long incidentally learned trustworthiness lasts, despite the fact that this question carries important implications for interpreting this effect. If this trust learning is short-lived and easily disrupted by an intervening task, then this suggests that it reflects an active, online monitoring of in-the-moment statistical contingencies, concerned only with short-term interactions. If, on the other hand, this effect can survive interference, it suggests a mechanism that is actively feeding such short-term monitoring into durable, long-term representations of interaction partners.

Consider the Haxby, Hoffman, and Gobbini (2001) model of face processing, which proposes two separate streams of information when viewing faces; one through the fusiform gyrus and anterior temporal lobe that appears to encode intransient stable features of the faces such as identity and physical appearance, and a second that projects more dorsally through the superior temporal sulcus (STS) and processes more transient aspects of the face such as expression and gaze direction. That there is some communication between these two streams is evident from our previous research. That is, the specific property

of face identity is associated with particular patterns of gaze behaviour, resulting in changes of face trustworthiness.

Trustworthiness can be considered a stable property of person identity, and hence one might predict it will be stable over time. That is, once a person is encoded as less trustworthy, such information should be available for future encounters with that person. However, note that the learning of the association between face identity and patterns of gaze takes place while the face is irrelevant and ignored while participants undertake a different task. Previous work has shown little awareness of this learning (Bayliss & Tipper, 2006; Rogers et al., 2014), and in an unpublished study identical to the first experiment to be reported here, participants could not explicitly recall whether particular face identities had always looked towards or away from targets. Hence the lack of explicit awareness of the face-gaze relationship might reflect weak and transient memories.

To further test the possibility for stable representations of trust, we also manipulate properties of the above model. As noted, incidental learning during gaze cueing relies not only on attention orienting evoked by eye gaze but also on trial-invariant aspects of the face such as identity recognition. When learning about the face identities from their gaze cueing behaviour, these distributed systems must share information such that untrustworthy behaviour can be linked to the individual expressing it. In previous studies, the faces used have been unfamiliar, and so the identity representations that serve as the anchors for these trust representations are less stable, which may compromise the strength of these associations. More familiar faces have been shown to have more stable neural representations (Eger, Schweinberger, Dolan, & Henson, 2005) and that

these neural representations are related to better behavioural performance in face identification tasks (Weibert & Andrews, 2015). As such, we propose that the association between the face identity and patterns of gaze behaviour incidentally learned while ignoring the face will be facilitated if the face representation is more familiar.

That is, with more familiar faces there is a pre-existing representation of that person's identity, and so when learning new information about them such as consistently misleading gaze, the gaze behaviour can more easily be associated with a specific individual. As noted, this new procedure is in contrast to all previous studies of this effect where gaze is associated with an unknown individual. As such, we predict that increased familiarity of the face stimuli will produce more durable memories.

Finally, a related issue that could influence the stability of memory is how the durability of the effect may be affected by the nature of the information to be remembered. Strachan et al. (2016) showed that a decrease in trust for invalid faces was a more robust result than was an increase in trust for valid faces, and this supports previous research demonstrating memory advantages for cheaters over co-operators (Bell et al., 2012; Buchner, Bell, Mehl, & Musch, 2009). As an example, Bayliss and Tipper (2006) showed that after incidental learning of trust via patterns of eye-gaze, participants subsequently reported that the invalidly cueing low-trust faces had been presented more often. Hence we might expect to observe more stable memory for invalid faces in the current studies.

To briefly preview our findings, in Experiment 1 we replicate the original effect using two sets of scalar ratings immediately before and immediately after five blocks of gaze cueing, and once again find evidence of incidental trust

learning. In Experiment 2, we add an additional task between the gaze cueing and final trustworthiness ratings to see if this can disrupt the effect. We keep this task brief (around five minutes) and show no faces during this period, to ensure that the interference is minimal. Evidence of trust learning is weaker, but cannot be conclusively dismissed following this brief interference.

Having disrupted (but not eliminated) the effect in Experiment 2, Experiment 3 includes a familiarisation procedure to explore whether being more familiar with the face identities at the point of gaze cueing can lead to a clearer and more durable effect. More durable memory was indeed observed with richer identity representations prior to the incidental learning of gaze behaviour.

Finally, we take the identity familiarisation task in Experiment 3 and in Experiment 4 examine whether incidentally learned trustworthiness can survive a break where participants leave the laboratory for an hour. We observe that even with an hour of real-world interference, traces of this learning persist.

Experiment 1

In Experiment 1 we replicate the incidental learning of trust outlined by previous studies (Manssuer et al., 2016; Manssuer et al., 2015; Strachan et al., 2016) by using scalar ratings of trustworthiness at the beginning and end of five blocks of gaze cueing.

Methods

Participants. 32 volunteers from the University of York (25 female, mean age 20.09) participated in this study in exchange for payment or course credit. Due to abnormal response accuracy, one participant was removed from analysis, and one participant was removed for retaining less than 70% of their trials after

RT filters were applied. This left 30 participants in the analysis.

All participants in all experiments described in this study provided written consent and the research was given ethical approval by the Departmental Ethics Committee of the University of York Department of Psychology.

Stimuli. Target stimuli for the object categorisation task were kitchen and garage object images used in Bayliss and Tipper (2006). There were 13 unique objects in each category (kitchen/garage) and these appeared in both left and right orientations. All stimuli were coloured in blue. In total there were 52 individual images used in the experiment. Face stimuli were taken from the Karolinska Directed Emotional Faces (KDEF) stimulus set (Lundqvist, Flykt, & Öhman, 1998), and included sixteen images; eight male and eight female. These faces were initially selected by eye from a figure in the Supplementary Material of Oosterhof & Todorov (2008), in which the faces from this set are plotted along six judgement dimensions. The faces used were all taken from the centre (1SD from the intersection of all six dimensions) of this plot, so the faces used in our experiments were, compared with the rest of the KDEF set, as close to neutral trait judgements as possible.[1]

These faces were split into two groups, which would appear as either 100% valid or 100% invalid cues in the experiment (counterbalanced across participants). The eyes of each face were manipulated using Adobe PhotoShop CS6 to generate faces where the eye gaze was either straight ahead, left, or right. Unaltered images were used for the trustworthiness ratings.

The study was run on an Intel Core i5 PC with a 21.5” monitor. The experiment was presented using E-Prime 2.0 software with a white background

throughout and the resolution set to 1024x768 pixels. Participants were sat approximately 60cm from the display, and during trustworthiness ratings the face stimuli had a visual angle of 19.29˚ horizontally and 20.97˚ vertically, while during gaze-cueing the face stimuli had a visual angle of 13.36˚ horizontally and 14.93˚ vertically.

Design and procedure. Participants were told that they would be asked to perform an object categorisation task on images of objects that appeared on the left or right side of the screen, and to respond with whether these were garage or kitchen objects. They were also told that the central face images were irrelevant and to be ignored. Before the experiment participants were allowed to study printed versions of the kitchen/garage items, in order to familiarise themselves. This was done firstly to ensure that participants knew what each object was, and secondly to ensure that early responses from the first trial block were not confounded by uncertainty as to the object categories of the targets.

Each trial began with a 600ms fixation cross in the centre of the screen, which was then replaced by a face showing a direct gaze for 1,500ms. The face then shifted gaze either to the left or the right for 500ms before the target stimulus appeared on either the same (valid) or opposite (invalid) side of the gaze direction. For each participant, the cueing behaviour of faces was set such that each face would provide a valid or invalid cue 100% of the time. The target stimulus remained either until the participant's response was logged or until 2,500ms had passed, following which participants received feedback from an error tone that would sound if an incorrect response were logged. The face then shifted back to direct gaze for another 1,000ms. A blank screen followed for 500ms before the next trial began. The trial structure is shown in Figure 1.

The object categorization responses were the H key and the space bar of a keyboard, chosen because the H key appears directly above the space bar on QWERTY keyboards and this direction was orthogonal to the possible location of the target. Participants were instructed to respond with their index finger on the H key and thumb on the space bar. For half the participants, H represented kitchen objects, while for the other half it represented garage objects.

In total, there were five blocks of 32 trials each, and each face appeared twice in each block, once gazing left and once right (10 times in total across the experiment; five left, five right, but always either valid or invalid depending on the identity). The order of faces was randomised, as was the order of target objects, the side that the target appeared, and the order of valid and invalid trials.

At the beginning and the end of the experiment, participants were shown all 16 faces (unmanipulated original images) in a random order and asked to rate how trustworthy they found them. A calibration screen would appear with the question, “How TRUSTWORTHY do you think this person is?” with the word ‘START’ written beneath. Participants had to click the word ‘START’ to progress the trial, after which the face would appear for 1,000ms. Following this the face would disappear and a screen with an uninterrupted rating scale appeared. Participants were instructed to click along the scale at the point that corresponded to how trustworthy they thought the person was. The scale recorded responses between -100 and +100, calculated by the distance from the centre of the line of the participants’ mouse click – responses to the left of the centre of the line were coded as negative, while those to the right were coded as

positive (these were indicated on the screen with a − and + sign at either end of the scale). A schematic representation of the procedure is shown in Figure 1.

[Figure 1 approximately here]

Data analysis.

We analysed the data using a linear mixed effects modelling approach. Previous research, such as that by Strachan et al. (2016) has used factorial ANOVAs with participant as a within-subjects identifier. However, there was also a possibility that any effects we found could have been driven by certain stimuli. For example, it could be the case that not all faces produce gaze cueing and/or not all faces are equally well remembered when learning trust. For this reason we decided to run analyses that would be able to control for both subject- and materials-level random effects.

To do this, we used linear mixed effects models, using the *lme4* package in *R.* Initially, we modelled the maximum random structure – that is, we generated a purely random model without any fixed factors, only random factors (subject and stimulus identity). This allows us a clear comparison for other models, as any differences between this model and later ones must be due to the fixed factors and not simply the design of the experiment.

For analysis of RTs, we generated a maximum random structure that controlled for both subject and stimulus identity, and compared this with a model that contained cue validity as a fixed factor, using the *anova* function in *R* (see Baayen, Davidson, & Bates, 2008). This is conceptually similar to running a paired-samples t-test on valid and invalid trials. In this experiment, neither the maximum random structure for RTs nor the validity as fixed factor model for accuracy rates would converge with both subject- and identity-level variations in

validity controlled, which indicated that the model was over-fitting the data. In this case it is customary to remove features of the model until it converges – in this instance, we managed this by removing the validity | subject term from all models.

When analysing trustworthiness judgements in this and all subsequent experiments, this maximum random structure (which we hereafter term the null model), would not converge when all repeated-measures factors (time | subject; validity | subject; time | identity) were included, so we removed the time | identity random factor.

Trustworthiness ratings were compared across time and validity with subject and stimulus identity (identity) as error terms. In addition to the null model, we generated a model with time as a fixed factor (time-only), one with validity as a fixed factor (validity-only), one with both time and validity as fixed factors (time + validity) and one with an interaction modelled between the two factors (time x validity). The time-only model and the validity-only model were each compared with the null model using the *anova* function. The time + validity and time x validity models were also compared to statistically examine the interaction between the two fixed factors.

In order to more closely examine the nature of any interactions, we also ran targeted analyses on the pre- and post-experiment ratings separately. These involved building a null model as above, which modelled the pre- or post-experiment ratings as a function of the random variables, and a validity model, which included validity as a fixed factor. Comparing these models separately for pre- and post-experiment ratings is conceptually similar to performing planned contrasts with a typical ANOVA.

Results

[Figure 2 approximately here]

Gaze cueing. The RT and accuracy results of Experiment 1 are shown in Figure 2 and Table 1, respectively. RTs were faster to valid trials (*M* = 754.84ms, *s.d.* = 93.28ms) than to invalid trials (*M* = 799.90ms, *s.d.* = 114.70ms), and fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs ($\beta = 46.41$, SE = 7.82, $\chi^2(1) = 20.89$, $p < .001$). This improvement was not seen for accuracy scores ($\beta = -0.03$, SE = 0.09, $\chi^2(1) = 0.128$, $p = 0.721$).[2]

[Table 1 approximately here]

Trustworthiness ratings.

[Figure 3 approximately here]

The trustworthiness ratings in Experiment 1 are shown in Figure 3a. We conducted a linear mixed-effects analysis controlling for variance in both subject and the stimulus materials used in the experiment, and compared different models using the *anova* function in R.

Adding time to the null model did not make it fit the data significantly better ($\beta = -2.06$, SE = 1.95, $\chi^2(1) = 1.13$, $p = 0.288$), but the null model did fit significantly better when validity was included ($\beta = -6.77$, SE = 1.78, $\chi^2(1) = 12.96$, $p < .001$). Finally, the interaction model (time x validity) fit the data significantly better than did the full model (time + validity), where both factors were modelled but without an interaction ($\beta = -11.34$, SE = 3.27, $\chi^2(1) = 11.99$, $p < .001$).

Separate analyses comparing the null and validity models for both pre-experiment and post-experiment ratings separately found that adding validity as a fixed factor did not explain variance in the pre-experiment ratings any better

(β = -1.01, SE = 2.08, $\chi^2(1) = 0.24$, p = 0.625), but it did fit the post-experiment ratings significantly better (β = -12.41, SE = 2.92, $\chi^2(1) = 15.73$, p <.001). This indicates that face validity had a clear effect on ratings at the end, but not the beginning of the experiment, as expected.

Discussion

The results of Experiment 1 replicate those of previous studies (Bayliss, Griffiths, & Tipper, 2009; Bayliss & Tipper, 2006; Manssuer et al., 2016; Manssuer et al., 2015; Strachan et al., 2016). Observing somebody's gaze movements automatically triggers a shift in attention to the same location and results in faster identification of objects at that location as compared to objects not gazed at. The association between direction of gaze (valid or invalid) and face identity appears to be encoded: even though the face was irrelevant to the task, individuals who looked away from the target object (invalid cues) were trusted less.

Having replicated the original effect in Experiment 1, we now move on to explore the key question of this study, which is how long this effect can survive a period of interference. In Experiment 2, we introduce a brief distraction task between the final block of gaze cueing and the final trustworthiness ratings.

Experiment 2

Methods

Participants. 30 participants (21 female, mean age 20.63) volunteered for this study.

Stimuli, design and procedure. This experiment was identical to Experiment 1 in every way except that participants performed a number of filler

tasks between the gaze-cueing procedure and the final trustworthiness ratings (see Figure 4). All other details were identical.

[Figure 4 approximately here]

Filler task. The main filler task that participants completed involved looking at 12 pairs of videos. Each video showed a household object in the centre of the screen, then a hand reached out from either the bottom or top of the screen, picked it up, and moved it out of view. Participants watched two such videos (each pair showed matched but distinct objects, i.e. two glasses, two mugs, etc.) and then reported which object they had preferred.

The filler task involved attending to stimuli and making judgements, but included no faces so as not to interfere with any possible face-specific memory processes. The choices that participants made are not of interest to this study and so are not reported; the important detail from this filler task is that participants spent approximately 05:39 minutes completing it.

Data analysis. All details of data analysis are identical to those outlined in Experiment 1, with the exception that in this experiment no participants were removed on the basis of pre-processing filters.

In this experiment, the null model for accuracy scores would not converge with the maximum random structure modelled, so we removed the validity | subject random term. The RT models converged with the maximum random structure with no need to remove terms.

Also in this experiment, the validity-only model of trustworthiness ratings would not converge with the defined random structure, and so we removed the time | subject slope from all of the models to help direct comparison.

Results

Gaze cueing. The RT and accuracy results of Experiment 2 are shown in Figure 2 and Table 1, respectively. RTs were faster to valid trials (*M* = 837.49ms, *s.d.* = 159.39ms) than to invalid trials (*M* = 863.72ms, *s.d.* = 155.01ms). Adding validity to the null model of RTs significantly improved the fit (β = 25.49, SE = 9.36, $\chi^2(1) = 7.41$, p = 0.006), but the same was not the case for accuracy rates (β = -0.12, SE = 0.70, $\chi^2(1) = 0.03$, p = 0.864).

Trustworthiness ratings. The changes in trustworthiness ratings for the faces in Experiment 2 are shown in Figure 3b. In this experiment, there were small changes in the expected directions for both invalid and valid faces, but these tended to be smaller in magnitude than those in Experiment 1.

We conducted a linear mixed-effects analysis controlling for variance in both subject and the stimulus materials used in the experiment. Adding time to the null model significantly improved the fit (β = -3.66, SE = 1.68, $\chi^2(1) = 4.74$, p = 0.029), as did including validity (β = -5.88, SE = 1.69, $\chi^2(1) = 10.29$, p = 0.001). Finally, the interaction model (time x validity) fit the data better than did the full model (time + validity) and this approached significance (β = -6.12, SE = 3.35, $\chi^2(1) = 3.35$, p = 0.067).

Separate analyses comparing the null and validity models for both pre-experiment and post-experiment ratings separately found that adding validity as a fixed factor did not explain variance in the pre-experiment ratings any better (β = -2.83, SE = 2.39, $\chi^2(1) = 1.40$, p = 0.237), but it did fit the post-experiment ratings significantly better (β = -8.94, SE = 4.45, $\chi^2(1) = 3.96$, p = 0.047). This indicates that face validity had an effect on ratings at the end, but not the beginning of the experiment, as expected.

Discussion

Experiment 2 aimed to explore whether a brief period of interference could disrupt the pattern of trust learning observed in Experiment 1. While the effect was more fragile, the a priori predicted pattern of trust changes was observed.

There are trends for susceptibility to decay when an interference task is introduced. One question, then, is what is the cause of this susceptibility to decay. We propose that familiarity with the faces used as stimuli may be the deciding factor – in Experiment 2, the only exposure participants have to the faces is the initial pre-experiment ratings, which provides only a superficial opportunity to encode these identities. This means that when participants experience helpful or misleading gaze cues, they have to build representations of these identities from the bottom up. It may be that this effect is more durable when cueing validity information is added to a pre-existing representation of these identities – that is, when the faces are more familiar.

In Experiment 3, we explore this question by including an additional task at the beginning of the experiment designed to increase participants' familiarity with the faces – we use the same faces as in Experiments 1 and 2 to avoid any confounds of different stimuli, and to avoid any pre-existing expectations of trustworthiness that might arise with using naturally familiar stimuli such as famous faces.

Experiment 3

Experiment 3 replicates Experiment 2 but adds an additional familiarisation task to the beginning of the procedure to trigger greater familiarity with the face stimuli used. That is, it employs procedures developed by Andrews, Jenkins, Cursiter, and Burton (2015) where faces are viewed from different angles and

express different emotions in a same-different face matching task. Such encoding has been shown to significantly improve face recognition performance.

Methods

Participants. 32 participants (21 female, mean age 20.17) volunteered for this study in return for a mixture of course credit and payment. Due to runtime errors, 2 participants' data were not collected, and so this left 30 participants for inclusion.

[Figure 5 approximately here]

Stimuli, design and procedure. This experiment was identical to Experiment 2 in every way except that participants also performed a face-matching task at the beginning of the experiment in order to allow for deeper encoding of the KDEF faces. Participants were shown images of all sixteen identities that varied in their head orientation (full-left, half-left, half-right or full-right) and emotion (happy, angry, disgusted, surprise, afraid or sad) – these were unaltered images from the KDEF stimulus set (Lundqvist et al., 1998) and so were presented with an off-white/brown background, rather than the plain white background that was used for the images in the gaze-cueing and trust-rating portions. Participants made same/different judgements of the identities of face pairs, responding with a button press of S if the two images showed the same person, and D if they showed different people. Image pairs showed the same identity on 25% of trials, and a written feedback screen appeared after each trial reporting either “Correct”, “Incorrect” or “No response detected”, depending on what response was logged (see Figure 5). During the course of a trial a fixation cross was presented for 500ms, followed by the two images either side of a fixation for 1,500ms, and finally the feedback screen for 1,000ms.

It was expected that the variability in these images and the nature of the identity judgement task would prompt participants to encode viewpoint- and emotion-independent identity representations of the individuals. Such a task has been used before to good effect (Andrews et al.), as participants reconcile within-identity variability in face photographs to develop a richer representation of the individual.

Data analysis. All analyses were the same as in Experiment 2. In this experiment, the RT null model would not converge until both the validity | subject and validity | identity error terms were removed. Both models of accuracy rates converged with the full maximum random structure with no need to remove any random terms.

For trustworthiness ratings, the validity-only, time + validity, and time x validity models would not converge with the defined random structure, and so we removed both the time | subject and time | identity slopes from all models.

Results

Gaze cueing. The RT and accuracy results of Experiment 3 are shown in Figure 2 and Table 1. RTs were faster to valid trials (M = 781.27ms, *s.d.* = 126.00ms) than to invalid trials (M = 824.27ms, *s.d.* = 116.97ms). Adding validity to the null model of RTs significantly improved the fit ($\beta = 45.26$, SE = 8.44, $\chi^2(1) = 28.61$, $p < .001$), but the same was not the case for accuracy rates ($\beta = 0.45$, SE = 0.72, $\chi^2(1) = 0.40$, $p = 0.530$).

Trustworthiness ratings. The results of Experiment 3 are shown in Figure 3c. We conducted a linear mixed-effects analysis controlling for variance in both subject and stimulus identity. Adding time to the null model did not make it fit the data significantly better ($\beta = 1.85$, SE = 1.72, $\chi^2(1) = 1.17$, $p = 0.280$), but

it did fit significantly better when validity was included (β = -6.97, SE = 1.70, $\chi^2(1) = 16.42$, p <.001). Finally, the interaction model fit the data significantly better than did the full model, where both factors were modelled but without an interaction (β = -13.88, SE = 3.38, $\chi^2(1) = 16.78$, p <.001).

Separate analyses comparing the null and validity models for both pre-experiment and post-experiment ratings separately found that adding validity as a fixed factor did not explain variance in the pre-experiment ratings any better (β = -0.03, SE = 2.09, $\chi^2(1) = 0.00$, p = 0.990), but it did fit the post-experiment ratings significantly better(β = -13.91, SE = 2.65, $\chi^2(1) = 27.93$, p <.001). This indicates that face validity had a clear effect on ratings at the end, but not the beginning of the experiment, as expected.

Discussion

Experiment 3 explored whether initial familiarity with the face identities used in the gaze cueing led to stronger learning of trust, and this appears to be the case given the clear pattern of changes in trustworthiness found.

These results suggest that, while with unfamiliar faces this effect appears to be somewhat susceptible to interference, increased familiarity with the faces can make this incidental learning more resilient to decay over a short period of interference. This of course raises the logical question: if this effect can now survive a brief (around 5 minute) period of interference, could it survive longer periods? To explore this, in Experiment 4 we replace the 5-minute filler interference task with an hour-long break during which participants were sent away from the laboratory. This provides a naturalistic interference, as participants were given no instructions about what to do during that time, and so

means that if we still see evidence of the effect after this time that this gaze cueing manipulation can lead to particularly durable changes in trustworthiness.

Experiment 4

Methods

Participants. 30 participants (24 female, mean age 19.93) volunteered for this study. All participants returned after one hour at the agreed upon time and so all were included in the analyses.

Stimuli, design and procedure. This experiment was identical to Experiment 3, as it included a face familiarisation task at the beginning, followed by an initial trustworthiness rating and then five blocks of gaze cueing. After this, however, participants were sent away for an hour and instructed to return at a certain time, at which point they would complete the second trustworthiness ratings and receive their compensation.

Data analysis. All data analysis procedures were the same as in previous experiments.

In this experiment, both models of RTs converged with the full maximum random structure with no need to remove any random terms. However, the accuracy rates null model would not converge until both the validity | subject and validity | identity error terms were removed.

For trustworthiness ratings, the validity-only, time + validity, and time x validity models would not converge with the defined random structure, and so we removed the time | identity slope from all models.

Results

Gaze cueing. The RT and accuracy results of Experiment 4 are shown in Figure 2 and Table 1. RTs were faster to valid trials (M = 807.84ms, *s.d.* =

134.22ms) than to invalid trials (*M* = 834.97ms, *s.d.* = 133.42ms). Adding validity to the null model of RTs significantly improved the fit (β = 27.74, SE = 8.63, $\chi^2(1) = 8.63$, p = 0.003), but the same was not the case for accuracy rates (β = 0.32, SE = 0.74, $\chi^2(1) = 0.19$, p = 0.665).

Trustworthiness ratings. The changes in trustworthiness ratings for the faces in Experiment 4 are shown in Figure 3d. Adding time to the null model did not make it fit the data significantly better (β = -1.59, SE = 1.63, $\chi^2(1) = 0.95$, p = 0.329), nor did it predict the data significantly better when validity was included (β = -2.86, SE = 1.87, $\chi^2(1) = 2.34$, p = 0.126). However, the interaction model fit the data significantly better than did the full model, where both factors were modelled but without an interaction (β = -6.77, SE = 3.23, $\chi^2(1) = 4.40$, p = 0.036).

Separate analyses comparing the null and validity models for both pre-experiment and post-experiment ratings separately found that adding validity as a fixed factor did not explain variance in the pre-experiment ratings any better (β = 0.52, SE = 2.58, $\chi^2(1) = 0.04$, p = 0.839), but it did fit the post-experiment ratings significantly better (β = -6.25, SE = 2.71, $\chi^2(1) = 5.21$, p = 0.022). This indicates that face validity had a clear effect on ratings at the end, but not the beginning of the experiment, as expected.

Discussion

Experiment 4 explored if the effect described in Experiment 3 could survive an hour of naturalistic interference (that is, exposure to non-laboratory conditions). Indeed trust associated with particular face identities is relatively stable, surviving an hour of real-world interference. This is the first evidence that gaze cueing behaviour can lead to lasting changes in judgements of trust, and points to a mechanism for monitoring and storing social information about potential

interaction partners that can be accessed and used long after exposure to trait-diagnostic information has ended (in this case, valid or invalid gaze cues).

Cross-Experiment Analysis

[Figure 6 approximately here]

In order to see how an interference task and familiarity level impacted the overall effect, the results of Experiments 1, 2, 3, and 4 were combined. We then conducted a linear mixed-effects analysis controlling for variance in both subject and the stimulus materials used. In this analysis, the null model included the maximum random structure and no features had to be removed to allow for convergence. The outcome variable was now *change* in trustworthiness over the course of the experiments (as such, time was no longer a factor and was now a property of the measured variable). Fixed factors were validity (valid/invalid) and experiment (1-4). The null model was compared to validity-only and experiment-only models separately, and then an interaction model (validity x experiment) was compared with a non-interaction model (validity + experiment).

Adding validity to the null model significantly improved the fit ($\beta = 9.68$, $SE = 2.35$, $\chi^2(1) = 16.06$, $p < .001$), but including experiment did not ($\beta = 0.71$, $SE = 0.95$, $\chi^2(1) = 0.56$, $p = 0.456$). The interaction model of validity x experiment did not fit the data better than did the model where both factors were modelled but without an interaction ($\beta = -0.78$, $SE = 2.11$, $\chi^2(1) = 0.14$, $p = 0.708$), indicating that incidental learning of trust was largely similar across experiments.

General Discussion

Even when ignoring a face its pattern of eye-gaze behaviour can be learned, subsequently influencing ratings of the face's trustworthiness: the key question explored here concerned the stability of this incidental associative learning process. After replicating the basic effect in Experiment 1, we showed in Experiment 2 that with minimal interference we could reduce the effect, although we did not completely eradicate the pattern of trust learning.

We go on to show in Experiment 3 that by including a familiarisation task at the beginning of the experiment, the effect can now convincingly survive the same interference that weakens it in Experiment 2. And finally in Experiment 4 we show that traces of this trust learning can persist an hour after gaze cueing has ended. The change scores of initial to final trust ratings from each experiment are shown together in Figure 6.

This is the first study that investigates how long this incidental learning can last, and we show that it can be durable and somewhat resilient to interference. While Experiments 2 and 4 tend to show weaker learning effects than we see in Experiments 1 and 3, the overall profile of results persists, and we are confident that although these interference tasks do appear to weaken the effect, it nonetheless survives. This is supported by the fact that a cross-experiment analysis found that modelling an interaction of validity and experiment when predicting changes in trustworthiness to faces did not fit the data significantly better than modelling no interaction.

Of course there could be other interfering tasks that are more disruptive of incidentally learned trustworthiness. For example, re-presenting the faces used in the experiment in a task where no gaze cueing was observed is likely to

cause extinction of prior learning (c.f. Rogers et al., 2014). In contrast, it is likely that exposure to faces not presented in the experiment might not disrupt prior learning. For example, during the one-hour interference task where participants left the laboratory (Experiment 4) they were exposed to other faces as they moved around the campus, and the effect survived. Clearly more formal studies of the stability of incidentally learned trust will be worthwhile.

We initially considered two possibilities: that this incidental trust learning might be short-lived, and reflect in-the-moment monitoring of statistical contingencies, or that these traces might be integrated into a longer and more durable representation. We find evidence supporting the latter interpretation, and taken with previous research using a similar paradigm we begin to see a complex underlying mechanism for social learning emerge. This suggests that at some point during gaze cueing (likely around the fifth presentation of the face identity, c.f. Manssuer et al., 2015) information about the cueing behaviour of faces is transferred to a longer term and more durable representation that feeds into a network that focuses on intransient aspects of identity recognition.

It is important to note that, on the basis of the current data, we cannot make any claims that this trust learning is *implicit*, in that it occurs completely outside of conscious awareness, only that it is *incidental*, in that participants learn about a feature they are instructed to ignore. However, the possibility of this learning being implicit – or at least not consciously driven and maintained – seems plausible given reaction time results from previous studies using gaze cueing that used the same design as the current study (Manssuer et al., 2016; Manssuer et al., 2015; Strachan et al., 2016). That is, faces are always either valid or invalid, and so if trust learning reflected the degree to which a participant was

aware of the manipulation, one might expect this to be reflected in a reduced cueing cost later in the experiment compared with the beginning. However, no evidence has emerged to support this recovery (including the current study, see footnote 2). Furthermore, although no explicit measures of awareness of eye-gaze behaviour is reported here, there is some evidence from unpublished data (Strachan & Tipper), which used a task that matched the procedure of Experiment 1, where participants failed to explicitly recall previous eye-gaze behaviour.

Hence our work contrasts with a number of previous approaches where trust is manipulated in more explicit ways. First, previous research has shown impressions of trust are based on the physiognomy of the face. These physical cues in the face can be manipulated where people's judgements can be altered by changing the physical features of even familiar face identities (Keating, Randall, & Kendrick, 1999) and can be reliably simulated using image-based analysis (Vernon, Sutherland, Young, & Hartley, 2014). This trust based on the physical properties of a face contrasts with the associative learning of gaze behaviour we study. For example, trust decisions based on the physical properties of the face are made extremely rapidly, with people making reliable decisions after only 100ms exposure (Willis & Todorov, 2006), and ERP measures showing sensitivity to trustworthiness similarly early (Marzi, Righi, Ottonello, Cincotta, & Viggiano, 2012). In contrast, Manssuer et al. (2015) showed that EEG measures detected trust after 1000ms of face viewing in the incidental learning tasks described here where there are no manipulations of the physical properties of faces.

Second, other approaches show that gaze-cueing is susceptible to top-down control, such as manipulating beliefs about what a face can see (Teufel, Alexis, Clayton, & Davis, 2010) or about the agency of the face (Wiese, Wykowska, Zwickel, & Muller, 2012). That is, such knowledge about a face significantly influences the patterns of gaze cueing effects. In contrast, we have observed repeatedly that even though trustworthiness of a face changes through gaze-target reliability, the actual cueing of attention by the eye-gaze is unaffected, likely due to the fact that learning and cueing occur simultaneously.

And third, Falvello, Vinson, Ferrari, and Todorov (2015) had participants explicitly associate short descriptions of trustworthy and non-trustworthy behaviour with different faces. They found this explicit association of trust behaviour with individual faces showed an impressive capacity – participants could reliably learn over 400 associations of faces with behavioural vignettes in a single session. Thus far we have not examined the numbers of individual faces that can be associated with patterns of eye-gaze while the faces are irrelevant and to-be-ignored, but we suspect our implicit procedures would not have such large capacity. Furthermore, Falvello et al. (2015) also found similar learning of place images that were learnt in the same way as faces, suggesting that this learning recruited a global affective mechanism. On the other hand, when Manssuer et al. (2016) used a similar cueing paradigm to the current study to measure learning, they found that participants did not learn about non-face images (arrows) in the same way as faces.

One final thing to note about the results of all four experiments is that, while the change scores for valid faces are somewhat varied across experiments, the same is not true for invalid faces – the reduction in trust for invalid faces is

consistent and somewhat stable across all four experiments (see Figure 6). This lends support to our hypothesis that there may be differences in how traces of learning survive between valid and invalid faces, as it appears that memory for invalid faces is more stable and appears to survive the interference that reduces the signal in Experiments 2 and 4.

This finding supports previous literature that shows memory advantages for cheaters or deceivers over co-operators (Bell et al., 2012; Buchner et al., 2009; Suzuki & Suga, 2010). As people's default expectation of others is for them to co-operate rather than deceive, invalid gaze cues provide a clear error signal that results in a stronger and lasting memory for the interaction partners involved.

In conclusion, we have reported the results of four experiments that examine for the first time the question of how durable cueing-induced changes in trustworthiness are, and we show that although the effect tends to deteriorate over time, it is still surprisingly resilient and traces of it do survive, particularly in the form of decreased trust towards invalid faces. With more familiar faces these effects can be seen up to an hour after cueing exposure has ended. Taken together these results point to a mechanism for building robust, lasting representations of the identities of deceptive or unhelpful interaction partners, even when not explicitly attending to them while focused on a different task.

Footnotes

1. This assumption of matched ratings has been corroborated by previous work (Strachan et al., 2016), which found that the two groups into which face images were split did not differ significantly in terms of trust ratings at the beginning of the experiment, as judged by 100 people across 5 experiments: Group A, M = -2.71, *s.d.* = 13.04; Group B, M = 0.85, *s.d.* = 6.82; $t(14)$ = - 0.68, p = 0.506.
2. We are mindful of the fact that, as faces were always either valid or invalid within a gaze-cueing procedure, each face then serves as a reliable cue of the subsequent location (e.g. by the end of the experiment, if an invalid face looks left then the participant could arguably learn and anticipate this to know that the target will then appear on the right). This is not a primary concern of this paper, but in order to address this potential concern, across all experiments in this paper we examined the gaze-cueing results broken down across the five experimental blocks (see also Mannsuer et al, 2015; 2016; Strachan et al, 2016). We report the results of a 2 (validity: valid/invalid) x 5 (block: 1-5) repeated measures ANOVA looking at the RTs, which found a main effect of validity ($F(1,119)$ = 81.43, p <.001, η_G^2 = 0.01) and a main effect of block (Greenhouse Geisser corrected: $F(2.91,346.11)$ = 154.32, p <.001, η_G^2 = 0.20), but no interaction between the two (Greenhouse Geisser corrected: $F(3.42,407.41)$ = 1.88, p = 0.123, η_G^2 = 0.00).

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly journal of experimental psychology (2006)*, 1-19. doi:10.1080/17470218.2014.1003949

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412. doi:10.1016/j.jml.2007.12.005

Bayliss, A. P., di Pellegrino, G., & Tipper, S. P. (2004). Orienting of attention via observed eye gaze is head-centred. *Cognition, 94*, B1-10. doi:10.1016/j.cognition.2004.05.002

Bayliss, A. P., Frischen, A., Fenske, M. J., & Tipper, S. P. (2007). Affective evaluations of objects are influenced by observed gaze direction and emotional expression. *Cognition, 104*, 644-653. doi:10.1016/j.cognition.2006.07.012

Bayliss, A. P., Griffiths, D., & Tipper, S. P. (2009). Predictive gaze cues affect face evaluations: The effect of facial emotion. *Eur J Cogn Psychol, 21*, 1072-1084. doi:10.1080/09541440802553490

Bayliss, A. P., Paul, M. A., Cannon, P. R., & Tipper, S. P. (2006). Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic Bulletin and Review, 13*, 1061-1066.

Bayliss, A. P., & Tipper, S. P. (2005). Gaze and arrow cueing of attention reveals individual differences along the autism spectrum as a function of target context. *Br J Psychol, 96*, 95-114. doi:10.1348/000712604X15626

Bayliss, A. P., & Tipper, S. P. (2006). Predictive gaze cues and personality judgments: Should eye trust you? *Psychol Sci, 17*, 514-520. doi:10.1111/j.1467-9280.2006.01737.x

Bell, R., Buchner, A., Erdfelder, E., Giang, T., Schain, C., & Riether, N. (2012). How specific is source memory for faces of cheaters? Evidence for categorical emotional tagging. *J Exp Psychol Learn Mem Cogn, 38*, 457-472. doi:10.1037/a0026017

Buchner, A., Bell, R., Mehl, B., & Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evolution and Human Behavior, 30*, 212-224. doi:10.1016/j.evolhumbehav.2009.01.004

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze Perception Triggers Reflexive Visuospatial Orienting. *Visual Cognition, 6*, 509-540.

Droulers, O., & Adil, S. (2015). Perceived gaze direction modulates ad memorization. *Journal of Neuroscience, Psychology and Economics, 8*, 15-26.

Eger, E., Schweinberger, S., Dolan, R., & Henson, R. (2005). Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *Neuroimage, 26*(4), 1128-1139.

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The Robustness of Learning about the Trustworthiness of Other People.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review, 5*, 490-495.

Frischen, A., & Tipper, S. P. (2006). Long-term gaze cueing effects: Evidence for retrieval of prior states of attention from memory. *Visual Cognition, 14*, 351-364. doi:10.1080/13506280544000192

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2001). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223-233.

Keating, C. F., Randall, D., & Kendrick, T. (1999). Presidential Physiognomies: Altered Images, Altered Perceptions. *Political Psychology, 20*, 593-610. doi:10.1111/0162-895X.00158

Langton, S. R. H., & Bruce, V. (1999). Reflexive Visual Orienting in Response to the Social Attention of Others. *Visual Cognition, 6*, 541-567. doi:10.1080/135062899394939

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces. *Stockholm, Sweden: Karolinska Institute*.

Manssuer, L. R., Pawling, R., Hayes, A. E., & Tipper, S. P. (2016). The role of emotion in learning trustworthiness from eye-gaze: Evidence from facial electromyography. *Cognitive Neuroscience*.

Manssuer, L. R., Roberts, M. V., & Tipper, S. P. (2015). The late positive potential indexes a role for emotion during learning of trust from eye-gaze cues. *Social Neuroscience*.

Marzi, T., Righi, S., Ottonello, S., Cincotta, M., & Viggiano, M. P. (2012). Trust at first sight: evidence from ERPs. *Soc Cogn Affect Neurosci*. doi:10.1093/scan/nss102

Rogers, R. D., Bayliss, A. P., Szepietowska, A., Dale, L., Reeder, L., Pizzamiglio, G., . . . Tipper, S. P. (2014). I Want to Help You, But I Am Not Sure Why: Gaze-

Cuing Induces Altruistic Giving. *Journal of Experimental Psychology: General*. doi:10.1037/a0033677

Strachan, J. W. A., Kirkham, A. J., Manssuer, L. R., & Tipper, S. P. (2016). Incidental learning of trust: Examining the role of emotion and visuomotor fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Suzuki, A., & Suga, S. (2010). Enhanced memory for the wolf in sheep's clothing: facial trustworthiness modulates face-trait associative memory. *Cognition, 117*, 224-229. doi:10.1016/j.cognition.2010.08.004

Teufel, C., Alexis, D. M., Clayton, N. S., & Davis, G. (2010). Mental-state attribution drives rapid, reflexive gaze following. *Attention, perception & psychophysics, 72*, 695-705. doi:10.3758/APP.72.3.695

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E3353-3361. doi:10.1073/pnas.1409860111

Weibert, K., & Andrews, T. J. (2015). Activity in the right fusiform face area predicts the behavioural advantage for the perception of familiar faces. *Neuropsychologia, 75*, 588-596.

Wiese, E., Wykowska, A., Zwickel, J., & Muller, H. J. (2012). I see what you mean: how attentional selection is shaped by ascribing intentions to others. *PLoS One, 7*, e45391. doi:10.1371/journal.pone.0045391

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol Sci, 17*, 592-598. doi:10.1111/j.1467-9280.2006.01750.x

Table 1. Mean accuracy rates (% correct) in response to valid and invalid trials across five blocks in Experiments 1-4(*with standard deviations*).

		Experiment			
Validity		**1**	**2**	**3**	**4**
Valid	Mean	91.92	91.46	91.38	90.54
	s.d.	*5.75*	*7.35*	*4.85*	*6.08*
Invalid	Mean	91.75	91.67	91.29	90.79
	s.d.	*5.90*	*6.73*	*5.22*	*5.75*

Figure captions

Figure 1. The basic paradigm used in this experiment. Participants first completed trustworthiness ratings of all faces, then five blocks of gaze cueing where each face would either appear as an invalid (male face, top row) or valid face (female face, bottom row). They then completed an additional trustworthiness rating at the end.

Figure 2. Bar chart showing reaction times in response to valid (white bars) and invalid trials (grey bars) in Experiment 1, with no interference or initial familiarisation; Experiment 2, with brief interference but no initial familiarisation; Experiment 3, with brief interference and initial familiarisation; and Experiment 4 with an hour-long break and initial familiarisation. Error bars show standard error. $^{**}p<.01$ $^{***}p<.001$.

Figure 3. Line graphs showing trustworthiness ratings as a function of time (First Rating vs. Second Rating) of valid (dotted lines) and invalid faces (solid lines) in (a) Experiment 1, with no interference or initial familiarisation; (b) Experiment 2, with brief interference but no initial familiarisation; (c) Experiment 3, with brief interference and initial familiarisation; and (d) Experiment 4, with an hour-long break and initial familiarisation. Error bars show standard error.

Figure 4. Schematic of the paradigm of Experiment 2, with the addition of the interference paradigm between the gaze-cueing and final trustworthiness ratings. This same interference task was used in Experiment 3.

Figure 5. Schematic of the familiarisation task participants completed at the very beginning of Experiments 3 – they were shown two images of faces and asked to judge if they were the same or different identities. The paradigm was the same for Experiment 4 except that the 2AFC object preference task introduced in

Experiment 2 was replaced with an hour away from the lab. Feedback was provided for incorrect responses.

Figure 6. Bar chart showing change in trustworthiness (calculated by subtracting first ratings from second ratings) for valid (white bars) and invalid faces (grey bars) in Experiment 1, with no interference or initial familiarisation; Experiment 2, with brief interference but no initial familiarisation; Experiment 3, with brief interference and initial familiarisation; and Experiment 4 with an hour-long break and initial familiarisation. Error bars show standard error.