OPEN MIND

Discoveries in Cognitive Science

an open access journal

Check for updates

Citation: Guha, R., Arias-Sarah, P., & Aucouturier, J.-J. (2026). The Social Transfer Function: How Dynamic Predictions of Facial Consequences Drive Judgements of Social Contingency. *Open Mind: Discoveries in Cognitive Science.* Advance publication. https://doi.org/10.1162/OPMI.a.400

DOI:
https://doi.org/10.1162/OPMI.a.400

Supplemental Materials:
https://doi.org/10.1162/OPMI.a.400

Received: 3 June 2025
Accepted: 28 September 2026

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Rudradeep Guha
r.guha@outlook.com

Copyright: © 2026 Rudradeep Guha, Pablo Arias-Sarah, and Jean-Julien Aucouturier. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

REPORT

The Social Transfer Function: How Dynamic Predictions of Facial Consequences Drive Judgements of Social Contingency

Rudradeep Guha[1], **Pablo Arias-Sarah**[2,3,4], **and Jean-Julien Aucouturier**[1]

[1]Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000 Besançon, France
[2]School of Psychology and Neuroscience, University of Glasgow, UK
[3]Department of Cognitive Science, Lund University, Sweden
[4]Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LPNC, Grenoble, France

Keywords: social contingency, temporal response functions, backchanneling, social cognition

Uncorrected Proof

ABSTRACT

Human social interactions display a complex choreography of dynamic, non-verbal signals such as nods and blinks. Yet, little is known about how these cues contribute to the detection of social contingency, i.e., how exactly does one know that two people are interacting with one another? To answer this question, we developed an experimental paradigm where observers had to recognize genuine and fake contingency in naturalistic, dyadic interactions. With two independent experiments (N = 206), we show that perception of contingency is robust even with severely degraded interactional stimuli and that observers causally rely on the link between a speaker's speech and the listener's motion. Crucially, inference about appropriate contingency is driven by temporally aligned, dynamic predictions rather than average quantity of movement or any lagged correlation between speech and motion. To model this perceptual process, we introduce the concept of a 'social transfer function': a pre-learned schema that generates predictions of a listener's speech-contingent response and, based on how well-matched the predictions are to observed behaviour, is able to predict observer judgments of genuine contingency. Taken together, these results provide mechanistic insights into the audiovisual features that contribute to perception of social contingency, and open new ways to identify markers of contingency in people with disorders of consciousness, autism and social anxiety.

INTRODUCTION

Q1

Human social interactions are complex, dynamic processes that require quick and efficient coordination in real-time. They require individuals to detect multimodal signals, integrate them with high-level cognitive inferences and plan and produce appropriate responses within milliseconds. Thus, the ability to detect contingent behaviour (i.e., recognizing that one signal is the social consequence of another) is fundamental to social interaction (Coey et al., 2012; Dale et al., 2013; Hermans et al., 2022). This ability is thought to develop early in infancy, primarily through caregiver-child interactions, where the child engages in goal-directed behaviour to draw attention or elicit contingent behaviour from caregivers (Brazelton et al., 1975; Goldberg, 1977; Rochat, 2001). In the influential "television" paradigm in

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

developmental psychology (Murray, 1985), 2-month old infants display distress when interaction with their mothers is switched from a real-time video interaction to a pre-recorded video. This suggests that infants possess the ability to process expressive signals from an interlocutor and how they should depend on their own signals. In adulthood, recognition of social contingency is often considered a prerequisite to high-level social abilities such as joint attention (Mundy & Newell, 2007), turn-taking in conversation and theory of mind (Frith & Frith, 2012).

In spoken conversations, contingent social behaviour often manifests itself through the phenomenon of backchanneling (Brunner, 1979; Knudsen et al., 2020). Backchanneling cues[1] encompass a wide range of physical behaviours like nods, blinks and smiles, or non-lexical utterances like hmm and uh-huh which are signifiers of engagement or acknowledgement that can be deployed quickly and efficiently in conversations (Holler, 2022). Most research into backchannelling signals has either looked at them in the context of emotion expression, or at the variety of semantic and pragmatic functions they play in conversations (Bavelas & Chovil, 2018). For instance, head nods display multiple meanings such as agreement, acknowledgement, greetings and emphasis (Poggi et al., 2010) while the temporal structure of blinks is important with short and long blinks signalling end-of-turn and understanding respectively (Hömke et al., 2017). Studies have also shown that perception of social contingency is influenced by more low-level processes, such as the detection of biological motion (Scaliti et al., 2023; Schultz & Frith, 2022). This ties in neatly with developmental work on social contingency, with a seminal study showing predisposition towards biological motion even in 2-day-old infants (Simion et al., 2008). Preferential attention to biological motion not only develops early, but is also found in many different species (Blake, 1993; Larsch & Baier, 2018; Regolin et al., 2000). Neurophysiological and neuroimaging studies showing overlapping brain regions for processing biological motion and visual social cognition (Pavlova, 2012; Pelphrey et al., 2005) suggest that the ability to detect biological motion and preferential orientation towards it supports more high-level social abilities. This is further supported by findings that preference for biological motion is disrupted in individuals with disorders associated with social disfluencies such as autism spectrum disorder (Klin et al., 2009; Todorova et al., 2019). Moreover, recognition of biological motion is preserved even when stimuli are masked or stripped down to minimal representations of motion (Neri et al., 1998), when individuals' motion perception is impaired (Jordan et al., 2002), and even when biological motion is represented in a non-visual modality (Bidet-Caulet et al., 2005). Taken together, these studies show that mechanisms underlying recognition of biological motion are evolutionarily well-preserved, develop early, are robust, provide the foundation for recognition of cues and signals like facial expressions, and thus are central to the perception of social contingency.

To provide mechanistic insight into the perception of social contingency, a common strategy has been to degrade stimuli. A popular method is the use of point-light displays, i.e., minimal representations of body movement through a set of points representing the joints, which allow isolating the stimulus properties that facilitate detection of interaction. For instance, a study found that synchronous movement of noisy point-light displays of two agents

[1] In this study, we use the term 'backchanneling cues' in a low-level perceptual sense, as the ensemble of visual cues from a listener's behaviour while a speaker speaks, and that are available for a third-party observer to process. We do not primarily make a distinction between such cues based on their function (e.g., linguistic or emotional), semantics or underlying generative processes (e.g., voluntary or not). However, see Discussion sections below for how our results may depend on such aspects.

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

enhanced visual detection of the target agent compared to when movement was asynchronous (Neri et al., 2006). In another study, point-light displays of two musicians either improvising together or playing solo were spliced together to create genuine and fake musical interactions (Moran et al., 2015), with observers able to recognize genuine interactions even in the absence of music, or musical expertise (for the related question of recognizing biological motion in a single body, see also Nackaerts et al., 2012). Such studies paint a picture of the general ability for 'interpersonal predictive coding', by which observers use the actions of one agent to predict both the content and temporality of a second agent's actions (Manera et al., 2011). However, while point-light stimuli allow quantifying the spatial coordination between interacting bodies and how it correlates with observer decisions, they do not easily translate to the vocal and facial features observed in real-world conversations (Takarae et al., 2021). More importantly, although these studies illustrate a predictive route to social contingency perception by showing its sensitivity to, e.g., time shifts or individual differences (Manera et al., 2011, 2013), they do not operationalize this mechanism with a concrete model that can be used to make experimental predictions. Simply put, what exact temporal prediction, of which specific expressive signal, has to break down before I—the observer of an interaction—decide that it isn't genuine?

To address this question, we introduce a computational modeling paradigm, the 'social transfer function', which assumes that observers possess a schema (or dynamic representation in the sense of Freyd, 1987) of contingent interactions acquired over time by observation and participation, and which can generate real-time predictions of the temporal dynamics of an agent's response to the speech of another agent (Figure 1). At the algorithmic level of explanation (Marr, 2010), we instantiate such a transfer function using temporal response functions

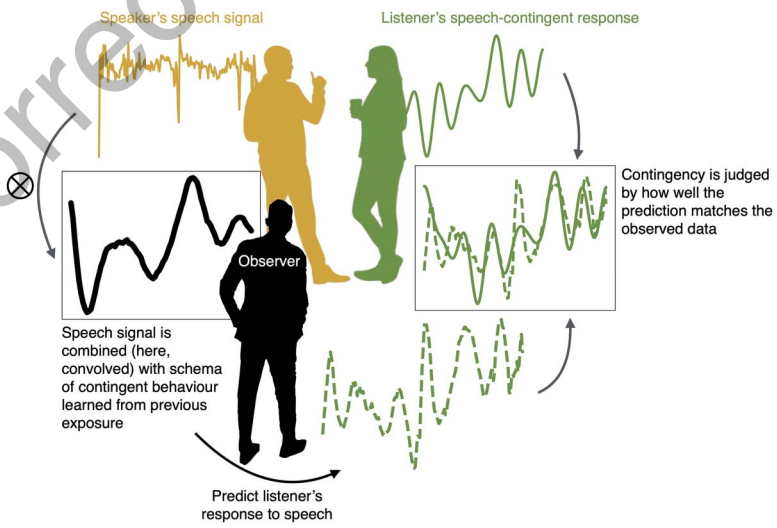


Figure 1. The "social transfer function" computational modeling paradigm. To operationalize how external observers judge the contingency of a social interaction, we propose that observers possess a schema of contingent interactions acquired over time by observation and participation. In this paper, we model such a 'transfer function' using a temporal response function (TRF), i.e., a pre-trained impulse response that is convolved with an agent's speech signal such that another agent's response to the speech signal in terms of the temporal dynamics of their backchanneling cues can be matched against the transfer function's real-time predictions to quantify how contingent the interaction appears to be. In the following, we investigate whether social transfer functions learned from a dataset of speed-dating interactions can predict observer performance, as well as what specific facial features or regions they use to detect genuine and fake interactions.

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

(TRFs; Crosse et al., 2016) which assume[2] that the system can be represented by an impulse response H that is convolved with the input to generate the output ($Y = H \circledast X$). In other words, we propose that the observer of an interaction between A and B uses something akin to a pre-trained TRF to generate the likely output of B as a response to A (in our algorithmic specification, $H \circledast A$), and that this predicted output is then matched against the observed signal to quantify how contingent the interaction appears to be.

To test this mechanism, we collected a corpus of video recordings of naturalistic speed-dating interactions (Arias-Sarah et al., 2024), and extracted segments from the videos that were 'one-sided' (i.e., where only one person was speaking while the other just listened and back-channeled). We then created genuine and fake extracts by replacing the video recording of the real listener with another in half of the trials. In two successive behavioural experiments (the first, in-lab with $N = 18$, and the second, online with $N = 188$), observers were asked to discriminate genuine vs fake (i.e., non-contingent) interactions. We investigate whether social transfer functions learned from that dataset can predict observer performance better than a simpler model based on average quantity of movement or lagged correlation between speech and the listener's speech-contingent non-verbal responses; whether they can explain what exact facial features observers use for processing contingency; and whether they predict how observers would perform when parts of the stimuli are masked.

RESULTS

Observers Can Recognize Genuine Contingency Without Any Visual Information From the Speaker

Participants were presented with 10 s recordings of extracts from naturalistic, opposite-sex dyadic interactions where only one person was talking (speaker) and the other was silent but backchanneling (listener). Among the 198 unique extracts presented, half were 'fake', i.e., the video of the actual listener was replaced by that of some other random listener. Apart from genuine and fake contingency, the amount of available multimodal information was also manipulated: in the audio-video (A-V) modality, the speaker could only be heard and the listener could only be seen; in the video-video (V-V) modality, both the speaker and the listener could be seen but neither could be heard; and in the audiovisual-video (AV-V) modality, the speaker could be seen and heard while the listener could only be seen (Figure 2). Participants had to watch videos of interactions and rate them as genuine or fake (1-interval, 2-alternative forced choice).

Participants performed significantly above chance at discriminating genuine vs. fake interactions, $t(17) = 10.23$, $p < .001$, $d = 2.46$. Performance was markedly stronger when speaker behaviour was presented with audio (A-V block: $d' = 0.71$; AV-V block: $d' = 0.68$) than in video-only (V-V: $d' = 0.25$), which was lower than both A-V, $t(17) = 4.31$, $p < .001$, $d = 1.46$, and AV-V, $t(17) = 4.11$, $p < .001$, $d = 1.26$. There was no performance difference between the A-V and AV-V blocks, $t(17) = 0.33$, $p = .74$, $d = 0.08$ (Figure 2B). On the whole, this pattern of results was consistent with the idea that observers in this task mostly relied on speech-contingent non-verbal responses of the listener to judge the genuineness of interactions.

[2] The algorithmic choice of TRFs carries strong assumptions, namely, linearity and time-invariance, which should be evaluated empirically and discussed theoretically, while the more general computational notion of "social transfer function" does not entail such assumptions—see Discussion sections below.

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

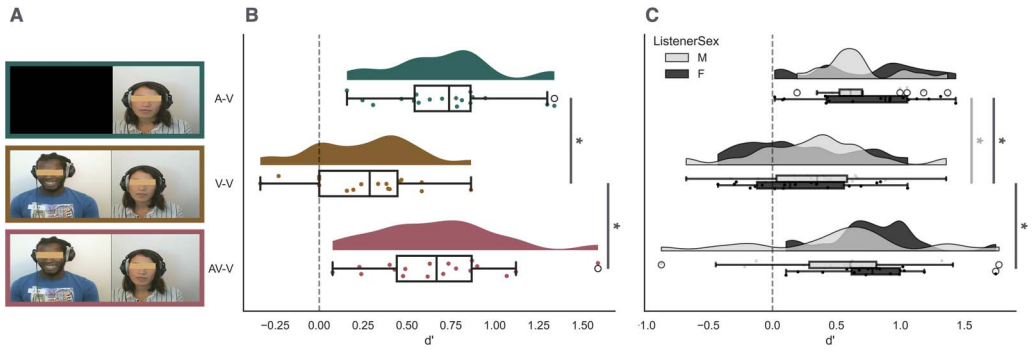


Figure 2. Observers can recognize genuine interactions even with degraded stimuli, e.g., even when only visual signals from the listener are visible. (A) Observers were presented audiovisual extracts from speed-dating interactions in which only one person was talking while the other was silent but backchanneling. Trials were presented in three possible 'modalities': audio-video (A-V, top) in which the speaker could be heard but not seen (i.e., their video recording replaced by a black screen), and the listener could be seen but not heard (i.e., their audio recording replaced by silence); video-video (V-V, middle), in which both the speaker and the listener could be seen but not heard; and audiovisual-video (AV-V, bottom) in which the speaker could be seen and heard while the listener could only be seen. (B) Sensitivity (d') over participants was significantly above chance in all modalities, with better performance in A-V and AV-V compared to V-V. (C) Stratifying the data on the basis of listeners' sex revealed similar performance in the AV-V and V-V modalities for interactions with male listeners, while performance in A-V and V-V modalities remained significantly different. Performance in female-listener interactions was better in both A-V and AV-V modalities compared to V-V. Box-plots are marking median values, inter-quartile range (IQR) and data points within 1.5 IQR, with * indicating statistical significance at the .05 alpha level (paired *t*-tests). Horizontal bars over the eyes were added in this image to protect the anonymity of the participants and were not displayed during the experiment.

Uncorrected Proof

To investigate if the sex of the listener influenced participant performance, we compared d' across blocks for subsets of data containing trials with either male or female listeners only. Results showed that trials with female listeners showed the same pattern as seen over all trials, i.e., significant difference in d' for both A-V vs. V-V, $t(17) = 3.61$, $p = .002$, $d = 1.17$, and AV-V vs. V-V trials $t(17) = 4.94$, $p < .001$, $d = 1.38$. However, trials with male listeners showed significant difference in d' only for A-V vs. V-V trials, $t(17) = 2.87$, $p = .011$, $d = 0.92$, suggesting that being able to see a female speaker interfered with observers' ability to recognize a genuine interaction (Figure 2C).

Temporal Response Function Dynamics Suggest Listener Motion Tends to Peak During Pauses or Low-Intensity Periods in Speech

To model the temporal dynamics of the relationship between speech and listeners' contingent non-verbal responses, we used the system identification technique of *temporal response functions* (TRFs; Crosse et al., 2016). To use as the speech feature, we estimated the perceived loudness of speech in a given interaction by processing it with a computational model of the auditory nerve (Zilany et al., 2014). To quantify listeners' contingent responses, we estimated the motion of a listener by computing the optical flow over consecutive video frames. A TRF was trained over the subset of all genuine A-V trials, with speech intensity as input and listener motion as output.

The dynamics of the TRF, with its initial negative trajectory until ~1–1.5 s suggested an increase in listener motion during pauses or periods of low speech intensity (since TRF inputs and outputs were *z*-scored, low speech intensity corresponded to negative values, and thus, convolution with negative values in the TRF yields positive values) that declined after 1 s (Figure 3A). To investigate whether the dynamics of the speech-motion relationship are influenced by the sex of the listener, we trained separate TRFs for interactions with male

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

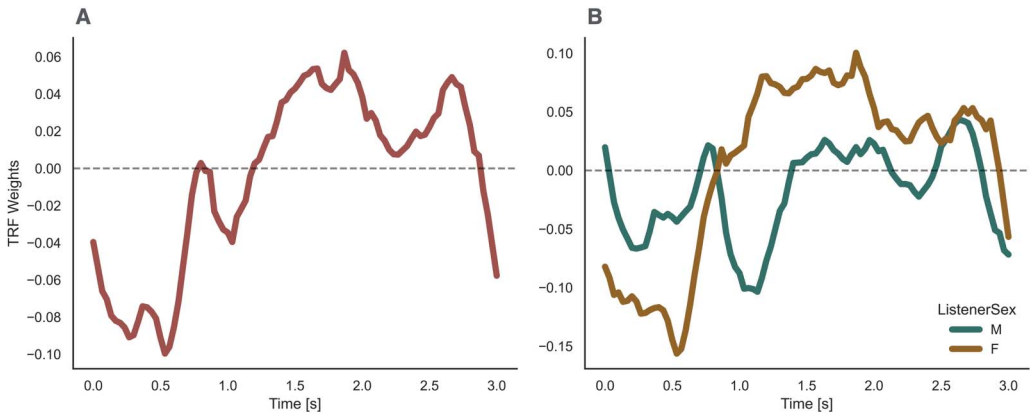


Figure 3. TRF dynamics suggest listener motion tends to peak during pauses or low-intensity periods in speech. (A) The listener-motion temporal response function (TRF) over all trials not only captured the differential dynamics of listener motion in genuine vs. fake interactions, but also allowed statistically significant prediction of observer decisions of the genuineness of interactions. The dynamics of the TRF, with its initial negative trajectory until ~1–1.5 s suggest increase in listener motion during pauses or periods of low speech intensity that declines after 1 s. (B) Training separate TRFs on interactions with male and female listeners reveals faster responses from female (with early negative peak at ~0.5 s) compared to male listeners (early negative peak at ~1 s). *Note.* Estimation of TRFs incorporates ridge regression to avoid overfitting and uses cross-validation over a set of trials for hyperparameter search, which does not easily allow computing, e.g., individual TRFs per trial, and consequently, confidence intervals for the TRFs. TRFs have been smoothed using a Savgol-Golay filter for visualization.

Uncorrected Proof

and female listeners. These specific listener-sex TRFs show differential dynamics, with the earlier negative peak for the female-listener TRF suggesting faster responses by female listeners by approximately 500ms compared to male listeners (Figure 3B).

Genuine and Fake Interactions Are Characterized by Differential Temporal Dynamics of Speech-Contingent Listener Motion

Using the TRFs estimated from genuine interactions, we predicted listener motion for both genuine and fake interactions. The match between actual and predicted listener motion (hereafter, TRF fit) was evaluated using Pearson's correlation coefficient *r*. This allowed us to test whether genuine and fake interactions were physically different, in terms of the temporal dynamics of the speech-motion relationship. To do so, we ran two separate analyses.

First, we compared the TRF fits between genuine and fake interactions, and found that genuine trials had statistically larger TRF fits than fake trials, $t(198) = 2.59$, $p = .011$, $d = 0.36$. As a confirmatory analysis, we fit a linear mixed model with the formula $Fit \sim Condition + (1|Dyad)$ to account for the effect of the same dyad contributing several trials. It yielded a similar *t*-statistic (2.61), but with a singular fit due to low Dyad variance, thus confirming that the *t*-test was appropriate. Second, we further evaluated whether TRF fits were computationally sufficient to accurately discriminate between genuine and fake trials. We trained a machine-learning classifier on the TRF fits as the input feature and 'genuine'/'fake' as class labels. A nested 10-fold stratified cross-validation (inner loop: hyperparameter tuning; outer loop: generalization estimate) with a linear SVM yielded above-chance classification accuracy ($M = 0.61$, $SD = 0.13$). A one-sample *t*-test confirmed that performance exceeded chance-level (0.5), $t(9) = 2.61$, $p = .027$, $d = 0.82$. Taken together, both streams of evidence suggest that the genuine and fake interactions in our task were indeed different with respect to how much they matched pre-learned dynamic predictions of listener motion.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

Perception of Genuine Contingency is Driven by The Dynamics of Listener Motion Rather Than its Average Quantity or (Lagged) Correlation With Speech

To test whether the TRF fits of trials predict observers' perception of genuineness, we fit generalized (logistic) linear mixed models with the formula Response ~ TRFfit + Intensity + xCorr + (1|Subject) + (1|Dyad). This allowed us to test whether human observers' behaviour, in terms of their binary responses, could be predicted by the particular speech-to-motion dynamics described by the TRF, rather than the average quantity of listener motion or its lagged correlation with speech. Generalized linear mixed models showed statistically significant relationships only between observer responses and TRF fits, and not average quantity of motion or any lagged correlation between speech intensity and listener motion (Table 1, Full model).

To see if the differential TRF dynamics observed for male and female listener trials (Figure 3B) translate to differential predictability of observer responses, we fit the GLMM Response ~ MaleFit * ListenerSex + FemaleFit * ListenerSex + (1|Subject) + (1|Dyad), where MaleFit and FemaleFit were the male-listener and female-listener TRF fits of trials respectively, and ListenerSex was the categorical variable encoding the sex of the listener (M/F) in a trial. Results showed that the female-listener TRF only predicted observer responses to female-listener trials but not male-listener trials (Table 1, Sex-stratified model). On the other hand, the male-listener TRF didn't predict observer responses to either male-listener or female-listener trials. The results suggest that observers may employ different TRFs to detect contingency depending on the sex of the listener.

Uncorrected Proof

Table 1. Observer perception of genuineness is driven by the dynamics of listener motion rather than its average quantity or (lagged) correlation with speech. In the full model, i.e., Response ~ Fit + Intensity + xCorr + (1|Subject) + (1|Dyad), Response is the observers' binary responses indicating whether they thought the interaction was genuine (1) or fake (0), Fit is the correlation between actual and predicted listener motion, Intensity is the average quantity of motion over the duration of the interaction, and xCorr is the cross-correlation between the speech feature and actual listener motion. Results indicate that observer responses are driven by the TRF fit, i.e., the specific dynamic relationship between speech and listener motion described by the TRF. In the sex-stratified model, i.e., Response ~ MaleFit * ListenerSex + FemaleFit * ListenerSex + (1|Subject) + (1|Dyad), MaleFit and FemaleFit are TRF fits of the TRFs trained on interactions with male and female listeners respectively, and ListenerSexM is a categorical variable encoding the sex of the listener in a trial. The results suggest that observer responses are driven by motion largely in female-listener interactions, with the lack of significance for male-listener TRF fits perhaps indicating a weaker effect or less expressiveness in male listeners.

Full model			Sex-stratified model		
Predictor	β	p	Predictor	β	p
Fit	0.231	<.001*	MaleFit	−0.006	.948
Intensity	−0.062	.355	FemaleFit	0.289	.005*
xCorr	0.078	.242	ListenerSexM	0.103	.445
			MaleFit:ListenerSexM	0.161	.301
			FemaleFit:ListenerSexM	−0.163	.314

* indicates $p < .05$.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

TRFs Can Help Isolate Low-Level Features (Beyond Listener Motion) That Physically Differ Between Genuine and Fake Interactions and That Drive Perception of Contingency

To obtain more fine-grained insight into the specific facial features driving perception of contingency, we broke down listener motion into the activity of specific facial action units (AUs; Ekman & Friesen, 1978). We extracted 18 AUs, namely, AU01, AU02, AU04, AU05, AU06, AU09, AU10, AU12, AU14, AU15, AU17, AU23, AU24, AU25, AU26, AU28, AU43, and Nods, and trained separate TRFs with the same speech input as before and each of the 18 AUs as output, across the subset of genuine A-V trials.

Of the 18 tested AUs, genuine trials had statistically larger TRF fits than fakes along 8 AUs: 5 of them mouth-related – AU10 ($t(189.1) = 2.39$, $p = .017$, $d = 0.34$), AU12 ($t(193.4) = 2.39$, $p = .017$, $d = 0.34$), AU25 ($t(195.8) = 2.32$, $p = .021$, $d = 0.33$), AU26 ($t(195.9) = 2.43$, $p = .015$, $d = 0.34$), AU28 ($t(190.1) = 2.05$, $p = .041$, $d = 0.29$); 2 eye-related: AU02 ($t(194.8) = 2.02$, $p = .043$, $d = 0.28$), AU06 ($t(192.1) = 2.61$, $p = .009$, $d = 0.37$); and nods ($t(195.3) = 2.48$, $p = .013$, $d = 0.35$). This suggested that the genuine and fake stimuli in our task differed most apparently on listener nods and mouth reactions such as smiling. We trained a machine-learning classifier on the AU TRF fits of trials as the input features and 'genuine'/'fake' as class labels to evaluate whether TRF fits were computationally sufficient to accurately discriminate between genuine and fake trials. The same nested 10-fold stratified cross-validation procedure as before with a linear SVM yielded above-chance classification accuracy ($M = 0.60$, $SD = 0.12$), with a one-sample t-test confirming that performance exceeded chance-level, $t(9) = 2.51$, $p = .033$, $d = 0.79$.

Finally, we tested whether observers' responses could be predicted by the TRF fits of AUs, the average quantity or intensity of AU activity over the duration of the interaction, and the lagged correlation between speech and the activity of a given AU. Fitting the GLMM Response ~ TRFfit + Intensity + xCorr + (1|Subject) + (1|Dyad) per AU revealed significant effects of TRF fits for AU17 ($\beta = 0.13$, $p = .037$), AU25 ($\beta = 0.16$, $p = .019$) and nods ($\beta = 0.16$, $p = .014$). Significant effects of average intensity were found for AU01 ($\beta = -0.16$, $p = .021$), AU02 ($\beta = -0.15$, $p = .031$), AU24 ($\beta = 0.21$, $p = .004$), AU25 ($\beta = -0.24$, $p = .001$), AU26 ($\beta = -0.19$, $p = .008$) and nods ($\beta = -0.13$, $p = .046$). For the lagged correlation between speech and AU activity (xCorr), only AU01 ($\beta = 0.15$, $p = .028$) and nods ($\beta = 0.17$, $p = .013$) reached significance.

Follow-Up Study

To investigate which specific parts of a listener's face drive observer perception of contingency, we employed a more conclusive, causal test. First, we selected a subset of 40A-V interactions from the previous study to control the difficulty of the task. Then, we created dynamic visual masks that isolated specific parts of a listener's face while hiding everything else (Figure 4). This yielded 3 versions of each of the 40 extracts: where the speaker's audio was played over a video that featured either the complete face area ('Original' condition), only the eye region ('Eyes' condition), or only the mouth region ('Mouth' condition). 188 participants were recruited online and assigned to one of the three conditions in a between-subject design, such that each condition had around 60 participants. Participants had to watch videos of interactions and rate them as genuine or fake (1-interval, 2-alternative forced choice).

Replicating the results of the previous study, we found that performance was significantly above chance for the full-information videos in the Original condition ($d' = 0.21$), $t(62) = 3.71$, $p < .001$, $d = 0.46$. Performance was also above chance for the Eyes condition ($d' = 0.24$),

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

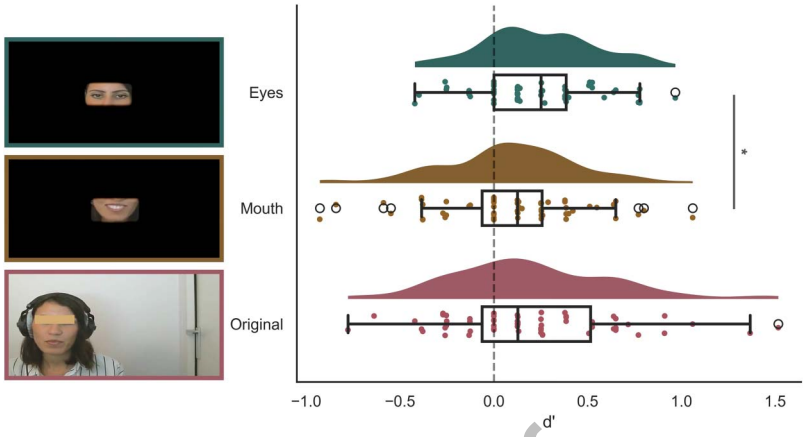


Figure 4. Observers can recognize genuine contingency even when only the eyes or the mouth of a listener is visible. Left: A-V trials from the previous study were manipulated by creating dynamic visual masks that isolated specific parts of the listener's video while hiding everything else. This yielded 3 different versions of each trial where the speaker's audio was played over a video that featured either only the eye region ('Eyes', top), only the mouth region ('Mouth', middle) or the complete face area ('Original', bottom). Right: Sensitivity (d') was significantly better for participants in the Eyes condition than in the Mouth condition. Viewing trials in the Original condition, i.e., the full non-masked videos, conferred no performance advantage over the conditions that involved masking. Horizontal bars over the eyes were added in this image to protect the anonymity of the participants.

$t(60) = 6.05$, $p < .001$, $d = 0.77$, with no difference from Original, $t(110.6) = 0.46$, $p = .645$, $d = 0.08$, but significantly greater than in the Mouth condition, $t(125.1) = 2.59$, $p = 0.011$, $d = 0.45$. Performance in the Mouth condition was not significantly above chance ($d' = 0.08$), $t(66) = 1.84$, $p = .069$, $d = 0.22$, but it also was not significantly lower than in the Original condition, $t(120.9) = -1.72$, $p = .087$, $d = 0.30$ (Figure 4).

Table 2. Observer perception of genuineness is driven by the dynamics of listener motion more strongly when the eyes are visible. The GLMM Response ~ TRF fit * Block + Intensity + xCorr + (1|Subject) + (1|Dyad) investigated how observer perception of social contingency, encoded by their binary response to whether an interaction was genuine, was influenced by TRF fits (and how they interact with what part of the face is visible), average motion quantity and lagged correlation between speech and listener motion. Results revealed that the dynamics of listener motion drives observer responses only in the 'Eyes' block, i.e., when only the eyes of the listener are visible. While there was no main effect of Block or Intensity, lagged correlation between the speech of the speaker and listener motion showed a significant effect.

	β	p
TRF fit	0.121	.036*
Blockmouth	−0.02	.965
Blockoriginal	−0.09	.225
Intensity	0.061	.141
xCorr	0.084	.035*
TRF fit:Blockmouth	−0.138	.015*
TRF fit:Blockoriginal	−0.035	.568*

* indicates $p < .05$.

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

We used generalized linear mixed models to test whether participant responses correlated with listener-motion TRF fits, i.e., the correlation between actual listener motion and that predicted by the TRF estimated from genuine A-V trials in Study 1. Fitting the GLMM Response ~ TRF fit * Block + Intensity + xCorr + (1|Subject) + (1|Dyad) on the data revealed that TRF fit significantly predicted responses in the 'Eyes' block, that this effect was significantly reduced in the 'Mouth' block, and that the 'Original' block did not differ from the 'Eyes' block (Table 2). While motion intensity did not reach significance, the lagged correlation between speech and listener motion also showed a significant, albeit weaker effect. Overall, these results suggested that speech-contingent mouth dynamics did not drive observer judgments of social contingency, and that it was driven to a large extent by the speech-contingent dynamics of the eye region.

DISCUSSION

Previous research has shown that social contingency detection is a robust human ability, but very little is known about which cues and signals contribute to the detection of contingency and how. In this article, we introduced a novel behavioural paradigm in which observers had to recognize genuine contingent behaviour in recorded video interactions. We manipulated both the contingency (genuine or fake) and the nature or amount of information available to observers, either by withholding information from specific modalities (Study 1) or by masking parts of listeners' faces (Study 2). Our results showed that observers were consistently able to recognize genuine interactions even with severely degraded interactions, causally relied on the link between a speaker's speech and a listener's motion, and that this link was driven by time-aligned, dynamic predictions rather than average quantity of movement or its lagged correlation with speech. In both studies, observer perception of social contingency was well-predicted by a computational model that evaluated the agreement of observed data with the predictions of speech-contingent responses by a pre-learned 'social transfer function'. The finding that observers possess highly redundant models of contingency allowing them to exploit partial information and adapt to sub-optimal interactional contexts is at odds with other forms of facial signalling which often critically depend on the availability of a single cue to make inferences (e.g., eye information for fear recognition, Adolphs et al., 2005, mouth/nose information for disgust Pavlova et al., 2023), and is consistent with the idea that social contingency detection may be an early developmental stepping stone towards such high-level social inferences (Frith & Frith, 2012).

Are Observers Able to Recognize Genuine Social Contingency?

The fact that participants performed above chance at the task confirms that the detection of social contingency is a robust ability. Given that even 4-week old infants are sensitive to disruption in social contingency (Happé & Frith, 2014), above chance performance in our task might seem trivial. But the contingency manipulation in this study produced relatively subtle, asynchronous behaviour rather than plain interruptions. Robust sensitivity to such ecological variations suggests that contingency is a graded evaluation built on cumulative evidence of synchronized or desynchronized behaviour. Interestingly, participants did not perform better with AV-V compared to A-V trials, and performed worst of all in V-V trials. This pattern of results appears at odds with the literature showing multimodal signals playing a facilitatory role in social-cognitive judgments like emotion recognition and mimicry (Krumhuber et al., 2023). For instance, Drijvers and Holler (2023) show that observers are better at shadowing speech when presented with multimodal (i.e., both audio and visual) information than when only hearing the audio—an effect that persists even if the speaker's mouth is blurred. A

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

possible explanation for the contrasting results is that the AV-V and V-V modalities in this study require observers to spatially divide their attention between two simultaneous streams of data (the speaker's and the listener's). The A-V modality, however, requires processing the alignment of speech with a single visual stream (comparable to the task in Drijvers & Holler, 2023), thus yielding no performance gain upon adding the speaker's video. It remains an open question whether preference for the amount of multimodal information depends on the timescale of the signals: fast (within milliseconds) for backchannels, and possibly slower for the kind of joint action explored in previous dyadic visual tasks (Moran et al., 2015; Neri et al., 2006).

Do Observers Use Social Transfer Functions to Judge Whether an Interaction is Contingent?

Estimating 'social transfer functions' from genuine interactions, we were able to capture the temporal dynamics of the relationship between a speaker's speech and a listener's motion. The specific dynamics of the listener motion TRF suggests that listener motion tends to occur during pauses or low-intensity periods of speech, peaking at 0.5 s after the onset of a pause. After approximately 1–1.5 s, the weights gradually become positive, suggesting a finite temporal window beyond which a listener's motion tends to decrease. This is confirmed by a wealth of studies showing that listeners produce non-verbal responses close to grammatical completion and/or pauses in the speaker's speech (Bravo Cladera, 2010; Clancy et al., 1996; Dittmann & Llewellyn, 1969; Hadar et al., 1985; Schegloff, 1982; Truong et al., 2011) rather than overlapping with speech (White, 1989). Speech intensity, in particular, appears to be an important feature, with falling intensity a reliable indicator for inviting backchannels (Gravano & Hirschberg, 2011) such that removing intensity dynamics negatively impacts the performance of backchannel prediction models (Inoue et al., 2025). The suppression of listener motion during speech and tendency to occur during breaks or pauses also echoes evidence of motor suppression during periods of high cognitive load followed by its release once the load subsides (Coupal et al., 2025; Huber et al., 2023). Importantly, we found that genuine and fake trials did differ in terms of how well they matched 'social transfer function' predictions of listener motion and AUs. While this does not imply that observers actually use that information, we presented two separate streams of evidence to support it. First, we show that TRF fits provide sufficient information for a machine classifier to reach human-level performance. While this is not conclusive evidence that observers use the same cues, it provides an important proof-of-possibility that the cues could support such an inference (for similar arguments, see De Boer & Kuhl, 2003; Goupil & Aucouturier, 2021; Piazza et al., 2017). And second, we found that observers' perception of genuineness correlated with TRF fit of both listener motion and some specific mouth-related AUs over and beyond their average quantities and their lagged correlation with speech.

Are Social Transfer Functions Fixed, Stable Representations?

We observed some discrepancies in the perception of contingency depending on whether the listener was male or female. First, when observers could see a female speaker (i.e., male listener trials), performance in the AV-V modality decreased to the extent that it was statistically similar to the poor performance seen in the V-V modality (trials with no audio). Second, separate TRFs trained on interactions with male and female listeners revealed differential dynamics, with female listeners producing faster responses compared to males by approximately 0.5 s. Finally, only the female-listener motion TRF showed specificity, i.e., was predictive of observer responses only for female-listener interactions. These stronger effects in female-listener interactions might be because the male-listener interactions happened to be

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

more difficult (since trial difficulty was not controlled for), but could also reflect lower expressiveness of male listeners compared to females (Dimberg & Lundquist, 1990; Korb et al., 2015; Kring & Gordon, 1998) resulting in relatively poor performance as well as greater response latencies (see Holland, O'Connell, & Dziobek, 2021 for a meta-analysis of facial mimicry, empathy and emotion recognition and how they are moderated by sex-differences).

Are Specific Regions of the Face More Salient for Perception of Contingency?

While listener motion proved to be a robust, important cue, it was a more holistic feature. To obtain fine-grained understanding of the facial features underlying perception of contingency, we broke down listeners' speech-contingent behaviour in terms of action units, which suggested it was mostly mouth-related features that physically differed in genuine and fake interactions and were used by observers to judge genuineness. However, these results were descriptive and did not provide a formal test of causality (Casadevall & Fang, 2008). For instance, it could be that genuine trials contain other cues that are actually used by observers either at locations (e.g., pupil size Goswami et al., 2020; Hess & Petrovich, 2014; Kret, 2018) or dynamical scales not captured by AUs or our methodology. To address these limitations, we adopted a causal manipulation design by masking different parts of the face and found that the ability to detect genuine contingency was significantly better when observers could only see a listener's eyes than when they could only see their mouth. We also found no statistical evidence in the original condition to suggest any performance improvement on providing participants with complete face information. This not only suggests that no other facial cues besides the eye provide any discriminating information for contingency in this specific task, but also that participants did not utilize the interaction between the eyes and the mouth to any avail. In other words, dynamic predictions of eye and mouth activity constitute redundant cues/signals for detecting social contingency, a property that contrasts with other types of facial inferences typically utilizing a dynamic and complementary hierarchy of signals over time (Jack et al., 2014).

This pattern of results was supported by evidence that the fits of the Study 1 motion TRF correlated with, and predicted observer responses, rather than average intensity of motion or its lagged correlation with speech. The results strongly suggest that dynamic predictions of facial consequences in the eye-region of listeners constitute a mechanism for third-party observers judging social contingency. The greater importance of the eyes, compared to the mouth, for detecting social contingency is consistent with the literature showing that eye gaze is a key channel through which listeners signal engagement (Conty et al., 2006; Kendon, 1967), attention (Argyle et al., 1994; Langton et al., 2000) and mutual understanding (Cummins, 2012; Hömke et al., 2018, 2025; Wahn et al., 2022), facilitate turn-taking (Hirvenkari et al., 2013; Jokinen et al., 2013) and infer mental states (Baron-Cohen & Cross, 1992; Cañigueral & Hamilton, 2019). Importantly, it has also been shown that it is the temporal dynamics of gaze behaviour, rather than simply its presence, that drives perception of social contingency (Pfeiffer et al., 2012). Better performance in the eyes block could thus be because observers benefit from combining motion information with other gaze signals, a finding that is consistent with a study demonstrating preferential use of gaze over movement kinematics and resorting to kinematics only when gaze signals were unavailable (Trujillo et al., 2018). Taken together, our results provide further evidence of the crucial role of the eyes in the perception of social contingency that provides scaffolding for more high-level social-cognitive inferences.

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

Versatility of Social Transfer Functions

In this work, we implemented 'social transfer functions' using temporal response functions (TRFs; Crosse et al., 2016). Though they make strong assumptions about the system (more below), they offer a particularly parsimonious representation of a predictive model of contingent responses: namely, an impulse response with which the input (here, a speaker's speech) is convolved to generate the output (a listener's non-verbal response). Our social transfer function paradigm not only confirmed past findings, but also captured speech-contingent temporal dynamics in a precise, quantitative and testable manner. Taken together, the pattern of results suggests that pre-learned models that enable the dynamic prediction of speech-contingent responses of listeners constitute a mechanism by which third-party observers judge social contingency—which we proposed as 'social transfer functions'.

The concept of a social transfer function, as implemented here with TRFs, provides an operational mechanism for the general ability of 'interpersonal predictive coding', by which observers use the actions of one agent to predict both the content and temporality of a second agent's actions (Manera, Del Giudice, et al., 2011). This renders them quite useful as analytical or generative tools. Analytically, how well their predictions fit observed data provides a way to quantify the typicality of listeners' response dynamics. This could be used to quantify abnormal conversational dynamics (e.g., turn-taking patterns in schizophrenia, Lucarini et al., 2024), or detect AI-generated videos (Y. Li et al., 2018). Combined with facial animation techniques in virtual avatars (Yu et al., 2012) or real-life videos (Arias-Sarah et al., 2024), social transfer functions can also be used to generate novel stimuli possessing specific dynamics for experimental control or to improve synthetic media (e.g., manipulate perceived contingency of deep-faked conversations in human-computer interaction, Kaate et al., 2023).

A related question is whether social transfer functions are uniquely related to judgements of social contingency. While this work has focussed on the detection of social contingency, we believe that social transfer functions reflect more general mechanisms and support other social-cognitive inferences that rely on dynamic predictions of conversational signals. For instance, storing separate pre-learned transfer functions for interactions between familiar and unfamiliar agents (Gráczi & Bata, 2010), interactions involving agreement (Müller et al., 2022) and those characterized by enjoyment (H. Z. Li et al., 2010) would facilitate recognition of which is more likely to be occurring. By providing a parsimonious representation of conversational dynamics, social transfer functions appear promising as a way to study how these constructs are signalled, and to model how they are recognized by observers. Different social transfer functions for different constructs also implies the presence of flexible definitions of contingent dynamics that may help construct high-level social judgements. From a clinical perspective, social transfer functions could also provide valuable insights into disorders affecting conversational dynamics such as autism spectrum disorders (Wehrle et al., 2024), parental depression (Smith et al., 2023) and disorders of consciousness (Hermann et al., 2018), both by characterizing how they are produced, and how they are evaluated as abnormal by caregivers.

Finally, while this study operationalized social transfer functions using speech intensity as input and listener motion or action units as outputs, the paradigm developed here is agnostic to the specific features chosen, i.e., one can substitute in any acoustic feature—pitch, stress, or even linguistic surprisal, any listener signal as output, and can be extended naturally to multivariate inputs or nonlinear impulse response variants that could capture threshold-like listener dynamics.

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

Limitations of Social Transfer Functions

Linearity. It is unknown how social transfer functions are cognitively represented, or even if they are at all (for a related discussion of "dynamic mental representation" of object or human motion, see Freyd, 1987). While we argue that convolution with an impulse response is a parsimonious way to represent an input-output mapping, it remains an experimental question whether human observers indeed encode and decode interactions this way. Impulse responses make strong assumptions about the linearity of a system (Keesman, 2011) and thus fail to represent a potentially large class of listener behaviour that may have non-linear, threshold-like qualities. It would be interesting to compare impulse responses with more advanced transfer function learning techniques for a better understanding of what observers typically consider 'contingent'. More generally, there are other possible computational and/or cognitive architectures for learning a conditional distribution $p(y/x)$ between input and output that do not use the formalism of transfer functions, such as discrete rules (e.g., detection of a low-pitch region late in an utterance, Poppe et al., 2010) or conditional random fields (Morency et al., 2008), and may be alternative mechanisms that future work could evaluate.

Time and Context Invariance. The "time-invariant" aspect of social transfer functions assumes that there is an average first-order relation between agents in real-time interactions which can be learned and tested against new data. In other words, we assume that observers' expectations of the dynamics of backchanneling behaviour is independent of absolute time and that recognition of social contingency is thus driven by applying the same internal representation to observed behaviour even as, e.g., a listener's engagement decreases and modulates the dynamics of their contingent signals. This also ties in to the concept of context-invariance. The social transfer functions estimated in this study are embedded in the very specific context of speed-dating interactions, i.e., 'first-time' encounters usually demanding a reasonable amount of affiliation and politeness, or even shyness. This excludes a lot of variance driven, for example, by familiarity and contexts allowing disagreement and argument. Similarly, it remains an open question how the dynamics of social transfer functions differ and how they are learned and modified in different cultures. For instance, there is debate about whether conversational conventions that are not shared across cultures (e.g., how much feedback one is expected to give) contribute to misunderstanding or stereotyping (White, 1989). It is thus entirely possible that our social transfer function predictions would not correlate with observer ratings of contingency when applied to other types of interactions or cultures. But we argue that they at least provide a testable, operational mechanism, with good parsimonious properties (storing a TRF is more memory-efficient than storing examples of input-output pairs or a conditional probability distribution—a computational 'trick' also exploited in so-called convolutional deep learning architectures, Mallat, 2016).

MATERIALS AND METHODS

Study 1

Participants. $N = 18$ (male = 14; $M = 25.8$, $SD = 10.04$) native French speakers participated in the study. Participants were recruited from the Master's program at SUPMICROTECH (Besançon, France).

Speed Dating Corpus. Stimuli used in this work were extracted from a corpus of video recordings of naturalistic speed-dating interactions, which we collected as part of a larger project (Arias-Sarah et al., 2024). $N = 31$ French-speaking participants (male = 15; mean age = 22 [20–27]) were part of the dataset collection. All participants were heterosexual, single, and

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

were willing to participate in a real speed-dating experiment where they would have the option to potentially connect with their partners at the end of the experiment.

Participants were paired into M/F dyads such that each male interacted with each female participant within that session. Each dyad had a 4-minute conversation over a video-conferencing platform, while seated in a windowless cubicle. The conversations were entirely unscripted: we instructed participants to talk about any conversation topic they wanted with their interacting partner for the whole duration of the interaction. We equipped participants with Beyerdynamic DT770 Pro headphones and recorded all interactions with Logitech C920 webcams at 30 frames per second. We organized data collection in batches of eight participants. For each batch, four males and four females interacted with each other, following a round-robin design (Kenny et al., 2020). We collected 4 batches of 8 participants in total. One female participant was absent in one of the sessions. Thus, we collected a total of 60 interactions from 31 different participants. The dataset collection was approved by the Institut Européen d'Administration des Affaires (INSEAD) IRB. In accordance with the American Psychological Association Ethical Guidelines, all participants gave their informed consent and were debriefed and informed about the purpose of the research after the experiment.

Stimuli. From recorded conversations in the Speed Dating corpus, we extracted 305 segments lasting around 10 seconds (M = 10.01 [5–26]) in which only one person was talking while the other was silent and only displayed backchanneling cues like nods, smiles and blinks. From these, 99 segments were selected as genuine interactions. From the remaining 206 segments, 99 'fake' interactions were created by pairing recordings of speakers with some other random listener, i.e., not the listener the speaker was originally talking to. This resulted in a total of 198 interactions (99 genuine and 99 fake). Each interaction was unique, meaning that every single interaction had a different audio and video track and was not repeated across conditions or blocks.

Finally, for each genuine and fake interaction, we created 3 presentation 'modalities' of the same extract: audio-video (hereafter: A-V) in which the speaker could be heard but not seen (i.e., their video recording replaced by a black screen), and the listener could be seen but not heard (i.e., their audio recording replaced by silence); video-video (V-V), in which both the speaker and the listener could be seen but not heard; and audiovisual-video (AV-V), in which the speaker could be seen and heard while the listener could only be seen (Figure 2A).

Procedure. Participants were presented 3 blocks of 66 video trials, each block containing trials from one of the A-V, V-V and AV-V modalities. Blocks, and trials within blocks, were presented in random order, with short self-paced breaks in between. No interactions were repeated, meaning that a given extract did not have a genuine and fake 'version' but were completely separate interactions. After each video extract, participants were asked to report whether they thought the interaction was genuine (1-interval, 2-alternative forced choice).

Social Transfer Functions. To model how well genuine/contingent interactions matched predictions of the temporal dynamics of a listener's speech-contingent non-verbal responses to a speaker's speech, we used the system identification technique of temporal response functions (TRFs; Crosse et al., 2016). First, we estimated the time series of perceived loudness from the speech in a given interaction by processing the speech extract with a computational model of the auditory nerve (Zilany et al., 2014) designed to reproduce features of loudness compression of the human auditory system (a technique suggested to improve TRF modeling in Lindboom et al., 2023 and Benghanem et al., 2024).

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

We used 2 different features to capture the listeners' responses: motion and action units (AUs; Ekman & Friesen, 1978). Listener motion was extracted from video recordings using optical flow computed with the Farneback algorithm as implemented in the OpenCV Python library (Bradski & Kaehler, 2000), which estimates the motion of each pixel between consecutive video frames. The mean magnitude was computed for each frame, yielding a continuous time series representing the overall movement of the listener. Both the speech feature and the listener's motion time-series were synchronized at the same frame rate and standardized. Using the ridge regression method as implemented in the mtrfpy toolbox (Bialas et al., 2023), a TRF was trained on the subset of genuine A-V interactions to model the dynamic relationship between speech and listener motion found in ecological social interactions. To isolate the specific facial features contributing to perception of contingency, we broke down listener motion into the activity of specific AUs. We extracted the time series of the activity of 18 facial action units (AU01: inner brow raiser, AU02: outer brow raiser, AU04: brow lowerer, AU05: upper lid raiser, AU06: cheek raiser, AU09: nose wrinkler, AU10: upper lip raiser, AU12: lip corner puller, AU14: dimpler, AU15: lip corner depressor, AU17: chin raiser, AU23: lip tightener, AU24: lip pressor, AU25: lip part, AU26: jaw drop, AU28: lip suck, AU43: eyes closed, Pitch: head nods; i.e., 6 eye and 11 mouth-related AUs) from the listener's video, using the Py-feat library (Cheong et al., 2023). A separate TRF was trained for each AU.

Once trained, the TRFs allowed predicting a listener's motion or AU activity in response to a specific speaker's speech by convolving the input speech with the TRF, based on the regularities it learnt from the data on which it was trained (i.e., genuine, ecological interactions). In any given interaction, the match between the time-series predicted by convolution with TRF and the actual listener's time-series was evaluated using Pearson's correlation coefficient *r*.

Statistical Analyses. Participant performance was quantified using the *d'* sensitivity index and tested for statistical difference from chance-level (0) with one-sample *t*-tests, and for differences across modalities (within-participant) with paired *t*-tests (3 levels: A-V, V-V, AV-V).

To evaluate whether genuine and fake trials physically differed in terms of how well they matched TRF predictions, we compared Fisher z transformed Pearson correlation coefficients between the predicted and actual listener-motion time-series (hereafter: *TRF fit*) of genuine and fake trials with two-sample *t*-tests. To evaluate whether TRF fits were computationally sufficient to accurately discriminate between genuine and fake trials, we trained a support vector machine (SVM) classifier with a linear kernel which took the vector of the listener-motion TRF fits as the solitary input feature and a binary classification of a trial as being genuine or fake as output. We used a nested 10-fold stratified cross-validation procedure where the inner loop was used for hyperparameter (C and Gamma) optimization by grid-search, and the outer loop was used to obtain a generalization estimate of the classifier performance. Classification scores over the 10 folds were tested for statistical significance with a one-sample *t*-test against chance-level (0.5). SVM training was implemented using the scikit-learn Python package.

To test whether the TRF fits of trials predict observers' perception of genuineness, we fit generalized (logistic) linear mixed models with the formula: Response ~ Fit + Intensity + xCorr + (1|Subject) + (1|Dyad). The dependent variable, Response, was the binary, yes/no observer response regardless of whether the trial was genuine or not. The predictor Fit represented the TRF fit of a given trial, while Intensity was the average magnitude of actual listener motion over the total duration of the trial. Finally, to have a stronger control than simple quantity of motion, we added xCorr as a predictor, representing the maximum absolute value of the cross-correlation coefficient between speech and listener motion at any lag. Average motion quantity was meant to control for the contingency perception driven by simple visual patterns

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

or the mere presence of listener features rather than the dynamics. xCorr, on the other hand, was included to account for responses driven by general, time-shifted correlation of speech with the listener feature instead of following the specific dynamics described by the TRF. Random intercepts of observer (Subject) and the dyad to which a speaker-listener pair belonged (Dyad) were also included. While random slopes of the predictors would have been appropriate, the relatively low sample size of this study meant that they yielded singular fits. The GLMM analysis was conducted with the lme4 package in R (Bates et al., 2015).

Analyses with listener AUs instead of motion were similar to those described above, with only a few differences. First, since a separate TRF was trained for each AU, separate *t*-tests were conducted for each AU comparing TRF fits in genuine and fake trials. Second, the SVM was fed the TRF fits of all 18 AUs as features, and consequently, allowed us to use SVM weights as a rudimentary but, in the case of linear kernels, conservative indication of feature importance (Guyon & Elisseeff, 2003). And finally, as with the *t*-tests, a separate GLMM was fit for each AU. Given the exploratory nature of this investigation and the complex correlational structure between AUs, no multiple-comparisons correction was applied to either the *t*-tests or GLMMs (see Supplementary Material for correlations between AUs and how correcting for multiple comparisons impacts the observed p-values).

Study 2

Participants. We recruited $N = 188$ participants through Prolific in a between-subject design with approximately 65 participants in each of the three conditions – $N_{Eyes} = 61$ ($M = 31.5$, $SD = 9.76$, male = 39), $N_{Mouth} = 67$ ($M = 31.1$, $SD = 10.66$, male = 39) and $N_{Original} = 63$ ($M = 30.8$, $SD = 10.77$, male = 36). Participants gave their informed consent and were compensated at a standard rate. An a priori power analysis conducted using G*Power (Faul et al., 2009) found a minimum sample size of 64 (per group) to obtain 80% power for detecting a medium effect at $\alpha = .05$.

Stimuli. Stimuli for Study 2 were selected as a subset of stimuli from Study 1, to control the difficulty of the task more formally. Because V-V stimuli were not recognized accurately in Study 1, and AV-V trials did not provide any performance advantage over A-V, Study 2 was restricted to A-V stimuli. To select the subset, we classified the 66 A-V trials of Study 1 as hits, misses, correct rejections or false alarms based on the most frequent decision made by Study 1 participants and selected 30 stimuli controlled for difficulty in each of the four signal-detection categories, resulting in a total of 120 A-V stimuli.

Trials were further manipulated by creating dynamic visual masks that isolated specific parts of the face in the listener's video while hiding everything else (Figure 4). We used DaVinci Resolve by Blackmagic Design to track a manually-specified rectangle centred either on the eye or mouth region in the video recordings and manipulated the outside of the rectangle at zero pixel intensity. This yielded 3 different versions of each AV-V trial where the speaker's audio was played over a video that featured either the complete face area ('Original', same as Study 1), only the eye region ('Eyes' condition), or only the mouth region ('Mouth' condition).

Procedure. Participants were presented with 40 stimuli in one of the three conditions in a between-subject design. In each condition, the task was the same as in Study 1, where participants had to watch videos of interactions and rate them as genuine or fake (1-interval, 2-alternative forced choice). A previous version of this task was piloted with 20 offline participants and a within-subjects design as opposed to between-subjects, but was changed due

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

to the discovery of order effects. The procedure was preregistered at https://aspredicted.org/fsp7-g7jw.pdf.

ACKNOWLEDGMENTS

The work was funded by ANR SOUNDS4COMA and conducted in the framework of the EIPHI Graduate school (ANR-17-EURE-0002 contract), and UK Research and Innovation HORIZON-MSCA-2021-PF-01 to Pablo Arias-Sarah.

AUTHOR CONTRIBUTIONS

Project was formulated by JJA, PAS, and RG. Analysis was conducted by RG, JJA and PAS. All authors wrote the manuscript.

Q4 Q5

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

Q6 **REFERENCES**

Q7 Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, *433*(7021), 68–72. https://doi.org/10.1038/nature03086, PubMed: 15635411

Argyle, M., Cook, M., & Cramer, D. (1994). Gaze and mutual gaze. *British Journal of Psychiatry*, *165*(6), 848–850. https://doi.org/10.1017/S0007125000073980

Arias-Sarah, P., Bedoya, D., Daube, C., Aucouturier, J.-J., Hall, L., & Johansson, P. (2024). Aligning the smiles of dating dyads causally increases attraction. *Proceedings of the National Academy of Sciences*, *121*(45), e2400369121. https://doi.org/10.1073/pnas.2400369121, PubMed: 39467124

Baron-Cohen, S., & Cross, P. (1992). Reading the eyes: Evidence for the role of perception in the development of a theory of mind. *Mind & Language*, *7*(1–2), 172–186. https://doi.org/10.1111/j.1468-0017.1992.tb00203.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bavelas, J., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture*, *17*(1), 98–127. https://doi.org/10.1075/gest.00012.bav

Benghanem, S., Guha, R., Pruvost-Robieux, E., Lévi-Strauss, J., Joucla, C., Cariou, A., Garavet, M., & Aucouturier, J.-J. (2024). Cortical responses to looming sources are explained away by the auditory periphery. *Cortex*, *177*, 321–329. https://doi.org/10.1016/j.cortex.2024.05.018, PubMed: 38908362

Bialas, O., Dou, J., & Lalor, E. C. (2023). mTRfpy: A python package for temporal response function analysis. *Journal of Open Source Software*, *8*(89), 5657. https://doi.org/10.21105/joss.05657

Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlupt, P. (2005). Listening to a walking human activates the temporal biological motion area. *NeuroImage*, *28*(1), 132–139. https://doi.org/10.1016/j.neuroimage.2005.06.018, PubMed: 16027008

Blake, R. (1993). Cats perceive biological motion. *Psychological Science*, *4*(1), 54–57. https://doi.org/10.1111/j.1467-9280.1993.tb00557.x

Bradski, G., & Kaehler, A. (2000). OpenCV. *Dr. Dobb's Journal of Software Tools*, *3*(2), 1–81.

Bravo Cladera, N. (2010). Backchannels as a realization of interaction: Some uses of mm and mhm in spanish. In D. Koike, & L. Rodriguez-Alfano (Eds.), *Dialogue in Spanish: Studies in functions and contexts* (pp. 137–156). John Benjamins Publishing Company. https://doi.org/10.1075/ds.7.08bra

Brazelton, T. B., Tronick, E., Adamson, L., Als, H., & Wise, S. (1975). Early mother-infant reciprocity. In R. Porter, & M. O'Connor (Eds.), *Ciba foundation symposium 33-parent-infant interaction* (pp. 137–154). Elsevier. https://doi.org/10.1002/9780470720158.ch9

Brunner, L. J. (1979). Smiles can be back channels. *Journal of Personality and Social Psychology*, *37*(5), 728–734. https://doi.org/10.1037/0022-3514.37.5.728

Cañigueral, R., & Hamilton, A. F. d. C. (2019). The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in Psychology*, *10*, 560. https://doi.org/10.3389/fpsyg.2019.00560, PubMed: 30930822

Casadevall, A., & Fang, F. C. (2008). Descriptive science. *Infection and Immunity*, *76*(9), 3835–3836. https://doi.org/10.1128/IAI.00743-08, PubMed: 18625736

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., & Chang, L. J. (2023). Py-feat: Python facial expression analysis toolbox. *Affective Science*, *4*(4), 781–796. https://doi.org/10.1007/s42761-023-00191-4, PubMed: 38156250

Clancy, P. M., Thompson, S. A., Suzuki, R., & Tao, H. (1996). The conversational use of reactive tokens in english, Japanese, and Mandarin. *Journal of Pragmatics*, *26*(3), 355–387. https://doi.org/10.1016/0378-2166(95)00036-4

Coey, C. A., Varlet, M., & Richardson, M. J. (2012). Coordination dynamics in a socially situated nervous system. *Frontiers in Human Neuroscience*, *6*, 164. https://doi.org/10.3389/fnhum.2012.00164, PubMed: 22701413

Conty, L., Tijus, C., Hugueville, L., Coelho, E., & George, N. (2006). Searching for asymmetries in the detection of gaze contact versus averted gaze under different head views: A behavioural study. *Spatial Vision*, *19*(6), 529–545. https://doi.org/10.1163/156856806779194026, PubMed: 17278526

Coupal, P., Zhang, Y., & Deroche, M. (2025). Reduced eye blinking during sentence listening reflects increased cognitive load in challenging auditory conditions. *Trends in Hearing*, *29*, 23312165251371118. https://doi.org/10.1177/23312165251371118, PubMed: 40910460

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604, PubMed: 27965557

Cummins, F. (2012). Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language*

Uncorrected Proof

and Cognitive Processes, *27*(10), 1525–1549. https://doi.org/10.1080/01690965.2011.615220
Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. In *Psychology of learning and motivation* (Vol. 59, pp. 43–95). Elsevier. https://doi.org/10.1016/B978-0-12-407187-2.00002-2
De Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134. https://doi.org/10.1121/1.1613311
Dimberg, U., & Lundquist, L.-O. (1990). Gender differences in facial reactions to facial expressions. *Biological Psychology*, *30*(2), 151–159. https://doi.org/10.1016/0301-0511(90)90024-Q, PubMed: 2285765
Dittmann, A. T., & Llewellyn, L. G. (1969). Body movement and speech rhythm in social conversation. *Journal of Personality and Social Psychology*, *11*(2), 98–106. https://doi.org/10.1037/h0027035, PubMed: 5778352
Drijvers, L., & Holler, J. (2023). The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, *30*(2), 792–801. https://doi.org/10.3758/s13423-022-02178-x, PubMed: 36138282
Ekman, P., & Friesen, W. V. (1978). *Facial action coding system*. Environmental Psychology & Nonverbal Behavior.
Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149, PubMed: 19897823
Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review*, *94*(4), 427–438. https://doi.org/10.1037/0033-295X.94.4.427
Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449, PubMed: 21838544
Goldberg, S. (1977). Social competence in infancy: A model of parent-infant interaction. *Merrill-Palmer Quarterly of Behavior and Development*, *23*(3), 163–177.
Goswami, M., Manuja, M., & Leekha, M. (2020). Towards social & engaging peer learning: Predicting backchanneling and disengagement in children. *arXiv preprint*. https://doi.org/10.48550/arXiv.2007.11346
Goupil, L., & Aucouturier, J.-J. (2021). Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, *212*, 104661. https://doi.org/10.1016/j.cognition.2021.104661, PubMed: 33756151
Grászi, T., & Bata, S. (2010). The effect of familiarization on temporal aspects of turn-taking: A pilot study. *Acta Linguistica Hungarica*, *57*(2–3), 307–328. https://doi.org/10.1556/ALing.57.2010.2-3.7
Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601–634. https://doi.org/10.1016/j.csl.2010.10.003
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.
Hadar, U., Steiner, T. J., & Clifford Rose, F. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, *9*(4), 214–228. https://doi.org/10.1007/BF00986881
Happé, F., & Frith, U. (2014). Annual research review: Towards a developmental neuroscience of atypical social cognition. *Journal of Child Psychology and Psychiatry*, *55*(6), 553–557. https://doi.org/10.1111/jcpp.12162, PubMed: 24963529
Hermann, B., Goudard, G., Courcoux, K., Valente, M., Labat, S., Despois, L., Bourmaleau, J., Richard-Gilis, L., Faugeras, F., Demeret, S., Sitt, J. D., Naccache, L., & Rohaut, B. (2018). "Doc-feeling": A new behavioural tool to help diagnose the minimally conscious state. *bioRxiv*. https://doi.org/10.1101/370775
Hermans, K. S. F. M., Kirtley, O. J., Kasanova, Z., Achterhof, R., Hagemann, N., Hiekkaranta, A. P., Lecei, A., Zapata-Fonseca, L., Lafit, G., Fossion, R., Froese, T., & Myin-Germeys, I. (2022). Capacity for social contingency detection continues to develop across adolescence. *Social Development*, *31*(3), 530–548. https://doi.org/10.1111/sode.12567
Hess, E. H., & Petrovich, S. B. (2014). Pupillary behavior in communication. In *Nonverbal behavior and communication* (pp. 327–348). Psychology Press.
Hirvenkari, L., Ruusuvuori, J., Saarinen, V.-M., Kivioja, M., Peräkylä, A., & Hari, R. (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PloS One*, *8*(8), e71569. https://doi.org/10.1371/journal.pone.0071569, PubMed: 23951192
Holland, A. C., O'Connell, G., & Dziobek, I. (2021). Facial mimicry, empathy, and emotion recognition: A meta-analysis of correlations. *Cognition & Emotion*, *35*(1), 150–168. https://doi.org/10.1080/02699931.2020.1815655, PubMed: 32924798
Holler, J. (2022). Visual bodily signals as core devices for coordinating minds in interaction. *Philosophical Transactions of the Royal Society B*, *377*(1859), 20210094. https://doi.org/10.1098/rstb.2021.0094, PubMed: 35876208
Hömke, P., Holler, J., & Levinson, S. C. (2017). Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, *50*(1), 54–70. https://doi.org/10.1080/08351813.2017.1262143
Hömke, P., Holler, J., & Levinson, S. C. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS One*, *13*(12), e0208030. https://doi.org/10.1371/journal.pone.0208030, PubMed: 30540819
Hömke, P., Levinson, S. C., Emmendorfer, A. K., & Holler, J. (2025). Eyebrow movements as signals of communicative problems in human face-to-face interaction. *Royal Society Open Science*, *12*(3), 241632. https://doi.org/10.1098/rsos.241632, PubMed: 40078914
Huber, S. E., Martini, M., & Sachse, P. (2023). Task-synchronized eye blink modulation neither requires visual stimulation nor active motor response and is modulated by task predictability. *International Journal of Psychophysiology*, *187*, 1–10. https://doi.org/10.1016/j.ijpsycho.2023.01.011, PubMed: 36773888
Inoue, K., Lala, D., Skantze, G., & Kawahara, T. (2025). Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 7171–7181). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.367
Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064, PubMed: 24388852
Jokinen, K., Furukawa, H., Nishida, M., & Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(2), 1–30. https://doi.org/10.1145/2499474.2499481
Jordan, H., Reiss, J. E., Hoffman, J. E., & Landau, B. (2002). Intact perception of biological motion in the face of profound spatial deficits: Williams syndrome. *Psychological Science*, *13*(2),

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

162–167. https://doi.org/10.1111/1467-9280.00429, PubMed: 11934001

Kaate, I., Salminen, J., Jung, S.-G., Almerekhi, H., & Jansen, B. J. (2023). How do users perceive deepfake personas? Investigating the deepfake user perception and its implications for human-computer interaction. In *Proceedings of the 15th Biannual Conference of the Italian Sigchi Chapter* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3605390.3605397

Keesman, K. J. (2011). *System identification: An introduction*. Springer Science & Business Media. https://doi.org/10.1007/978-0-85729-522-4

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*(1), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4, PubMed: 6043092

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2020). *Dyadic data analysis*. Guilford Publications.

Klin, A., Lin, D. J., Gorrindo, P., Ramsay, G., & Jones, W. (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, *459*(7244), 257–261. https://doi.org/10.1038/nature07868, PubMed: 19329996

Knudsen, B., Creemers, A., & Meyer, A. S. (2020). Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, *11*, 593671. https://doi.org/10.3389/fpsyg.2020.593671, PubMed: 33240183

Korb, S., Malsert, J., Rochas, V., Rihs, T. A., Rieger, S. W., Schwab, S., Niedenthal, P. M., & Grandjean, D. (2015). Gender differences in the neural network of facial mimicry of smiles–an rtms study. *Cortex*, *70*, 101–114. https://doi.org/10.1016/j.cortex.2015.06.025, PubMed: 26211433

Kret, M. E. (2018). The role of pupil size in communication. is there room for learning? *Cognition & Emotion*, *32*(5), 1139–1145. https://doi.org/10.1080/02699931.2017.1370417, PubMed: 28857664

Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology*, *74*(3), 686–703. https://doi.org/10.1037/0022-3514.74.3.686, PubMed: 9523412

Krumhuber, E. G., Skora, L. I., Hill, H. C., & Lander, K. (2023). The role of facial movements in emotion recognition. *Nature Reviews Psychology*, *2*(5), 283–296. https://doi.org/10.1038/s44159-023-00172-1

Langton, S. R., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50–59. https://doi.org/10.1016/S1364-6613(99)01436-9, PubMed: 10652522

Larsch, J., & Baier, H. (2018). Biological motion as an innate perceptual mechanism driving social affiliation. *Current Biology*, *28*(22), 3523–3532. https://doi.org/10.1016/j.cub.2018.09.014, PubMed: 30393036

Li, H. Z., Cui, Y., & Wang, Z. (2010). Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. *International Journal of Psychological Studies*, *2*(1), 25. https://doi.org/10.5539/ijps.v2n1p25

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7). IEEE. https://doi.org/10.1109/WIFS.2018.8630787

Lindboom, E., Nidiffer, A., Carney, L. H., & Lalor, E. C. (2023). Incorporating models of subcortical processing improves the ability to predict eeg responses to natural speech. *Hearing Research*, *433*, 108767. https://doi.org/10.1016/j.heares.2023.108767, PubMed: 37060895

Lucarini, V., Grice, M., Wehrle, S., Cangemi, F., Giustozzi, F., Amorosi, S., Rasmi, F., Fascendini, N., Magnani, F., Marchesi, C., Scoriels, L., Vogeley, K., Krebs, M.-O., & Tonna, M. (2024). Language in interaction: Turn-taking patterns in conversations involving individuals with schizophrenia. *Psychiatry Research*, *339*, 116102. https://doi.org/10.1016/j.psychres.2024.116102, PubMed: 39089189

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions: Series A, Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150203. https://doi.org/10.1098/rsta.2015.0203, PubMed: 26953183

Manera, V., Becchio, C., Schouten, B., Bara, B. G., & Verfaillie, K. (2011). Communicative interactions improve visual detection of biological motion. *PloS One*, *6*(1), e14594. https://doi.org/10.1371/journal.pone.0014594, PubMed: 21297865

Manera, V., Del Giudice, M., Bara, B. G., Verfaillie, K., & Becchio, C. (2011). The second-agent effect: Communicative gestures increase the likelihood of perceiving a second agent. *PLoS One*, *6*(7), e22650. https://doi.org/10.1371/journal.pone.0022650, PubMed: 21829472

Manera, V., Schouten, B., Verfaillie, K., & Becchio, C. (2013). Time will show: Real time predictions during interpersonal action perception. *PloS One*, *8*(1), e54949. https://doi.org/10.1371/journal.pone.0054949, PubMed: 23349992

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Moran, N., Hadley, L. V., Bader, M., & Keller, P. E. (2015). Perception of 'back-channeling'nonverbal feedback in musical duo improvisation. *PLoS One*, *10*(6), e0130070. https://doi.org/10.1371/journal.pone.0130070, PubMed: 26086593

Morency, L.-P., de Kok, I., & Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *International workshop on intelligent virtual agents* (pp. 176–190). https://doi.org/10.1007/978-3-540-85483-8_18

Müller, P., Dietz, M., Schiller, D., Thomas, D., Lindsay, H., Gebhard, P., André, E., & Bulling, A. (2022). Multimediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th Acm International Conference on Multimedia* (pp. 7109–7114). Association for Computing Machinery. https://doi.org/10.1145/3503161.3551589

Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current Directions in Psychological Science*, *16*(5), 269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x, PubMed: 19343102

Murray, L. (1985). Emotional regulation of interactions between two-month-olds and their mothers. *Social Perception in Infants*, 177–197.

Nackaerts, E., Wagemans, J., Helsen, W., Swinnen, S. P., Wenderoth, N., & Alaerts, K. (2012). Recognizing biological motion and emotions from point-light displays in autism spectrum disorders. *Plos One*, *7*(9), e44473. https://doi.org/10.1371/journal.pone.0044473, PubMed: 22970227

Neri, P., Luu, J. Y., & Levi, D. M. (2006). Meaningful interactions can enhance visual discrimination of human agents. *Nature Neuroscience*, *9*(9), 1186–1192. https://doi.org/10.1038/nn1759, PubMed: 16936721

Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature*, *395*(6705), 894–896. https://doi.org/10.1038/27661, PubMed: 9804421

Pavlova, M. A. (2012). Biological motion processing as a hallmark of social cognition. *Cerebral Cortex*, *22*(5), 981–995. https://doi.org/10.1093/cercor/bhr156, PubMed: 21775676

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026

Pavlova, M. A., Moosavi, J., Carbon, C.-C., Fallgatter, A. J., & Sokolov, A. N. (2023). Emotions behind a mask: The value of disgust. *Schizophrenia*, *9*(1), 58. https://doi.org/10.1038/s41537-023-00388-3, PubMed: 37709796

Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements. *Cerebral Cortex*, *15*(12), 1866–1876. https://doi.org/10.1093/cercor/bhi064, PubMed: 15746001

Pfeiffer, U. J., Schilbach, L., Jording, M., Timmermans, B., Bente, G., & Vogeley, K. (2012). Eyes on the mind: Investigating the influence of gaze dynamics on the perception of others in real-time social interaction. *Frontiers in Psychology*, *3*, 537. https://doi.org/10.3389/fpsyg.2012.00537, PubMed: 23227017

Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology*, *27*(20), 3162–3167. https://doi.org/10.1016/j.cub.2017.08.074, PubMed: 29033333

Q9 Poggi, I., D'Errico, F., & Vincze, L. (2010). Types of nods. the polysemy of a social signal. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association.

Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20–22, 2010. Proceedings 10* (pp. 146–158). Springer. https://doi.org/10.1007/978-3-642-15892-6_16

Regolin, L., Tommasi, L., & Vallortigara, G. (2000). Visual perception of biological motion in newly hatched chicks as revealed by an imprinting procedure. *Animal Cognition*, *3*(1), 53–60. https://doi.org/10.1007/s100710050050

Rochat, P. R. (2001). Social contingency detection and infant development. *Bulletin of the Menninger Clinic*, *65*(3), 347–360. https://doi.org/10.1521/bumc.65.3.347.19847, PubMed: 11531131

Scaliti, E., Pullar, K., Borghini, G., Cavallo, A., Panzeri, S., & Becchio, C. (2023). Kinematic priming of action predictions. *Current Biology*, *33*(13), 2717–2727. https://doi.org/10.1016/j.cub.2023.05.055, PubMed: 37339628

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing Discourse: Text and Talk*, *71*(93).

Schultz, J., & Frith, C. D. (2022). Animacy and the prediction of behaviour. *Neuroscience and Biobehavioral Reviews*, *140*, 104766. https://doi.org/10.1016/j.neubiorev.2022.104766, PubMed: 35798127

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, *105*(2), 809–813. https://doi.org/10.1073/pnas.0707021105, PubMed: 18174333

Smith, N. A., McDaniel, V. F., Ispa, J. M., & McMurray, B. (2023). Maternal depression and the timing of mother–child dialogue. *Infant and Child Development*, *32*(1), e2389. https://doi.org/10.1002/icd.2389

Takarae, Y., McBeath, M. K., & Krynen, R. C. (2021). Perception of dynamic point light facial expression. *American Journal of Psychology*, *134*(4), 373–384. https://doi.org/10.5406/amerjpsyc.134.4.0373

Todorova, G. K., Hatton, R. E. M., & Pollick, F. E. (2019). Biological motion perception in autism spectrum disorder: A meta-analysis. *Molecular Autism*, *10*, 49. https://doi.org/10.1186/s13229-019-0299-8, PubMed: 31890147

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A kinect study. *Cognition*, *180*, 38–51. https://doi.org/10.1016/j.cognition.2018.04.003, PubMed: 29981967

Truong, K. P., Poppe, R., Kok, I. d., & Heylen, D. (2011). A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proceedings Interspeech 2011* (pp. 2973–2976) International Speech Communication Association.

Wahn, B., Schmitz, L., Kingstone, A., & Böckler-Raettig, A. (2022). When eyes beat lips: Speaker gaze affects audiovisual integration in the mcgurk illusion. *Psychological Research*, *86*(6), 1930–1943. https://doi.org/10.1007/s00426-021-01618-y, PubMed: 34854983

Wehrle, S., Vogeley, K., & Grice, M. (2024). Backchannels in conversations between autistic adults are less frequent and less diverse prosodically and lexically. *Language and Cognition*, *16*(1), 108–133. https://doi.org/10.1017/langcog.2023.21

White, S. (1989). Backchannels across cultures: A study of Americans and Japanese. *Language in Society*, *18*(1), 59–76. https://doi.org/10.1017/S0047404500013270

Yu, H., Garrod, O. G. B., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics*, *36*(3), 152–162. https://doi.org/10.1016/j.cag.2011.12.002

Zilany, M. S. A., Bruce, I. C., & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *Journal of the Acoustical Society of America*, *135*(1), 283–286. https://doi.org/10.1121/1.4837815, PubMed: 24437768

Uncorrected Proof

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.400/2632257/opmi.a.400.pdf by guest on 22 October 2026