The social transfer function: how dynamic predictions of facial consequences drive judgements of social contingency

Rudradeep Guha[a,*], Pablo Arias Sarah[b,c], Jean-Julien Aucouturier[a]

[a] *Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000, Besançon, France*
[b] *School of Psychology and Neuroscience, Glasgow University, Glasgow, UK*
[c] *Department of Cognitive Science, Lund University, Sweden*

Abstract

Human social interactions abound with time-aligned multimodal information such as nods and eyeblinks, 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? We developed a novel experimental paradigm in which observers discriminate between video recordings of genuine and fake dyadic interactions based solely on the interplay between the speaker's speech and/or facial expressions and the listener's facial backchanneling cues. Using a combination of computational modeling using temporal response functions (TRFs) and behavioral data from two independent experiments (N=206), we show that observers perform above chance when recognizing genuine social interactions; that, to do so, they causally rely on the link between the speaker's speech and the listener's mouth and eye information; and that this inference is driven by time-aligned, dynamic predictions rather than average quantities of movement. In both experiments, judgements of social contingency are well-predicted by a computational model that evaluates the agreement of observed data with the output of a pre-learned "social transfer function" that dynamically predicts the facial consequences of a given speech signal. These results provide mechanistic insights into the features that contribute to perception of social contingency, and could potentially be used to identify markers of contingency in people with disorders of consciousness, autism and social anxiety.

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

*Corresponding author
Email address: r.guha@outlook.com (Rudradeep Guha)

1. Introduction

Our daily social interactions are dynamic and complex endeavours that require quick and efficient coordination in real-time. Given the need to continuously detect multimodal signals, integrate them with high-level cognitive inferences and produce appropriate responses within milliseconds, 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 of developmental psychology, 2-month-old infants show distress when interacting with their mothers via a pre-recorded video but not when engaged in a genuine real-time interaction (Murray, 1985), suggesting that they possess the ability to jointly process the expressive signals of the interlocutor and their own and how they should depend on one another. In adulthood, recognition of social contingency is often considered a prerequisite to higher-level social behaviour such as joint attention (Mundy and Newell, 2007), turn-taking in conversation and theory of mind (Frith and Frith, 2012).

In spoken conversations, contingent social behaviour often manifests itself through the phenomenon of backchanneling (Brunner, 1979, Knudsen et al., 2020). Backchanneling cues can be non-verbal, encompassing 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 deployed quickly and efficiently in conversations. Most research into facial signals has either looked at them in the context of emotion expression or the variety of semantic and pragmatic functions they play in conversations (Bavelas and Chovil, 2018). For instance, the temporal structure of blinks is important with short and long blinks signalling end-of-turn and understanding respectively (Hömke et al., 2017) while eyebrow raises purportedly serve to emphasise information (Flecha-García, 2010). However, comparatively few studies have investigated what specific backchanneling cues contribute to the detection of social contingency, and how.

To provide mechanistic insights into the perception of social contingency, a common strategy has been to degrade stimuli and only keep point-light displays of participant figures, and investigate which properties of these stimuli facilitate their detection. For instance, one such study found enhanced visual detection of a target agent within noisy point-light displays of two agents when the dyads were moving synchronously as compared to asynchronously despite the irrelevance of synchrony to the task (Neri et al., 2006). In another study, point-light displays of two musicians who were either improvising together or playing solo were spliced together to generate genuine and fake musical interactions (Moran et al., 2015), and participants were able to tell them apart 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 a 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., 2011a). 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 vocal and facial features such as those observed in real-world conversations (Takarae et al., 2021). Yet, we know that observers possess the uncanny ability to match the time-aligned dynamics of such cues to external stimuli (e.g., heartbeat: Galvez-Pol et al. (2022)). 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., 2013, 2011b), they do not attempt to operationalize this mechanism in a concrete model which can be used to make experimental predictions. What exact temporal prediction, of which specific expressive signals, 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 of contingent interactions, acquired over time by observation and participation, and which can generate real-time predictions of the temporal dynamics of a backchanneling cue of an agent in response to the speech of another agent (Figure 1). At the algorithmic level, we instantiate such a 'transfer function' using temporal response functions (TRFs; Crosse et al. (2016)), which assume linearity and time-invariance of the system and can therefore be represented by an impulse response H that is convolved with the input to generate the output ($Y = H \circledast X$). When observing A talking to B, we essentially propose that observers utilize something akin to 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 backchanneled). We then created genuine and fake extracts by replacing the real listener with another in half of the trials and extracted the time-aligned time-series of the vocal and facial cues of each participant using state-of-art signal processing algorithms. In two successive behavioural experiments (N=18 and N=188), human 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; 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.

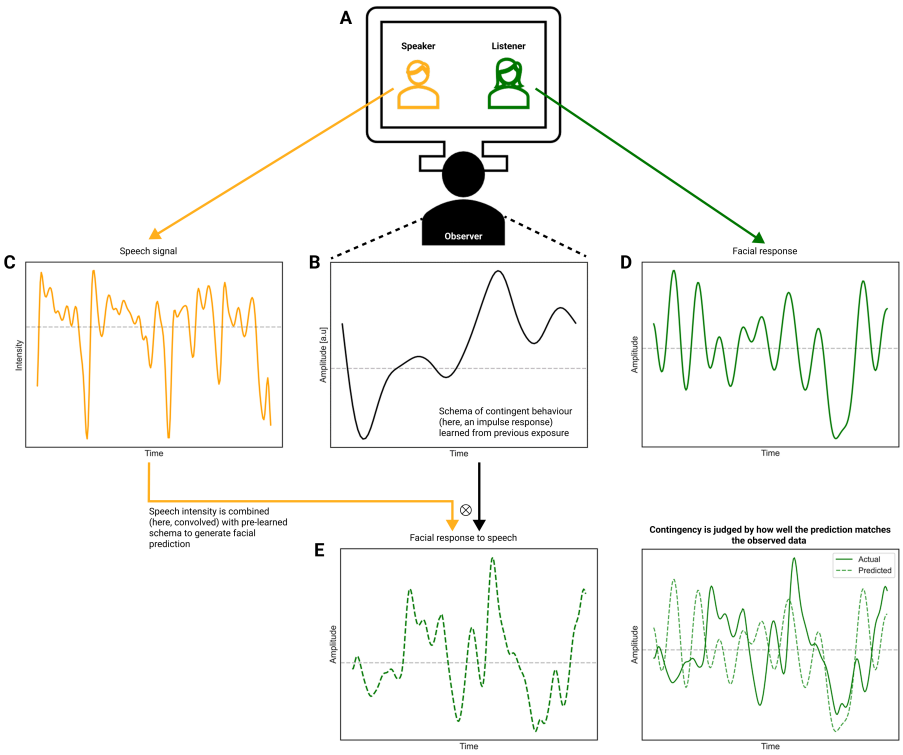

Figure 1: **The "social transfer function" computational modeling paradigm:** to operationalize how external observers judge the contingency of a social interaction (**A**), we propose that observers possess a schema of contingent interactions acquired over time by observation and participation (**B**). In this paper, we instantiate such a 'transfer function' using a temporal response function (TRF), i.e. a pre-trained impulse response that can be convoluted with the input signal to generate real-time predictions (**E**) of the temporal dynamics of an agent's backchanneling cue in response to the speech of another agent (**C**). This predicted output can then matched against the observed facial signal (**D**) 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 facial features they use to detect genuine and fake interactions.

2. Study 1

In Study 1 (conducted in the lab), we ask participants to discriminate between genuine and fake audiovisual interactions assembled from a dataset of ecological speed-dating conversations and explore whether their ratings are consistent with a social transfer function model predicting a listener's backchanneling cues from a speaker's speech.

2.1. Materials and Methods

2.1.1. 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. A priori power analysis using G*Power (Faul et al., 2009) determined a required sample size of 21 for 70% power for a medium effect at a significance level of $\alpha = .05$.

2.1.2. Stimuli

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).

Dataset participants: N=31 French-speaking participants (male=15; mean age=22 [20-27]) were part of the dataset collection. All participants were heterosexual, single, and 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.

Dataset procedure: 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.

Stimuli: From recorded conversations in the speed-dating dataset, we extracted n=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. 'Fake' interactions were created by putting together the recording of the original speaker with that of another listener, i.e. not the listener the speaker was actually talking to. This resulted in n=198 extracts (99 genuine and 99 fake).

Finally, for each genuine and fake interaction, we created 3 presentation 'modalities' of the same extract: one audio-video (thereafter: 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); one video-video (V-V), in which both speaker and the listener could be seen but not heard; and one audiovisual-video (AV-V) in which the speaker could be seen and heard while the listener could only be seen (Figure 2).

Dataset ethics: 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.

2.1.3. 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). Participant performance at the task was quantified using the d' sensitivity index.

2.1.4. Social transfer functions

To model how well genuine/contingent interactions matched a prediction of the temporal dynamics of the listener backchannel (facial cues) in response to the speaker's behaviour (speech), we used a combination of automated speech/face analysis and the system identification technique of *temporal response functions* (TRFs; Crosse et al. (2016)). First, we estimated the time series of perceived loudness from the speaker's speech in a given interaction, by computing the RMS intensity of the vocal signal on successive 100ms windows and processing it 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)). Then, we extracted the time series of 11 facial action units (AUs) (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. 1 eye, 1 head and 9 mouth-related AUs) from the listener's video, using the Py-feat library (Cheong et al., 2023), in such a way that both vocal and facial time series were synchronized at the same frame rate. Finally, for every AU, we trained a separate temporal response function (TRF) to model the transfer function that converts the speaker's speech into the listener's facial behaviour. TRFs were trained only on the subset of trials corresponding to genuine interactions, in order to model the dynamical relation between speech

and face that is found in ecological social interactions. The TRFs were trained using the ridge regression method as implemented in the mtrfpy toolbox (Bialas et al., 2023).

Once trained, a TRF allows predicting an observer's backchanneling response (a series of AU intensity) to a specific speaker's speech (a series of speech intensity), by convolving the input speech with the TRF, based on the regularities it managed to learn from the dataset. In any given interaction, the match between the series predicted by convolution with TRF and the actual observer's times series can be evaluated using Pearson's correlation coefficient r between the two time series.

2.1.5. Statistical analyses

Participant performance was tested for statistical difference from chance level (d'=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 the prediction of the TRF model, we compared Pearson correlation coefficients between the predicted and actual facial AU series (thereafter: *TRF fit*) between groups of genuine and fake trials with two-sample t-tests, corrected for multiple comparisons across the 11 AUs under consideration. In addition, an SVM classifier was trained on the TRF fits and used to predict whether a trial was genuine or fake to provide further evidence for physical differences between genuine and fake trials.

To test whether the TRF fits of trials predict observers' decision of genuineness, we regressed individual observer ratings on each trial (binary: 0/1) using a generalized (logistic) linear model (GLM) with a random effect on the observer (`response` $\sim$ `TRF fit + intensity + (1|observer)`). GLM analysis was performed with the pymer4 package (Jolly, 2018).

2.2. Results

Participants performed significantly above chance at discriminating genuine vs fake interactions (d' =0.53, $t(17) = 10.23, p < .001$). 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, smaller than A-V:$t(17) = 4.31, p < .001$; and AV-V:$t(17) = 4.11, p < .001$). There was no performance difference between the A-V and AV-V blocks ($t(17) = 0.33, p = 0.74$). On the whole, this pattern of results was consistent with the fact that observers in this task mostly relied on matching the facial features of the listener with the vocal features of the speaker.

We then tested the hypothesis that genuine and fake trials physically differed in terms of how well they matched a prediction of the temporal dynamics of the backchanneling cues of the listener in response to the speech of the speaker. To do so, we trained individual TRFs that linked the speaker's speech intensity with the listener's backchanneling signals, for every action unit (AU), across the subset of 99 genuine trials, and then compared the distribution of TRF

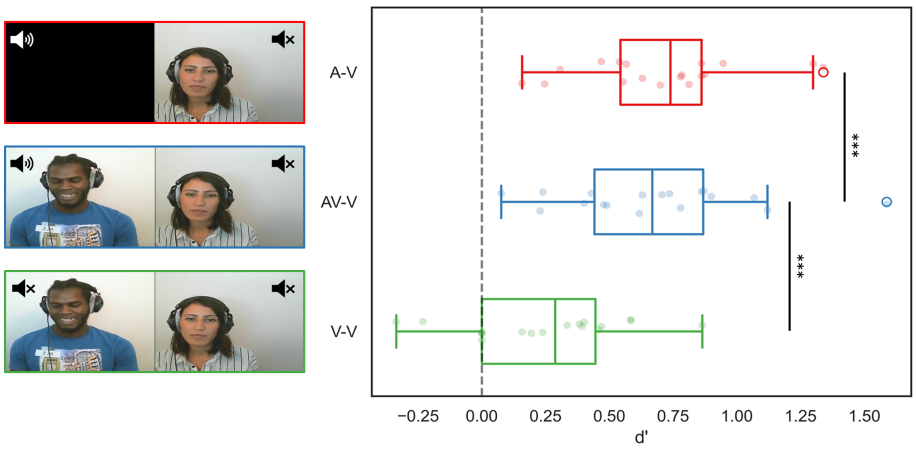

Figure 2: **Study 1. Left:** Observers were presented audiovisual extracts from speed-dating interactions in which only one person was talking while the other was silent and only displayed backchanneling cues. Trials were presented in three possible 'modalities': audio-video (A-V, **top**) in which the 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); audiovisual-video (AV-V, **middle**) in which the speaker could be seen and heard while the listener could only be seen; and video-video (V-V, **bottom**), in which both the speaker and the listener could be seen but not heard. **Right:** Sensitivity (d') over participants was significantly above chance in all modalities, with better performance in A-V and AV-V compared to V-V. Box-plot marking median values, inter-quartile range (IQR) and data points within 1.5 IQR. *** marks statistical significance at the 0.001 alpha level (paired t-tests)

fit between genuine and fake trials. Of the tested AUs, genuine trials had statistically larger TRF fits than fakes along 4 (all of them mouth-related) of them (AU12: ($t(196) = 2.78, p < .05$), AU25: ($t(196) = 2.62, p < .05$), AU26: ($t(196) = 2.63, p < .05$), AU28: ($t(196) = 2.92, p < .05$), as well as for head nods ($t(196) = 2.6, p < .05$). This suggested that the genuine and fake stimuli in our task indeed differed with respect to how much they matched pre-learned dynamic predictions of backchanneling, most apparently on listener nods and mouth reactions such as smiling.

We also tested whether such physical information was computationally sufficient to accurately discriminate genuine and fake trials, by training a machine-learning classifier on the trial's TRF fit. 5-fold cross-validation was used to train an SVM classifier on a predefined set of parameter values to find the optimal parameters, which were then used to fit the final classifier on the training set. On the testing set, the SVM reported a classification accuracy of 58%.

Finally, we tested whether human observers' behaviour was consistent with this information by predicting observed responses from both dynamic and static quantities of motion. Generalized linear models were statistically significant for the dynamic TRF fit (but not their static quantities) only for AU25 ($\beta = -0.75, p_{corrected} < .05$) and AU43 ($\beta = -0.78, p_{corrected} < .05$). This suggests that participants behaved as if they used dynamic prediction for cues in both the mouth area, as predicted above, as well as the eyes.

Observing the dynamics of the AUs used to discriminate between genuine and fake contingent behaviour (AU25 and AU43) reveals that both TRFs contain strong early negative components around 300ms, and that their peak activations are offset by around 1s, with AU25 peaking early at ∼1s and around ∼2s for AU43. We also see AU43 activity being inhibited for almost the entire duration of AU25 activation (shaded area in Figure 5).

2.3. Discussion

Study 1 investigated observers' ability to detect contingent behaviour in dyadic interactions. We manipulated trials such that they contained varying amounts of multimodal signals and found that participants performed above chance in all modalities, with the best results when observing a speech-to-face configuration. Finally, we tested whether genuine trials could be recognized, both by humans and machines, based on dynamic "transfer-function" predictions of backchanneling and found that they predicted observer ratings over and beyond what could be predicted by static quantities of motion, based on the listeners' mouth and eye action units.

The fact that participants performed above chance at the task confirms that detecting social contingency is a robust human ability, one that is plausibly used as a building block for higher-level social cognitive functions (Frith and Frith, 2012). Observing the absence or asynchrony of interactive responses in a conversation could be considered the third-person equivalent of the classic 'still face' paradigm of developmental psychology, in which adults interacting with infants are asked to freeze and cease to respond for a set period. It was shown that infants from around 4 weeks show sensitivity to such disruptions (Happé

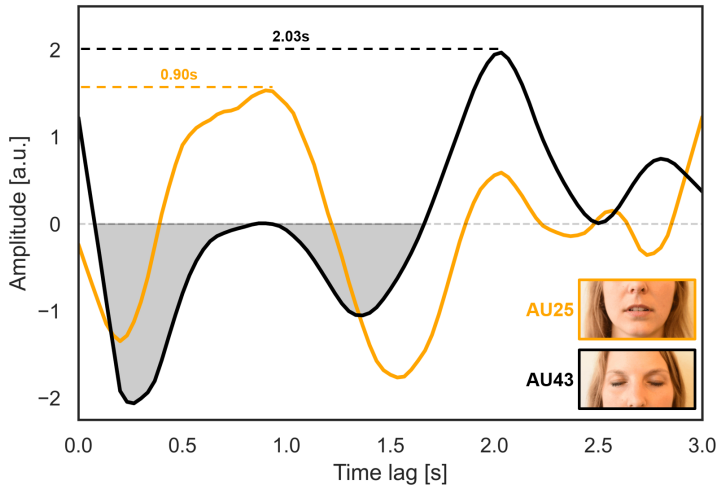

Figure 3: **The expected facial dynamics of social contingency:** Two temporal response functions (TRFs) allowed statistically significant prediction of observer decisions of genuineness, based on both a mouth- (AU25, *lip part*, orange) and an eye-related action unit (AU43, *eyes closed*, black). Comparison of these TRFs, or impulse responses (x-axis: time, y-axis: amplitude), reveal different expected timings for contingent facial responses in each of these AUs. Both TRFs contain strong early negative components around 300ms but their peak activations are offset by 1s, with AU25 peaking early at ∼1s and AU43 peaking around ∼2s. We also see AU43 activity being inhibited for almost the entire duration of AU25 activation (shaded area).

and Frith, 2014), and it may therefore only appear logical that adults should also perform well at a similar task. In the present task though, the manipulated contingencies were not plain interruptions but rather desynchronized behaviour in which backchanneling from one conversation was paired with another unrelated conversation. The robust sensitivity of adults to such ecological variations suggests that contingency is a graded evaluation built on cumulative evidence of synchronized or desynchronized behaviour. It should be noted, however, that the good performance achieved in this experimental paradigm (mean d'=0.53) should not be taken as a psychophysical measure of sensitivity, as fake interactions were paired "as found" in the speed-dating dataset, and may vary in terms of the perceptual evidence in favour of contingency or the lack thereof. Study 2 will attempt to replicate these results in a dataset with more controlled task difficulty.

In our task, participants performed worse in the silent V-V modality than in the other two and they did not perform more accurately when provided the speaker's video (AV-V) in addition to its recorded speech (A-V). This pattern of results appears at odds with a large literature suggesting a facilitating effect of multimodal signals in social cognitive judgements such as emotion recognition or mimicry (Krumhuber et al., 2023). For instance, a study with a similar paradigm investigated whether multiple modalities in face-to-face dyadic interactions facilitate or interfere with language processing (Drijvers and Holler, 2023). To test this, they had 30-second extracts of a speaker talking to their conversation partner uninterrupted and presented the trials in three conditions: audiovisual (AV), audiovisual + mouth blurred (AB), and audio only (AO). Participants were better at shadowing speech when they received multimodal signals suggesting that they had a facilitatory effect and did not increase cognitive load. Results in the present paradigm are likely explained by the fact that the task required comparing two simultaneous streams of data (a speaker's and a listener's) from a third-person perspective. In such a situation, simultaneous video modalities (AV-V, VV) require spatially dividing one's attention among the two ongoing streams (looking left, looking right) leading to difficulties processing cues of asynchrony between the two. On the other hand, the A-V modality requires processing the alignment of sound with a single video stream which is comparable to judging multimodal signals from a single talking head, and may therefore lead to better performance (and no advantage upon further adding the speaker's video information). It is interesting to ponder whether such cognitive limitations in e.g. judging the contingency between two concurrent visual streams may have lead to the development of abilities that favour the detection of speech-to-face over face-to-face coordination and whether the preference for one modality or another depends on the timescale of the coordination: fast (milliseconds) for facial backchannelling, plausibly slower for other types of joint action explored in previous dyadic visual tasks (Neri et al., 2006, Moran et al., 2015).

TRF analysis of the speaker's speech loudness and the listener's facial action units revealed that genuine interactions were characterized by systematic 'social transfer functions', predominantly at mouth action units (AUs 12, 25, 26 and

28) and head nods. TRFs peaked between 1.5-2s for the majority of mouth AUs, and at ∼2.5s for head nods (Figure 1-D), which suggests slower dynamics for the latter. This pattern of results is consistent with previous descriptions of the dynamics of backchanneling in the non-verbal behaviour literature (Hömke et al., 2018, Boudin et al., 2024). Moreover, the dynamics of the AUs important for perceiving contingency (AU25 and AU43) reveal the inhibitory behaviour of blinks until the offset of AU25. It is possible that blinks that would normally have occurred are suppressed, suggesting that blinks could function as an index of the end of an expression.

The fact that genuine and fake trials differ in how well they match TRF fits for these AUs does not imply, of course, that observers actually use that information to do the task. Here, we have presented two separate streams of evidence that speak to this question. First, we used a machine classifier to show that TRF fit provides sufficiently discriminating information to reach similar levels of performance as human observers. While such machine arguments do not conclusively indicate that observers use the same cues, they provide an important proof-of-possibility that these cues would support such an inference if they did (for similar arguments, see e.g. (Goupil and Aucouturier, 2021, Piazza et al., 2017, De Boer and Kuhl, 2003)). Second, we found that observer judgements of genuineness, regardless of correctness, correlated with TRF fit, over and beyond static quantities of motion at AUs 25 and 43. While such correlations suggest that trials that match dynamic predictions of facial consequences are the same trials that observers also judge more likely to be genuine, they remain descriptive and do not provide a formal test of causality (Casadevall and Fang, 2008). For instance, it could be that while genuine trials indeed contain TRF-predictable eye or mouth backchanneling, they also provide other cues either at locations (e.g. pupil size (Hess and Petrovich, 2014, Kret, 2018, Goswami et al., 2020)) or at dynamical scales that are not captured by AUs and the TRF methodology used here. Consequently, perhaps it is this latter information that influences observer ratings. Study 2 below will provide a more causal test of the influence of the eye or mouth region in the perception of contingency by using dynamic masks to prevent observers from processing information in these regions.

Finally, the current analysis left some ambiguity as to what exact cues are used by observers in the task: while machine classifiers suggest that genuine and fake trials did not differ in terms of eye-TRF fit (but only in terms of mouth predictions), both mouth- and eye-TRF fits correlated with human observer ratings. Because of the correlational nature of these results, this may indicate a number of relations between these variables: e.g., that both face regions are in fact discriminative and utilized, but in a way that is not captured by our automated AU analysis; that only mouth information is useful but that observers are also biased to use eye information (even if counterproductive); that mouth and eye predictions are ecologically correlated in the dataset, etc. While all of the relations can in principle be explored by further correlational analysis, Study 2 below will address the question more conclusively by presenting stimuli that only contain one or the other type of information to a new, larger sample of participants. If eye-TRF fit is not discriminative, then performance should

collapse when presented with eye-only trials.

3. Study 2

Study 1 established that observers were able to discriminate between backchanneling in genuine and fake interactions, and showed that dynamic predictions of the facial consequences of speech based on prelearned "social transfer functions" (i.e. TRF fit) in the mouth and eye were consistent with such judgements. It potentially provides a mechanism explaining the detection of social contingency in human observers (but also leaves ambiguity about whether both mouth and eye information is actually utilized and/or useful).

Study 2 aims to replicate these results, and provide a more conclusive causal test of this hypothesis, by presenting a new, larger sample of participants with stimuli manipulated with dynamical visual masks to present only eye or mouth-area dynamic information. In addition, Study 2 also controls the baseline difficulty of the task by selecting equal numbers of correctly and incorrectly-recognized stimuli (based on the ratings of Study 1 participants).

3.1. Materials and Methods

Participants: We recruited N=188 participants through Prolific in a between-subject design with approximately 65 participants in each condition ($N_{eyes} = 61, male = 39, M = 31.53, SD = 9.76; N_{mouth} = 67, male = 39, M = 31.17, SD = 10.66; N_{original} = 63, male = 36, M = 30.84, SD = 10.77$). 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 the minimum sample size required in each group to be $n = 64$ to obtain 80% power for detecting a medium effect at $\alpha = .05$.

Stimuli selection: 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 n=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 n=30 stimuli controlled for difficulty in each of the four signal-detection categories, resulting in a total of 120 A-V stimuli (see Supplemental Information for detail).

Stimulus manipulation: 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 the DaVinci Resolve software (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 ("eye" condition), or only the mouth ("mouth" condition).

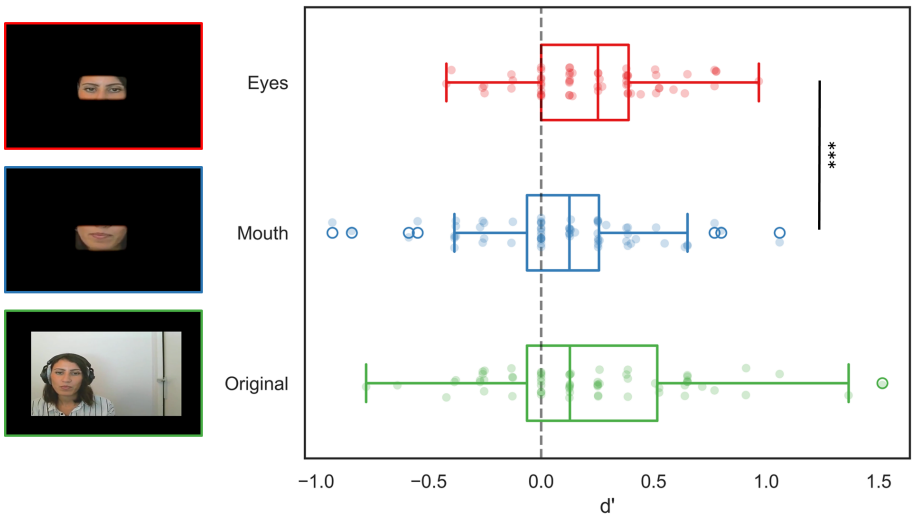

Figure 4: **Study 2. Left:** AV-V trials from Study 1 were manipulated by creating dynamic visual masks that isolated specific parts of the listener's video while hiding everything else. This yielded 3 different versions of each trial where the speaker's audio was played over a video that featured either the complete face area ('Original', **bottom**), only the eye region ('Eyes' condition, **top**), or only the mouth ('Mouth' condition, **middle**). **Right:** Sensitivity (d') was significantly better for participants in the Eyes condition than in the Mouth condition and viewing the trials in the Original condition, i.e. the full non-masked videos, conferred no performance advantage over the conditions with manipulation.

Procedure: Participants were presented with 40 stimuli in either one of the three conditions, between subjects (eyes: N=61; mouth: N=67; original: N=63). In each condition, the task was the same as in Study 1 with participants watching the videos and rating each interaction as either genuine or fake (1-interval, 2-alternative forced choice). A previous version of this task was piloted with n=20 offline participants and a within-subjects design as opposed to between-subjects, but was changed due to the discovery of order effects.

3.2. Results

Results replicated the results of Study 1, with performance significantly above chance for the original, full-information videos ($d' = 0.21, t(62) = 3.70, p < .001$). Performance was also above chance for the eyes condition ($d' = 0.24, t(60) = 6.06, p < .001$), with no difference from original videos ($t(122.0) = 0.46, p = 0.65$), but significantly greater than in the mouth condition ($t(126.0) = 2.57, p < .05$). The mouth condition was not significantly above chance ($d' = 0.09, t(66) = 1.84, p = .07$), but it wasn't significantly lower than the original condition either ($t(128.0) = -1.73, p = 0.09$).

We further reproduced the TRF analysis of Study 1 in the original condition. Because the masking in both manipulated conditions rendered Py-feat unable to detect faces for subsequent AU analysis, we only analysed stimuli in the original

condition. We used generalized linear models to test whether participant responses in the original condition correlated with the TRF fit and average intensity of AU25 and AU43 (`response` $\sim$ `AU25 fit + AU25 intensity + AU43 fit + AU43 intensity + (1|participant)`) and found only AU25 TRF fit ($\beta = 1.39, p < .001$) and AU43 TRF fit ($\beta = 0.70, p < .001$) to be significant predictors.

3.3. Discussion

By adopting a causal manipulation design in which we isolated either eye or mouth information in a more controlled subset of stimuli from Study 1, Study 2 provided a strong test of observers' use of information in the eye and mouth regions and provided causal evidence that participants can use either eye or mouth-region information to judge social contingency in conversations.

In addition, we 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 and (to a lesser extent) the mouth provide any discriminating information for contingency (replicating the only 2 Bonferroni-corrected AU predictors in Study 1), but also that participants did not utilize the *interaction* between the eye and mouth to any avail. This suggests that dynamic predictions of eye and mouth activity constitute redundant cues/signals for the aim of detecting social contingency, a property that contrasts with other types of facial inferences which typically utilize a dynamic and complementary hierarchy of signals over time (Jack et al., 2014).

Moreover, Study 2 replicated the results seen in Study 1 in that the TRF fit of both AU25 and AU43 correlated with participant ratings in the original condition and static intensity information did not. Taken together, this pattern of results strongly suggests that dynamic predictions of facial consequences in both the eye and mouth regions of listeners constitute a mechanism for third-party observers judging social contingency.

In particular, Study 1 left some ambiguity about whether dynamic eye information was used or even useful. Results in Study 2 established that it was indeed the case and that eye-only performance was significantly better than looking only at the mouth. This result is therefore consistent with TRF predictions in Study 1 and in the 'original' condition of Study 2, but not with the physical comparisons and machine classifications of Study 1 which showed that stimuli only differed on mouth-AU predictions. One reason might be that the action unit detector used in this study only provides data for one eye-related AU (AU43), but several different AUs for the mouth. Subsequently, the model may fail to capture important from the eyes (e.g. gaze direction - (Conty et al., 2006, Cañigueral and Hamilton, 2019, Wahn et al., 2022)), which our human observers are instead able to exploit.

4. General Discussion

While previous research has repeatedly shown that detecting contingency in conversational backchanneling is a robust human ability, and that it is likely an important precursor to developing higher-level social cognitive skills, very little is known about what specific backchanneling cues contribute to the detection of social contingency, and how. In this article, we introduced a novel behavioural paradigm in which participants were asked to identify genuine contingent behaviour in recorded video interactions. We manipulated both the contingency (genuine or fake) and the nature of information present in the interactions, either through different audio-visual modalities (Study 1) or by masking parts of the listeners' faces (Study 2). Consistent between the two studies, our results showed that observers perform above chance when recognizing genuine social interactions; that, to do so, they causally rely on the link between the speaker's speech and the listener's mouth and eye information; and that this inference is driven by time-aligned, dynamic predictions rather than average quantities of movement. In both experiments, judgements of social contingency are well-predicted by a computational model that evaluates the agreement of observed data with the output of a pre-learned "social transfer function" that dynamically predicts the facial consequences of a given speech signal.

The fact that, across two experiments and two independent samples of participants (N=18 and N=180), we found replicated evidence that participants were above chance at discriminating fake from genuine backchanneled interactions, even when severely degraded to contain only part of the face, confirms that social contingency detection is a robust social-cognitive capacity in adult observers - and that our paradigm is a robust task to study this capacity. In particular, observers were able to do the task even when the speaker's face was masked (Study 1, A-V condition) and showed no drop in performance when the listener's video only featured a small rectangle of dynamic information around the eye, or the mouth region (Study 2). This suggests that observers have developed highly redundant models of contingency that can exploit partial information and are therefore adaptive to a variety of interactional circumstances. This is at odds with other forms of facial signalling such as the inference of emotional expressions, which often critically depend on the availability of one single cue to disambiguate alternative 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 higher-level forms of social inferences.

Both studies found repeated evidence that to detect contingency, observers relied on dynamic predictions of facial consequences. This was manifest in 3 types of correlational analyses showing that such predictions discriminated genuine from fake trials; that they provided enough information for a machine classifier to do the task; that participant responses correlated with how well backchanneling signals in the eye and mouth AUs were predicted dynamically, but not with their average activity. Study 2 also provided causal manipulations

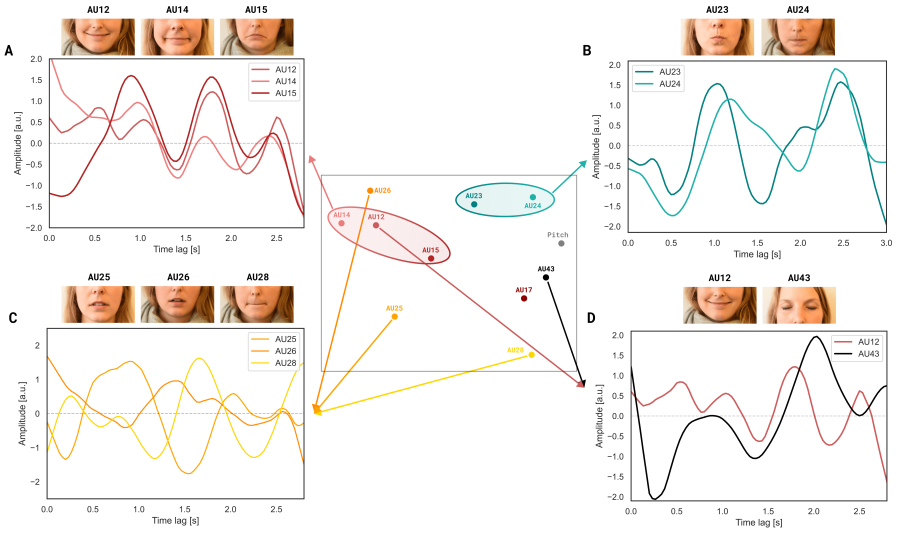

Figure 5: **The complex choreography of facial expressions for social communication.** TRFs for each AU projected onto low-dimensional Euclidean space using multidimensional scaling such that the similarity/distance between each in high-dimensional input space is maintained. (**A and B**) highlight that the dynamics of mouth-related AUs converge to a similar temporal structure, while (**C**), on the other hand, reveals the distinct but sequential nature of the activations of AUs 25, 26 and 28 shining a light on the complex choreography involved in the composition of facial expressions for social communication. (**D**) shows the dynamics of AU12 (smile) and AU43 (blink) with inhibition of blinks prior to a smile followed by blink onset simultaneously with smile offset.

to confirm that information restricted to these two face regions was sufficient to do the task. Taken together, this pattern of results strongly suggests that pre-learned models that enable the dynamic prediction of facial consequences in the eye and mouth regions of listeners constitute a mechanism by which third-party observers judge social contingency - which we proposed as 'social transfer functions'.

In this work, we implemented such social transfer functions using temporal response functions (TRFs; Crosse et al. (2016)). While they make strong modeling assumptions on the system (more below), TRFs offer a particularly parsimonious representation of what a predictive model of backchanneling could look like: namely, an impulse response to which the speaker's input speech is convolved to generate the listener's facial output. Once trained, social TRFs can be compared e.g. across action units, or dyads/individuals. Here, using training data from an ecological dataset of speed-dating conversations, we were able to derive TRFs for predictive action units (AUs) in both the eye and mouth regions.

Apart from contingency, the TRFs obtained also provide insight into the temporal dynamics of action units in naturalistic conversations. We show that TRFs can capture the dynamics of facial expressions allowing the grouping of

expressions or AUs that have similar dynamics such as those of AUs 12, 14 and 15 with comparable peak latencies at around 1.8s (Figure 5A), as well as AUs 23 and 24 with both containing multiple peaks at 1 and 2.5s (Figure 5B). The remarkably similar temporal structure of some of these AUs could potentially be difficult to disentangle and confound the outputs of the increasingly popular automated AU detection models. Other than grouping AUs based on similar dynamics, TRFs also allow comparisons between the chronometry of AUs that are supposed to be physiologically related - for example, Figure 5C reveals that AU25, AU26 and AU28 are activated sequentially over the course of 1s peaking at 0.9s, 1.4s and 1.7s respectively. Comparing AUs such as these with varying temporality can also shine a light on interesting relationships between them as in the case of AU12 (smile) and AU43 (blink) (Figure 5D) where we see a blink being inhibited prior to the onset of a smile and ultimately occurring not quite after the smile offset but simultaneously with it, an observation in line with the literature on the temporal coordination of smiles and blinks (Trutoiu et al., 2013, Rupenga and Vadapalli, 2016). More generally, we show that observing the dynamics of AUs with this methodology can facilitate the discovery of 'groups' of seemingly disparate AUs and provide insights into how the complex choreography of facial expressions compose meaningful social signals.

Beyond their descriptive interest, social TRFs are also useful as analytical or generative tools. Analytically, how well TRF predictions fit observed data (as implemented here e.g. with Pearson's correlation) provides a way to quantify the realism/typicality of backchannel dynamics. This could be used to quantify abnormal conversational dynamics (e.g. turn-taking patterns in schizophrenia - (Lucarini et al., 2024)), or as a security measure to detect forged AI videos (Li et al., 2018). If combined with modern facial animation techniques in avatars (Yu et al., 2012) or real-life videos (Arias-Sarah et al., 2024), social TRFs can also be used to generate novel stimuli that have specific dynamics, either for experimental control (e.g. animate the eyes as if they were driven by the mouth TRF) or to improve synthetic media (e.g., manipulate the perceived contingency of deep-faked conversations in human-computer interaction - (Kaate et al., 2023)).

More generally, 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 et al., 2011b). While the mechanism in this study provides accurate predictions of e.g., how much a trial appears genuine, or what part of a trial is used for such inference, this study leaves several important questions unanswered. First, we trained TRFs from an ecological dataset of interactions assuming the existence of a similar schema of conversational contingency in observers developed through previous exposure and participation in social interactions. It remains an open question how this learning may operate, and how plastic it may be to e.g. changing interactional cultures. For instance, there is debate about whether backchanneling conventions that are not shared across cultures (e.g. how much feedback one is expected to give) contribute to misunderstanding or stereotyp-

ing, such as being too impatient or, contrarily, unresponsive (White, 1989), and this process of cross-cultural adaptation to what constitutes appropriate contingency can be framed as a social transfer-function learning problem.

Second, it is also unknown how social transfer functions are cognitively represented, or even if they are at all. While we argue here that convolution with an impulse response is a particularly parsimonious form for representing an input-output mapping (a computational 'trick' also exploited in so-called convolutional deep learning architectures (Mallat, 2016)), it remains an experimental question whether human observers indeed encode and decode interactions this way. In particular, impulse responses make strong assumptions on the input-output phenomenon (namely, linearity and time-invariance, (Keesman, 2011)), and therefore fail to represent a potentially large class of backchanneling behaviour that may have e.g. non-linear, threshold-like qualities. It would be interesting to compare TRFs with more advanced transfer function learning techniques in their ability to predict contingency judgements, to achieve better understanding of what human 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 these may also be alternative mechanisms that future work could evaluate.

Finally, while this work has focussed on the detection of social contingency, it is interesting to question whether social transfer functions also support other types of social-cognitive inferences that rely on dynamic predictions of conversational backchanneling, such as judging familiarity (Gráczi and Bata, 2010), agreement (Müller et al., 2022) or even enjoyment (Li et al., 2010) in interactions. By providing a parsimonious representation of conversational dynamics which can be learned from each individual, social transfer functions such as TRFs appear promising as a way to study both how these constructs are signalled in ecological behaviour, and to model how they are detected by observers. From a clinical perspective, social transfer functions may also provide valuable insights into disorders affecting conversational dynamics such as autism spectrum disorders (ASD) (Wehrle et al., 2024) or parental depression in caregiver-child interactions (Smith et al., 2023) and disorders of consciousness (Hermann et al., 2018).

5. Acknowledgments

Work funded by ANR SOUNDS4COMA and conducted in the framework of the EIPHI Graduate school (ANR-17-EURE-0002 contract).

References

R. Adolphs, F. Gosselin, T. W. Buchanan, D. Tranel, P. Schyns, and A. R. Damasio. A mechanism for impaired fear recognition after amygdala damage.

Nature, 433(7021):68–72, 2005.

P. Arias-Sarah, D. Bedoya, C. Daube, J.-J. Aucouturier, L. Hall, and P. Johansson. Aligning the smiles of dating dyads causally increases attraction. *Proceedings of the National Academy of Sciences*, 121(45):e2400369121, 2024.

J. Bavelas and N. Chovil. Some pragmatic functions of conversational facial gestures. *Gesture*, 17(1):98–127, 2018.

S. Benghanem, R. Guha, E. Pruvost-Robieux, J. Levi-Strauss, C. Joucla, A. Cariou, M. Gavaret, and J.-J. Aucouturier. Cortical responses to looming sources are explained away by the auditory periphery. *Cortex*, 2024.

O. Bialas, J. Dou, and E. C. Lalor. mtrfpy: A python package for temporal response function analysis. *Journal of Open Source Software*, 8(89):5657, 2023.

A. Boudin, S. Rauzy, R. Bertrand, M. Ochs, and P. Blache. How is your feedback perceived? an experimental study of anticipated and delayed conversational feedback. *JASA Express Letters*, 4(7), 2024.

T. B. Brazelton, E. Tronick, L. Adamson, H. Als, and S. Wise. Early mother-infant reciprocity. In *Ciba Foundation Symposium 33-Parent-Infant Interaction*, pages 137–154. Wiley Online Library, 1975.

L. J. Brunner. Smiles can be back channels. *Journal of personality and social psychology*, 37(5):728, 1979.

R. Cañigueral and A. F. d. C. Hamilton. The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in psychology*, 10: 560, 2019.

A. Casadevall and F. C. Fang. Descriptive science. *Infection and immunity*, 76 (9):3835–3836, 2008.

J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang. Py-feat: Python facial expression analysis toolbox. *Affective Science*, 4(4):781–796, 2023.

C. A. Coey, M. Varlet, and M. J. Richardson. Coordination dynamics in a socially situated nervous system. *Frontiers in human neuroscience*, 6:164, 2012.

L. Conty, C. Tijus, L. Hugueville, E. Coelho, and N. George. Searching for asymmetries in the detection of gaze contact versus averted gaze under different head views: a behavioural study. *Spatial vision*, 19(6):529–546, 2006.

M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor. The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, 10: 604, 2016.

R. Dale, R. Fusaroli, N. D. Duran, and D. C. Richardson. The self-organization of human interaction. In *Psychology of learning and motivation*, volume 59, pages 43–95. Elsevier, 2013.

B. De Boer and P. K. Kuhl. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4):129–134, 2003.

L. Drijvers and J. Holler. The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, 30(2):792–801, 2023.

F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

M. L. Flecha-García. Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in english. *Speech communication*, 52(6):542–554, 2010.

C. D. Frith and U. Frith. Mechanisms of social cognition. *Annual review of psychology*, 63(1):287–313, 2012.

A. Galvez-Pol, S. Antoine, C. Li, and J. M. Kilner. People can identify the likely owner of heartbeats by looking at individuals' faces. *cortex*, 151:176–187, 2022.

S. Goldberg. Social competence in infancy: A model of parent-infant interaction. *Merrill-Palmer Quarterly of Behavior and Development*, 23(3):163–177, 1977.

M. Goswami, M. Manuja, and M. Leekha. Towards social & engaging peer learning: Predicting backchanneling and disengagement in children. *arXiv preprint arXiv:2007.11346*, 2020.

L. Goupil and J.-J. Aucouturier. Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, 212:104661, 2021.

T. Gráczi and S. Bata. The effect of familiarization on temporal aspects of turn-taking: a pilot study. *Acta Linguistica Hungarica*, 57(2-3):307–328, 2010.

F. Happé and U. Frith. Annual research review: Towards a developmental neuroscience of atypical social cognition. *Journal of Child Psychology and Psychiatry*, 55(6):553–577, 2014.

B. Hermann, G. Goudard, K. Courcoux, M. Valente, S. Labat, L. Despois, J. Bourmaleau, L. Richard-Gilis, F. Faugeras, S. Demeret, et al. "doc-feeling": a new behavioural tool to help diagnose the minimally conscious state. *bioRxiv*, page 370775, 2018.

K. S. Hermans, O. J. Kirtley, Z. Kasanova, R. Achterhof, N. Hagemann, A. P. Hiekkaranta, A. Lecei, L. Zapata-Fonseca, G. Lafit, R. Fossion, et al. Capacity for social contingency detection continues to develop across adolescence. *Social Development*, 31(3):530–548, 2022.

E. H. Hess and S. B. Petrovich. Pupillary behavior in communication. In *Nonverbal behavior and communication*, pages 327–348. Psychology Press, 2014.

P. Hömke, J. Holler, and S. C. Levinson. Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, 50 (1):54–70, 2017.

P. Hömke, J. Holler, and S. C. Levinson. Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS one*, 13(12):e0208030, 2018.

R. E. Jack, O. G. Garrod, and P. G. Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192, 2014.

E. Jolly. Pymer4: connecting r and python for linear mixed modeling. *Journal of Open Source Software*, 3(31):862, 2018.

I. Kaate, J. Salminen, S.-G. Jung, H. Almerekhi, and B. J. Jansen. 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*, pages 1–12, 2023.

K. J. Keesman. *System identification: an introduction*. Springer Science & Business Media, 2011.

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

B. Knudsen, A. Creemers, and A. S. Meyer. Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, 11:593671, 2020.

M. E. Kret. The role of pupil size in communication. is there room for learning? *Cognition and Emotion*, 32(5):1139–1145, 2018.

E. G. Krumhuber, L. I. Skora, H. C. Hill, and K. Lander. The role of facial movements in emotion recognition. *Nature Reviews Psychology*, 2(5):283–296, 2023.

H. Z. Li, Y. Cui, and Z. Wang. Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. *International journal of psychological studies*, 2(1):25, 2010.

Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018.

E. Lindboom, A. Nidiffer, L. H. Carney, and E. C. Lalor. Incorporating models of subcortical processing improves the ability to predict eeg responses to natural speech. *Hearing research*, 433:108767, 2023.

V. Lucarini, M. Grice, S. Wehrle, F. Cangemi, F. Giustozzi, S. Amorosi, F. Rasmi, N. Fascendini, F. Magnani, C. Marchesi, et al. Language in interaction: turn-taking patterns in conversations involving individuals with schizophrenia. *Psychiatry Research*, 339:116102, 2024.

S. Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

V. Manera, C. Becchio, B. Schouten, B. G. Bara, and K. Verfaillie. Communicative interactions improve visual detection of biological motion. *PloS one*, 6(1):e14594, 2011a.

V. Manera, M. Del Giudice, B. G. Bara, K. Verfaillie, and C. Becchio. The second-agent effect: communicative gestures increase the likelihood of perceiving a second agent. *PLoS One*, 6(7):e22650, 2011b.

V. Manera, B. Schouten, K. Verfaillie, and C. Becchio. Time will show: real time predictions during interpersonal action perception. *PloS one*, 8(1):e54949, 2013.

N. Moran, L. V. Hadley, M. Bader, and P. E. Keller. Perception of 'backchanneling'nonverbal feedback in musical duo improvisation. *PLoS One*, 10 (6):e0130070, 2015.

L.-P. Morency, I. De Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *International workshop on intelligent virtual agents*, pages 176–190. Springer, 2008.

P. Müller, M. Dietz, D. Schiller, D. Thomas, H. Lindsay, P. Gebhard, E. André, and A. Bulling. Multimediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7109–7114, 2022.

P. Mundy and L. Newell. Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5):269–274, 2007.

L. Murray. Emotional regulation of interactions between two-month-olds and their mothers. *Social perception in infants*, pages 177–197, 1985.

E. Nackaerts, J. Wagemans, W. Helsen, S. P. Swinnen, N. Wenderoth, and K. Alaerts. Recognizing biological motion and emotions from point-light displays in autism spectrum disorders. *PLOS One*, 7(9), e44473, 2012.

P. Neri, J. Y. Luu, and D. M. Levi. Meaningful interactions can enhance visual discrimination of human agents. *Nature neuroscience*, 9(9):1186–1192, 2006.

M. A. Pavlova, J. Moosavi, C.-C. Carbon, A. J. Fallgatter, and A. N. Sokolov. Emotions behind a mask: the value of disgust. *Schizophrenia*, 9(1):58, 2023.

E. A. Piazza, M. C. Iordan, and C. Lew-Williams. Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology*, 27(20):3162–3167, 2017.

R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen. Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 146–158. Springer, 2010.

P. R. Rochat. Social contingency detection and infant development. *Bulletin of the Menninger Clinic*, 65(3: Special issue):347–360, 2001.

M. Rupenga and H. B. Vadapalli. Investigating the temporal association between eye actions and smiles. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6. IEEE, 2016.

N. A. Smith, V. F. McDaniel, J. M. Ispa, and B. McMurray. Maternal depression and the timing of mother–child dialogue. *Infant and child development*, 32 (1):e2389, 2023.

Y. Takarae, M. K. McBeath, and R. C. Krynen. Perception of dynamic point light facial expression. *The American Journal of Psychology*, 134(4):373–384, 2021.

L. C. Trutoiu, J. K. Hodgins, and J. F. Cohn. The temporal connection between smiles and blinks. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

B. Wahn, L. Schmitz, A. Kingstone, and A. Böckler-Raettig. When eyes beat lips: speaker gaze affects audiovisual integration in the mcgurk illusion. *Psychological Research*, 86(6):1930–1943, 2022.

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

S. White. Backchannels across cultures: A study of americans and japanese1. *Language in society*, 18(1):59–76, 1989.

H. Yu, O. G. Garrod, and P. G. Schyns. Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3):152–162, 2012.

M. S. Zilany, I. C. Bruce, and L. H. Carney. Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America*, 135(1):283–286, 2014.