Doctoral Thesis from Université Marie et Louis Pasteur

The Social Transfer Function
A system-identification approach to studying social contingency

Author:
Rudradeep GUHA

Supervisors:

Jean-Julien AUCOUTURIER	DR CNRS, FEMTO-ST
Pablo ARIAS-SARAH	Lecturer, School of Psychology and Neuroscience, University of Glasgow

Reviewers:

Judith HOLLER	Associate Professor, Donders Institute for Brain, Cognition & Behaviour, Radboud University
Etienne THORET	CR CNRS, Institut de Neurosciences de la Timone, Aix-Marseille University

Examiners:

Catherine PELACHAUD	DR CNRS, ISIR, UPMC
Louise GOUPIL	CR CNRS, LPNC, Université Grenoble Alpes
Renaud SÉGUIER	Professeur, IETR, CentraleSupélec

Thèse de doctorat de l'ED 37: Sciences Physiques pour l'Ingénieur et Microtechniques (SPIM)
Defended in Besançon on 17 DECEMBER 2025

Abstract

The ability to detect social contingency, i.e. recognizing a specific behaviour as the social consequence of another, is believed crucial in the development of social cognition as a mechanistic prerequisite for abilities such as joint attention, turn-taking and theory of mind (Frith and Frith, 2012). Yet, while the related ability for perceiving biological motion in a single individual is now relatively well-understood (Neri et al., 1998), questions remain about how the brain infers temporal causality of socially meaningful motion in two-person interactions: what exact temporal prediction, of which specific expressive signal, has to break down before the observer of a social interaction decides that the interaction isn't quite right?

In this thesis, we develop the concept of a 'social transfer function' — a cognitive representation that, we propose, allows observers of interactions to predict how speech signals will influence facial backchanneling over time. Operationalized using simple algorithms adopted from the system-identification literature, these social transfer functions provide a computational model of what constitutes social contingency. We show that 1) observers can reliably discriminate between genuine and manipulated contingent behaviour even when stimuli are severely degraded and 2) they causally rely on the link between a speaker's speech and signals from a listener's mouth and eye regions. Furthermore, we characterize observers' internal representations of contingent smiles in response to speech as social transfer functions and, using reverse correlation, extract them in a data-driven manner. We conclude that social transfer functions can be used to operationalize third-party observers' internal representations of social contingency, such that they can be learned from data, make predictions and can be tested quantitatively. This work opens up new avenues for research by quantifying conversational dynamics and grants the ability to investigate these dynamics across different conversational contexts, cultures and in disorders affecting communication like autism spectrum disorder.

Résumé

La capacité à détecter la contingence sociale, c'est-à-dire à reconnaître qu'un comportement est la conséquence sociale d'un autre comportement, est considérée comme cruciale dans le développement de la cognition sociale. En effet, cette capacité est un mécanisme nécessaire pour des fonctions cognitives telles que l'attention conjointe, la tenue de tour de parole ou la théorie de l'esprit (Frith and Frith, 2012). Cependant, alors que les mécanismes qui permettent de reconnaitre une agentivité biologique dans le mouvement d'un seul individu sont désormais relativement bien compris (Neri et al., 1998), nous savons encore très peu sur la manière dont le cerveau perçois la causalité temporelle socialement significative dans les interactions entre deux personnes : quelle exacte prédiction temporelle, de quel signal expressif précisément, doit être réfutée pour que l'observateur d'une interaction sociale décide qu'elle n'est pas tout à fait correcte ?

Dans cette thèse, nous développons le concept de «fonction de transfert sociale », une représentation cognitive qui, nous le proposons, permet aux observateurs d'interactions de prédire comment les signaux d'un agent influenceront le backchanneling de l'autre agent au fil du temps. Opérationnalisées à l'aide d'algorithmes simples tirés de la littérature automaticienne sur l'identification des systèmes, ces fonctions de transfert sociales fournissent un modèle computationnel de ce qui constitue la contingence sociale. Nous montrons que 1) les observateurs peuvent distinguer de manière fiable les comportements contingents authentiques des comportements manipulés, même lorsque les stimuli sont fortement dégradés, et 2) qu'ils s'appuient de manière causale sur le lien entre la parole d'un locuteur et les signaux provenant de la bouche et des yeux d'un auditeur. En outre, nous caractérisons les représentations internes des observateurs des sourires contingents en réponse à la parole comme des fonctions de transfert sociales et, à l'aide d'une technique de "corrélation inverse", nous les extrayons de façon purement "basée données". Nous concluons que les fonctions de transfert sociales peuvent être utilisées pour opérationnaliser les représentations internes de la contingence sociale des observateurs, qu'elles peuvent être apprises à partir de données, et qu'elle permettent de faire des prédictions qui peuvent être

testées quantitativement. Ce travail ouvre de nouvelles perspectives de recherche en quantifiant la dynamique conversationnelle et permet ainsi d'étudier comment cette dynamique est modulée dans les troubles affectant la communication, tels que les troubles du spectre autistique, ainsi que dans différentes cultures et différents contextes conversationnels.

Acknowledgements

First of all, I would like to thank my doctoral advisors, JJ and Pablo, for their unwavering support throughout the PhD and the words of encouragement despite the (many) mistakes I made. JJ made this PhD an unexpectedly stress-free experience by being there for it all, from small things like helping draft emails to big things like preventing nervous breakdowns before presentations (and reviewing the thesis manuscript at 2 in the morning while on summer vacation). Pablo was always available to help, provided valuable feedback and his overall excitement and enthusiasm for the work was infectious.

I am very grateful to my thesis committee for agreeing to review this work and, through their own research, being an important source inspiration.

My thanks also go to all my past and present colleagues here at FEMTO-ST – Paul, Aynaz, Coralie, Paige, Zhenxing and Manaoj.

Most of all, I would like to thank my parents and my brother who have worked hard and sacrificed so much for me to have the opportunity to do a PhD in the first place.

Contents

List of Figures

Chapter 1

Social Cognition

The study of human cognition - how the mind and brain make sense of the world and of itself - has a rich history of scholarship spanning multiple disciplines and theoretical frameworks. It has relied on successive paradigms to understand mental processes, including computational approaches that viewed the mind as an information processor, symbolic models that emphasized rule-based representations, and connectionist frameworks inspired by neural networks (Rumelhart et al., 1986; Varela, 1996). More recently, cognition has come to be regarded from the perspective of systems science (Thompson, 2010), which emphasizes the dynamic, emergent nature of cognitive phenomena. From this perspective, cognition can be understood as an emergent property of a system composed of more granular systems such as those underlying perception, attention, memory and action planning. The study of cognition thus becomes a mereology, or the study of part-whole relationships (Pessoa, 2023). In other words, the system is understood not only through the study of each part independently, but also their relationship with each other and how the interactions between them shape the system.

This is especially apparent in the case of social cognition, where the systems-science perspective can be used to characterize individuals as 'systems', with interactions between them playing a fundamental role in their development and how they experience the world. While a wealth of research has deepened our understanding of the cognition of individuals, understanding how it is shaped by social interactions requires new methodological approaches that can capture the temporal and dynamic nature of social interactions. The work presented in this thesis addresses this need through the development of 'social transfer functions', drawing from system-identification and control-engineering concepts to model mental representations of interactional dynamics.

1.1 The nature of social cognition

Social cognition has witnessed an increase in attention and, as a consequence, has shifted the locus of cognition away from the individual and towards the dynamic networks formed between individuals through social interactions (Schilbach et al., 2013; Redcay and Schilbach, 2019). While social cognition was initially studied in terms of individuals' perception of social cues, e.g. how an observer decodes emotions from facial expressions (Jack and Schyns, 2015), modern accounts conceptualize social cognition as an emergence or synergy of multiple systems with two dominant accounts: *enaction* and the *'we-mode' theory*. According to enactivists, cognition is a relational process or a way to make sense of the environment and the world around us by interacting with it, and so social interactions function as 'participatory sense-making' (De Jaegher and Di Paolo, 2007). Importantly, enactivists argue that social interactions are autonomous systems that arise from such participatory sense-making and that they cannot be reduced to the properties and contributions of the individuals. On the other hand, while the proponents of *we-mode* agree with the claim of cognition being irreducible and an interactive relational process, they reject the idea of social interaction as simply sense-making but *with others*. Instead, they argue for the existence of differences between making sense of the world *with* others and making sense of others within the world. They argue that "sociality is not just co-presence" (Gallotti and Frith, 2013), i.e. that social interactions activate cognitive representations which are not available in settings involving individual action or involving other individuals but requiring no intention of acting jointly and engaging in interactive behaviour. This so-called *we-mode* encodes representations that remain latent until a socially interactive context is encountered, meaning that, despite being a property of individuals, it cannot be understood in a first-person or isolated context.

Despite differences, both accounts advocate for a greater role for social cognition in the study of individual cognition, claiming that an account of perception or cognition which ignores the effects of social interactions is incomplete (Gallagher, 2009). The claim is supported by an increasing number of studies showing increased perceptual sensitivity for detecting noisy patterns at locations where perception is shared with another (Seow and Fleming, 2019) or impaired proactivity of gaze patterns towards the motor target of the other when it is out of the other's reach (but within reach of the participants themselves) implying the adoption of shared spatial representations in social situations that then influence individual perception (Costantini et al., 2012). For instance, in a dot perspective task, participants tasked with counting the number of dots they can see in an image of a room with dots on the walls and an avatar facing one of the walls respond significantly faster when the avatar sees the same number of dots, i.e. shares the same perspective (Samson et al., 2010). Such social influence on perception and cognition, also termed co-perception (Deroy et al., 2024) is theorized to ground the conscious perception of individuals in an objective shared reality (Frith, 2025).

1.2 Predictive coding in social cognition

While these accounts of social cognition attempt to elucidate the nature of the social cognitive system, a comprehensive account also requires understanding the mechanisms of its component systems. Consider a relatively simple dyadic interaction with a speaker and a listener. As the speaker talks, the listener processes the acoustic features of the speaker's speech, which are then promoted to high-level areas of the brain where the semantic content is parsed. This occurs in parallel with visual processing of the speaker's facial expressions, gaze direction and body posture (Holler and Levinson, 2019). The incoming multi-modal information is combined to generate the content of the listener's response and the appropriate facial cues (for instance, to signal the intention to interrupt) by recruiting language-production and motor-planning areas. These processes also interact with other processes, such as memory and attention, while still constantly integrating the ongoing stream of information from the speaker. All of this happens, on average, in under 400ms (Heldner and Edlund, 2010). How is such large complexity resolved at such disproportionately small temporal scales, given, for instance, latencies in the order of 600ms for word production alone (Indefrey and Levelt, 2004)? This apparent discrepancy can be explained by *predictive coding*, a framework that interprets the perceptual system as a hierarchical generative model aiming to minimize prediction errors (Rao and Ballard, 1999; Friston, 2003). Predictions generated internally by high-level areas of the brain are passed down the cortical hierarchy to low-level sensory areas, where mismatches between the prediction and the data produce so-called “prediction errors". These prediction errors then ascend the cortical hierarchy to modify and improve the model generating the prediction. Predictive coding eschews the notion of the brain trying to find a mapping between the external environment and internal states in favour of the notion of the brain attempting to *infer* the external environment from its effects on internal states. Perception can thus be thought of as a “controlled hallucination" (Clark, 2015; Millidge et al., 2021). The illusory nature of the external world or the causes of sensory signals is particularly evident in the case of social interactions, where the external environment being modelled includes constructs that are not directly perceptible, like the mental states and intentions of others. From a predictive coding perspective, the small latencies for producing complex behaviours in social interactions may be facilitated by mental representations of others' behaviour in terms of their beliefs, intentions and knowledge (Koster-Hale and Saxe, 2013), a capacity referred to as theory of mind (Premack and Woodruff, 1978). However, this raises the issue of infinite regress: i.e. the dyadic interaction would require the speaker to model the brain of the listener and the listener to model the speaker's brain, in turn implying that the speaker needs a model of the listener that includes their model of the speaker and so on ad infinitum.

One solution to this representational problem casts the interacting agents as coupled dynamical systems (Rulkov et al., 1995), meaning that if the interacting agents share similar generative models of social behaviour then knowing the state of one allows predicting the states of the other, irrespective of how distinct the state trajectory of each agent is (Friston and Frith, 2015). Dynamical coupling (or, as Friston calls it, *generalised synchrony*) thus reformulates the problem of inferring mental states of others from their actions to the more tractable problem of inferring our own mental states given some observed action and transposing them onto the other. With each participant in an interaction minimizing the prediction error between the self and the other, their generative models gradually become similar, arguably leading to the development of mutual understanding (Friston and Frith, 2015; Mayo and Shamay-Tsoory, 2024). The ability to use the actions of one agent to predict both the content and temporality of a second agent's actions, referred to as 'interpersonal predictive coding' (Manera et al., 2011), is sensitive to individual differences as well as small changes in the timing of communicative actions (Manera et al., 2013, 2011). Casting cognitive processes in both isolated and social contexts as predictive-coding dynamical systems has enabled the re-conceptualization of disorders like schizophrenia (Fletcher and Frith, 2009; Adams et al., 2013; Okruszek et al., 2018, 2019), alexithymia (Palser et al., 2018) and social anxiety disorder (Gerrans and Murray, 2020) (see Smith et al., 2021 for a comprehensive review). Neurodevelopmental disorders affecting communication and social interactions, such as autism spectrum disorder (ASD) in particular, have seen predictive-coding flavoured characterizations. For instance, studies show a lack of flexibility in processing violations of expectations and learning from dynamic noisy stimuli in individuals with ASD (Van de Cruys et al., 2014), as well as a sub-optimal increase in weighting (or precision) of prediction errors generated by bottom-up processes compared to predictions generated by top-down processes (Haker et al., 2016), implying a greater focus on the details of the environment or stimuli instead of on the high-level meaning conveyed by them.

1.3 Development of social contingency

Apart from demonstrating the versatility of predictive coding as a unifying framework for understanding the mechanisms of perceptual and social processes in the brain, these findings also highlight one of the key aspects of social interactions - namely, *social contingency* or the behaviour of one agent contingent upon the behaviour of another. Because the modelled environment is populated by a dense network of social signals, often with distinct dynamics (e.g. frequent, rapid blinks contrasted with the slower unfolding of smiles or nods), developing better generative models requires efficient extraction of socially relevant cues from a constant stream of multimodal signals, integrating them with high-level inferences and producing appropriate responses within milliseconds. Isolating the relevant cues and detecting social contingency (from a predictive-coding perspective, the dynamical coupling between signals), thus becomes fundamental to social interaction (Coey et al., 2012; Dale et al., 2013). The facilitatory effect of social contingency in interactions has been demonstrated by studies showing enhanced comprehension in individuals listening to dialogues than when listening to monologues (Fox Tree, 1999; Branigan et al., 2011) and better learning performance through live lectures involving interactions (and contingency) with the lecturer compared to pre-recorded lectures (De Felice et al., 2021).

This ability to detect and process social contingency begins to develop from infancy through caregiver-child interactions, with 2-month-old infants displaying social expectations and awareness of their caregivers as communicative partners (Rochat, 2001). This is well-established by several influential paradigms in developmental psychology, like the 'still-face' paradigm (Fig. 1.1) in which 2-month-old infants display increased gaze aversion and general negative affect when the caregiver becomes unresponsive and maintains a neutral expression (Tronick et al., 1978). In a similar 'double video' paradigm, infants again show distress when interacting with their mothers via a pre-recorded video but not when en-

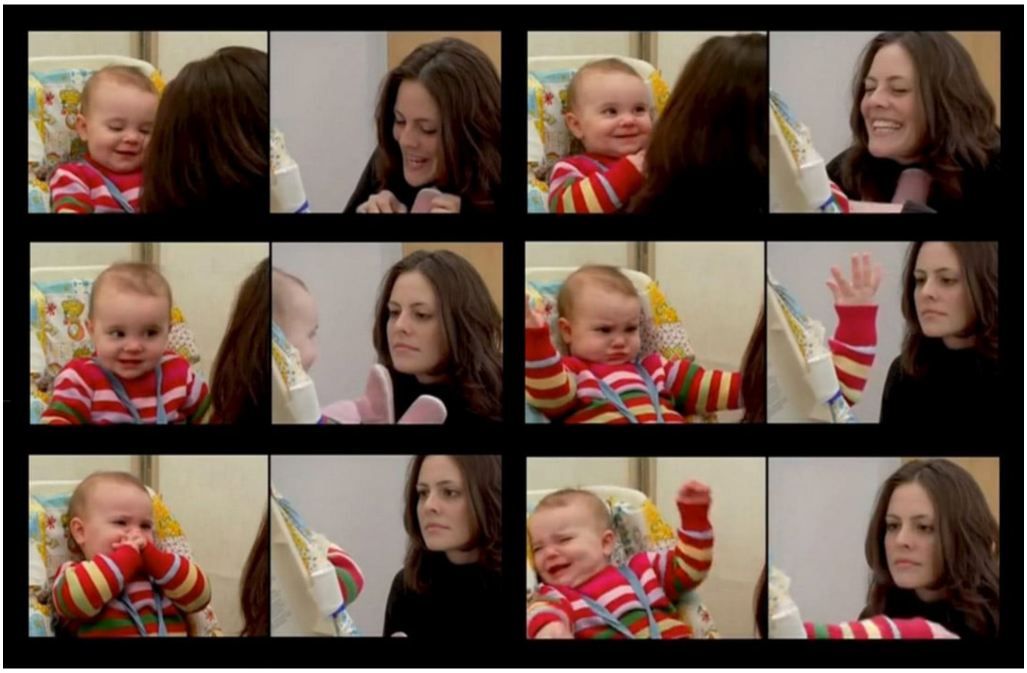


Figure 1.1: **The still-face experiment.** In a seminal study demonstrating infants' ability to detect social contingency, when caregivers engaged in normal interactions with their children suddenly become unresponsive and adopt a 'still face', infants are quick to perceive this change and display increased negative affect and gaze aversion.

Figure based on work by Tronick et al. (1978) and adapted from Save the Children's Resource Center (2022).

gaged in a genuine real-time interaction (Murray, 1985), suggesting that they possess the ability to jointly process expressive signals generated by themselves, the interlocutor, and how they should depend on one another. By 7-9 months, infants in still-face experiments begin to take an intentional stance and attempt to re-engage their caregivers through actions aimed at attracting attention (Striano and Rochat, 1999). At ages between 7 and 12 months, infants also start to represent increasingly complex social concepts, manifesting in preference for pro-social behaviour (Hamlin et al., 2007), attribution of intentions to others' actions and prediction of future actions based on those intentions (Kovács et al., 2010; Brandone, 2015). Older infants at around 24 months also display observational causal learning (Meltzoff et al., 2012), i.e. learning causality by observing the behaviour of others and initiating interventions to generate similar outcomes. Infants display such learning in a social context by learning the causal connection between one person shaking an object and another person producing a marble and subsequently performing the correct intervention to procure a marble for themselves (Waismeyer and Meltzoff, 2017). The capacity for understanding cause-effect relationships is significant because contingency detection is a necessary but not sufficient condition for developing a sense of agency and attributing agency to others in the environment (Beier and Carey, 2014). Social cognitive abilities continue to develop through adolescence, with late adolescents performing better than early and mid- adolescents at detecting contingency or the coordination of dynamic real-time behaviour in interacting dyads (Hermans et al., 2022).

1.4 Social contingency and psychopathology

In developmental psychology, social-cognitive processes are often considered to be developmentally associated through direct causal links, such that the processes develop sequentially (i.e. the development of one process is necessary for the development of the other). This model of developmental cascade (Masten et al., 2005), combined with the long developmental course of social competencies, has prompted significant research into connections between dysfunctions of pre-social behaviour in early stages of life and subsequent social atypicality. Autism, often characterized by disruptions in social interactions, has received particular interest. In studies investigating preferential attention towards biological motion, arguably a precursor to attributing intentions to others and engaging in typical social interactions, children between 2 and 7 years old with autism fail to recognize point-light displays of biological motion and instead focus on non-social physical contingencies in the stimuli that are ignored by the control group (Klin et al., 2009; Annaz et al., 2012). Children with autism also display impaired ability to evaluate and interpret whether others' social behaviour is appropriate in a given context (Mazza et al., 2017), a crucial component of processing social cues and selecting appropriate responses, which themselves constitute a significant proportion of the complexity of a socially contingent interaction. Likewise, adults with high-functioning autism also do not show any significant improvement at detecting point-light displays of communicative gestures over non-communicative gestures, with behavioural evidence pointing towards impairments in predictive-coding mechanisms as being responsible (Von Der Lühe et al., 2016). Other studies with neurotypical populations also demonstrate how social contingency between infants and their caregivers influences attachment styles later in life. One study shows that both low and high levels of contingency can lead to insecure attachment and that there is an optimal level that facilitates the development of a secure attachment style (Beebe et al., 2010). This is supported by a longitudinal study showing that children of postnatally depressed mothers (often characterized by disruption in contingent behaviour) are at significantly greater risk of depression (Murray et al., 2011), while another study shows high gaze focus and indistinguishable affect in both contingent and non-contingent interactions in infants of mothers suffering from postnatal depression as compared to infants of non-depressed mothers (Skotheim et al., 2013). The evidence highlights how the socio-cognitive abilities of children are affected by atypical situations and environments consisting of disturbed or insufficient social contingency, and how such deficiencies may cascade into psychopathology later in life (see Happé and Frith (2014) for a review of atypical social cognition across different stages of childhood and Matyjek et al. (2025) for a review of differences in social behaviours in autistic and neurotypical populations).

1.5 Signals and cues of social contingency

Assessing the quantity and quality of social contingency in a systematic manner first requires isolating the signals and cues that make up a contingent interaction. Despite stereotypical characterizations of speakers and listeners as active senders and passive receivers, respectively, everyday conversations require rapid and regular signals from listeners to speakers. To facilitate active contingent interactions, listeners often produce short, non-verbal and verbal feedback signals called *backchannels* (Yngve, 1960).

Non-verbal backchannels, including facial expressions, nods and eye gaze, have begun to receive significant interest. For instance, though smiles are commonly conceived of as emotional expressions (Barrett et al., 2019), they also serve a pragmatic purpose by contextualizing semantic content as ironic or humorous (Bavelas and Chovil, 2018). Likewise, listeners also produce 'feedback smiles', although these are typically shorter to be unobtrusive (non-feedback expressions in general last approximately 1s longer than feedback expressions - Jensen, 2015). Like smiles, nods are also versatile signals whose meaning varies with their temporal organization (Poggi et al., 2010); slow nods in coordination at delays of approximately 600ms are thought to be a form of mimicry promoting social affiliation, while fast nods are usually employed as backchannels by listeners to signal engagement or agreement (Hale et al., 2020). Perhaps surprisingly, even subtle cues like blinks provide feedback to speakers, with two types of blinks (short and long) proposed as serving different purposes. While short blinks occur reliably at the end of turns where speakers typically look to the listener for feedback, the more infrequent long blinks provide more high-level feedback like understanding (Hömke et al., 2017). The duration of blinks is a significant factor, with a 400ms change in a listener's blink duration resulting in the shortening of the speaker's response by the order of several seconds, despite the speaker failing to consciously register the signal (Hömke et al., 2018). In addition to non-verbal backchannels, interactions also feature verbal backchannels, which are usually broken down into three categories: non-lexical, phrasal and substantive. Non-lexical feedback consists of utterances like 'hmm' and 'uh-huh' and are semantically meaningless but signal engagement. Phrasal backchannels such as 'really?' and 'are you serious?', though semantically meaningful, are interpreted as signals of acknowledgement. Substantive backchannels are the most semantically rich and involve repetition, conversational repair, summary statements and clarifying questions. Together, these verbal and non-verbal backchannels help develop a shared grounding between individuals and create interactions out of “collective monologues" (Piaget, 2005).

Though backchanneling frequencies and the cues from the speaker that elicit those backchannels vary across individuals (Blomsma et al., 2024) and cultures (Li et al., 2010), backchannels remain ubiquitous (Heinz, 2003). Due to their semantic simplicity (allowing them to be deployed and processed quickly and efficiently), they are believed to play a crucial role in the procedural coordination of interactions by facilitating simultaneous speech planning and processing without substantially increasing cognitive load or causing interference (Knudsen et al., 2020). This is supported by studies showing less backchanneling (both verbal and non-verbal) in individuals with autism than in neurotypical individuals (Rifai et al., 2022) and lower frequency and variability in backchanneling behaviour in dyadic interactions between individuals with autism compared to neurotypical controls (Wehrle et al., 2024).

Backchannels, which are associated with listeners or addressees in an interaction, also have a counterpart on the speaker's side: co-speech gestures (Kendon, 1996). But unlike backchannels, co-speech gestures are tightly coupled with speech in terms of their temporal coordination and the information they seek to convey by, for instance, accompanying the phrase “and it was pushed away" with the forward motion of hands. Early theories classified such gestures as either self-directed (i.e. for the benefit of the speaker themself by facilitating lexical access) or listener-directed (i.e. possessing a communicative purpose) but recent work shows mimicry of co-speech gestures in face-to-face interactions helps achieve mutually shared understanding, thereby suggesting that the two theories may not be mutually exclusive (Holler and Wilkin, 2011).

1.6 Methods for studying social contingency

The near-constant flow of these signals, overlapping in the domains of time, frequency and function, renders the study of social contingency rather difficult (Vinciarelli et al., 2009). Perhaps unsurprisingly, methods for investigating social signals in experimental settings have traditionally attempted to reduce the complexity of naturalistic social interactions to a more manageable degree. For instance, in a perceptual crossing experiment, social contingency is distilled down from its multi-modal nature to a more minimalist setting involving only interaction through haptic feedback (Auvray et al., 2009). Another common strategy has been to degrade stimuli and use point-light displays (i.e. minimal representations of body movements with a smaller set of landmarks often displayed as white points on a dark background) to investigate the specific stimulus properties that facilitate their detection. For instance, one study showed enhanced visual detection of a target agent within noisy point-light displays of two agents when they moved synchronously than when they moved asynchronously (Neri et al., 2006). Another study shows that participants are able to discriminate between point-light displays of two musicians who are either improvising together or playing solo, and that this ability remains consistent even in the absence of any music or musical expertise (Moran et al., 2015) (for the related question of recognizing biological motion in a single body, see Nackaerts et al., 2012). 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).

To this end, there has been a growing body of research attempting to develop real-time manipulation of multimodal signals. One review (Arias et al., 2021) highlights the difficulties associated with manipulating specific acoustic features in speech and identifies a few constraints

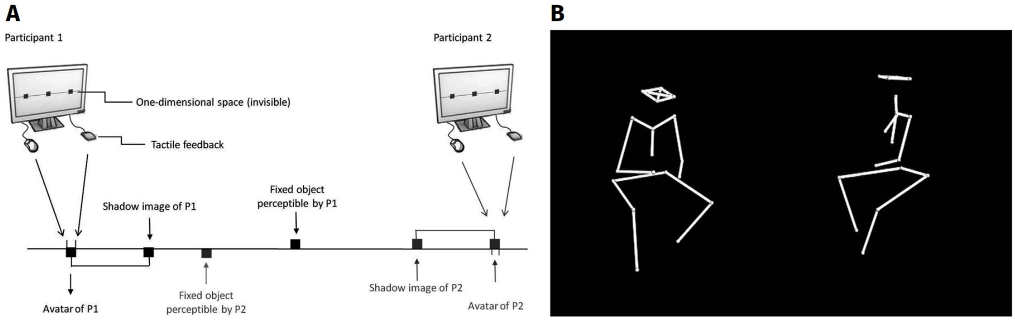


Figure 1.2: **(A) The Perceptual Crossing experimental paradigm.** Pairs of blindfolded participants are placed in separate rooms and interact in a common virtual one-dimensional perceptual space. Each participant moves a cursor (an avatar representing her body) along a line and receives a tactile stimulus to the free hand when encountering something on the line. Participants are asked to click a mouse button when they perceive the presence of the other participant. Apart from each other, participants can encounter a static object or a displaced "shadow image" of the partner that is strictly identical with respect to shape and movement characteristics. Therefore, the only difference between the partner and their shadow image is that the former can at the same time perceive and be perceived, i.e., that there can be live dyadic interactions. A solution to the task has to rely at least partially on performing and detecting a live interaction.
Figure and description adapted from Auvray and Rohde (2012).
(B) An example of a still image taken from a video point-light display of a real musician duo in Moran et al. (2015)

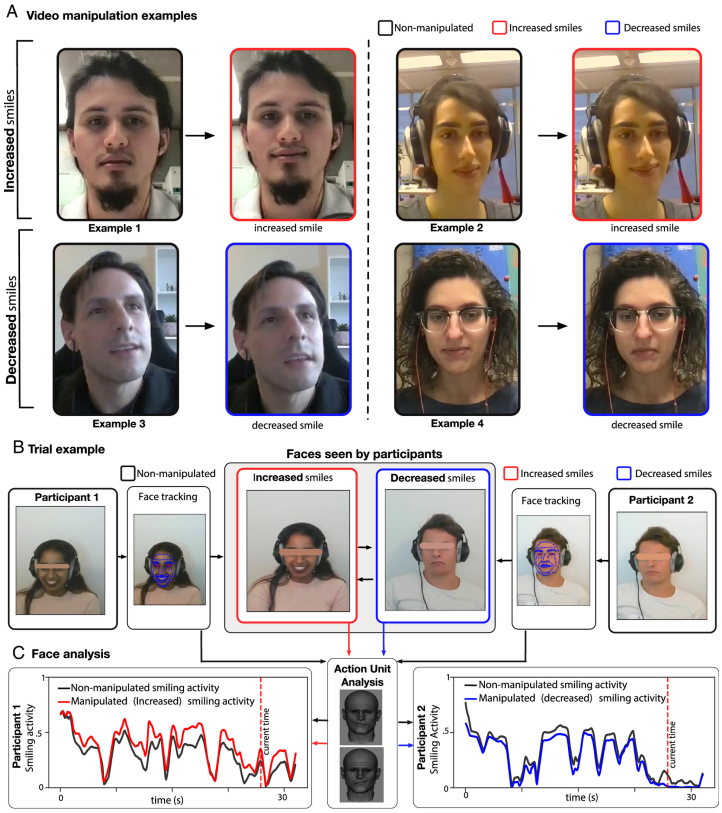


Figure 1.3: **Real-time manipulation of smiles using Ducksoup.** (A) Face manipulation examples. Nonmanipulated faces (black) and the corresponding increased (red frame) and decreased (blue frame) smile manipulation examples. (B) Schematics of the experimental paradigm and facial expression analysis. Participant 1 nonmanipulated (black frame) face is tracked and manipulated to increase her smile (red frame). In parallel, participants' 2 original facial expression (black frame) is tracked and manipulated to decrease his smile (blue frame). Participants only see the manipulated videos of their interacting partners and not their own (shaded gray box); the bar over the face is to preserve anonymity. (C) After the experiments, we use video recordings to extract participants' manipulated (red and blue) and nonmanipulated (black) smiling activity over time. Note that the manipulation only changes the time series on the y-axis and by a small amount, i.e., the manipulation is a static shift in smiling activity levels. The horizontal red bar indicates the moment in the interaction when pictures were taken.
Figure and description adapted from Arias-Sarah et al. (2024).

for the use of voice transformation algorithms in experimental settings, with a key one being the ability to generate real-time transformations. A recent study introduced an online experimental platform called DuckSoup for real-time face transformations (Arias-Sarah et al., 2024). The platform was used to causally investigate the effect of smile manipulations on the emergence of romantic attraction in real-time 'speed dating' interactions. The manipulations were either aligned (both smiles were either increased or decreased) or misaligned (smile of one was increased while the other's was decreased) and occurred without the participants' knowledge (Fig. 1.3). Results indicated that increasing the smile of the other person caused participants to think that person was more attracted to them and that increasing the smiles of both at the same time led to greater attraction between them and increased their perception of the conversation quality. This paradigm opens up new avenues of research by going beyond the correlational nature of traditional interaction analyses and circumventing issues related to low ecological validity associated with virtual avatars. However, the question of how individuals represent the dynamics of specific social signals remains an open question. For instance, in response to what specific cues and at what latencies does the perception of a smile shift from being affiliative to signalling romantic attraction? Though such tools now allow the manipulation of contingency in real-time, *how* specific social signals need to be manipulated to make them appear contingent remains mostly unknown.

Roadmap

In this chapter, we cast interacting individuals as coupled systems attempting to find a shared representation of the world and described the dominant theories about how this common ground is converged upon. Furthermore, we detailed the signals involved in everyday social interactions and how maladaptive learning and recognition of those may manifest as psychopathology. Finally, we highlighted some of the experimental paradigms used to study social cognition, with recent examples leading to the possibility of manipulating or even controlling vocal and facial signals in real-life interactions. However, what appears missing from this state-of-the-art is a way to quantify how such signals should be controlled to "illuminate the black box" (Brinberg et al., 2025) of socially contingent dynamics.

In this thesis, we propose using a simple concept from control engineering, the impulse response, to provide a model that is able to both analyse and synthesize dynamic contingency in social interactions. Impulse responses, or as we call them more generally, "social transfer functions", operationalize the mechanism of interpersonal predictive coding in a way that can be used to make experimental predictions.

In the next chapter (Chapter 2), we will turn to the field of control engineering and review a few fundamental concepts of system identification like transfer functions, highlight its growing presence in cognitive science and elucidate several ways in which it has already contributed to answering open questions in neuroscience and cognitive science.

In the remainder of the thesis, we show that social transfer functions allow making predictions of how observers rate social contingency in natural interactions (Chapter 3) and, provided they rely on appropriate models of facial expressions (Chapter 4), can be used to probe listeners' internal representations of contingency (Chapter 5).

Chapter 2

System Identification in Control Engineering and Cognitive Science

The framework of *interpersonal predictive coding* (Manera et al., 2011) casts interacting agents as coupled dynamical systems that aim to obtain a shared representation or mutual understanding of the world. Put simply, a dynamical system is a system whose state x evolves according to some specific rule $\dot{x} = f(x)$. When dynamical systems x_1 and x_2 are coupled, the state of one system can be used to predict the state of the other such that $\dot{x_1} = f_1(x_1, x_2)$ and $\dot{x_2} = f_2(x_1, x_2)$. However, in an interactive context, such coupling requires that individuals possess similar internal representations of interactive dynamics (i.e., social contingency), allowing them to 1) generate appropriate contingent behaviour in response to incoming stimuli and 2) predict others' responses and initiate repair or modulation of signals in case of a mismatch between the actual and predicted response. While there is a wealth of research, detailed in Chapter 1, documenting the form, frequency and timing of interaction outputs in response to inputs (e.g. the typical timing of a smile), we lack a mechanistic formulation of the generative mechanisms responsible for those outputs (e.g. the precise dynamics of a smile in response to some specific dynamics of incoming speech). Such a generative mechanism would allow for experimental predictions by, for instance, measuring how well the smiling behaviour in a given interaction matches expected dynamics, as well as provide a testable cognitive model for how human observers perceive social contingency. One way to build such a model, provided that one has access to recordings of both input and output signals, is to use *system identification*, a class of methods with origins in control engineering. This chapter presents a brief introduction to system-identification methods, with a focus on the notion of linear, time-invariant systems and impulse responses, how they have been used in cognitive science so far, and how their use in social cognition can help formulate interaction dy-

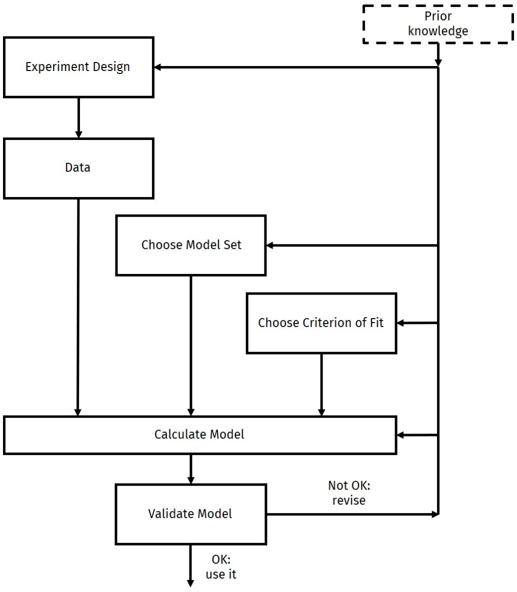


Figure 2.1: The system identification loop, adapted from Ljung et al., 1987, visualizes the process of calculating a model as involving many iterations over different sets of candidate models and the criteria for evaluating them until a model's properties can be satisfactorily validated.

namics (i.e. to model interacting pairs as a single system) in terms of a *social transfer function*.

2.1 System identification

System identification is an approach to modelling 'black-box' systems, which are characterized by a lack of access to and understanding of their dynamics, meaning that these systems cannot be modelled (simply, or at all) using first principles. The system-identification approach, being data-driven, involves experimenting on the system, observing the consequent input-output relationships and then inferring a model by analysing those relationships. Not unlike the classical machine-learning procedure, the process is often characterized by its 4 parts: observed input and output data $\{u(t), y(t)\}_{t=1...T}$, a set of candidate models and their parameters θ, a criterion of fit and the validation of the chosen model. It can thus be expressed as "finding that model in the candidate set that best describes the data, according to the criterion, and then evaluating and validating that model's properties" (Ljung, 1998) (Fig. 2.1).

2.1.1 General Formulation

Consider a general input-output dynamical system of the form

$$y(t) = G(z, \theta)u(t) + v(t) \tag{2.1}$$

where $G(z, \theta)$ is the system's transfer function, $y(t)$ is the measured output signal, $u(t)$ is the measured input signal, and $v(t)$ is a non-measurable disturbance term. If we make the following assumptions about the system:

- stable: $G(z)$ is rational with all poles in the left half-plane
- linear autoregressive: current output depends linearly on a finite number of past outputs and inputs
- time invariant: if input u produces output y, then a time shift in the input $u(t - \tau)$ produces output $y(t - \tau)$
- finite-dimensional: there exist a finite number of coefficients in input and output

then the system equation can be written equivalently as a finite difference equation:

$$\begin{aligned} y(t) + a_1 y(t-1) + a_2 y(t-2) + \ldots + a_{N_a} y(t - N_a) = \\ b_1 u(t-1) + b_2 u(t-2) + \ldots + b_{N_b} u(t - N_b) + v(t) \end{aligned} \tag{2.2}$$

While many systems do not satisfy the linear time-invariant (LTI) assumption, typical system identification approaches consider deviations from this framework as likely being small and, in any case, contaminated by unknown disturbance signals which are expected to be captured by $v(t)$ (i.e. $v(t)$ incorporates both measurement and modelling error). In a typical modelling framework called prediction-error identification (Ljung, 1998), the disturbance $v(t)$ is modelled as a zero-mean stationary stochastic process of the form:

$$v(t) = H(z)e(t) \tag{2.3}$$

where $H(z)$ is a (stable, linear, time-invariant, finite-dimensional) transfer function, and $e(t)$ is zero-mean white noise.

In order to define sets of candidate models to estimate $G(z)$ and $H(z)$, a popular way (one adopted by the Matlab `System Identification Toolbox` - Ljung, 1995) is to parametrize $G(z,\theta)$ and $H(z,\theta)$ in terms of fractions of polynomials in z^{-1} with the notation $z^{-k}u(t) = u(t-k)$ leading to the most generic formulation called the Box-Jenkins model (Box and Jenkins, 1976):

$$y(t) = \frac{B(z^{-1})}{A(z^{-1})}u(t) + \frac{C(z^{-1})}{D(z^{-1})}e(t) \tag{2.4}$$

where

$$\begin{aligned} A(z,\theta) &= 1 + a_1 z^{-1} + \ldots + a_{N_a} z^{-N_a} \\ B(z,\theta) &= b_1 + b_1 z^{-1} + \ldots + b_{N_b} z^{-N_b} \\ C(z,\theta) &= 1 + c_1 z^{-1} + \ldots + c_{N_c} z^{-N_c} \\ D(z,\theta) &= 1 + d_1 z^{-1} + \ldots + d_{N_d} z^{-N_d} \end{aligned} \tag{2.5}$$

where a_i and b_i correspond to the coefficients in Equation 2.2. In the following, we refer to the model parameters θ as the set of coefficients $\{a_1, \ldots, a_{N_a}, b_0, \ldots, d_{N_d}\}$.

2.1.2 Common Model Structures

An important aspect of system identification is to constraint parameters in $G(z,\theta)$ and $H(z,\theta)$, leading to a variety of model structures, that are frequently applied in identification problems. The most important ones are listed in Table 2.1, by order of decreasing generality. For instance, the ARX (Autoregressive with Exogenous Input) model structure makes the assumption that all models in the set have a common denominator in $G(z,\theta)$ and $H(z,\theta)$.

The choice for a specific model structure can be based on prior information about the process to be modelled (e.g. knowledge about where disturbance signals enter the system process) but also have practical consequences on the estimation process for θ. As will be described below, in the particular case of FIR (finite impulse response) models, the parameter estimation procedure can be formulated as a linear regression problem, which is very appealing from a computational point of view.

Model structure	$G(z,\theta)$	$H(z,\theta)$
Box-Jenkins	$\frac{B(z^{-1})}{A(z^{-1})}$	$\frac{C(z^{-1})}{D(z^{-1})}$
ARMAX	$\frac{B(z^{-1})}{A(z^{-1})}$	$\frac{C(z^{-1})}{A(z^{-1})}$
ARX	$\frac{B(z^{-1})}{A(z^{-1})}$	$\frac{1}{A(z^{-1})}$
Output error	$\frac{B(z^{-1})}{A(z^{-1})}$	1
FIR	$B(z^{-1})$	1

Table 2.1: Common model structures used to identify linear time-invariant systems, listed by decreasing generality. *Table adapted from (Bombois and Van den Hof, 2006).*

2.2 Finite Impulse Response Systems (FIRs)

2.2.1 Formulation as a Linear Regression Problem

The FIR model, one of the several possible model structures (Table 2.1), has the property that the model's output is a linear function of the unknown parameters $\theta = \{b_k\}_{k=1\ldots N_b}$:

$$y(t) = \sum_{k=1}^{N_b} b_k u(t-k) + v(t),\ t = 0, 1, 2... \tag{2.6}$$

A consequence of this linearity is that the least-squares identification criterion defined on the prediction error is a quadratic function in θ (Ljung et al., 1987). As a result, there is an analytical expression for the optimal parameter θ_0 that minimizes the criterion, which can be obtained using the classical linear regression procedure, as:

$$\theta = (S^T S)^{-1} S^T y \tag{2.7}$$

$$S = \begin{bmatrix} u(1) & 0 & 0 & \cdots & 0 \\ u(2) & u(1) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u(T) & u(T-1) & u(T-2) & \cdots & u(T-N_b) \end{bmatrix} \tag{2.8}$$

where S is a matrix whose columns are the time-lagged versions of the system's input $u(t)$ up to the system's order N_b, $y(t)$ is the actual system output, and T is the total number of samples collected in the training data.

2.2.2 Regularization

The expression of FIR system identification as a linear regression problem is computationally attractive because it allows incorporating regularization terms in the optimization criteria. Standard least-squares linear regression is known to produce unstable/unreliable coefficient estimates in situations of high dimensionality and multicollinearity (i.e. when regressors are correlated). This situation is typically the case in system-identification contexts where the independent variables $u(t-k)$ are the numerous (order N_b) time-shifted samples extracted from an input time-series. Such cases also typically involve input signals that have their own internal dynamics and successive samples cannot be considered independently and identically distributed (i.i.d.) - providing one reason why classical system identification procedures use experimental measurements involving white-noise (Marmarelis and Naka, 1972). In situations where multicollinearity cannot be avoided (e.g. when identifying systems from ecological inputs that do not allow easy experimental control over autocorrelation), modern machine-learning extensions to linear regression (Pillonetto et al., 2014) add a penalty term (also called regularization) to the least-square loss function, which can, e.g. be proportional to the absolute value (L_1 norm):

$$Loss_{L_1} = MSE + \lambda \sum_{i=1}^{N} |w_i| \tag{2.9}$$

or the squared value (L_2 norm) of the coefficients

$$Loss_{L_2} = MSE + \lambda \sum_{i=1}^{N} w_i^2 \tag{2.10}$$

where MSE is the traditional mean-squared-error term, w_i are the coefficients (weights) of the model, N is the number of regressors and λ is a regularization parameter controlling the strength of regularization typically learned through cross-validation on a separate validation set. Regression with L_1 regularization is also called Lasso regression, and encourages sparsity in the model (i.e. some coefficients may become exactly zero). L_2 regularization, also called ridge regression, specifically penalizes larger weights and is often preferred in system identification contexts (Pillonetto et al., 2014). Apart from these, other machine-learning extensions are also commonly applied to the problem of estimating FIR models, such as boosting (a step-wise procedure that increments the w_i

of the variable that's most correlated with e.g. the current residuals at that iteration leading to sparse estimates that are perhaps comparable to Lasso) or subspace pursuit (see Kulasingham and Simon, 2022 for a review and empirical comparisons).

2.2.3 Assumptions Made by FIR models, in Theory and Practice

In theory, FIR models make some rather unrealistic assumptions about system behaviour. First, they impose the constraint of linearity. Being a linear function of the inputs implies that, for instance, doubling the input would double the output. Second, the lack of an $A(z^{-1})$ term means that past output has no influence on future outputs, i.e. its input-output behaviour is time-invariant. Finally, $H(z, \theta) = 1$ means that no parameters are used to model the behaviour or characteristics of the disturbance term, which is assumed to be independent and identically distributed (i.i.d).

In practice, however, the non-linear components of a system can be considered relatively small (e.g. second-order) with white-noise residuals capturing their variance reasonably well. Provided with some knowledge of the system, non-linearities can also be added to the input signal itself (Lindboom et al., 2023). Moreover, extensions to FIR models, such as regularization and boosting can compensate for some of the suboptimal assumptions, particularly in terms of avoiding overfitting to input dynamics and multicollinearity in the regressors. Finally, in the context of cognitive science, the simplicity of FIR models is often helpful for understanding complex cognitive processes and, on the basis of that understanding, developing and using more complex models. Importantly, knowledge of the degree to which cognitive phenomena is captured by simple, first-order linear approximations is in itself an interesting empirical question (Marmarelis and Naka, 1972). In the remainder of this chapter, we highlight how FIR models and their variants, despite the strict theoretical assumptions, have become increasingly prevalent in the study of cognition.

2.3 System Identification in Cognitive Science

In system-identification terms, the internal representation of a cognitive process can be thought of as a system that can be expressed as some combination of experimentally controlled inputs and behavioural or (neuro)physiological responses as outputs. The goal then is to compute some form of transfer function from the observed data to obtain a complete characterization of the internal representation, which can be used to make predictions for novel inputs and thus provide mechanistic control in experimental settings. As seen above, one simple type of transfer function for system identification is the FIR model, which assumes that a system's output is the convolution of the input with a fixed impulse response. Although FIR models make strong and often theoretically unrealistic assumptions about the system (notably, time-invariance), their formulation as a linear-regression problem makes them both computationally attractive and, in practice, a reasonable option for high-dimensional, noisy phenomena such as those encountered in behavioural or neurophysiological data.

2.3.1 Reverse correlation

One of the most common system-identification techniques in cognitive science is reverse correlation, based broadly on the principle that noise can be used to study black-box systems (Wiener and Masani, 1958). Reverse correlation has its origins in neurophysiology, where it was used to study biological systems by characterizing sensory neurons as stimulus-response transducers driven by external stimulation. In a landmark study, catfish retinal neurons were probed using Gaussian white-noise modulated light intensity stimuli with the resulting neuronal spikes recorded as output (Marmarelis and Naka, 1972). Random stimuli that elicited a response were averaged to derive Wiener kernels (essentially impulse responses; Schetzen, 2006), which could be used to make quantitative predictions about the behaviour of the neurons (Fig. 2.2). Other studies investigating neurons in the visual cortex of non-human animals utilized reverse correlation to estimate neuronal receptive fields by cross-correlating Gaussian white noise inputs with the corresponding outputs (Reid and Shapley, 1992; DeAngelis et al., 1993; McLean et al., 1994). A similar study reduces the dimensions of the input space to obtain significantly improved signal-to-noise ratio by replacing Gaussian white noise stimuli with inputs sampled from a sub-space constructed based on *a priori* knowledge of cell properties (Ringach et al., 1997). This modification is particularly important in the context of the adoption of reverse correlation in psychophysics, where highly noisy stimuli and a large number of trials both often have a detrimental effect on the behavioural responses

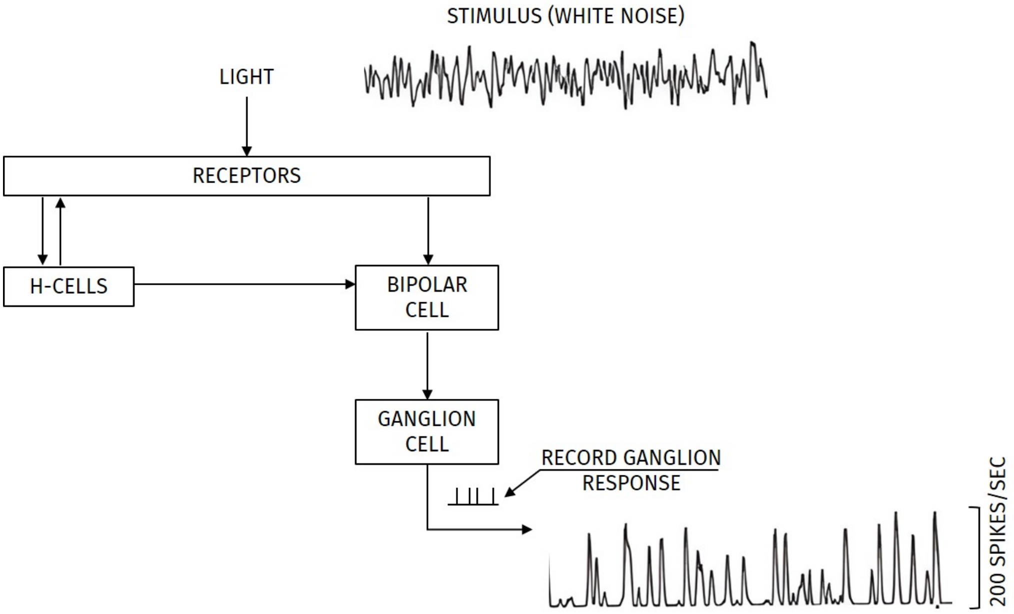


Figure 2.2: **White-noise analysis of retinal ganglion cell activity.** A block diagram depicting the main signal-flow pathways in the vertebrate retina and a light stimulus with the resulting ganglion cell response. Other stimulus-response pairs can be chosen experimentally (after Marmarelis and Marmarelis, 1978). *Figure and description adapted from Marmarelis (2004).*

of human participants.

In psychophysics, the reverse correlation procedure usually involves a 2-alternative forced-choice task to avoid bias and record binary participant responses to hundreds of pairs of randomly perturbed stimuli. At the end of the reverse correlation procedure, the average noisy stimuli chosen by the participant is used to compute a model called the *classification image*, which shows the stimulus features that drive behaviour. While dynamic stimuli with continuous inputs and outputs are modelled as impulse responses based on the framework described in the previous sections, binary outputs reduce the system equation to a simple scalar product between the input and the model. In this case, rather than an impulse response, the model is considered a template or a classification image that represents what the input should look like in order to generate a response. Corresponding to the description of impulse responses, this *linear observer model* assumes linearity, time-invariance (i.e. the same template is always used so that a given input always yields the same output)

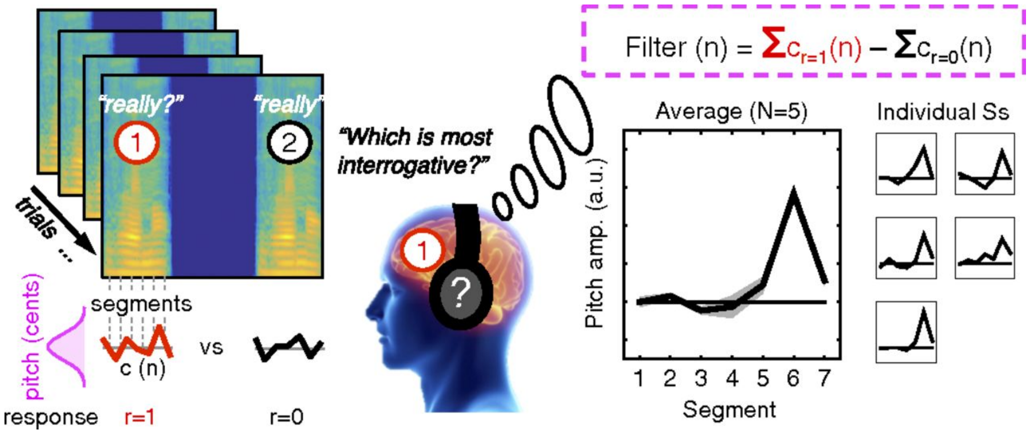


Figure 2.3: **Accessing mental representations of interrogative prosody by using reverse correlation.** To validate the paradigm used in this study, we examined prosodic prototypes related to the evaluation of interrogative vs. declarative utterances. **(Left)** Utterances of the same word "vraiment" ("really") were digitally manipulated to have random pitch contours $c(n)$. Participants were presented pairs of manipulated words and judged which was most interrogative. **(Right)** Prosodic mental representations, or prototypes, were computed as the mean pitch contour of the voices perceived as interrogative ("really?"), minus those judged declarative ("really."). As predicted, the prototypes associated with interrogative judgments showed a clear pitch increase at the end of the second syllable, which was observable both in averaged and in individual prototypes. amp., amplitude.

Figure and description adapted from Ponsot et al. (2018).

and contains a noise term representing the 'internal noise' of the system. Psychophysics studies have applied reverse correlation to investigations of internal representations of more high-level cognitive processing. For instance, one study reverse correlates random perturbations of the spectral features of the vowel [a] to extract internal representations of smiling speech (Ponsot et al., 2018). Another study with random pitch profiles of the word "*Hello*" demonstrates that the internal representation of trustworthiness in speech is characterized by low average pitch that increases towards the end of the utterance, while that of dominance is characterized by even lower average pitch that decreases at the end (Ponsot et al., 2018). In the visual domain, studies using reverse correlation (or its parent method called 'bubbles' - Gosselin and Schyns, 2001) highlight the regions of the face driving judgements of trustworthiness and dominance (Dotsch and Todorov, 2012), internal self-representations of individuals and how they correlate with personality traits like self-esteem (Moon et al., 2020; Maister et al., 2021), and the cultural differences in internal representations of emotions (Jack et al., 2012).

2.3.2 Impulse Responses in Electrophysiology

Over the years, system identification techniques have received increasing attention from researchers using methods like fMRI and, in particular, electroencephalography (EEG). The marked interest in applying system identification to EEG is driven in large part by the limitations of traditional analysis methods like event-related potentials (ERPs; Handy, 2005). Traditional ERP analysis segments EEG signals into epochs that are time-locked to many identical repeats of the same stimulus, and averages them. One limitation of this method is that it is predicated upon the assumption that stimuli are presented almost instantaneously, as in the case of image stimuli. When this assumption holds, the time-locked ERP can be thought to directly reflect the sequence of cognitive events triggered by the stimulus in the form of successive electrophysiological components (N1, P2, P3, etc.). However, when the stimulus itself extends over a non-negligible period of time and overlaps with the timings of ERP components (e.g. measuring P300 time-locked to the onset of a 1-second sound recording, Benghanem et al., 2024), the measured output is best thought of as a continuous integration (or convolution) of the ongoing stimulus with some unknown to-be-determined component.

Relatedly, ERP analysis also forces stimuli to be presented as isolated events, often as sequences with silent inter-stimulus intervals. The need for many repetitions of a single stimulus to obtain ERPs means that the stimuli become less ecological and more importantly, make the resulting ERPs less generalizable to other stimuli. Together, these factors limit the applicability of ERPs for studying a wide range of cognitive mechanisms that take continuous streams of information as inputs. For instance, cortical responses to natural continuous speech often overlap in time and show dynamic tracking or entrainment to acoustic features of speech (Ahissar et al., 2001; Abrams et al., 2008; Lalor and Nidiffer, 2025), implying that average ERPs might obscure the precise dynamics of brain responses. So while ERPs have been invaluable for identifying components involved in the processing of sounds and individual words, using them to investigate the cognitive mechanisms underlying language comprehension and processing of continuous speech has proven to be difficult. For these reasons, alternative methods to ERPs are increasingly being proposed.

One such method is VESPA (Visually Evoked Spread Spectrum Response Potential - Lalor et al., 2008). This method circumvents the need for repeated presentations of discrete visual stimuli to evoke visually evoked potentials by instead using random, continuously changing visual properties as input and corresponding brain signals as output. Using least squares estimation, an impulse response $w(\tau)$ is estimated from the input-output pairs, such that its convolution with the continuous input signal can then be used to generate the output. Such an impulse response can be thought of as a generalization of visually evoked potentials, without being restricted by the number of image stimuli that can feasibly be presented. The counterpart of VESPA in the auditory domain, the AESPA (Lalor and Foxe, 2010), also tackles the issue of continuous stimuli by estimating an impulse response from natural, continuous speech input (specifically, its amplitude envelope) and the corresponding EEG response (Fig. 2.4A). The ability of these system-identification methods to handle continuous, dynamic stimuli has allowed the investigation of more complex cognitive phenomena like the cocktail party effect (Cherry, 1953). In a seminal study, impulse responses were estimated for the mapping between neural data of participants simultaneously listening to 2 different speakers and the amplitude envelopes of each speech signal (O'sullivan et al., 2015). The impulse responses of attended and unattended speech were then used to reconstruct the amplitude envelope of each. Results showed that when using the impulse response of attended speech, the correlation between actual attended speech and its reconstruction was higher than with unattended speech. Similarly,

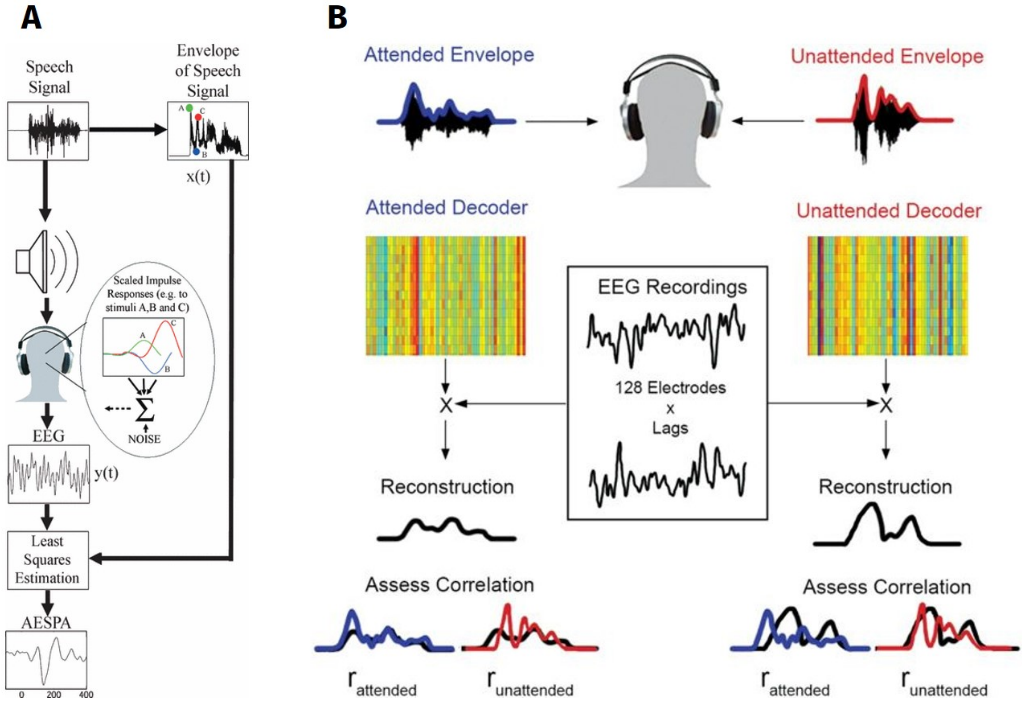


Figure 2.4: **(A) The AESPA method.** The speech signal is presented to the subject concurrent to electrophysiological (e.g. EEG) data being recorded. The amplitude envelope of the speech signal is calculated by determining the RMS of the audio signal values occurring in the time frame of each sample of the neural data. It is assumed that the neural data consists of a convolution of the amplitude envelope of the speech signal with an unknown impulse response function, plus noise. Given the recorded data, knowledge of the speech signal and accurate synchronization between the two, this impulse response function, known as the AESPA, can be estimated using least-squares estimation.

Figure and description adapted from Lalor and Foxe (2010).

(B) Using system identification to decode selective attention in a Cocktail Party paradigm. Data from all electrode channels are decoded simultaneously to give an estimate of the amplitude envelope of the input speech stream. The correlation between this reconstruction and both the attended and unattended speech streams is then calculated for each trial.

Figure and description adapted from O'sullivan et al. (2015).

when using the impulse response of unattended speech, its reconstruction showed greater correlation with the actual unattended speech than attended speech (Fig. 2.4B). In another study, similar regression methods were used to estimate STRFs (spectro-temporal receptive fields) to reconstruct the auditory spectrogram of *Another Brick in the Wall, Part 1 (by Pink Floyd)* from high-frequency activity in intracranial EEG recordings of participants listening to the song (Bellier et al., 2023). Interestingly, though these models do not boast particularly impressive prediction accuracies (highest correlations with observed data approximately 0.1 and 0.3 in O'sullivan et al., 2015 and Bellier et al., 2023, respectively), their ability to handle continuous stimuli and not require a large number of repeated trials has meant an increase in popularity of FIRs and its many variants, perhaps reflecting the acceptance of models "guided by *usefulness* rather than *truth*" (Ljung et al., 1987).

2.3.3 Impulse Responses as Surrogate Models

A more conceptual (rather than methodological) limitation of ERPs is that while they establish the influence of stimulus characteristics on cortical responses, it is difficult to draw conclusions about the specific cause or realizer of the responses. In other words, differences in cortical responses could either entail different cognitive processes (i.e. caused by differential cognitive evaluation) or the same cognitive process simply reflecting the physical differences in the stimuli. This stimulus-to-process inference problem is somewhat mitigated in studies involving explicit tasks where behavioural responses are collected in conjunction with ERPs because they allow the comparison of ERP differences with corresponding behavioural evaluations (Castro et al., 2020; Nussbaum et al., 2022). However, the problem becomes particularly intractable when investigating implicit effects, especially in clinical settings involving unresponsive patients like those with disorders of consciousness (DoCs). For instance, suppose a comatose patient displays differences in ERPs in response to speech sounds with varying emotional prosody. Since there is no corresponding behavioural evidence, can we conclude that the differential ERPs are a marker of preserved emotional perception, or do they simply indicate that observed differences are solely the result of the auditory periphery treating different prosodic patterns as distinct cortical inputs? The latter precludes any high-level cognitive differentiation and confounds subsequent evaluations of patient prognosis. Thus, what is needed is a way to model responses to different stimuli *as if* they were generated by the same cognitive process, so that a mismatch between the model's prediction and the observed data can lend support to the alternative two-process interpretation involving differential cognitive processing.

In a project that is somewhat ancillary to the broader theme of this PhD manuscript, we collaborated with neurologists at Sainte-Anne Hospital in Paris to address the inference problem by leveraging system identification using FIRs. We investigated one of the more salient examples of prioritized auditory processing in the brain, namely, the perceptual bias towards looming compared to receding sounds. Studies show that compared to receding sounds, looming sounds are consistently overestimated as being louder (Neuhoff, 1998; Ponsot et al., 2015) and faster (Rosenblum et al., 1987; Schiff and Oldak, 1990) and activate a wider network of regions subserving auditory spatial perception and attention (Bach et al., 2008; Seifritz et al., 2002). However, it is possible that evidence for different brain responses could be driven by the dramatically different temporal profiles of looming and receding sounds, i.e. per the inference problem, evidence of different outputs does not constitute evidence for different processes.

To address this, we collected EEG data during an oddball paradigm to elicit ERP-like components in response to deviant stimuli with both dynamic (looming and receding) and constant level (flat) differences to the standard in the same participants (N=18). We then used FIR system identification with ridge regression (called TRF in the mTRFpy Python package - Bialas et al., 2023) to model single-participant ERP responses to flat deviants, and used the model to predict the effect of the same mechanism on looming and receding stimuli. The idea was that the modelled impulse response was a description of the cognitive mechanism underlying processing of flat-intensity sounds. Consequently, if sounds with looming and receding intensities involve different mechanisms, as suggested by the literature, model predictions for those sounds would be less accurate. Interestingly, we found that the impulse response for flat-intensity sounds explained 45% and 33% of the variance of observed responses to looming and receding sounds (Fig. 2.5D). Essentially, cortical responses to looming and receding sounds were generated by the same cognitive mechanism as flat-intensity sounds, and the observed differences in their responses were the sole consequence of their particular physical morphology getting amplified and integrated by peripheral auditory mechanisms. The study was published in Cortex (Benghanem et al., 2024), and a follow-up study with the same collaborators is in preparation using the

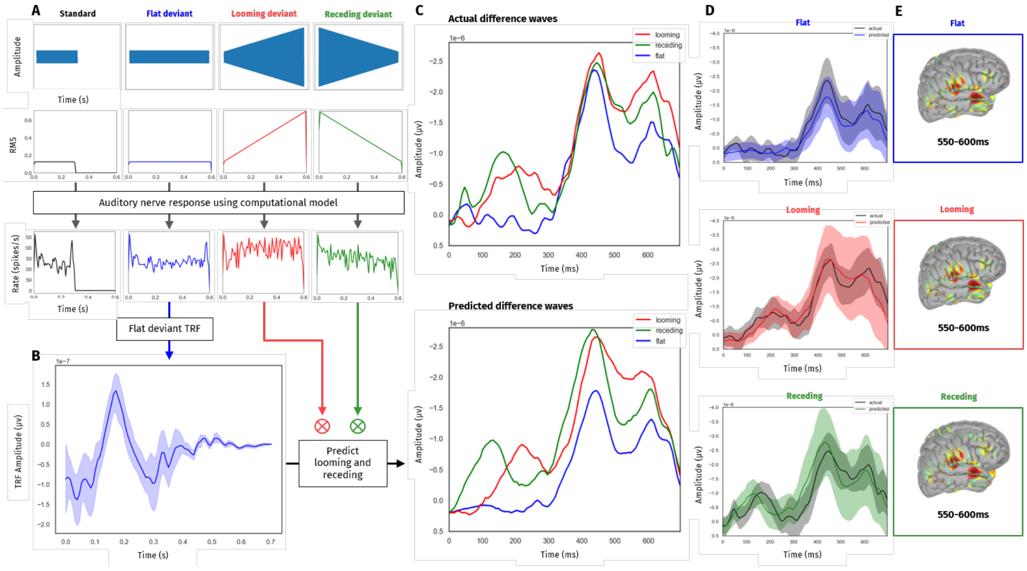


Figure 2.5: **Cortical asymmetries between looming and receding sounds are explained away by the auditory periphery. (A)** We collect EEG mismatch negativity (MMN) data in response to deviant stimuli that had either dynamic (looming and receding) or constant (flat) level differences to the standard. **Top:** Waveform and RMS intensity profiles of all stimuli. **Bottom:** Corresponding simulated auditory nerve response of the stimuli using a computational model of the inner ear Zilany et al. (2014). **(B)** Taking the simulated auditory nerve response of flat-intensity sounds (blue) as input and the corresponding ERP difference wave as output, we estimate an FIR as a description of the cognitive mechanism for processing flat-intensity sounds. **(C)** Predictions of this cognitive mechanism for looming (red) and receding (green) sounds are computed as a convolution of the simulated auditory nerve response of those sounds with the FIR to show striking similarities between the predictions and observed data (looming: 45%; receding: 33% explained variance). **Top:** Grand average of the observed difference waves (deviant minus standard) at the Fz sensor. **Bottom:** Predicted difference waves according to the FIR estimated from flat-intensity sounds. **(D)** Cluster permutation tests at the Fz sensor between observed and predicted responses to flat (blue), looming (red), receding (green) sounds show no statistically significant differences. **(E)** Source localization in the right lateral cortical surface for observed responses to flat (top), looming (middle) and receding (bottom) sounds show statistically similar generation of sources.

same methodology but with prosodic differences in stimuli.

Roadmap

In this chapter, we reviewed the basics of system identification and highlight how a few of those concepts (e.g. reverse correlation and FIRs) have been utilized in cognitive science. Armed with the theoretical background of social cognition and system identification, the remainder of the thesis will present work at the intersection of the two.

In Chapter 3, we attempt to model third-party observers of social interactions as FIRs and use them to isolate the facial features driving observers' perception of social contingency.

In Chapter 4, we use the system identification technique of reverse correlation to probe machine-learning based black-box models (of the kind used in Chapter 3 for extracting the signals upon which our FIRs were modelled). By extracting their internal representations, we attempt to 'explain' the model in terms of the specific input features responsible for generating some output.

Finally, the experimental work culminates in Chapter 5, where we combine insights from the preceding chapters and use classical reverse correlation *on* FIRs to obtain third-party observers' internal representations of social contingency in a data-driven manner.

Chapter 3

Observer perception of social contingency: Corpus Analysis

In Chapter 1, we reviewed evidence from the social cognition literature suggesting that the perception of social contingency, and the dynamics of non-verbal signals in general, relies on interpersonal predictive coding that is sensitive to, for instance, how facial signals of one can be predicted by facial or vocal signals of the other (Manera et al., 2013, 2011). However, there are discrepancies within the existing literature in terms of available evidence at the *computational* and *algorithmic* levels of inquiry (Marr, 2010). While there is evidence showing which computations are performed by the social cognition system and why (i.e. at the computational level), the processes and representations used for those computations (i.e. at the algorithmic level) remain unclear. In other words, we do not know if and how people internally represent the coupled dynamics of interacting agents or how a specific social signal needs to be manipulated to make it appear contingent.

In Chapter 2, we reviewed a few fundamental concepts of system identification, a class of methods with origins in control engineering, that allow data-driven modeling of input-output systems as parametric transfer functions. While system-identification methods have some history in cognitive science, notably via the technique of reverse correlation and, more recently, temporal response functions (**?**), they have received relatively little attention in the modeling of social cognition.

In this chapter, we introduce a computational modeling paradigm, the 'social transfer function', which assumes that observers possess a schema (or a dynamic representation; Freyd, 1987) of contingent interactions, conceivably acquired over time by observation and participation, and which can generate real-time predictions of the temporal dynamics of one agent's facial signals in response to the speech of another agent (Fig. 3.2). At the algorithmic level of explanation, we instantiate such a 'transfer function' using the temporal response functions (TRFs; **?**), which assumes that the system can be represented by an impulse response H that is convolved with the input to generate the output ($Y = H \circledast X$). 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. As noted in Chapter 2, the algorithmic choice of impulse responses/TRFs carries strong assumptions of linearity and time-invariance which should be evaluated empirically and discussed theoretically. However, while the more general computational notion of a "social transfer function" does not necessarily entail such assumptions, we will return repeatedly to the question of the validity of these assumptions throughout the thesis (see Chapter 6).

To test this mechanism, we analysed a corpus of video recordings of naturalistic speed-dating interactions, previously recorded by our collaborators (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 video recording of the real listener with another in half of the trials. In two successive behavioural experiments (the first in-lab, N=18; the second, online and preregistered, N=188), human observers were asked to discriminate genuine vs fake (i.e. non-contingent) interactions. We investigated whether social transfer functions learned from the dataset could predict observer performance better than a simpler model based on the average quantity of movement, their ability to determine the specific facial features used by observers to detect contingency and their ability to predict observer performance when stimuli were degraded by masking different regions of faces.

3.1 Study 1: Are listeners' facial expressions alone sufficient for social contingency perception?

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. It is important to note that in this study, as well as the rest of this thesis, we use the term 'backchanneling cues' in a low-level perceptual sense to refer to

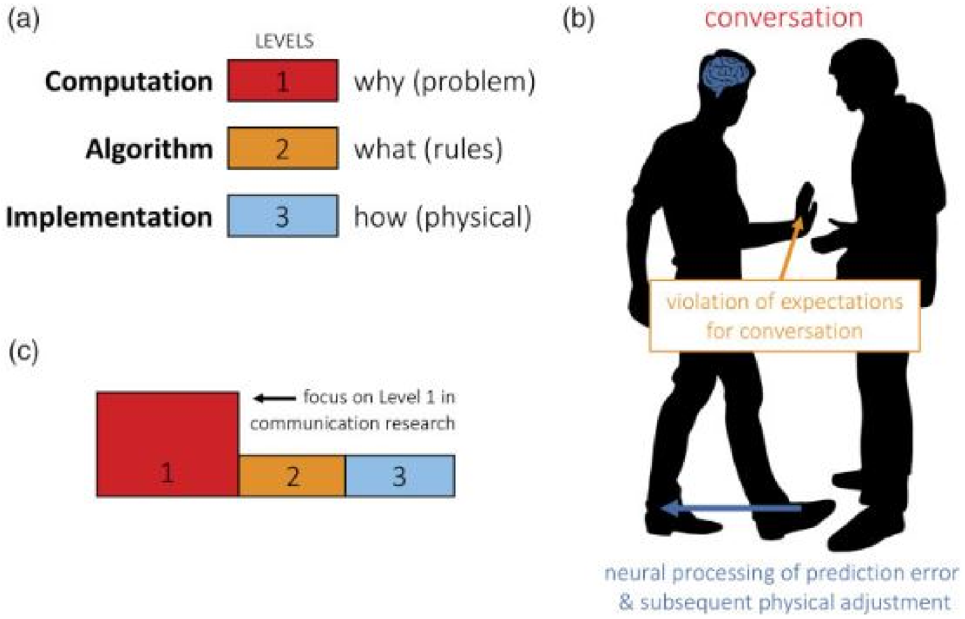

(a)
LEVELS
Computation
1
why (problem)
Algorithm
2
what (rules)
Implementation
3
how (physical)
(b)
conversation
violation of expectations for conversation
neural processing of prediction error & subsequent physical adjustment
(c)
1
2
3
focus on Level 1 in communication research

Figure 3.1: **Marr's framework as applied to a dyadic interpersonal interaction.** (a) The levels as represented in Krakauer et al. (2017). (b) Communication example: A message sender attempts to share information with a conversational partner while following a set of socially constructed rules that govern the interpersonal interaction (goal). Here, the interaction partner stands too close, which triggers a violation between expectation and actuality, also known as a prediction error (algorithmic realization) that is cognitively processed in the brain and results in a physical adjustment of the body positioning (physical implementation). (c) Historically, communication research emphasizes the computational level of communication phenomena.

Figure and description adapted from Huskey et al. (2020).

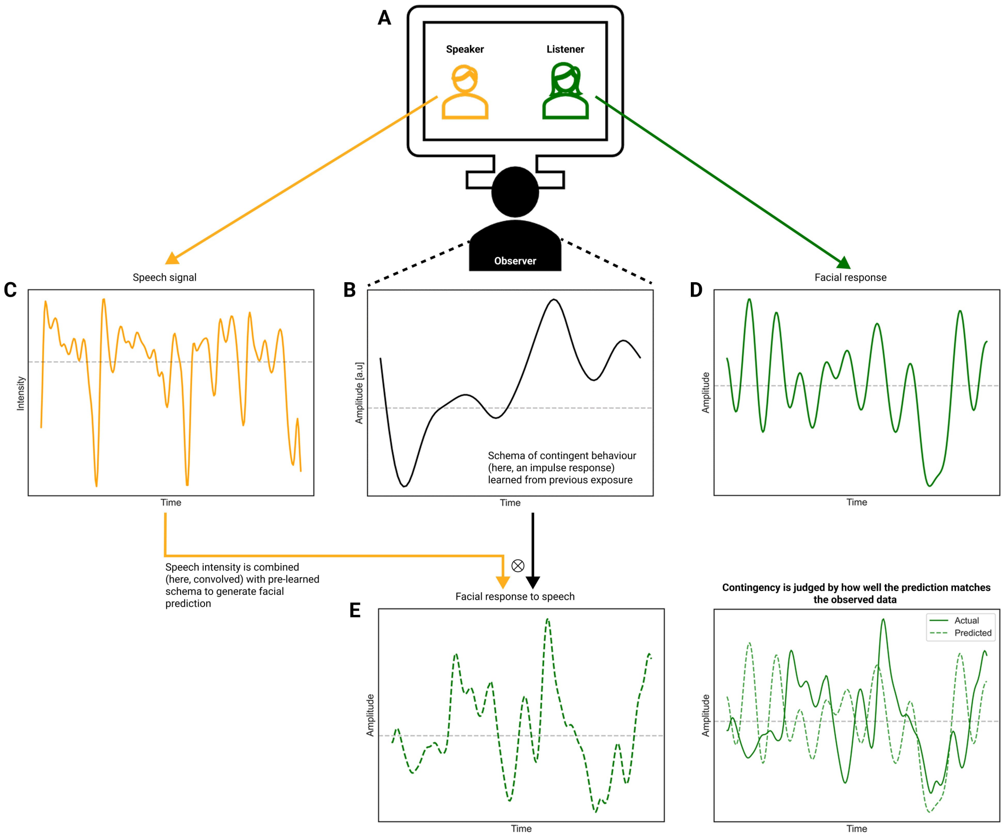

A
Speaker
Listener
Observer
C
Speech signal
Intensity
Time
B
Amplitude [a.u]
Schema of contingent behaviour
(here, an impulse response)
learned from previous exposure
Time
D
Facial response
Amplitude
Time
Speech intensity is combined
(here, convolved) with pre-learned
schema to generate facial
prediction
⊗
E
Facial response to speech
Amplitude
Time
Contingency is judged by how well the prediction matches
the observed data
Actual
Predicted
Amplitude
Time

Figure 3.2: **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 thesis, 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 **(C)** such that another agent's response to the speech signal in terms of the temporal dynamics of their backchanneling cues **(D)** can be matched against the transfer function's real-time predictions **(E)** 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 the facial features used to detect genuine and fake interactions.

the ensemble of visual cues from a listener's behaviour while a speaker speaks that is 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). Discussion sections in the rest of this thesis will comment on how our results may depend on such aspects.

3.1.1 Materials and Methods

Speed Dating Corpus

Stimuli used in this work were extracted from a corpus of video recordings of naturalistic speed-dating interactions, which was collected as part of an earlier 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 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 from 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 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 pairing 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 the 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 (Fig. 3.3).

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).

Procedure

Participants were presented with 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', i.e. the genuine and fake extracts 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 was quantified using the d′ sensitivity index.

Social transfer functions

To model how well genuine/contingent interactions matched a prediction of the temporal dynamics of the listener's facial responses 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; **?**). 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) activity (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 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 (Section 2.2.2), as implemented in the mTRFpy toolbox (Bialas et al., 2023).

Once trained, a TRF allows predicting an observer's facial response (a time-series of AU intensity or, equivalently, visual occurrence probability) to a specific speaker's speech (a time-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 time-series predicted by convolution with the TRF and the actual observer's times series can be evaluated using Pearson's correlation coefficient r between the two time series.

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.

To evaluate whether TRF fits were computationally sufficient to accurately discriminate genuine and fake trials, we trained a support vector machine (SVM) classifier with a linear kernel, which took the vector of 11 TRF fits as input and a binary classification of a trial as being genuine or fake as output. The dataset of 198 trials was divided into training and testing sets. The SVM hyperparameters (C and Gamma) were optimized by 5-fold cross-validated grid-search over the training set. The final model was then trained on the training set, and evaluated using classification accuracy on the testing set. We tested for statistical significance of the resulting accuracy against chance performance (0.5) using a binomial test. SVM training was implemented using the `scikit-learn` Python package. We used SVM weights as a rudimentary but, in the case of linear kernels, conservative indication of feature importance. (Guyon and Elisseeff, 2003).

Finally, to test whether the TRF fits of trials predict observers' decision of genuineness, over and beyond average AU intensity, 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)`), where `intensity` is the AU's average (i.e. static) intensity over the trial. GLM analysis was performed with the pymer4 package (Jolly, 2018).

3.1.2 Results

Participants performed significantly above chance at discriminating genuine vs fake interactions (d' =0.53, $t(17) = 10.23, p < .001$) with a large effect size ($d = 2.46$) as measured by Cohen's d. 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:$t(17) = 4.31, p < .001, d = 1.46$; and AV-V:$t(17) = 4.11, p < .001, d = 1.27$). There was no performance difference between the A-V and AV-V blocks ($t(17) = 0.33, p = 0.74$) (3.3). 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 differed in terms of how well listeners' backchanneling dynamics in response to speech matched TRF predictions of those dynamics. To do so, we

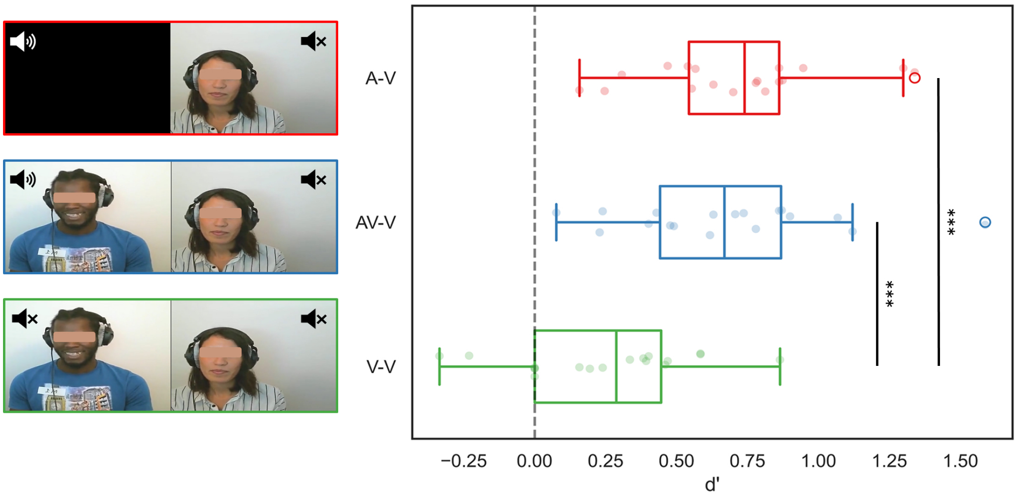


Figure 3.3: **Study 1. Left:** Observers were presented audiovisual extracts from speed-dating interactions in which only one person was talking while the other was silent and only displayed backchanneling cues. Trials were presented in three possible 'modalities': audio-video (A-V, **top**) in which 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).
Horizontal bars over the eyes were added in this image to protect the anonymity of the participants, but were not displayed during the experiment.

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 fits 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: ($t(196) = 2.78, p = .008, 0.38$), AU25: ($t(196) = 2.62, p = .01, d = 0.37$), AU26: ($t(196) = 2.63, p = .006, d = 0.39$), AU28: ($t(196) = 2.92, p = .02, d = 0.34$), as well as for head nods ($t(196) = 2.6, p = .01, d = 0.36$). 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 apparent on listener nods and mouth reactions such as smiling.

To evaluate whether TRF fits were computationally sufficient to accurately discriminate genuine and fake trials, we trained a machine-learning classifier on the TRF fits of trials as the input features and 'genuine'/'fake' as class labels (see Section 3.1.1). The SVM achieved a classification accuracy of 58% which exceeded chance performance (50%; binomial n=99, K=57, p=.025). The SVM weights assigned greater importance to AU12, AU25 and AU26 for discriminating between genuine and fake trials.

Finally, we tested whether human observers' behaviour, in terms of their binary responses, could be predicted by dynamic and static quantities of motion. Generalized linear models showed statistically significant relationships between observers' responses and dynamic TRF fit (but not their static quantities, i.e. average AU intensity over the trial) 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 TRFs 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 3.4).

3.1.3 Discussion

Study 1 investigated observers' ability to detect contingent behaviour in dyadic interactions. We manipulated trials such that they contained

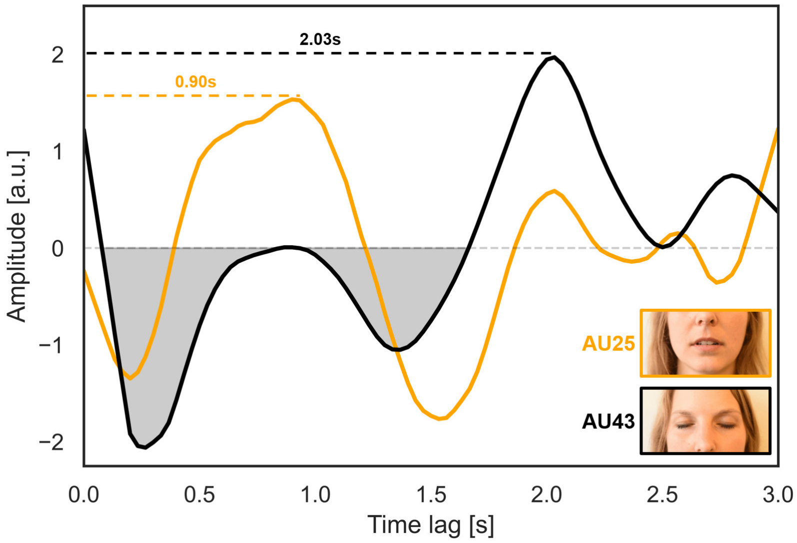


Figure 3.4: **The expected facial dynamics of social contingency.** Two temporal response functions (TRFs) allowed statistically significant prediction of observer decisions of genuineness, based on both a mouth- (AU25, *lip part*, orange) and an eye-related action unit (AU43, *eyes closed*, black). Comparison of these TRFs, or impulse responses (x-axis: time, y-axis: amplitude), reveals 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).

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 such as coordination and theory-of-mind (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 (see Section 1.3), 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 and Frith, 2014), and it may therefore only appear logical that adults should also perform well at a similar task. However, in the present task, the manipulated contingencies were not plain interruptions but rather desynchronized behaviour in which visual 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, the authors used 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. In contrast, 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 judging the contingency between two concurrent visual streams may have led 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, which would be compatible with the idea that blinks may function as an index of the end of an expression.

The fact that genuine and fake trials differ in how well they match the 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. The experimental paradigm introduced in Chapter 5 will also address the question of causality in more details.

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 and the low sample size, the relationship between the variables could indicate a number of different things. For instance, it could indicate that both face regions are in fact discriminative and utilized, but in a way that is not captured by our automated AU analysis, or that only mouth information is useful, but observers are also biased towards using eye information (even if counterproductively) or that both mouth and eye predictions are ecologically correlated in the dataset. 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.

3.2 Study 2: Eyes vs. Mouth - Is one sufficient for perceiving contingency?

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).

3.2.1 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 $\alpha = .05$.

Stimulus selection

Stimuli for Study 2 were selected as a subset of stimuli from Study 1, to control the difficulty of the task more formally. First, 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. Second, 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.

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 3.5). 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).

Procedure

Participants were presented with 40 stimuli in one of the three conditions in a between-subject design (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. The procedure was preregistered at https://aspredicted.org/fsp7-g7jw.pdf.

3.2.2 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. 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. Because the masking in both manipulated conditions rendered Py-feat unable to detect faces for subsequent AU analysis, we were unable to reproduce this analysis in the eye-only and the mouth-only stimuli.

3.2.3 Discussion

By adopting a design involving causal manipulation where 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 (to a lesser extent) 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 when participants were presented with complete face information. This not only suggests that no other facial cues besides the eye and the mouth provide any discriminating information for contingency in this specific task (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 (albeit on a controlled subset of the same stimuli) in that the TRF fit of both AU25 and AU43 correlated with participant ratings in the original condition while 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.

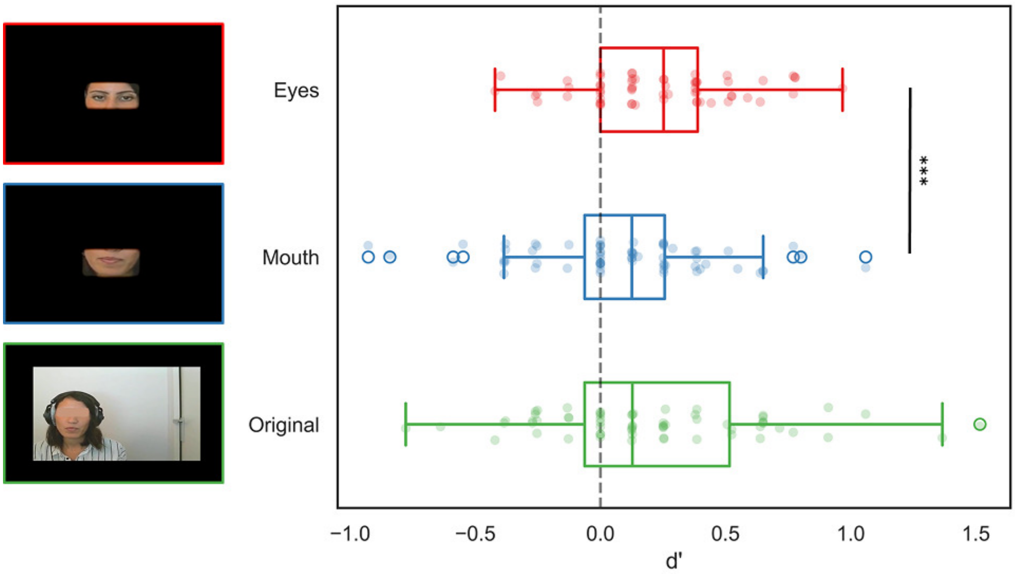


Figure 3.5: **Study 2. Left:** A-V trials from Study 1 were manipulated by creating dynamic visual masks that isolated specific parts of the listener's video while hiding everything else. This yielded 3 different versions of each trial where the speaker's audio was played over a video that featured either 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.

Horizontal bars over the eyes were added in this image to protect the anonymity of the participants.

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 while there are several AUs related to the mouth, the eye-related AUs are limited in terms of communicative information conveyed. Subsequently, the model may fail to capture important information from the eyes (e.g. gaze direction - Conty et al., 2006; Cañigueral and Hamilton, 2019; Wahn et al., 2022), which human observers are instead able to exploit.

3.3 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 chapter, 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=188), 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) or 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; and 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 to confirm that information restricted to these two face regions was sufficient to do the task (in the case of eyes), and no less accurate than control (in the case of mouth). 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 propose as 'social transfer functions'.

In this work, we implemented such social transfer functions using temporal response functions (TRFs). While they make strong modeling assumptions on the system (Chapter 2), 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, and showed that they embodied different expected timings for contingent facial responses: both TRFs contained early negative components around 300ms, suggesting either an expected deactivation of listener responses at the beginning of an utterance, or a fast positive reaction at the offset of an utterance. Both AUs also had strong positive peak activation, but at different peak locations (AU25: early, around 1sec; AU43: late, around 2sec). We also identified that AU43 activity was inhibited for almost the entire duration of AU25 activation, confirming the temporal complementarity of these cues.

Beyond AU25 and AU43, the wider range of TRFs obtained also provides insight into the temporal dynamics of action units in naturalistic conversations, regardless of their function for decoding contingency or other aspects. First, TRFs allow the grouping of expressions or AUs that have similar contingent dynamics, such as those of AUs 12, 14 and 15 with comparable peak latencies at around 1.8s (Figure 3.6A), as well as AUs 23 and 24 with both containing multiple peaks at 1 and 2.5s (Figure 3.6B). 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. Second, TRFs also allow comparisons between the chronometry of AUs that are thought to be physiologically related - for example, Figure 3.6C 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 3.6D) 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 composes meaningful social signals. In conclusion, the concept of a social transfer function, 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., 2011).

While the mechanism in this study provides accurate predictions of,

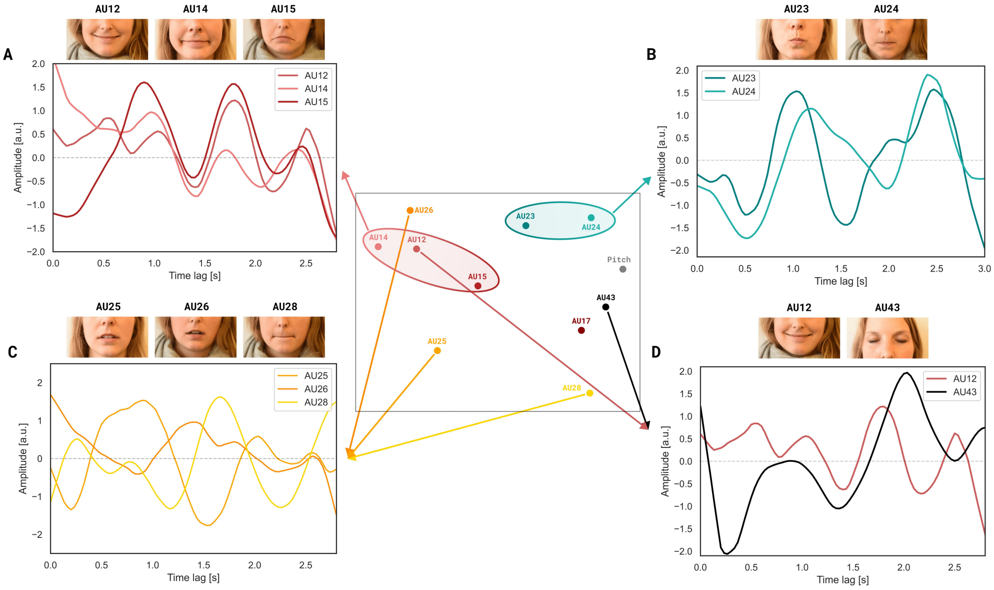


Figure 3.6: **The complex choreography of facial expressions for social communication.** TRFs for each AU projected onto low-dimensional Euclidean space using multidimensional scaling such that the similarity/distance between each in high-dimensional input space is maintained. **(A and B)** highlight 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.

e.g. the extent to which a trial appears genuine, or the parts of a face used to make such inferences, the results presented in this chapter leave several important questions unanswered. First, the nature of these results remains mostly correlational. While we showed that TRF agreement correlated with subjective ratings, it remains entirely possible that participants use entirely different cognitive representations and procedures to do the task - procedures whose outcomes happen to grossly correlate with those of our TRF algorithm (i.e. an issue at Marr's *algorithmic* level of inquiry; Marr, 2010). Second, our correlational evidence relied on machine-learning based estimations of AU activity. The black-box nature of these algorithms raises questions about whether the estimates reflect true activity in the eye or mouth region. For instance, Py-feat may use information from the mouth to estimate the activity of an eye-related AU, thus confounding comparisons of TRF predictability of a specific region over another. Results in this chapter indeed demonstrated some degree of ambiguity regarding which exact AU was predictive of contingency, with Study 1 suggesting that information related to contingency was mostly present in the mouth area, while Study 2 showed that eye-only performance was significantly better than looking only at the mouth. Thus, to ascertain the facial "modularity" of these results, one needs a way to probe *what exact visual information is used by machine-learning algorithms such as Py-feat* to estimate AU activity.

Roadmap

In this chapter, we introduce the 'social transfer function', an FIR model based on concepts and techniques discussed in Chapter 2. Using these social transfer functions, we show that observers detect social contingency by evaluating the alignment between a speaker's speech and a listener's facial responses and in particular, the dynamics of these responses rather than static movement quantities.

In the next chapter (Chapter 4), we will address both of these methodological needs at the same time. First, we use the system-identification technique of reverse correlation (section 2.3.1) to probe the facial modularity of Py-feat and critically reflect on some of the patterns of results presented in this chapter.

Second, the inferred reverse-correlation kernels also give us a way to parametrically control Py-feat AU activation in videos - a technique we will then exploit in Chapter 5 to provide a more causal, hypothesis-driven evidence as to whether observers actually use internal representations that can be characterized as impulse responses.

Chapter 4

Computational Interlude: Explaining Action-Unit Detection Models

In the previous chapter, we presented experimental evidence that observers' recognition of genuine social contingency relied on the link between a speaker's speech and signals from a listener's mouth and eye regions. Moreover, judgements of contingency were well-predicted by a computational model that characterized the mapping between acoustic features and facial signals as "social transfer functions", implemented here as impulse responses/TRFs.

However, despite consistent evidence across two separate experiments for the existence of such a link, its nature remained mostly correlational. To make concrete claims about the obtained model of contingent interactions, one would need a way to experimentally *manipulate* conversational dynamics. Such manipulations would allow one to formally test whether signals matching the predictions of the model produce the desired effect on observers' perception of contingency as opposed to when there is a mismatch (a process described as going from "merely descriptive to hypothesis-driven" in Casadevall and Fang, 2008). It is also possible that Py-feat, the model used to estimate action-unit (AU) activity in the listener's face in the previous chapter, estimates the activity of a mouth-related AU by incorporating, for instance, the activity of the eyes. This would confound any comparisons between the predictability of one region of the face over another in response to speech signals. Thus, what is needed is to ascertain the modularity of estimates of facial activity by probing what exact visual information is used by machine-learning algorithms such as Py-feat.

In this chapter, we present a "computational interlude" into the domain of AI explainability in an attempt to solve both of these methodological needs at the same time. First, we use the system-identification technique of reverse correlation (Section 2.3.1) to probe the modularity of two commonly used AU detection models, Py-feat (Cheong et al., 2023) and Openface (Baltrušaitis et al., 2016), and critically reflect on some of the patterns of results obtained in Chapter 3. Second, the inferred reverse-correlation kernels used for explainability also allow for parametric control over Py-feat's AU activation in videos. This will provide the methodological basis for investigating the causality of TRFs through reverse-correlation experiments with human participants in Chapter 5.

4.1 Explainable AI

The 'technology enactment theory' (Fountain, 2004) posits that perception of technology as being objective reflects only their physical capacity, i.e. the capabilities of their hardware and software. What is overlooked is the sociocultural influence of people and institutions on how the technology is used. In other words, technology may be objective in isolation, but its use by human agents inevitably imbues it with subjectivity through the transfer, deliberate or otherwise, of our preconceptions and biases. This has been particularly well-documented in research into bias in facial analysis systems. Studies show large variance in the performance of gender classification models depending on gender and ethnicity (Buolamwini and Gebru, 2018), bias against under-represented groups in the data, mainly black and Asian females (Serna et al., 2021), and faster degradation of precision for black faces compared to white faces in automatic face recognition (Majumdar et al., 2021). Widely used software like Google Cloud Vision, Microsoft Azure Computer Vision and Amazon Rekognition have been found to display gender bias in terms of the kind of images used to represent women and how they are labelled and categorized, with images of women containing significantly more annotations based on physical appearance than images of men (Schwemmer et al., 2020). A recent survey paper echoes the theory of enacted technology in its description of the various forms of bias in AI models (Mehrabi et al., 2021). The first, called 'data to algorithm' bias, is caused by imbalanced datasets or skewed representations of certain groups affecting model training and propagating to its outputs. These biased outputs then influence user behaviour, leading to 'algorithm to user' bias. The user bias is, in turn, reflected in the data they generate which is fed into AI models for further training, resulting in 'user to data' bias. This vicious cycle of bias has, in recent years, motivated a rise in mitigation efforts through more conscientious building of datasets and, especially importantly, the development of more explainable AI systems.

Explainability in AI is an attempt to describe AI systems and highlight any potential biases. Recently, two approaches to explaining the outputs of AI models have become prominent - sensitivity analysis (Simonyan et al., 2013), which quantifies the amount of change in each feature needed to affect model predictions, and layer-wise relevance propagation (LRP; Bach et al., 2015), which evaluates how much each feature contributes to the prediction. Explanations generated by these approaches are usually visualized with saliency maps, which highlight parts of the input that are discriminative with respect to a given class and, in so doing, provide a correlational mapping between the input and the model's prediction. For instance, LRP has been used to obtain neurophysiologically interpretable explanations for the classification of single-trial EEG by indicating the relevance of each point in high-dimensional spatio-temporal EEG data (Sturm et al., 2016). While techniques like LRP, among others like gradients (Simonyan et al., 2013) and excitation backprop (Zhang et al., 2018), backpropagate the relevance score of features through the layers of the neural network from output to input, others learn the weights of features by perturbing the inputs using, for instance, occlusion (Zeiler and Fergus, 2014) and observing the corresponding effect it has on model output. One such algorithm called RISE (Randomized Input Sampling for Explanation; Petsiuk et al., 2018) offers a generalizable approach by not requiring access to the model's internal features or parameters. Instead, it probes the model by randomly sub-sampling inputs and recording the resulting model outputs before generating a final saliency map, which is a linear combination of the random masks weighted by the corresponding output probabilities. Utilizing similar principles involving randomized inputs (but with additive noise instead of occlusion), reverse correlation was originally developed to probe neurons as black-box systems (see Chapter 2.3.1). However, recent studies have begun to explore its potential for generating explanations of AI models. For instance, using a reverse-correlation procedure, random perturbations of either the texture or shape of faces revealed a strong bias towards image texture compared to image shape in a CNN trained for facial recognition (Xu et al., 2018). Similar methodologies have also been used to demonstrate that a CNN tasked with distinguishing different facial identities places greater emphasis on the eyes, eyebrows and central face region (Tian et al., 2025) and to highlight the discrepancies between the features used by DNNs and humans to process facial identities (Daube et al., 2021). In a comprehensive account demonstrating its versatility (Thoret et al., 2021), reverse correlation is used to extract both discriminative and representative features of different kinds of classifiers across various tasks, namely, the classification of written numbers by a CNN, distinguishing speech from music with an SVM and classifying sleep stages from neurophysiological recordings.

4.2 Reverse Correlation with AU Detection Models

In this chapter, we propose to leverage reverse correlation to explain machine-learning based AU detection models. As already encountered in Chapter 3, action units (AUs) are distinct facial expressions characterized by specific muscle movements which can be combined to describe basic emotions by virtue of their modularity. By categorizing subtle expressions into distinct action units, a difficult-to-quantify high-dimensional feature set is transformed into a smaller set of high-level and, presumably, cognitively meaningful features. While the detection of AUs used to require manual hand-coding by trained FACS (Facial Action Unit Coding Scheme; Ekman and Friesen, 1978) experts, this obstacle was overcome by machine-learning-based AU detection models whose ease-of-use and accessibility allowed widespread application in facial expression recognition (Tian et al., 2001), emotion recognition (Zhi et al., 2021), studying social cognition in humans (Arias-Sarah et al., 2024), non-invasive pain detection without the need for self-reports (Bouazizi et al., 2025) and synthesis of facial expressions (Roesch et al., 2011; Fan et al., 2020; Zhao et al., 2021; Guo et al., 2023). Tasked with learning relevant information in the input data for maximising classification accuracy, these AU detection models boast high accuracy. Yet, it is unclear how and why a model classifies a facial expression in a certain way. Though there have been studies attempting explainability on emotion recognition in deep learning models (del Castillo Torres et al., 2022; Weitz et al., 2019), relatively little work has been conducted to investigate machine-learning models trained to predict lower-level facial features like AUs.

To address this gap, here we create thousands of random deformations on faces of different genders and ethnicities, and force AU detection models to score the deformed images as more or less likely to represent a given AU. We use psychophysical reverse correlation, a system identification technique already proposed to be useful for AI explainability (Thoret et al., 2021), to extract representational features of the model for each AU, i.e. the template against which the model compares a given configuration of facial features and judges it to be a specific AU. The templates (referred to as *kernels* in the reverse-correlation literature; Murray, 2011) of the AUs can then be compared to see if they are modular in image space. Here, we interpret modularity *qua* Fodor's 'modularity of mind' (Fodor, 1983) and focus particularly on *information encapsulation*, meaning that an AU detection model should not allow inflow of information from regions of the face that are not necessary to detect a particular AU. In other words, we ask: does the model only 'look at' the relevant facial features for classification, or does it, for instance, consider features around the mouth to classify an eye-related AU? We further explore this modularity using adversarial examples and compare the templates across different genders and ethnicities to uncover any bias. Finally, we use these templates to compute the contributions of individual AUs to make up basic emotional expressions and compare how these compositions in AU detection models differ from the theoretical composition as indicated in the FACS. Since models have been shown to display similar performance despite using largely different features (Arras et al., 2016, 2017; Lapuschkin et al., 2016), we perform reverse correlation with two commonly used AU detection models - Py-feat and Openface - to investigate any differences in the behaviour of different models.

4.3 Methods

4.3.1 Datasets

We obtained 128 high-resolution images of faces with neutral expressions from the Chicago Faces Database (Ma et al., 2015) and the London Set from the Face Research Lab (DeBruine and Jones, 2017). Images were selected such that they were balanced across genders (male, female; 2 x 64) and the four ethnic groups defined for this study (White, Black, West Asian, East Asian; 4 x 32) (Fig. 4.1).

4.3.2 Reverse-correlation stimuli

We used the FaceWarp module in CLEESE (Burred et al., 2019) to create 2000 randomly deformed images for each image in our dataset. Using a single photograph of one actor's resting face, we first used the video tracking software Mediapipe (Lugaresi et al., 2019) to extract the 2D coordinates of 468 facial landmarks on the actor's eyes, forehead, nose, mouth and chin; we then generated new random positions for the landmarks by

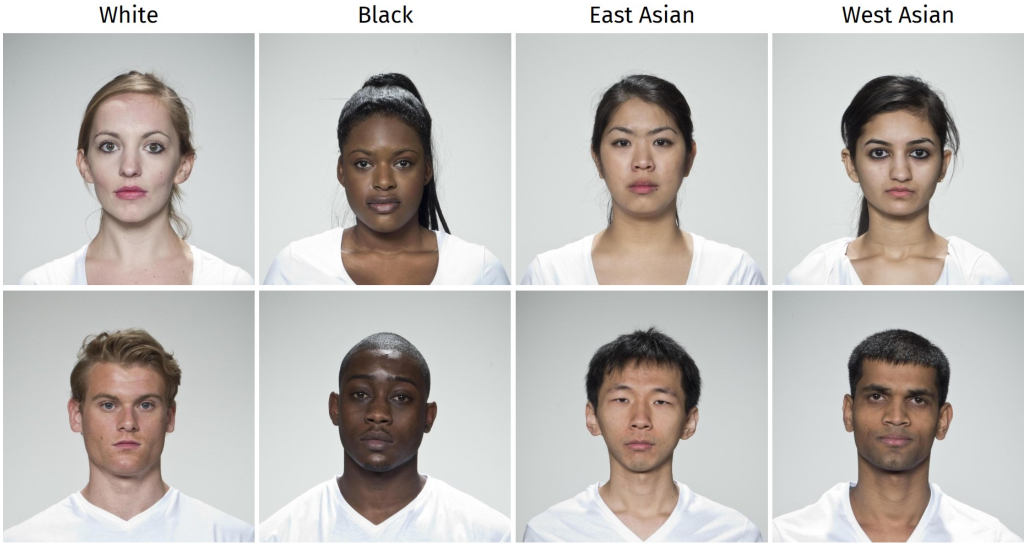


Figure 4.1: **Examples of facial identities from the 4 different ethnicities.** 128 such images were collected from the Chicago Faces Database (Ma et al., 2015) and the London Set from the Face Research Lab (DeBruine and Jones, 2017).

adding Gaussian noise to each of the (x,y) coordinates: $X_i^d = X_i + \mathcal{N}_i(0, \sigma)$ where $X_i = (x_i, y_i)$ are the original coordinates of the i^{th} landmark, X_i^d their newly-obtained deformed coordinates and $\mathcal{N}_i$ is a random sample from a truncated Gaussian distribution of mean $\mu = 0$, standard deviation σ, truncated at $\pm 2\sigma$. In order to obtain realistic deformations, we set σ empirically to be equal to $1/20^{th}$ of the pixel distance between the eyes of the original photograph (Fig. 4.2B). We then deformed the original photograph to match the position of the new landmarks, using a pixel mapping technique called rigid Moving Least Squares or MLS (Schaefer et al., 2006). MLS produces a function f that maps pixels in the non-deformed image to the deformed image, in a manner which transforms landmarks X_i to X_i^d, and smoothly interpolates all pixels in between (Fig. 4.2C). We then fill the resulting triangles using affine warping (Fig. 4.2D). Figure 4.2E illustrates some possible outputs of the procedure. The procedure was adapted from previous work by Arias et al. (2018) and Zaied et al. (2023). Using this procedure, a total of 265,000 images (128 facial identities x 2000 randomly deformed variants) were generated and used as stimuli for reverse correlation.

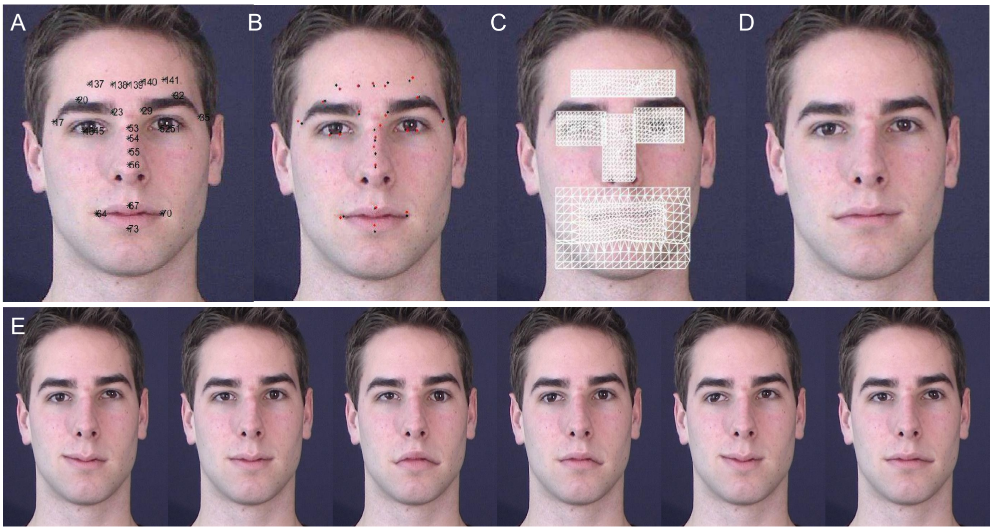


Figure 4.2: **Illustration of stimulus generation for AU explainability.** We generated facial reverse-correlation stimuli using a face deformation technique able to apply random perturbations of facial expressions in arbitrary face photographs. **(A)** We first use video tracking software to extract the 2D coordinates of facial landmarks on the actor's eyes, forehead, nose and mouth. The figure, adapted from Zaied et al. (2023), illustrates the procedure with 23 landmarks from the OpenFace video tracking software; the technique used in this thesis instead uses 468 landmarks from the MediaPipe software **(B)** We then generate new random positions (red) for each of the landmarks by adding gaussian noise to the (x,y) coordinates of original landmarks (black). **(C)** We then deform the original photograph in the neighborhood of each landmark, using a pixel mapping technique called rigid Moving Least Squares (MLS). **(D)** The resulting image is a smooth interpolation of pixels in between landmarks, mimicking a random facial expression. **(E)** Illustration of random manipulations obtained with this procedure. In the present work, we use these photographs as stimuli in two reverse-correlation experiments. *Figure and procedure adapted from Zaied et al. (2023).*

4.3.3 Procedure

Reverse correlation was then performed with two commonly used AU detection models, Py-feat (Cheong et al., 2023) and Openface (Baltrušaitis et al., 2016), for each facial identity separately. 2000 randomly deformed images of a facial identity were obtained from CLEESE and grouped into random pairs to get a total of 1000 trials. They were then fed into the AU detection models. The models assigned 'scores' to all AUs in both images in a trial and the image with a higher score for an AU was considered to be its choice or response. Reverse correlating the models' responses over all trials, we extracted a kernel for each AU of each facial identity using the classification-image technique (Murray, 2011). Specifically, for each AU, we computed the average random (x,y) displacement for each of the 468 landmarks in the 1000 images recognized as more representative of the AU, and subtracted the average random displacement of the images recognized as less representative. Kernels were then normalized by dividing them by the absolute sum of their values. For each facial identity, this procedure resulted in a 2×468 vector of (x,y) coordinates, representing the displacement to be applied to a given image in order to increase the probability of the resulting face being selected as more representative of a given AU (Fig. 4.3).

4.4 Results

4.4.1 AU detection models are not always modular

The normalized kernels of all 128 facial identities were averaged to obtain a single kernel for each AU. For each landmark in a kernel, we conducted one-sample t-tests against 0 (Bonferroni corrected across landmarks) on the deformation values along the x- and y-axis, such that only landmarks showing statistically significant deformations along either the x- or y-axis (or both) were retained. This resulted in a kernel containing the landmarks and the values by which to deform them in order for the model to detect the AU represented by the kernel in a face. The obtained kernels were then applied to images of the associated facial identities and fed into the model to obtain its scores for all AUs, in a process henceforth referred to as *decoding*. For each AU, we decoded the kernel (K^+) and its opposite (K^-), i.e. containing the same deformation values as in K^+ but multiplied by -1. Then, t-tests were conducted between the post-decoding AU scores of K^+ and K^- (over all facial identities), with

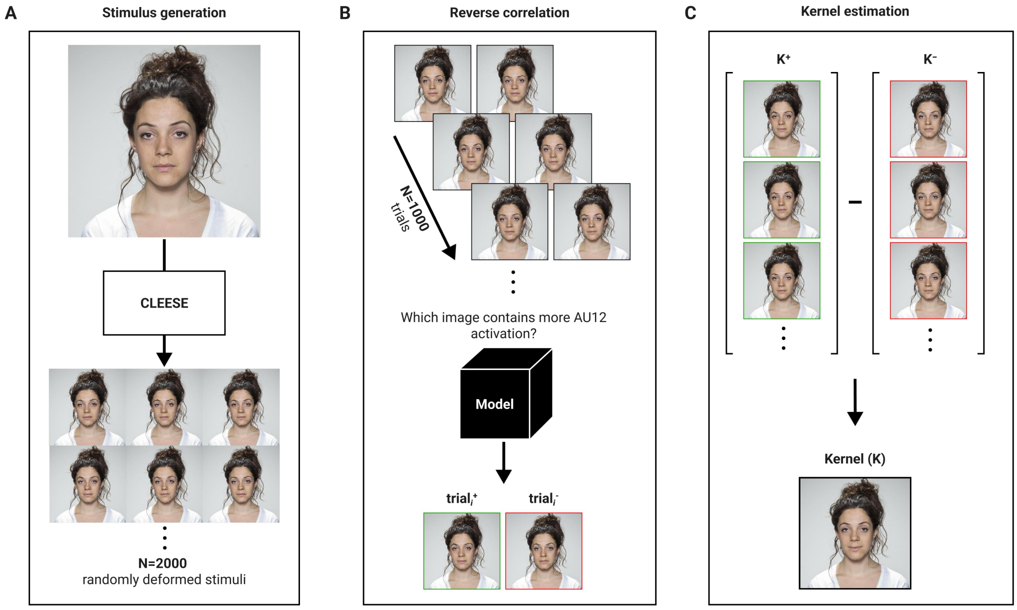


Figure 4.3: **The reverse correlation procedure for explaining AU detection models.** **(A)** Using the procedure described in Section 4.3.2, we create 2000 randomly deformed images for each of the 128 facial identities. **(B)** The reverse correlation procedure consists of assembling the randomly deformed stimuli into pairs and feeding them into the AU detection model to obtain its ratings of AUs. To obtain a binary 'response' as in typical reverse correlation experiments, we assign a value of 0 or 1 to a stimulus in the pair depending on which has a higher AU rating. **(C)** For each facial identity, and for each AU, we subtract the stimuli with positive responses (1) by the stimuli with negative responses (0) to obtain the kernel of that particular AU and facial identity.

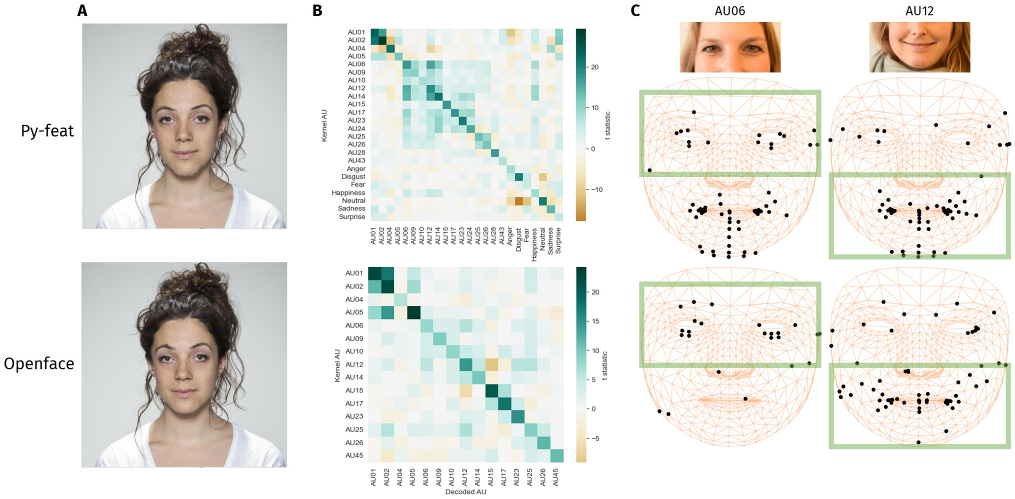


Figure 4.4: **Reverse correlation kernels show lack of modularity in AU detection. (A)** The AU12 (smile) reverse correlation kernels of Py-feat and Openface, visualized by applying them on a face, provide qualitative evidence for the convergence of kernels towards meaningful internal representations of the black-box models. **(B)** For most AUs, there are significant differences between model ratings of that AU when the kernel of that AU is applied as compared to when its opposite (or K^{-}) is applied. This suggests that the reverse correlation kernels can capture discriminative information about AUs, i.e. *which* specific facial features need to change and *how* for an expression to be classified as a particular AU. However, Py-feat's ratings of AU06 and AU12 appear to be correlated with each other, meaning that applying AU06 on a face changes Py-feat's rating of AU12 and vice versa. **(C)** This lack of modularity is confirmed by showing that Py-feat uses virtually the same landmarks to detect both AU06 and AU12 despite representing movements in different face regions (eyes and mouth, respectively). While Openface shows the expected pattern of results for AU06 by using mostly eye-related landmarks, it too appears to take advantage of eye-related landmarks to detect AU12.

the idea being that if the kernel did capture discriminative information about an AU, the difference between K^+ and K^- decoding scores would be large and consequently yield a higher T statistic.

We first visualized the AU12 (smile) kernels of both Py-feat and Openface (Fig. 4.4A) to obtain a qualitative picture of whether the reverse-correlation procedure was able to converge towards meaningful kernels. Quantitatively, we found that the reverse-correlation kernels of both Py-feat and Openface captured discriminative information about most AUs, as evident in the relatively large values of the T-statistic (dark green cells) along the diagonals of both heatmaps in Figure 4.4B. Essentially, this pattern implied that when the K^+ and K^- of a specific AU were decoded, only the ratings of that particular AU changed significantly while the others remained relatively constant. Interestingly, we observed a violation of this pattern in Py-feat, especially with AU06 (cheek raiser), AU12 (lip corner puller) and AU14 (dimpler), suggesting that a change in the intensity of AU06, for instance, induced a significant change in Py-feat's score of AU12. Concentrating on the two most non-modular AUs, we investigated which landmarks were significantly important for detecting AU06 (squinting of the eyes) and AU12 (smile) in Py-feat and Openface (Fig. 4.4C). In the case of AU06, we found that Py-feat, in addition to expected landmarks in the eye-region of the face, unexpectedly used a large number of landmarks around the mouth, while Openface mostly only used landmarks around the eyes. Remarkably, the landmarks used by Py-feat to detect AU06 were almost identical to those used to detect AU12. Contrary to expectations, both Py-feat and Openface demonstrated significant involvement of landmarks in the eye-region in addition to the mouth-region for the detection of AU12.

4.4.2 Kernels can be used to generate adversarial examples

To further investigate the influence of 'task-irrelevant' landmarks on AU ratings, we subjected the AU detection models to adversarial examples. From Py-feat's AU12 kernel, we isolated 3 landmarks: 280 (cheek), 285 (inner eyebrow) and 373 (lower left eye). These landmarks were chosen because they were found to significantly influence AU12 ratings in both Py-feat and Openface and, by virtue of being eye-related landmarks, were deemed less relevant for detecting the mouth-related AU12. Thus, we created 3 kernels containing a single landmark and the corresponding deformations along the x- and y-axis as specified by the AU12 kernel.

First, we separately applied each of the 3 single-landmark kernels

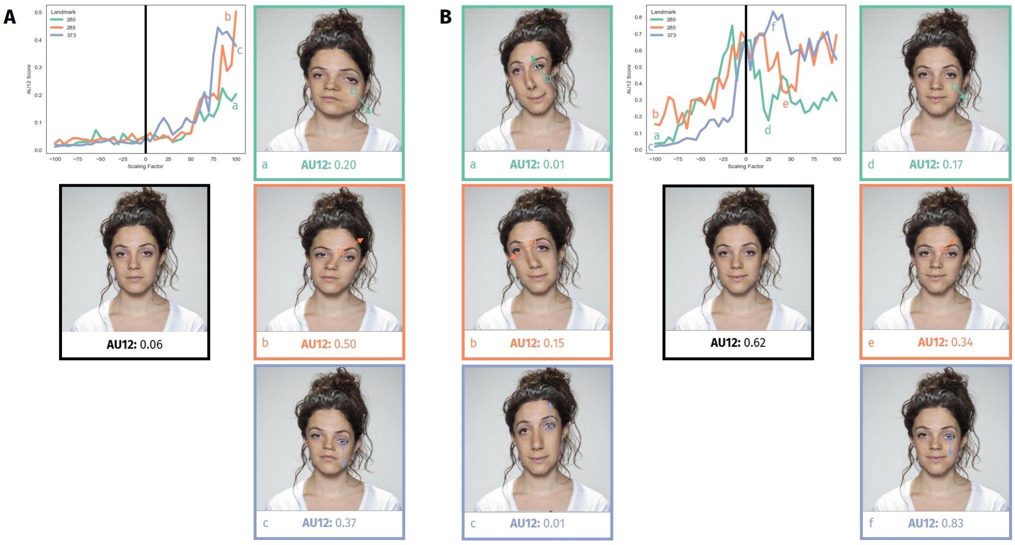


Figure 4.5: **Adversarial examples on Py-feat.** We isolated 3 landmarks in Py-feat's AU12 kernel found to have a significant impact on AU12 ratings despite not being in the mouth region - namely, 280 (left cheek, green), 285 (left inner eyebrow, orange) and 373 (lower left eye, blue). **(A)** On a neutral face, individual landmarks are deformed by varying magnitudes (i.e. by a scaling factor s, where $s \in [-100, -90, ..., 90, 100]$ with negative values representing deformation in the opposite direction). Deforming the landmarks by large positive values increases Py-feat's ratings of AU12 **(a, b, c)**, but the same is not observed for negative values of s. **(B)** A similar procedure is conducted on a face upon which the average AU12 kernel has been applied (i.e. a face already containing AU12). Large deformations by negative values of s dramatically decrease Py-feat's ratings of AU12 **(a, b, c)** despite changes occurring largely in the general shape of the head while leaving the smiles intact. Deforming landmarks by increasingly large positive values of s demonstrates significant variability in Py-feat's AU12 ratings with relatively small positive values producing large changes in AU12 ratings **(d, e, f)** despite, in most cases, very little discernible effects on the face.

to a neutral face for different values of the scaling factor s ($s \in [-100, -90, ..., 90, 100]$), i.e. the value by which to multiply the deformation values. None of the 3 landmarks affected AU12 ratings when their deformation values were scaled by negative values of s. However, with $s = 100$, Py-feat's AU12 ratings increased from 0.06 for the original neutral face to 0.2, 0.5 and 0.37 for landmarks 280, 285 and 373, respectively (Fig. 4.5A). The same procedure was repeated with the base figure containing a decoded AU12 kernel (i.e. already containing a smile). Large negative values of s dramatically decreased Py-feat's AU12 ratings of the faces from 0.62 to 0.01 for landmarks 280 and 373, and to 0.15 for landmark 285 (Fig. 4.5Ba, b and c). On the other hand, positive values of s resulted in a more volatile trend without any consistent increase or decrease in AU12 ratings. Even relatively small positive values of s that did not produce any discernible changes to the face resulted in a sharp decrease in AU12 rating from 0.62 to 0.17 for landmark 280 ($s = 25$) and an increase to 0.83 for landmark 373 ($s = 30$), while the more visible deformation of landmark 285 ($s = 45$) caused the AU12 rating to fall by half (Fig. 4.5Bd, e and f).

4.4.3 AU detection models do not show systematic gender or ethnicity bias

To investigate whether a model's internal representations of AUs vary with gender and/or ethnicity, we compared the kernels of the two groups of interest (e.g. male vs. female). We conducted t-tests between the deformation values of each of the 468 landmarks (Bonferroni corrected) over all faces in the two groups. This was done for every available AU in the model and visualized in the form of a heatmap showing, for each AU, whether there was at least one landmark with deformation values that were significantly different between the two groups being compared.

T-tests between the deformation values of landmarks as specified in the kernels of different genders and ethnicities revealed almost twice as many AUs with at least one significantly different landmark in Openface as compared to Py-feat. Compared to eye-related AUs, there are also almost twice as many mouth-related AUs with at least one significantly different landmark. Globally, the sparsity of the distribution of significant AUs and the lack of consistency across group comparisons in terms of which AUs were found to be significant did not indicate the presence of any systematic bias against a particular gender or ethnicity. Despite the total number of AUs/emotions with differences between groups, closer

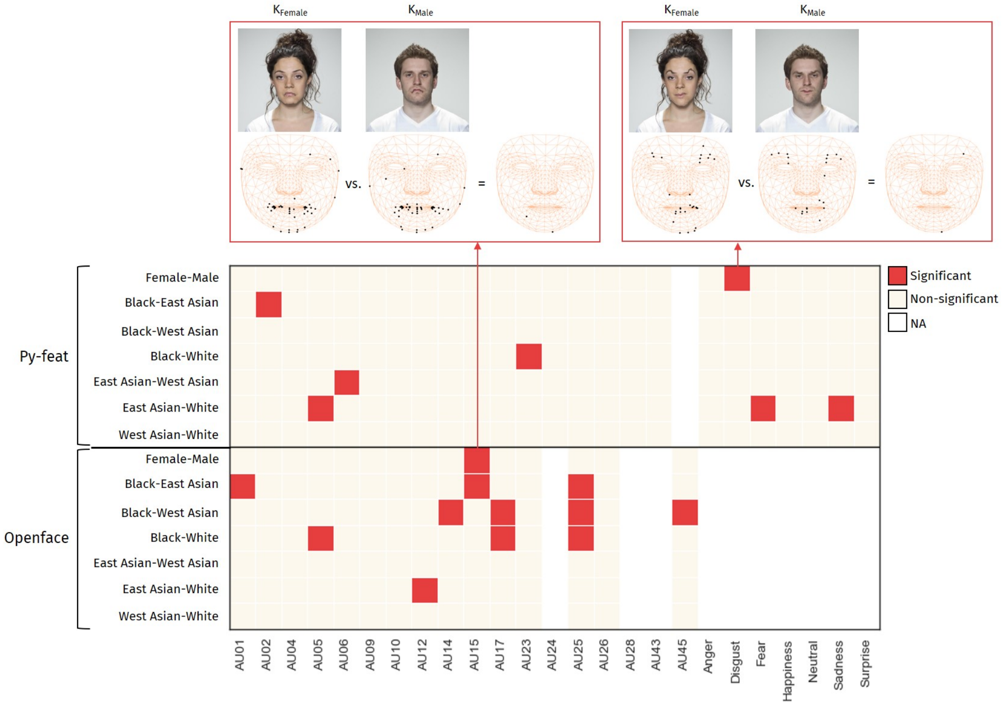


Figure 4.6: **Lack of systematic gender and ethnicity bias in Py-feat and Openface.** Conducting t-tests between deformation values of the kernels between groups reveals that no specific group comparison consistently yields a greater number of significantly different AUs (as seen in the sparse and inconsistent distribution of red cells in the heatmap). Examining significant AUs more closely shows that the significant differences are driven only by a few landmarks (2 on average and a maximum of 4 in one comparison for one AU). This is highlighted by visualizing the AU15 (top left) and disgust (top right) kernels of male and female faces and showing that there are only 2 significantly different landmarks in both. On the whole, the relatively similar spatial distribution of landmarks used by the models for different groups suggests that these models do not show systematic bias against any specific gender or ethnicity.

inspection revealed that most of these comparisons contained only one or two significantly different landmarks (with only a single comparison between West Asian and Black identities on AU45 showing a maximum of 4) (Fig. 4.6).

4.4.4 Kernels can quantify the composition of emotions in terms of AUs

Finally, we took the decoding scores of Py-feat's emotion kernels for all facial identities and used them to model the composition of basic emotions (happiness, sadness, surprise, fear, anger and disgust; Ekman and Friesen, 1971) in terms of AUs. We used ridge regression with 5-fold cross-validation to model the relationship between standardized decoding scores of AUs and the decoding score of the corresponding emotional expressions. Regularization parameters were optimized across a grid from 10^{-3} to 10^{3}, and coefficients were extracted to quantify each AU's contribution to emotion prediction while controlling for multicollinearity among AUs. The coefficients of the independent variables were taken to represent the extent of their contributions to the detection of the emotion specified as the dependent variable and visualized using radar plots. These plots were then compared qualitatively with the theoretical composition of emotions, i.e. the AUs whose combination is *supposed* to give rise to a given emotion as per the FACS.

Results showed that while the relationship between AUs and most of the emotions were relatively well-matched with theoretical expectations, there were a few discrepancies. While the composition of happiness showed the expected contributions of AU06 and AU12, it also showed a large contribution from AU02 (outer brow raiser). Similarly, both sadness and surprise demonstrated largely the same patterns as their theoretical compositions, but with the incongruous addition of AU25 (lips part). The composition of fear displayed lower contribution from AU04 (brow lowerer) and greater contribution from AU26 (jaw drop) than expected. On the other hand, the composition of anger and disgust deviated significantly from expectations, with anger showing contributions from a wide range of other AUs while disgust showed the greatest contribution from AU06, AU09 (nose wrinkler) and AU12 rather than the expected combination of AU09 and AU15 (lip corner depressor) (Fig. 4.7).

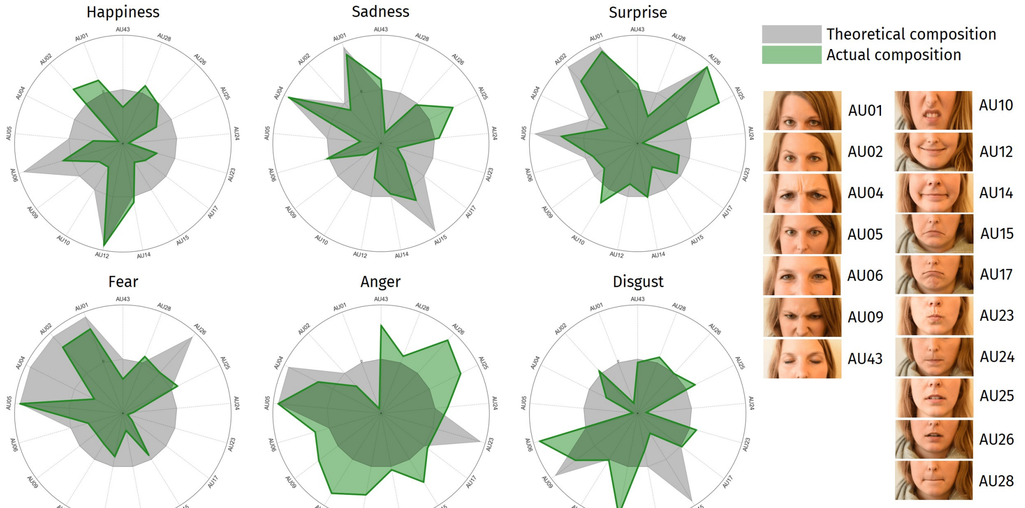


Figure 4.7: **Composition of emotions in terms of AUs.** Applying Py-feat's emotion kernels on the 128 facial identities, we model their emotion ratings with the corresponding ratings of all other AUs as predictors. Taking the regression coefficients (green) as the extent of contribution to the classification of a face as an emotion, they are compared against the theoretical composition of emotions (black) suggested by the FACS. Results reveal several discrepancies, like the involvement of AU01 and AU02 in happiness, AU25 in sadness and surprise and AU26 in fear. The compositions of anger and disgust, in particular, display significant deviance from the theoretical composition, with many different AUs involved in anger and AU06 and AU12 appearing to be the largest contributors to disgust.

4.5 Discussion

In this chapter, we leverage techniques from psychophysics, namely reverse correlation, to explain machine-learning action-unit detection models. We show that reverse-correlation kernels capture discriminative information about the features utilized by Py-feat and Openface to make judgements about the presence of AUs. Taking advantage of the discriminative information, we first highlight the lack of modularity in these models by showing that ratings of, e.g. AU12, are influenced by features in the eye-region in both models. Second, we use information from AU12 kernels to isolate some landmarks that ideally should not impact a model's ratings and show that the models can be sensitive to small changes that are seemingly imperceptible to humans. Furthermore, we demonstrate the method's potential to uncover gender or ethnicity differences in how the models internally represent facial expressions. While these differences exist, they are sparsely distributed across AUs and participant categories, and we did not find any systematic bias against one specific gender or ethnicity. Finally, we demonstrate that reverse-correlation kernels can be used to decompose high-level emotional facial expressions into their component AUs, with potential for achieving more fine-grained control over complex facial expressions and thus provide useful insight into how models internally represent these emotions in comparison to what can be expected theoretically.

The fact that kernels capture discriminative information about the AU detection models shows the viability of reverse correlation as a technique to explain machine-learning based models, a point also made by Thoret et al., 2021. Unlike some other algorithms for explainability, kernels computed from the reverse correlation procedure not only reveal *which* parts of the image affect model decisions but also *how* those parts need to be manipulated to induce that effect. Access to this information expands the utility of these kernels by allowing us to apply them to the original stimuli and probe the model again, in a process referred to in this study as decoding. Decoding the kernels thus provides a mechanism for quantifying how much of the information contained within them is actually discriminative or influential in driving the model's decisions. In further work, one could also potentially present the model's kernels as image stimuli to human participants, as well as participants' kernels as stimuli to the machine-learning models, and examine the differences between the features used by machines and humans.

The finding that Py-feat's AU06 kernel focuses heavily on landmarks around the mouth instead of the eyes and that the AU12 kernel focuses on the eyes in addition to the mouth shows that AU detection models do not always ensure modularity. This is significant because the FACS aims to break down complex social signals into combinations of independent, smaller units. If these smaller units are conflated with each other by AU detection models or are not treated modularly, AU activity inferred by machine-learning models is potentially confounded. For instance, using Py-feat to detect 'genuine' or Duchenne smiles, which are supposed to be a combination of AU06 and AU12, will likely produce incorrect results given the positive dependence of the two AUs in Py-feat and its use of virtually the same landmarks to detect both (Fig. 4.4C). The similar finding that both Py-feat and Openface detect AU12 using landmarks around the eyes in addition to those around the mouth poses similar problems. In the context of this thesis, these results provide a sobering look at the interpretations of Chapter 3, which showed some ambiguity as to whether participants used eye or mouth regions to process cues of social contingency. Because these were based on correlations with TRF-predictions of AU activity estimated by Py-feat, if AU activity is incorrectly quantified using similar facial landmarks, it could be that activity that was correlated for one AU is also correlated for other, unrelated AUs. Similar correlational confounds may plague a number of findings relying on machine-learning analysis of AUs in datasets, and highlight the need for experimental paradigms that provide stricter causal control on physical cues - something we address in Chapter 5. One possible cause for the lack of modularity could be the fact that they were trained on datasets containing naturalistic expressions. If the expressions were annotated for each AU independently and the distribution of the AUs is not independent, then the activity of one AU could be used as a shortcut to classify the other (Geirhos et al., 2020). Consequently, a model trained on a dataset featuring faces that only display parametric activation of a single AU, or with a random i.i.d. (independent and identically distributed) combination of AUs, might display more modularity.

In this chapter, the issue of modularity is further highlighted by evidence showing that AU detection models are sensitive to 'adversarial examples' featuring changes in ostensibly irrelevant regions of a face. We show that manipulating individual landmarks on the cheek, inner eyebrow, and lower left eye of a neutral face induces a large increase in Py-feat's rating of AU12 despite the lip corners actually moving downwards in some cases. Similarly, we find that manipulating these landmarks on an already smiling face also dramatically reduces or increases AU12 ratings depending on the direction of the manipulation. A few of these manipulations produced barely perceptible differences in the face, and yet yielded a significant change (decrease for the landmark on the cheek and increase for the landmark on the lower left eye) in the model's AU12 rating. Sensitivity to these seemingly "task-irrelevant" landmarks could reveal analytical strategies taken by the models to quantify task-relevant features. For instance, it is possible that faced with the task to learn a linear combination of landmark positions that correlates with AU12, models converge on estimating the distance between the lip corner and, e.g. the ipsilateral upper eyelid. Deformation of the eyelid could then lead the model to wrongly attribute it to a smiling expression. However, given that large deformations tend to create unnatural modifications to facial morphology, it might be anticipated that Py-feat, having been trained on naturalistic facial expressions, would fail to detect AU12 in images featuring unnatural face shapes. Although this could account for the reduced AU12 ratings observed when such deformations are applied to smiling faces, the persistence of elevated AU12 ratings when large deformations are applied to neutral faces remains noteworthy. This pattern suggests that, beyond global changes in face morphology, Py-feat also shows sensitivity to specific features being manipulated.

While there were examples of differences in AU detection across participant groups, we could not draw any conclusions about the presence of bias against a particular gender or ethnicity. For instance, while Openface displayed relatively consistent differences for AU15, AU17 and AU25 across group comparisons, closer examination revealed that only 2-3 landmarks actually differed in terms of how they needed to be deformed. The lack of large-scale differences in facial features suggests that the machine-learning models tested here do not exhibit the kind of social-cognitive stereotypes usually encountered in human participants (e.g. systematically associating one ethnicity with one personality trait Gingras et al., 2023). However, the presence of small-scale differences in a handful of landmarks might instead suggest a more granular algorithmic bias. This is borne out by the adversarial examples showing large changes in Py-feat's AU12 ratings as a result of deforming individual landmarks. This could potentially be investigated by comparing the effects of the same adversarial examples on different genders and ethnicities. It is also possible that the differences are the result of morphological differences in faces, depending on the gender or ethnicity. Differences in face morphology driving differential ratings of AUs and emotions raise the question of what constitutes bias. Is a model biased if its outputs are different for different groups? Or is it biased if it disregards differences and generates the same outputs for different groups (incidentally, corresponding to how the statistical term of bias is understood)? A recent study found that their seemingly effective algorithm to mitigate biased outputs in LLMs failed when the LLM was deployed in a different context (Ma et al., 2025). This context-dependence highlights the need for a more dynamic understanding of bias instead of ascribing it statically.

Discrepancies in the composition of emotions (in terms of AUs) in this study and the theoretical composition described by the FACS and the classical affective-science literature, if taken at face value, suggest issues in Py-feat's representations of the emotions, particularly for anger and disgust. However, studies have shown that while machine-learning models show high accuracies for classifying emotions in standardized datasets containing trained actors displaying prototypical posed emotions, their performance is more variable for emotional displays of non-standardized spontaneous emotional expressions (Dupré et al., 2020; Küntzler et al., 2021). Since Py-feat's models for emotion detection were trained on both posed and naturalistically elicited emotional expressions (Cheong et al., 2023), it is thus possible that Py-feat captures the true composition or at least a greater amount of the variability of emotional expressions (Jack et al., 2012). Indeed, several studies show that theoretically proposed AU patterns in emotional expressions are not backed up by empirical findings (Sato et al., 2019), with actors often either not displaying all the prototypical AUs or displaying AUs other than those predicted (Gosselin et al., 1995). Moreover, previous work also suggests that machine-learning classification of non-standardized portrayals of emotional expressions is often worse for negative emotions like anger and disgust than for happiness (Stöckli et al., 2018). This could be a potential explanation for the greater discrepancies between empirical and theoretical compositions of anger and disgust that we observe here.

Taken together, our results highlight the need for careful evaluation of the outputs of such black-box models. Indeed, the authors of the Py-feat toolbox acknowledge that the datasets on which their models are trained may be unbalanced and advise users to verify the outputs. However, the accessibility and ease-of-use of these tools, combined with the difficulty of manually verifying their outputs, often means that they are taken at face value. The tendency to highlight model performance on benchmarking datasets further adds to the issue by focusing on task performance statistics instead of providing more qualitative explanations of how such performance is achieved (Firestone, 2020). Moreover, while action units are a convenient decomposition of more complex facial expressions, their characterization as (still relatively high-level or rather, not sufficiently low-level) muscle movements can pose a problem, especially in experimental settings, where creating highly controlled stimuli is much sought after. Given that the FACS does not specify exactly how muscles need to be configured to 'make' a specific AU, it becomes difficult to use them as a basis for creating precise stimuli according to precise specifications. It is here that the kernels obtained from reverse correlation can come in and bridge the gap between low-level landmarks and high-level action units by mapping emotional expressions and AUs to their corresponding landmarks. As shown here with the decoding paradigm, kernels obtained from performing reverse correlation on machine-learning models can also function as filters for generative models. The ability to visualize these kernels on images allows them to be used as bases for creating images or videos of persons displaying a particular AU or emotional expression. This is particularly useful for generating synthetic media (for animators, for instance) as well as creating experimental stimuli containing fine-grained manipulations, something we now turn to in the next chapter.

Roadmap

In this chapter, we probed two commonly used black-box models for action-unit detection using the system-identification technique of reverse correlation. We highlight the need for more careful use and inspection of the outputs of these models, particularly in experimental settings, due to their lack of modularity and sensitivity to small, ostensibly irrelevant features and present a few examples of the different uses of reverse-correlation kernels as generative models.

In the next and final empirical chapter of this thesis (Chapter 5), we use the AU12 kernels of Py-feat and Openface to obtain a kernel containing only the landmarks important for both models in addition to excluding landmarks outside the mouth region. We then use the filtered kernel to generate smiles in videos such that the amplitude of smiling activity is scaled in accordance with the dynamics of random impulse responses. We then use these video stimuli to extract classification images of the impulse responses possessed by third-party observers to process social contingency. In doing so, we attempt to provide a more causal test of the hypothesis put forth in Chapter 3.

Chapter 5

Observer perception of social contingency (redux): A reverse-correlation experiment

In the work described in Chapter 3, we extracted a 'social transfer function' capturing the relationship between the speakers' speech and listeners' facial action units and showed that the extent to which it fit actual behaviour was associated with subjective ratings of contingency. However, these preliminary results had several limitations. First, the modelled interactions were embedded in a very specific 'speed dating' context that could potentially have altered the organization and timing of conversational phenomena like backchannels. Perhaps more fundamentally, while we argued that convolution with impulse responses was a parsimonious way of representing input-output mappings, the experimental evidence in Chapter 3 was largely correlational. Thus, whether observers encode and decode contingency in social interactions in this manner remains an open question.

In this chapter, we construct a novel experimental paradigm which aims to directly probe third-party observers' internal representations of social contingency (instead of inferring them from correlations between observers' judgements of contingency and predictions of transfer functions). Using reverse correlation, a psychophysics technique based on system-identification principles (see 2.3.1 and Chapter 4), we attempt to construct transfer functions of socially contingent smiles in response to speech in a data-driven manner. To do this, we use a modified version of the AU12 reverse-correlation kernels obtained from Py-feat and Openface in Chapter 4 to synthesize videos containing artificial smiling behaviour. We convolved different naturalistic speech extracts with (functionally) many random transfer functions. To avoid the constraints imposed by built-in assumptions inherent to a corpus study, we forced individuals to choose between two randomly varying temporal structures of smiles over a number of trials. In doing so, we are able to probe their perceptual mechanisms more directly and uncover how they internally represent the dynamics of socially contingent smiles.

Conceptualizing social transfer functions or impulse responses as cognitive mechanisms underlying social perception raises the question of abstraction or flexibility. Essentially, do social transfer functions encode abstract rules that can be applied flexibly to disparate speech inputs or do they encode input-specific rules resulting in different social transfer functions for different speech inputs? To investigate this, we first derive reverse-correlation kernels for 2 different input sentences and test for statistically-significant differences between them (Section 5.1); then, in a second validation study (Section 5.2), we test whether the kernels generalize to a larger number of possible sentences.

5.1 Study 1: Extracting observers' internal representations of contingent smiles

5.1.1 Methods

TRFWarp

To create stimuli for this study, we extended the functionality of CLEESE (Burred et al., 2019), an open-source Python toolbox that generates stimuli for reverse correlation experiments (and was also used in Chapter 4). The current version of CLEESE consists of two transformation "engines": *PhaseVocoder*, which creates random fluctuations around the acoustic feature contour of a given audio and *FaceWarp*, which creates random deformations of facial expressions in images. To create the random video stimuli required for this study, we utilized certain functionality of both and unified them within a new engine called *TRFWarp*.

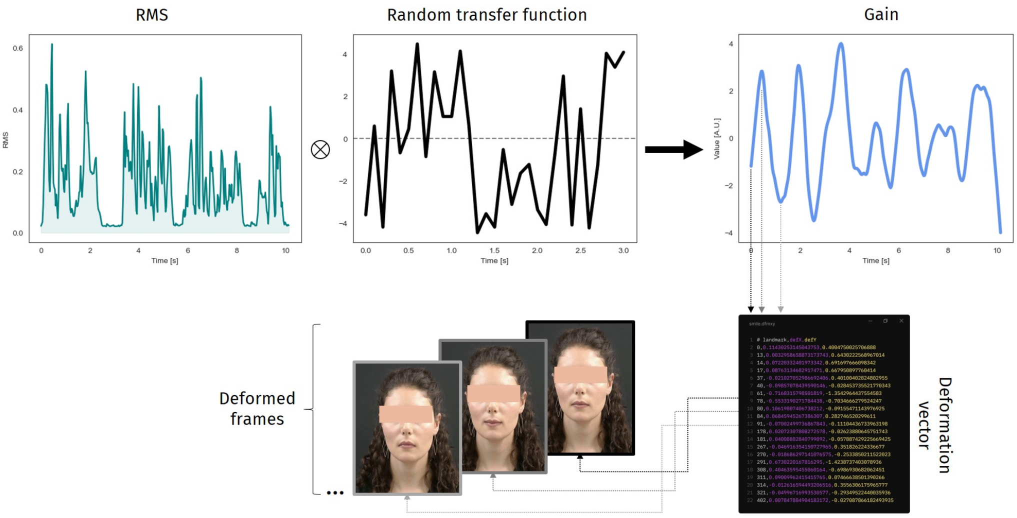
First, we reused the logics of *PhaseVocoder*'s stimulus design to generate a set of random breakpoint functions (BPFs) at specific time steps that define how the desired parameter varies over time. The parameter to transform may be constructed by linear interpolation between the breakpoints (ramp) or by having square signals with sloped transitions (square). The breakpoints are sampled from a Gaussian distribution, and the amplitude or intensity of the desired transformation is controlled by a standard-deviation parameter (with the ability to avoid extreme values by truncating the distribution by multiples of the standard deviation; see Burred et al. 2019 for details). While *PhaseVocoder* was originally intended for random transformation of pitch contours (i.e. randomizing stimuli directly in stimulus space), here we propose to use the random BPFs as random impulse responses and convolve them with an input signal to generate the stimuli.

Second, we reused functionality from CLEESE's *FaceWarp* engine to synthesize smiles in videos of faces. The AU12 (or smile) kernel obtained from reverse correlating AU detection models in Chapter 4 was provided as the deformation file containing a list of facial landmarks and their corresponding deformation vectors, i.e. vectors representing the magnitude and direction of shift of each landmark. To apply the deformations on any given image, CLEESE's *FaceWarp* engine uses tools like Mediapipe (Lugaresi et al., 2019) and dlib (King, 2009) to extract the coordinates of the landmarks specified by the deformation file. Provided with a set of Cartesian coordinates representing the locations of specific facial landmarks in the image, *FaceWarp* then performs linear warping of those locations as specified by the deformation vectors through rigid transformations using Moving Least Squares (Schaefer et al., 2006) - see Section 4.3.2 for details. This smile transformation is applied to videos by simply repeating the process for each frame in the provided video.

The *TRFWarp* engine integrates the above processes to generate video stimuli containing synthesized smiling activity in response to some speech, and whose dynamics are governed by a random impulse response (Fig. 5.1). From a software-engineering perspective, the main function in the engine takes an audio file and a configuration file as arguments. Users primarily interact with a configuration file where they can modify the parameters to generate the desired stimuli. Some key variables defined in the configuration file are:

- `mediaFile`: the path to the base video to be deformed
- `kernelFile`: the path to the deformation file to use to deform the video
- `trfDuration`: the duration of the randomly generated transfer function (in seconds)
- `window.len`: the window size to determine the spacing between successive breakpoints (in seconds)
- `std`: the standard deviation for defining the limits of the amplitude or intensity of the transfer functions

Provided with the requisite parameters, *TRFWarp* first extracts the RMS (intensity) of the audio and the individual frames of the video specified in `mediaFile`. It then proceeds to use the *PhaseVocoder* engine to



```
# landmark,defX,defY
0,0.11430253145043753,0.4004750025706888
13,0.0032958658873173743,0.6430222568967014
14,0.07220332401973342,0.691697666098342
17,0.08763134682917471,0.667950897760414
37,-0.021027052986692406,0.40100402824802955
40,-0.09857078439590146,-0.028453735521770343
61,-0.7168315798501819,-1.3542964437554583
78,-0.5533190271784438,-0.7034666279524247
80,0.10619807406738212,-0.09155471143976925
84,0.06845945267386307,0.282746520299611
91,-0.07002499736867843,-0.11104436733963198
178,0.02072307808272578,-0.02623880645751743
181,0.04008882840799892,-0.057887429225669425
267,-0.046916354150727965,0.351826224336677
270,-0.018686297141076575,-0.2533850211522023
291,0.6730220167816295,-1.4238737403078936
308,0.40463595455060164,-0.6986930682062451
311,0.09009962415415765,0.07466638501390266
314,-0.012616594493206516,0.3556306175965777
321,-0.04996716993530577,-0.29349522440035936
402,0.007847884904183172,-0.027087866182493935
```

Figure 5.1: **The TRFWarp procedure.** The engine extracts the RMS intensity from the provided audio file, generates a random impulse response whose points are sampled from a Gaussian distribution, and computes the convolution between the two to get an output time-series, which is interpreted as the *gain* of a transformation applied to the listener's face. Each time point in the gain is multiplied by a provided deformation vector. If the deformation vector defines a smile transformation, then multiplying it by gain values essentially means scaling the intensity of the smile. The scaled deformation vector is then applied to individual frames of a provided video file, which are then combined to give a video containing dynamic synthesized smiles driven by a linear combination of an audio input and a random transfer function.

generate a random transfer function of `trfDuration` seconds with random breakpoints at equally spaced intervals defined by `window.len`, and computes its convolution with the audio RMS to give the corresponding *gain* time series. Each value in gain is the factor (positive or negative) by which to scale smile intensity at that time instance by multiplying it with the set of deformation vectors stored in `kernelFile`. This scaled set of deformation vectors is then applied to the `mediaFile` frame at the corresponding time instance. Iterating over all gain values and applying the scaled deformation to the corresponding `mediaFile` frame, we combine the deformed frames into a video to obtain the desired stimulus (Fig. 5.1).

Stimuli

For this study, we selected two naturalistic speech extracts (*S04* and *S52*) lasting 10 seconds from the Speed Dating corpus used in Chapter 3. Both speech extracts featured different male voices originally addressing a female listener, and were chosen on the basis of their neutral semantic content and differences in terms of the number and timing of pauses within the speech (Fig. 5.4A). Stimulus *S04* was transcribed as *"euh moi j'aime beaucoup... j'aime beaucoup lire, apprendre des choses euh, j'aime beaucoup le cinéma... ouais j'aime beaucoup le cinéma... quoi d'autre.* (English: *"Umm, I really like... I really like reading, learning new things umm I really like movies... yeah, I really like movies... what else*."). Stimulus *S52* was transcribed as *"j'ai fait plusieurs euh... plusieurs sports différents... euh et non, non actuellement je, je fais un peu de tout*" (English: *"I've done several umm... several different sports... umm and no, no, currently I do a bit of everything*.")

The base video to be deformed, obtained with permission from a study by **?**, involved a female individual in 'resting state', i.e. maintaining a neutral expression while looking directly into the camera. To avoid any confounding effect of eye blinks on the perception of contingency while also preserving the subtle but important natural biological head motion (**?**), we extracted an approximately 3-second segment of the original video that did not contain blinks and extended it to match the length of the speech extract by reflecting the frames $(f_1, f_2 ... f_n | f_n, ... f_2, f_1 | f_1, f_2 ... f_n)$. Additionally, the RMS of the speech extracts were z-scored before convolution with random transfer functions to induce greater variance in the resulting gain, and re-scaled to be non-negative ($\tilde{g} = g - \min(g)$) following evidence from a pilot study showing significant bias against 'negative smiles'. Finally, the gains were smoothed using locally-weighted scatterplot smoothing (LOWESS;

Cleveland, 1979) to avoid rapid fluctuations in smiles, and truncated at a hand-picked value above which unnatural distortions begin to appear in the image. Since the duration of each trial was significantly longer than in typical reverse correlation experiments, we opted for a conservative experiment duration of 30 minutes, which reduced the likelihood of participant fatigue but also limited the number of trials we could present. Thus, we generated 240 video stimuli, which were assembled into random pairs to obtain a total of 120 trials. Whether a given number of trials is sufficient for a particular reverse correlation task is an empirical question and depends partly on difficult to anticipate cognitive aspects (such as the existence of a single unimodal sensory representation) or participant characteristics like consistency in decision-making (see e.g. Adl Zarrabi et al., 2024; Burred et al., 2019 for further discussions). We assess the suitability of the number of trials presented in this study for the given task in Section 5.1.2 below.

Participants

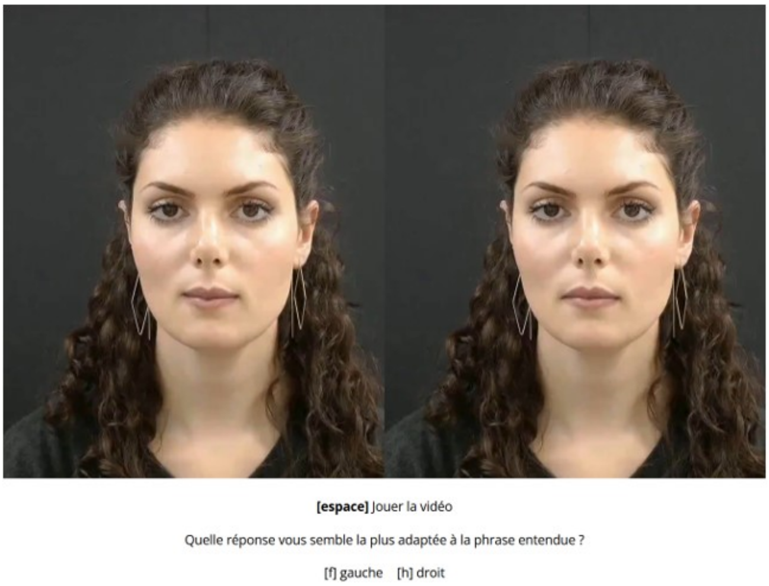
N = 27 (male: 19; M=35.2) native French speakers were recruited online on Prolific. Participants gave their informed consent and were compensated at a standard rate.

Procedure

The experiment was conducted using JONES, an online platform for reverse correlation experiments. Participants were presented with 120 trials in a 2AFC design, lasting approximately 30 minutes in total. There were an equal number of trials for the two speech extracts, i.e. 60 trials per speech extract. Each trial required participants to simultaneously watch a pair of videos of the same person with different smiling behaviour driven by random transfer functions while listening to the same speech extract. At the end of the trial (which could only be played once), participants were asked which of the two videos contained smiling behaviour that was most appropriate in response to the speech.

Statistical analysis

Using the classification image technique (Murray, 2011), we computed each participant's internal representation of the ideal contingent smile in response to each of the two speech extracts. Specifically, we subtracted the average gain of the stimuli that were not chosen from the average



[espace] Jouer la vidéo

Quelle réponse vous semble la plus adaptée à la phrase entendue ?

[f] gauche [h] droit

Figure 5.2: **Example of an experimental trial.** Stimuli containing smiling behaviour driven by random transfer functions were assembled side-by-side into random pairs. In each trial, participants watched the two videos simultaneously while listening to some speech. At the end of the trial, they were asked which of the two videos contained smiling behaviour that was most appropriate for the given audio.

gain of the chosen stimuli. The resulting *gain kernel* K_{Gain} (i.e. the kernel in stimulus space) was then normalized by dividing it by its root mean square. Similarly, by subtracting the average random transfer functions used to generate the gains of the non-chosen stimuli from the average transfer functions used to generate the gains of the chosen stimuli, we computed the transfer-function kernel (K_{TRF}), i.e. the kernel in TRF space - the rationale being that the participant's transfer-function kernel encodes the dynamics of contingent smiles and is responsible for generating gain predictions that are matched with the actual observed gain to determine contingency.

Given that the number of trials in this experiment was lower than in typical reverse correlation experiments, we performed some additional processing. First, we smoothed the K_{TRF} of each participant using LOWESS smoothing. Second, we computed participants' "internal noise" (Neri, 2010) or their level of consistency in applying an internal representation to repeated stimuli. As implemented in the open-source Python toolbox PALIN, this involves adding Gaussian noise (internal noise) to the responses of an idealized participant model, estimating the probability of its response bias for different standard deviations (between 0 and ± 5) of the additive noise and then finding the value of internal noise that minimizes the error between observed and predicted values for each participant's response bias. Because this experiment did not involve repeated (double-pass) trials, we used a noise-estimation method called 'Intercept' that was recently introduced by a colleague (ZARRABI, 2025). Five participants with internal noise greater than 2.5 standard deviations of the stimulus noise were excluded from part of the analysis, leaving 22 participants. While internal noise helped estimate consistency within individuals for a given speech extract, we further examined consistency between the two speech extracts based on the mean value of each sentence's K_{TRF}. As reported below (Fig. 5.5, we found evidence for bimodality in response strategies across participants. The dominant strategy with negative-valued K_{TRF} involved N=13 participants, and the alternative strategy with positive-valued K_{TRF} involved N=9 participants. Result subsections below indicate which of these groups of participants were analysed in each instance.

Finally, when testing for statistical significance of the kernels, the *S04* and *S52* K_{TRF} of participants were averaged and tested for significance at every time point with one-sample t-tests against 0. To test whether the K_{TRF} of the two speech extracts were statistically similar, we conducted paired t-tests between their values at every time point.

5.1.2 Results

Convergence of Kernels (N=27)

Correlating the final K_{TRF} and K_{Gain} of participants with their kernels after each trial enabled us to analyze whether the number of trials was sufficient for participants to converge to a final kernel. The results indicated, per expectations, that 120 trials were sub-optimal for convergence, particularly in the case of the K_{TRF} (Fig. 5.3), with no visible flattening or plateauing of the curves, suggesting the need for a greater number of trials for a clearer picture of observers' internal representations. This discrepancy between the rate of convergence between the K_{TRF} and K_{Gain} could reflect a relatively rapid agreement about the general occurrence of smiles during specific parts of the speech, while the precise onset and offset latencies of smiles require a greater number of trials. While future work should examine the impact of a larger number of trials, that impact

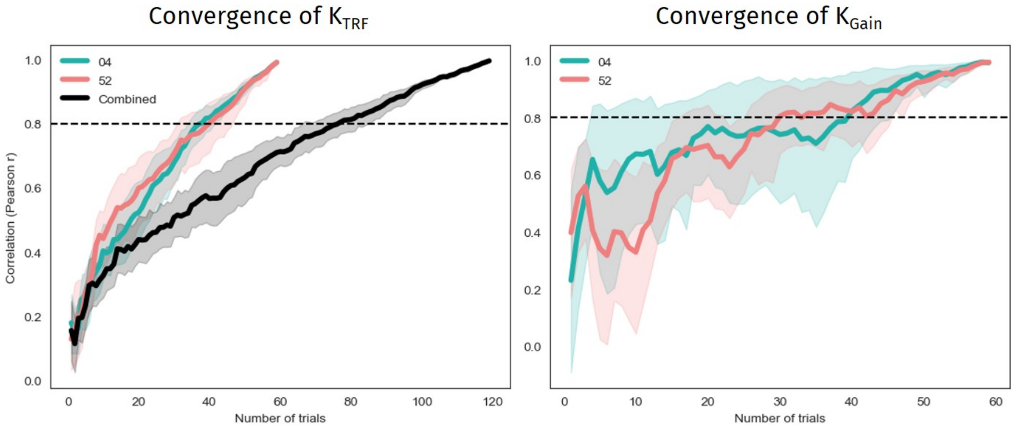


Figure 5.3: **Convergence of TRF and Gain kernels.** Each participant's final K_{TRF} and K_{Gain} is correlated with their kernel after each trial to see whether observers can converge to similar representations within the given number of trials. Despite the suboptimality of the number of trials, particularly in the case of K_{TRF}, the number of trials was not increased to maintain a reasonable experimental duration.

should also be weighted against the experiment duration, kept here at a reasonable 30 minutes.

The Dominant Strategy (N=13)

Analyzing the K_{Gain} of participants, we found that they were strikingly well-adapted to the distinct speech extracts. Both gain kernels showed participant preference for contingent smiles occurring during pauses in the speaker's speech, peaking around 1s (M=1.09, SD=0.4) after the pause onset, and returning to baseline upon the resumption of speech (Fig. 5.4B). Remarkably, even though the gain kernels of the two speech extracts were closely aligned with the distinctive characteristics of each and thus widely different from each other, their K_{TRF}s were relatively similar (Fig. 5.4C). Indeed, we found no statistically significant differences between the two, except at a solitary time point $t = 1$ (t(12)=-5.5, p=.01) (Fig. 5.4D), suggesting that the dynamics of contingent smiling behaviour, despite disparate inputs, might be encoded by sentence-independent representations. Taking advantage of the lack of difference between the K_{TRF} of the two sentences, we proceeded to focus on the combined kernel (Fig. 5.4E) for further analysis. The global negativity of the combined ker-

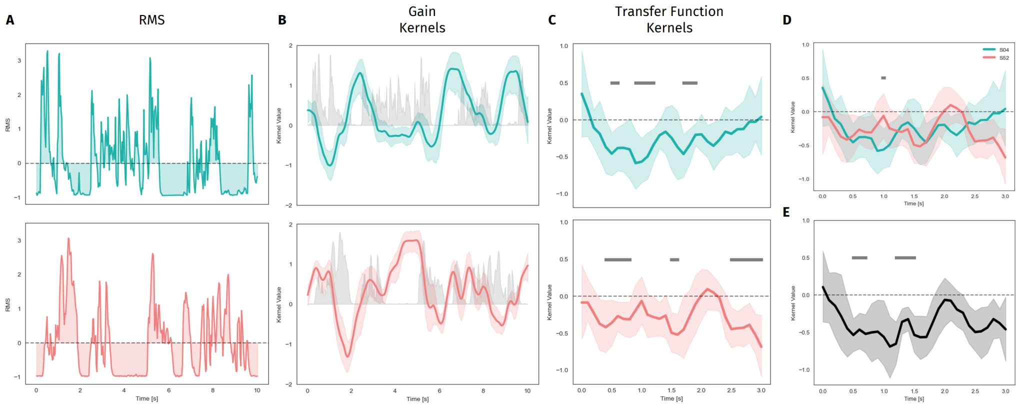


Figure 5.4: **Gain and TRF kernels obtained from the reverse correlation task in Study 1.** **(A)** Speech extracts were chosen based on differences in the timing of pauses on relatively neutral semantic content in both, with the speaker in S04 (**left**). **(B)** The gain kernels highlight participants' preference for contingent smiles occurring during pauses in the speaker's speech and peaking, on average, 1s after pause onset. **(C)** Transfer function kernels show statistically similar dynamics for different speech extracts (except at a solitary time point $t = 1$, t(12)=-5.5, p=.01, **(D)**). This similarity, together with distinct gain kernels, suggests efficient encoding of the temporal landscape of contingent smiles through parsimonious internal representations, characterized here as transfer functions. **(E)** Due to the lack of statistical differences between the transfer function kernels of the two sentences, a combined kernel was computed and utilized for further analysis.
Gray horizontal bars represent time points significantly different from 0.

nel suggests that when RMS is negative (as a result of the z-scoring), the corresponding smiles are positive. In other words, a K_{TRF} assigns low weights to high speech intensity and consequently, a smile occurring during speech is perceived as being less contingent than one occurring during a pause or a period of low speech intensity. Beyond its global negativity, the specific shape of the kernel points towards the onset of smiles occurring approximately 0.5s after the beginning of a pause, with the gradual return to 0 after approximately 2s (M=1.945, SD=0.66) suggesting a finite temporal window beyond which speech features cease to be relevant.

Bimodal strategy for contingency detection (N=27)

While the dominant strategy involved a negative kernel suggesting smiling behaviour during pauses, 4 of the 27 participants (33%) possessed K_{TRF} with the opposite positive pattern, i.e. smiling during speech and 'unsmiling' during pauses (Fig. 5.5A). Interestingly, we found that 9 participants alternated between a positive and a negative K_{TRF} depending on the speech extract. This suggested both between and within individual variability, taking the form of a bimodal strategy for perceiving contingency of listener smiles. Comparing the K_{Gain} and K_{TRF} of participants with opposed strategies, we found that this bimodality was reflected remarkably well in the kernels showing differences only in terms of their direction or polarity while leaving the temporal structure (i.e. latencies) largely intact (Fig. 5.5B and C). Moreover, we observed that almost all participants possessing large internal noise (>2.5) also had a tendency towards bimodal strategies, perhaps indicating a detrimental impact of switching strategies on decision stability during the task (especially since *S04* and *S52* trials were presented in random order rather than in separate blocks).

5.1.3 Corpus TRF vs. Reverse Correlation Kernel

We computed a smile TRF from genuine interactions in the Speed Dating corpus from Chapter 3, with RMS intensity of speakers' speech as input and the listeners' AU12 responses as output. This TRF, representing the production dynamics of contingent smiles found in naturalistic interactions, was then compared against the combined kernel of *S04* and *S52* representing the observer's internal representation/expectation of what these dynamics should be. Contrary to expectations, the Speed Dating

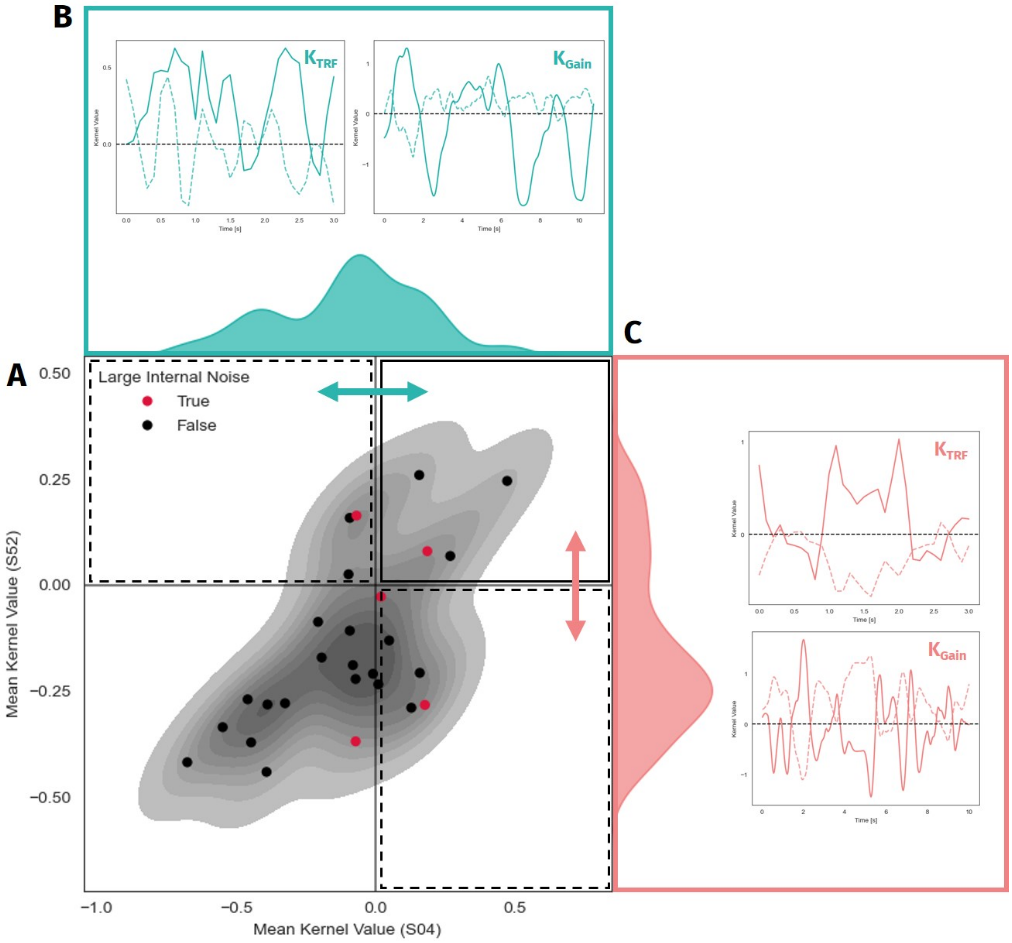


Figure 5.5: **Bimodality in participant strategies for social contingency perception. (A)** The distribution of the average values of observers' K_{TRF} reveals the existence of a 'dominant strategy' (lower left quadrant) consisting of a negative-value K_{TRF} for both speech extracts, suggesting a preference towards smiling during pauses. However, some observers also employ alternative strategies resulting in positive K_{TRF} for both speech extracts (upper right quadrant) or a combination of positive and negative K_{TRF} (upper left and lower right quadrants). A majority of the observers with such alternative strategies also display high levels (>2.5) of internal noise (red dots). **(B) and (C)** The *S04* (green) and *S52* (red) kernels of observers with positive K_{TRF} (solid lines) and those with negative K_{TRF} (dashed lines) highlight the bimodal nature of the observers' internal representations with similar dynamics that are opposed only in terms of their sign or direction.

TRF displayed an opposite pattern to that of the reverse correlation kernel (Fig. 5.6). These opposing polarities suggest that while observers in this study expect contingent smiles to occur during pauses in speech, actual listeners in the Speed Dating corpus tend to smile during speech. On the other hand, the onset and offset latencies of both were remarkably similar. This pattern of results was reminiscent of the roughly 1/3 of participants who had positive reverse-correlation kernels, and could suggest that these participants in fact used perceptual heuristics that more closely match production dynamics than the majority of observers.

We then examined whether the (majority, negative-valued) reverse-correlation kernel was better at predicting the subjective ratings of contingency collected in Chapter 3 than the AU12 production TRF (which was not found to be a significant predictor in Chapter 3, see Section **??**). To do so, we used the majority kernel (N=13) of *S04* and *S52* combined to predict listeners' AU12 responses to the RMS intensity of speech extracts in the Speed Dating corpus and calculated the correlation between the actual and predicted AU12 responses. The same process was repeated for the AU12 TRF computed from genuine interactions in the Speed Dating corpus. Using the two correlations as independent variables, we fitted a generalized linear mixed model (GLMM) with observer judgements of whether an interaction contained genuine or fake contingency in Chapter 3 as the binary dependent variable: `Response ~ TRF Correlation + Combined Kernel Correlation + (1|Subject)`. The results confirmed that the AU12 production kernel was not a significant predictor of observer judgements of contingency (speed-dating TRF: $\beta = 0.18, p = .66$), but neither was the reverse-correlation kernel ($\beta = 0.25, p = .57$).

While these results appear to indicate that the reverse-correlation kernels extracted here do not have ecological consequences when judging the contingency of interaction, attempting to validate the kernels obtained in this study against the AU12 TRF obtained in Chapter 3 presents two problems. First, the speed-dating TRF was learnt from genuine interactions, i.e. real-time first-person responses to changes in speech intensity, as opposed to the K_{TRF} evaluated from third-party observers' perception of contingency. Past literature has argued that social observation and social participation do not necessarily involve the same cognitive processes (Wilms et al., 2010; Tylén et al., 2012). Indeed, comparisons have shown fewer actual backchannels by participants in an interaction than expected by third-party observers (Heldner et al., 2013). Second, in Chapter 3, observers were presented with naturalistic interactions con-

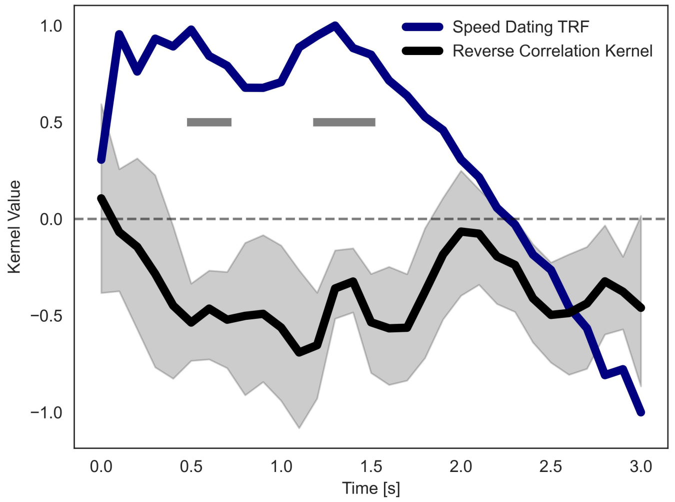


Figure 5.6: **Comparison of the combined kernel with the AU12 (smile) TRF obtained in Chapter 3.** In Chapter 3, we learned a TRF for the mapping between RMS and AU12 (smile) in genuine interactions. Comparing this TRF with the combined kernel of *S04* and *S52* showed similar onset and offset latencies in both. However, the polarity of the two representations were the opposite (with statistically significant differences between the two concentrated at the peaks as represented by the horizontal grey bars), meaning the smile TRF learned from the corpus encodes smiling behaviour during speech and 'unsmiling' during pauses while observers in the reverse correlation task converged on a representation encoding the opposite, i.e. not smiling during ongoing speech and smiling during pauses.

taining a wide variety of facial signals, including head nods, blinks and eye gaze, in addition to smiles, which were the only signals present in the stimuli in this study. Since the perception of contingency is modulated by a diverse range of signals and the interactions between them, it stands to reason that observers' expectations of the dynamics of contingent signals are likely to be affected when faced with unusually minimal interactions. Finally, one of the features of such contingent signals is that they are often replaceable, meaning that a head nod can be employed in place of a smile and vice versa (Ward and Tsukahara, 2000). Therefore, in impoverished 'interactions' like those in this study, the one solitary available signal might be co-opted to represent general contingent behaviour, something that may not be the case in more complex interactions.

5.2 Study 2: So, do observers prefer contingent smiles that match their internal representations?

In Study 1, we extracted reverse correlation kernels which were considered to be observers' internal representations of contingent smiles in response to speech. However, the relationship between a speaker's speech and a listener's smiles learned from naturalistic interactions in the Speed Dating corpus did not match the reverse correlation kernels and, in fact, demonstrated essentially the opposite pattern. Moreover, predictions of smiling behaviour generated by the reverse correlation kernels also did not show any significant correlation with observers' judgements of contingency in Chapter 3. To examine whether reverse-correlation kernels had any impact on perceived contingency in a more controlled task, we collected data from a new set of participants and tested the generalizability of the kernel on speech extracts that were not used in Study 1. To do so, we took the combined K_{TRF} of *S04* and *S52*, created several variants with different dynamic characteristics, and applied them to four additional speech extracts from the same dataset.

5.2.1 Methods

Stimuli

The combined kernel from Study 1 was thresholded on the basis of statistically-significant deviation from 0 at each time point using one-sample t-tests corrected for multiple comparisons, i.e. kernel values at time points where the t-statistic was below $\alpha_{corrected}$ were changed to 0. This kernel (K) was then used as the base kernel for creating 4 other variants - its opposite pattern $K_{inverse}$, a version time-shifted by 1 second $K_{timeShift}$, as well as 2 constant-valued kernels $K_{constant}$ and $K_{constantInverse}$ whose values were the mean values of K and $K_{inverse}$ respectively (Fig. 5.7A).

All 5 kernels were applied to 4 new speech extracts in addition to the 2 used in Study 1, for a total of 6, all lasting approximately 10 seconds and consisting of male speakers talking about different neutral topics (Fig. 5.7B). Videos were generated using the same deformation file as in Study 1 and combined into the following pairs: $(K, K_{inverse})$, $(K, K_{timeShift})$, $(K, K_{constant})$, $(K_{constant}, K_{constantInverse})$ and $(K_{constant}, K_{timeShift})$ to obtain 30 stimuli (5 kernels × 6 input sentences).

In light of the bimodal strategy employed by participants in Study 1, we investigated whether there was in fact a 'dominant strategy' by comparing K against $K_{inverse}$. To test whether the dynamics of K were meaningful, we compared it against the static kernel $K_{constant}$, which also allowed us to confirm the size of the temporal integration window suggested by K. The comparison of K with $K_{timeShift}$ was conducted in order to reveal whether observers were sensitive to changes in onset and peak latencies, and thereby validate the specificity of the dynamics described by the original kernel. We compared $K_{constant}$ and $K_{constantInverse}$ to investigate if observer preference or bimodality was affected by the absence of the specific dynamics of K, thus lending further support for it being meaningful. Finally, by comparing $K_{constant}$ with $K_{timeShift}$ we hoped to uncover whether changing the onset and offset latencies had a greater negative impact on observer perception of contingency than simply shifting the 'appropriate' dynamics in time.

Participants

N = 21 (male: 13; M=33.6) native French speakers were recruited online on Prolific. Participants gave their informed consent and were compensated at a standard rate. 3 participants were excluded from the analysis due to them taking approximately 60 minutes to complete the experiment, as opposed to the average of 31 minutes. Thus, a total of 18 participants remained for further analysis.

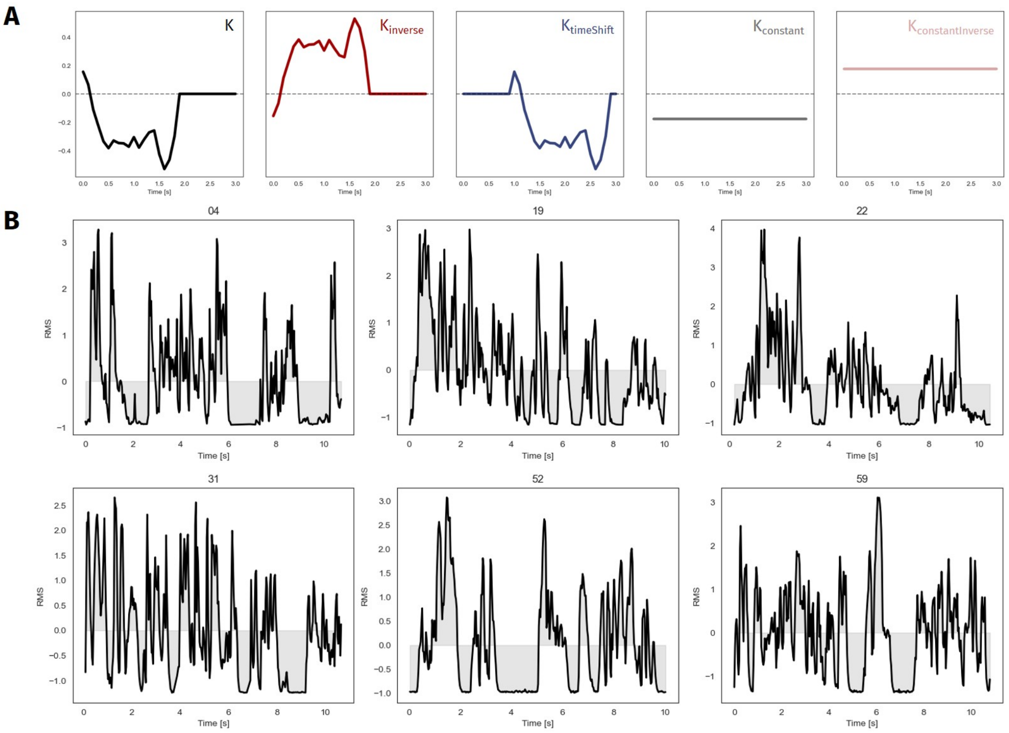


Figure 5.7: **Speech extracts and kernels used in the validation study. (A)** From the kernel obtained in Study 1, 5 variants were created, consisting of 3 dynamic kernels, K, $K_{inverse}$, $K_{timeShift}$ and 2 static kernels, $K_{constant}$, $K_{constantInverse}$. **(B)** A total of 6 different speech extracts were used for this study, all of them lasting ~10s and consisting of a male speaker talking about a neutral topic.

Procedure

Participants were presented with a total of 60 trials (30 unique trials repeated with the positions of the videos reversed, i.e. left or right) in an experiment lasting approximately 30 minutes. In a procedure similar to that in Study 1, each trial required participants to watch a pair of videos of the same person with different smiling behaviour driven by different K_{TRF} variants while simultaneously listening to a speech extract lasting approximately 10 seconds. They were asked to choose which of the two videos contained smiling behaviour that was most appropriate in response to the speech extract they heard. The pairs of kernels compared in the experiments were: K vs $K_{inverse}$, K vs $K_{timeShift}$, K vs $K_{constant}$, $K_{constant}$ vs $K_{constantInverse}$ and $K_{constant}$ vs $K_{timeShift}$; each with 6 sentences $times$ 2 repeats resulting in 12 trials per comparison.

5.2.2 Results

We conducted binomial tests to evaluate participant preference for one kernel over another over all trials for each pair of kernel variants. Consistent with the evidence of bimodality in Study 1, we found that 17% of the participants preferred $K_{inverse}$ and $K_{constantInverse}$ over K and $K_{constant}$, respectively. Despite this, the overall trend showed a strong preference for K over its opposite patten $K_{inverse}$ (p=.000) with K being chosen 63% of the time (Fig. 5.8A) and for K over $K_{timeShift}$ (p=.048), which consisted of the same pattern shifted forwards in time by 1s, with participants choosing K 56% of the time (Fig. 5.8B). Difference in preference between K and its static variant $K_{constant}$, albeit in the expected direction, was not statistically significant (p=.263) (Fig. 5.8C). Preference for $K_{constant}$ was only marginally greater (p=.088) than that for $K_{constantInverse}$, i.e. the static variant of $K_{inverse}$ with $K_{constant}$ chosen 56% of the time (Fig. 5.8D). Finally, no statistically significant difference in preference was found between $K_{constant}$ and $K_{timeShift}$ (p=.541).

Excluding the 3 participants that showed consistent preference for the opposite patterns of both K and $K_{inverse}$, the strength of evidence for participant preference for K over $K_{inverse}$ and $K_{timeShift}$ increased with K now being chosen 70% (p=.000) and 60% (p=.014) of the time respectively. Preference for $K_{constant}$ over $K_{constantInverse}$ was also significant for this group, with $K_{constant}$ being chosen 60% of the time (p=.000). The exclusion of these participants had negligible impact on participant preference in the comparisons between K and $K_{constant}$ and $K_{constant}$ and $K_{timeShift}$.

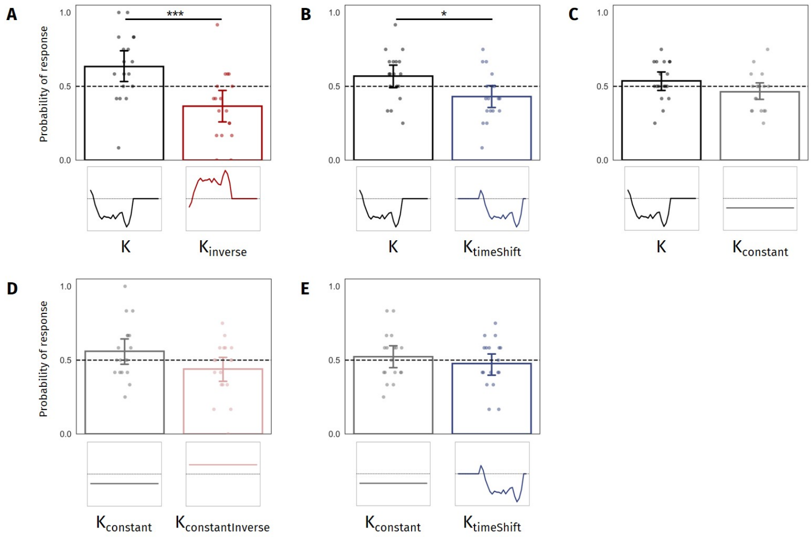


Figure 5.8: **Validation of combined K_{TRF} from Study 1.** **(A)** Participants showed significantly greater preference for smiles generated by the original kernel from Study 1 (K) compared to those generated by its opposite pattern ($K_{inverse}$). **(B)** Likewise, there was significant preference for K over its time-shifted version $K_{timeShift}$, albeit to a lesser extent. **(C)** However, when asked to choose between K and its static version $K_{constant}$, participants did not display any strong preference for one over the other. **(D)** While there was preference for $K_{constant}$ over $K_{constantInverse}$ (the static version of $K_{inverse}$), it was only marginally significant (p=.08). **(E)** Finally, no significant difference in preference was found between $K_{constant}$ and $K_{timeShift}$.

5.3 Discussion

While previous research has highlighted the role of smiles as backchanneling cues in conversation, it has, for the most part, relied on post-hoc analysis due to the difficulty of applying causal manipulations to social interactions. Consequently, the scope of any found evidence for the function and timing of such cues can often be limited by the inability to mechanistically test them. In Chapter 3, we attempted to remedy this using the system identification concepts of impulse responses/TRFs to operationalize a possible mechanism of social contingency perception in terms of the dynamics of backchannels. While they allowed testing quantitative predictions, and indeed, were found to partly correlate with observers' ratings of social contingency, their generalizability remained plagued with the same correlational limitations as in previous work. Moreover, in Chapter 4, we showed that estimates of AU activation by black-box models like Py-feat and Openface lack the desired modularity, suggesting that TRFs of some specific AU estimated with such data may well represent the tangled dynamics of several different AUs.

In the work described in this chapter, we developed a novel data-driven experimental paradigm to uncover the dynamics underlying the perception of social contingency. This paradigm allowed direct probing of perceptual mechanisms while carrying fewer assumptions than a corpus study. Combining system-identification concepts of linear time-invariant systems and transfer functions, we synthesized realistic smiles in videos in response to speech by generating random impulse responses and computing the time-course of those smiles as if they were governed by the convolution of the speech with the impulse response. Using reverse correlation, we then derived participants' internal representations of socially-contingent smiles in the form of impulse responses encoding the dynamics of listeners' smiles (Study 1). The generalizability of the obtained transfer function was then evaluated by creating variants on top of it, applying them to different speech, and testing them against each other (Study 2). Our results showed that while observers' internal representations of smile time series (i.e. gain kernels) were tailored to each distinct speech extract, the underlying transfer functions were reasonably similar. Furthermore, validation of the transfer function demonstrated that observers were quite sensitive to the dynamics, as evidenced by their preference for it over its opposite and time-shifted variants.

The fact that we did not find statistically significant differences between the impulse response kernels of the two speech extracts suggests that observers possess internal representations of the dynamics of contingent smiles (and potentially contingent signals in general) that are at least partially sentence-independent. This is supported by the finding that despite similar transfer function kernels, the kernels of actual smiling responses were well-adapted to the distinct speech extracts. For instance, differences in the timings and durations of pauses within the two speech extracts were captured quite accurately by predictions generated by the homogenous dynamics of the transfer functions. Since the variability and noise involved in everyday social interactions require a great degree of flexibility, it is perhaps unsurprising that representations of communicative behaviour in the brain reflect general, parsimonious patterns of abstract relations. The additional imperative to process such behaviour rapidly in conjunction with a multitude of other cognitive processes means that the organization of these abstract relations also needs to be computationally efficient. Organizing internal representations as transfer functions, possibly even something as simple as an impulse response, would in theory be coherent with those goals by condensing the relationship between, in this case, acoustic features of speech and corresponding smiling behaviour, into a more minimal structure. This parsimony allows efficient updating of representations in case of a mismatch between observed and predicted data, and importantly, reduces the cost of storing and maintaining many different variants of a representation where each one is specialized for a different context.

The shape of the combined transfer function of both speech extracts in Study 1 provides interesting insights into observers' preferred dynamics of contingent smiles in response to speech intensity. The negative trajectory of the kernel suggests that observers expect smiles to occur during pauses or periods of relative silence, with the specific dynamics implying a gradual increase in smile intensity until it peaks at around 1s after the onset of a pause, followed by a short period of relative stability lasting approximately another 1s before a gradual decrease coinciding with the onset of speech at the end of the pause. In essence, the transfer function kernel assigns negative weights to high speech intensity, thereby reducing observers' perception of smiles occurring during speech as being contingent and increasing it when smiles are aligned with periods of low speech intensity. Contrary to our expectations, the combined transfer function of both speech extracts did not match the AU12 (smile) TRF learnt from genuine interactions in the Speed Dating corpus, instead exhibiting the opposite pattern (smiles during speech instead of during pauses in speech), albeit with similar onset and offset latencies. One possible explanation for this could be the difference between social *interaction* and social *ob-*

servation, meaning that expectations of the timing of contingent smiles might differ for a listener engaged in an interaction as opposed to an observer of an interaction (Tylén et al., 2012). Another explanation concerns the relative paucity of signals in this study's stimuli compared to everyday social interactions, resulting in observers modifying their internal representations of the dynamics of contingent smiles. It is possible that the stimuli in Chapter 3's Study 1 also contained backchannels during pauses, but because they were ecological interactions, signals other than smiles could step in and fulfil the role of providing feedback during pauses. This is supported by recent accounts of facial expressions like the *behavioural ecology view of social displays* (BECV) which argues that facial expressions are tools used in social interactions to "signal our contingent next move to alter yours", i.e. as signals of contingent action as opposed to being behavioural representations of internal states (**?**). In this view, individual facial signals possess some degree of functional flexibility such that they can adopt different communicative roles in an interaction depending on the context or the task at hand. Thus, smiles in the Speed Dating corpus could have played a different role and consequently have had different dynamics, while in this study, the absence of any other signals meant that observers were forced to use smiles to infer contingency. This may also explain the bimodality of observers, with some of them perhaps not adapting their internal representations for the task at hand (involving severely degraded versions of everyday interactions), which in turn manifested in terms of high internal noise. Interestingly, past research in this area has shown that backchannel cues tend to occur mostly during short pauses lasting less than 1s rather than overlapping with speech (White, 1997), have higher frequency within pauses in speech (Truong et al., 2011) and occur after low pitch regions in the speaker's speech lasting at least 150ms (Ward and Tsukahara, 2000). The fact that these findings correspond to the transfer function kernel obtained in this study suggests that it represents the dynamics of backchannels in general rather than the dynamics of contingent smiles specifically. Here, we show that in addition to clustering around prosodic peaks (i.e. simple temporal coincidence) listener backchannels also incorporate temporal integration of preceding context.

In Study 2, comparisons between kernel variants with different dynamics and over additional speech extracts confirmed that smiling during speech negatively influences observers' perception of contingency of the smile. Moreover, observer preference for the base kernel over a variant that was shifted forward in time by 1s indicates that observers are quite sensitive to the temporal alignment of smiles with speech intensity and that any disruption to this alignment (namely, slowing down the facial reaction) decreases the likelihood of the smile being perceived as contingent. It should be noted that we only validated the kernel against a single time-shifted variant in the interest of maintaining a shorter experiment, but it could be extended to many different values of time-shift for a more robust estimate of the latencies involved in contingency perception. Interestingly, observers did not show a statistical preference for the base kernel over the static variant consisting of a constant value throughout the entire duration of the kernel. This suggests that while smiles that occur too late/slowly are judged as less contingent, early smiles that peak rapidly do not significantly affect perception of contingency as long as the causal temporal structure of $\Delta RMS \rightarrow \Delta Smile$ remains intact. More importantly, it implies that the *bandwidth*, representing the time lag at which speech intensity ceases to be of relevance, might not have a strong influence on judgements of contingency of smiles. Practically, the difference between the two kernels amounts to longer temporal integration for the constant variant such that the intensity of the smile might decrease with ongoing speech but not drop to 0, as would happen in the case of the base kernel. One possible explanation for the lack of difference between them could be that the bandwidth of 2s observed in the base kernel is driven by the lack of variance in pause duration in the two speech extracts and could have been different if observers were exposed to stimuli with longer pauses. It could, however, also be a way to account for short interruptions in pauses, i.e. if the speaker interrupts a pause very briefly, following which the pause resumes, smiles exactly matching this rapid onset-offset of intensity would presumably appear rather unnatural. Having a larger bandwidth or a longer temporal integration window might help counter this. A recent study investigating the effect of time delays on the appropriateness of backchannels showed that acceptability of backchannels decreased significantly for time delays beyond 1s (Boudin et al., 2024), corresponding neatly with our results showing observer preference for the base kernel as opposed to its time-shifted variant, as well as the dynamics of the base kernel showing decreasing relevance of input features approximately between 1.5-2s.

Taken together, these results provide evidence for the ability to characterize internal representations of contingency perception as 'social transfer functions' which, if cognitively represented in this manner, would be extremely economical in terms of computational efficiency and the cognitive cost of storage and maintenance. While we do not claim that contingency is necessarily represented in the brain as transfer functions or generate predictions through convolution with input signals, our results do suggest, *qua* Popper's falsifiability (Popper, 2005), that it is not an invalid assumption and can be used to draw valuable inferences. We show that by encoding the abstract rules governing the dynamics of contingent interactions, these social transfer functions, even when simplified to impulse responses, are able to generate appropriate predictions for disparate speech inputs. Moreover, the social transfer functions were obtained using purely data-driven methods. Investigating the multimodal dynamics of social interactions using reverse correlation allowed the exploration, and thus elimination, of a much larger subspace of possible transfer functions or internal representations than could have been achieved using traditional experimental paradigms.

Chapter 6

Discussion

6.1 Summary

Over the course of the experimental work presented in this thesis, we attempted to build up a coherent characterization of a social transfer function. The desire or need for such a characterization arose from a lack of clarity about the dynamics of social signals despite possessing significant insight into their timing and function. Here, we tackled the case of observer perception of social contingency or the ability to recognize the actions of one agent as responses to the actions of another, an ability that is thought to scaffold high-level social processes such as turn-taking and theory of mind. We cast interacting agents as coupled dynamical systems and used the system-identification concepts of impulse responses and transfer functions to model social contingency in terms of the relationship between a speaker's speech and a listener's facial signals.

In Chapter 3, we used a corpus of naturalistic interactions to demonstrate that third-party observers' ability to recognize contingency remained robust even when observers could only see either the eyes or the mouth of listeners. To supplement the behavioural results, we modeled separate social transfer functions, each describing the relationship between a given speech and a different facial signal from listeners. Comparing predictions of social transfer functions with observer judgements of contingency revealed that the latter were globally compatible with their being based on the dynamics (rather than average activity) of signals in the mouth-region over the eye-region.

In Chapter 4, representing a brief interlude in the exploration of social contingency, we used the system-identification technique of reverse correlation to probe machine-learning based action-unit detection models and explain their outputs in terms of the features used. Our results showed that such models do not always ensure modularity, i.e. they might use information from the mouth to judge activity around the eyes, and can be sensitive to small, barely perceptible changes in an image caused by shifting a single facial landmark. While no systematic bias was found against any gender or ethnicity, decomposition of basic emotional expressions into component action units also did not correspond exactly to expectations based on the literature.

In the final experimental chapter (Chapter 5), we combined methodological and theoretical insights from the previous chapters by synthesizing smiles in videos using the AU12 (smile) reverse correlation kernel from Chapter 4 such that their dynamics were governed by the convolution of speech intensity with random impulses. Combined with the Chapter 3 finding that observers could reliably recognize contingency even in the absence of a speaker's face, we created a reverse-correlation experiment to extract observers' social transfer functions of contingent smiles in response to speech intensity in a data-driven manner. We showed that these social transfer functions were reasonably speech-independent but nonetheless able to generate predictions of smiles that were well-aligned with the specific timing of speech and pauses in the two sentences we tested. Finally, the obtained social transfer function was validated in a follow-up study in which observers showed a clear preference for it over other variants that manipulated its average value and timing.

Taken together, our results show 1) the robustness of observers' ability to recognize social contingency, 2) highlight the specific regions of a listener's face used by observers to recognize contingency, and 3) determine the expected dynamics of a listener's contingent behaviour in response to speech. In doing so, we showcase how simple system-identification concepts can be deployed in versatile ways, ranging from probing machine-learning black-box models to operationalizing internal representations of cognitive abilities. Through this work, we develop an outline of the social transfer function as a possible parsimonious mechanism by which observers can internally represent social contingency and, more broadly, the mechanism of interpersonal predictive coding.

6.2 "All models are wrong..."

Beyond theoretical questions about the underlying mechanisms of social contingency, the work presented in this thesis also raises some interesting methodological questions which can be grossly summarized using the aphorism attributed to British statistician George Box: "all models are wrong, but some are useful" (Box, 1979). Of primary concern is the treatment of the relationship between speech and backchanneling behaviour as a linear time-invariant system when it is well-established that such conversational dynamics do not, in fact, satisfy these assumptions.

6.2.1 Linearity

Impulse responses make strong assumptions about 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. For instance, the simple FIR model used here could be replaced with more advanced system-identification model structures like the Box-Jenkins model (Section 2.1.2). Because the Box-Jenkins model accounts for past outputs to generate new outputs and can handle non-stationarity (i.e. changes in the average or baseline behaviour), it could better capture the dynamics of contingent behaviour in an interaction where, for example, the listener's engagement decreases over time and leads to differential frequency or timing of contingent responses. More generally, there are other possible computational and 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 might be interesting alternatives that future work could evaluate.

6.2.2 Time-invariance

The time-invariant aspect of our conception of the social transfer function assumes that there is an average first-order relationship between a speaker's speech and a listener's backchanneling behaviour, which can be learned and tested against new data. While an average first-order relationship can always be computed, it may not be predictive of anything if backchanneling is driven entirely by something else, like whether the sentence is a question or answer or whether the interacting individuals are in agreement or disagreement. Thus, the extent to which average behaviour represented by an FIR actually explains observer judgements of contingency remains an empirical question. 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 irrespective of the interactive context. Therefore, even though Chapter 3 shows that observer judgements are reasonably consistent with predictions by our time-invariant model, its problem of "context-invariance" remains unaddressed.

6.2.3 Context-invariance/Context-specificity

The FIRs learned in Chapter 3 and the speech extracts used to obtain FIRs in Chapter 5 are both embedded in the singular context of an introductory conversation during a speed-dating session involving, for the most part, relatively neutral semantic content. It is thus likely that the transfer functions we obtain are restricted to interactions within social contexts involving affiliation, politeness and perhaps even shyness. This excludes a lot of communicative variance driven, for instance, by familiarity and contexts which allow disagreement and arguments (e.g. political debates). Thus, it is entirely possible that while FIR predictions correlate with observer ratings of contingency for interactions in the Speed Dating corpus, they would not do so when applied to other kinds of interactions. Moreover, results in Chapter 3 and Chapter 5 assume that observers, through previous exposure and participation in social interactions, develop similar schema of conversational contingency. It remains an open question how this learning may operate and how plastic it may be to factors like 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 stereotyping as being too impatient or unresponsive (White, 1989). Importantly, our proposed concept of social transfer function is not intended as a mechanism to subtend backchanneling, i.e. we do not claim that a listener's facial signals in response to a speaker's speech are determined by linear, time-invariant impulse responses. Rather, we propose social transfer functions as a parsimonious way to encode third-party observers' perception or evaluation of interactive coupling.

It should be noted that in this thesis, we use the term 'backchanneling' in a relatively generic sense as the ensemble of non-verbal facial behaviour of a listener while a speaker speaks, without necessarily distinguishing it by its function (e.g. linguistic or emotional) or underlying cognitive processes (e.g. voluntary or not). Some may have explicit, voluntary communicative intent (e.g. nodding at the end of a statement to signal agreement - McClave, 2000), or implicate emotional contagion (e.g. smiling in response to a smile - Hess and Bourgeois, 2010), or may just be of a lower sensorimotor nature (e.g. periodic eye blinks coupled with sentence dynamics - Jin et al., 2018; Kobald et al., 2019; Nakano and Kitazawa, 2010). In judging social contingency, it is likely that third-party observers use all of these cues, though some perhaps more contextually than others.

6.3 "... but some are useful"

Despite the inherent limitations of our characterization of social transfer functions, through the work presented in this thesis, we attempt to address the question of whether they are a useful tool to model observer perception of social contingency. We argue that operationalizing perception of contingency as transfer functions provides a testable, operational mechanism with good parsimonious properties that allows making predictions and, combined with behavioural data, can help quantify social contingency.

6.3.1 Parsimony

One question raised by our findings is whether social transfer functions are uniquely related to judgements of social contingency, or to more general judgements: we believe the latter to be true. While this work has focused on the detection of social contingency, it is interesting to question whether social transfer functions, on different facial or bodily signals or at different temporal scales, also support other types of social-cognitive inferences that rely on dynamic predictions of conversational backchanneling. For instance, storing separate pre-learned TRFs for interactions between familiar and unfamiliar agents (Gráczi and Bata, 2010) would allow judgement of which is more likely to occur. Other examples would be agreement (Müller et al., 2022) or even enjoyment (Li et al., 2010). By providing a parsimonious representation of conversational dynamics which can be learned from each individual, social transfer functions could be promising as a way to study both how these constructs are signalled in ecological behaviour and to model how they are detected by observers. Using social transfer functions for different constructs would also imply the presence of flexible definitions of contingent dynamics represented by distinct transfer functions that help construct more high-level social judgements. Given the parsimonious nature of transfer functions, it is likely that maintaining different internal representations of contingency for different contexts permits greater flexibility in recognizing contingent behaviour while being more cognitively efficient considering that storing it is more memory-efficient than storing numerous examples of input-output pairs or a conditional probability distribution - a 'computational trick' also exploited in convolutional deep learning architectures (Mallat, 2016).

6.3.2 Versatility

Beyond their descriptive interest, social TRFs are also useful as analytical or generative tools. Analytically, how well the predictions of transfer functions fit observed data provides a way to quantify the realism/typicality of backchannel dynamics. This could be used to quantify atypical conversational dynamics often observed in disorders such as autism spectrum disorder (Wehrle et al., 2024), schizophrenia (Lucarini et al., 2024), parental depression in caregiver-child interactions (Smith et al., 2023) and disorders of consciousness (Hermann et al., 2018). Social transfer functions would not only allow a qualitative characterization of how the atypicality manifests in conversational dynamics but also help extract the features based on which they are evaluated as atypical by caregivers. For instance, future work could compare kernels obtained from reverse correlation experiments on clinical practitioners and naive observers to highlight the diagnostic features driving their perception of atypicality.

Apart from their utility in understanding psychopathology, social transfer functions could also function as a security measure to detect forged AI videos (Li et al., 2018). Combined with modern facial animation techniques in avatars (Yu et al., 2012) or real-life videos (Arias-Sarah et al., 2024), they can also be used to generate novel stimuli that have specific dynamics, either for experimental control (e.g. synthesizing gaze patterns as if they were driven by the transfer functions describing the dynamics of smiles) or to improve synthetic media (e.g. manipulating the perceived contingency of deep-faked conversations in human-computer interaction - Kaate et al., 2023).

Within the broader context of this thesis, having obtained a general social transfer function of contingent responses in a specific interactive context, future work could conduct reverse-correlation experiments involving random perturbations of that transfer function in different contexts to extract a more refined picture of if and how the dynamics vary with context. Similarly, the aforementioned problem of cultural differences in backchannel appropriateness could also be better understood by extracting and comparing the social transfer functions associated with different cultures.

6.4 Some final thoughts

In this thesis, we made an effort to achieve a mechanistic understanding of the cognitive processes underlying social cognition. To this end, we developed a methodological framework to study social contingency, a fundamental building block of everyday social interactions, in terms of the relationship between speakers' speech and the facial signals of listeners. While there exists a wealth of literature investigating the timing and function of social signals in interactions, we hoped to develop computational models that could be used to form testable hypotheses by predicting listener responses to speech and thus provide a way of explaining the complex input-output relationship.

With the experimental work presented here, we showed that our mathematical formalization of social contingency as social transfer functions can be quite useful in developing a clearer picture of the dynamics of contingent behaviour. However, we are also cognizant of the fact that the simplistic models used possess several limitations, which we detailed in this final chapter. The issues of linearity and time-invariance, in particular, highlight that social transfer functions lack the necessary computational capacity to explain the intrinsic non-linearities of social interactive behaviour. However, we still made a concerted effort to extend simple concepts and techniques based on the belief that inaccurate but interpretable models are perhaps more conducive to understanding than models that are accurate but opaque. We highlight this point using the 'computational interlude' in Chapter 4, where we discuss the recent explosion of explainability in AI as a result of the single-minded focus on model performance. As computational modeling becomes an increasingly important part of cognitive science, it is important to be mindful of this trade-off between complexity and interpretability in an area of study where the primary goal isn't simply approximating the outputs of cognitive mechanisms but to understand them.

Bibliography

[1] Daniel A Abrams, Trent Nicol, Steven Zecker, and Nina Kraus. Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience*, 28(15):3958–3965, 2008.

[2] Rick A Adams, Klaas Enno Stephan, Harriet R Brown, Christopher D Frith, and Karl J Friston. The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47, 2013.

[3] Aynaz Adl Zarrabi, Mélissa Jeulin, Pauline Bardet, Pauline Commère, Lionel Naccache, Jean-Julien Aucouturier, Emmanuel Ponsot, and Marie Villain. A simple psychophysical procedure separates representational and noise components in impairments of speech prosody perception after right-hemisphere stroke. *Scientific Reports*, 14(1):15194, 2024.

[4] Ralph Adolphs, Frederic Gosselin, Tony W Buchanan, Daniel Tranel, Philippe Schyns, and Antonio R Damasio. A mechanism for impaired fear recognition after amygdala damage. *Nature*, 433 (7021):68–72, 2005.

[5] Ehud Ahissar, Srikantan Nagarajan, Merav Ahissar, Athanassios Protopapas, Henry Mahncke, and Michael M Merzenich. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, 98(23):13367–13372, 2001.

[6] Dagmara Annaz, Ruth Campbell, Mike Coleman, Elizabeth Milne, and John Swettenham. Young children with autism spectrum disorder do not preferentially attend to biological motion. *Journal of autism and developmental disorders*, 42(3):401–408, 2012.

[7] Pablo Arias, Catherine Soladie, Oussema Bouafif, Axel Roebel, Renaud Seguier, and Jean-Julien Aucouturier. Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*, 11(3):507–518, 2018.

[8] Pablo Arias, Laura Rachman, Marco Liuni, and Jean-Julien Aucouturier. Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, 13(1):12–24, 2021.

[9] Pablo Arias-Sarah, Daniel Bedoya, Christoph Daube, Jean-Julien Aucouturier, Lars Hall, and Petter Johansson. Aligning the smiles of dating dyads causally increases attraction. *Proceedings of the National Academy of Sciences*, 121(45):e2400369121, 2024.

[10] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining predictions of non-linear classifiers in nlp. *arXiv preprint arXiv:1606.07298*, 2016.

[11] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. " what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12 (8):e0181142, 2017.

[12] Malika Auvray and Marieke Rohde. Perceptual crossing: the simplest online paradigm. *Frontiers in human neuroscience*, 6:181, 2012.

[13] Malika Auvray, Charles Lenay, and John Stewart. Perceptual interactions in a minimalist virtual environment. *New ideas in psychology*, 27(1):32–47, 2009.

[14] Dominik R Bach, Hartmut Schächinger, John G Neuhoff, Fabrizio Esposito, Francesco Di Salle, Christoph Lehmann, Marcus Herdener, Klaus Scheffler, and Erich Seifritz. Rising sound intensity: an intrinsic warning cue activating the amygdala. *Cerebral Cortex*, 18(1):145–150, 2008.

[15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[16] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[17] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered:

Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[18] Janet Bavelas and Nicole Chovil. Some pragmatic functions of conversational facial gestures. *Gesture*, 17(1):98–127, 2018.

[19] Beatrice Beebe, Joseph Jaffe, Sara Markese, Karen Buck, Henian Chen, Patricia Cohen, Lorraine Bahrick, Howard Andrews, and Stanley Feldstein. The origins of 12-month attachment: A microanalysis of 4-month mother–infant interaction. *Attachment & human development*, 12(1-2):3–141, 2010.

[20] Jonathan S Beier and Susan Carey. Contingency is not enough: Social context guides third-party attributions of intentional agency. *Developmental psychology*, 50(3):889, 2014.

[21] Ludovic Bellier, Anaïs Llorens, Déborah Marciano, Aysegul Gunduz, Gerwin Schalk, Peter Brunner, and Robert T Knight. Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLoS biology*, 21(8):e3002176, 2023.

[22] Sarah Benghanem, Rudradeep Guha, Estelle Pruvost-Robieux, Julie Lévi-Strauss, Coralie Joucla, Alain Cariou, Martine Gavaret, and Jean-Julien Aucouturier. Cortical responses to looming sources are explained away by the auditory periphery. *cortex*, 177:321–329, 2024.

[23] Ole Bialas, Jin Dou, and Edmund C Lalor. mtrfpy: A python package for temporal response function analysis. *Journal of Open Source Software*, 8(89):5657, 2023.

[24] Peter Blomsma, Julija Vaitonyté, Gabriel Skantze, and Marc Swerts. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16 (4):1158–1181, 2024.

[25] X. Bombois and P.M.J. Van den Hof. System identification sc4110. Lecture notes, available through Blackboard or Nextprint, January 2006. URL http://www.ampere-lab.fr/IMG/pdf/sc4110slides.pdf. Lecture slides.

[26] Mondher Bouazizi, Kevin Feghoul, Shengze Wang, Yue Yin, and Tomoaki Ohtsuki. A non-invasive approach for facial action unit extraction and its application in pain detection. *Bioengineering*, 12 (2):195, 2025.

[27] Auriane Boudin, Stéphane Rauzy, Roxane Bertrand, Magalie Ochs, and Philippe Blache. How is your feedback perceived? an experimental study of anticipated and delayed conversational feedback. *JASA Express Letters*, 4(7), 2024.

[28] George Box and GM Jenkins. Analysis: Forecasting and control. *San francisco*, 1976.

[29] George EP Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

[30] Amanda C Brandone. Infants' social and motor experience and the emerging understanding of intentional actions. *Developmental Psychology*, 51(4):512, 2015.

[31] Holly P Branigan, Ciara M Catchpole, and Martin J Pickering. What makes dialogues easy to understand? *Language and Cognitive Processes*, 26(10):1667–1686, 2011.

[32] Miriam Brinberg, Denise Haunani Solomon, Graham D Bodie, Susanne M Jones, and Nilam Ram. Using state space grids to quantify and examine dynamics of dyadic conversation. *Communication Methods and Measures*, 19(1):1–23, 2025.

[33] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[34] Juan José Burred, Emmanuel Ponsot, Louise Goupil, Marco Liuni, and Jean-Julien Aucouturier. Cleese: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PloS one*, 14(4):e0205943, 2019.

[35] Roser Cañigueral and Antonia F de C Hamilton. The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in psychology*, 10:560, 2019.

[36] Arturo Casadevall and Ferric C Fang. Descriptive science. *Infection and immunity*, 76(9):3835–3836, 2008.

[37] Maïté Castro, Fanny L'héritier, Jane Plailly, Anne-Lise Saive, Alexandra Corneyllie, Barbara Tillmann, and Fabien Perrin. Personal familiarity of music and its cerebral effect on subsequent speech processing. *Scientific reports*, 10(1):14854, 2020.

[38] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. Py-feat: Python facial expression analysis toolbox. *Affective Science*, 4(4):781–796, 2023.

[39] Edward Collin Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the acoustical society of America*, 25:975–979, 1953.

[40] Andy Clark. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015.

[41] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[42] Charles A Coey, Manuel Varlet, and Michael J Richardson. Coordination dynamics in a socially situated nervous system. *Frontiers in human neuroscience*, 6:164, 2012.

[43] Laurence Conty, Charles Tijus, Laurent Hugueville, Emmanuelle Coelho, and Nathalie 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.

[44] Marcello Costantini, Ettore Ambrosini, and Corrado Sinigaglia. Out of your hand's reach, out of my eyes' reach, 2012.

[45] Rick Dale, Riccardo Fusaroli, Nicholas D Duran, and Daniel C Richardson. The self-organization of human interaction. In *Psychology of learning and motivation*, volume 59, pages 43–95. Elsevier, 2013.

[46] Christoph Daube, Tian Xu, Jiayu Zhan, Andrew Webb, Robin AA Ince, Oliver GB Garrod, and Philippe G Schyns. Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, 2(10), 2021.

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

[48] Sara De Felice, Gabriella Vigliocco, and Antonia F de C Hamilton. Social interaction is a catalyst for adult human learning in online contexts. *Current biology*, 31(21):4853–4859, 2021.

[49] Hanne De Jaegher and Ezequiel Di Paolo. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the cognitive sciences*, 6(4):485–507, 2007.

[50] Gregory C DeAngelis, Izumi Ohzawa, and RD Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. ii. linearity of temporal and spatial summation. *Journal of neurophysiology*, 69(4):1118–1135, 1993.

[51] Lisa DeBruine and Benedict Jones. Face Research Lab London Set. 5 2017. doi: 10.6084/m9.figshare.5047666.v5. URL https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666.

[52] Guillermo del Castillo Torres, Maria Francesca Roig-Maimó, Miquel Mascaró-Oliver, Esperança Amengual-Alcover, and Ramon Mas-Sansó. Understanding how cnns recognize facial expressions: a case study with lime and cem. *Sensors*, 23(1):131, 2022.

[53] Ophelia Deroy, Louis Longin, and Bahador Bahrami. Co-perceiving: Bringing the social into perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 15(5):e1681, 2024.

[54] Ron Dotsch and Alexander Todorov. Reverse correlating social face perception. *Social Psychological and Personality Science*, 3(5):562–571, 2012.

[55] Linda Drijvers and Judith Holler. The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, 30(2): 792–801, 2023.

[56] Damien Dupré, Eva G Krumhuber, Dennis Küster, and Gary J McKeown. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *Plos one*, 15(4): e0231968, 2020.

[57] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17 (2):124, 1971.

[58] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[59] Yachun Fan, Feng Tian, Xiaohui Tan, and Housen Cheng. Facial expression animation through action units transfer in latent space. *Computer Animation and Virtual Worlds*, 31(4-5):e1946, 2020.

[60] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[61] Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117 (43):26562–26571, 2020.

[62] Paul C Fletcher and Chris D Frith. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58, 2009.

[63] Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

[64] Jane E Fountain. *Building the virtual state: Information technology and institutional change*. Rowman & Littlefield, 2004.

[65] Jean E Fox Tree. Listening in on monologues and dialogues. *Discourse processes*, 27(1):35–53, 1999.

[66] Jennifer J Freyd. Dynamic mental representations. *Psychological review*, 94(4):427, 1987.

[67] Karl Friston. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003.

[68] Karl Friston and Christopher Frith. A duet for one. *Consciousness and cognition*, 36:390–405, 2015.

[69] Karl J Friston and Christopher D Frith. Active inference, communication and hermeneutics. *cortex*, 68:129–143, 2015.

[70] Chris Frith. Sharing the world—a social aspect of consciousness. *Open Mind*, 9:814–824, 2025.

[71] Chris D Frith and Uta Frith. Mechanisms of social cognition. *Annual review of psychology*, 63(1):287–313, 2012.

[72] Shaun Gallagher. Two problems of intersubjectivity. *Journal of Consciousness Studies*, 16(6-7):289–308, 2009.

[73] Mattia Gallotti and Chris D Frith. Social cognition in the we-mode. *Trends in cognitive sciences*, 17(4):160–165, 2013.

[74] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[75] Philip Gerrans and Ryan J Murray. Interoceptive active inference and self-representation in social anxiety disorder (sad): Exploring the neurocognitive traits of the sad self. *Neuroscience of Consciousness*, 2020(1):niaa026, 2020.

[76] Francis Gingras, Daniel Fiset, Marie-Pier Plouffe-Demers, Andréa Deschênes, Stéphanie Cormier, Hélène Forget, and Caroline Blais. Pain in the eye of the beholder: Variations in pain visual representations as a function of face ethnicity and culture. *British Journal of Psychology*, 114(3):621–637, 2023.

[77] Frédéric Gosselin and Philippe G Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17):2261–2271, 2001.

[78] Pierre Gosselin, Gilles Kirouac, and Francois Y Doré. Components and recognition of facial expression in the communication of emotion by actors. *Journal of personality and social psychology*, 68(1):83, 1995.

[79] Mononito Goswami, Minkush Manuja, and Maitree Leekha. Towards social & engaging peer learning: Predicting backchanneling and disengagement in children. *arXiv preprint arXiv:2007.11346*, 2020.

[80] Louise Goupil and Jean-Julien Aucouturier. Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, 212:104661, 2021.

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

[82] Yanliang Guo, Xianxu Hou, Feng Liu, Linlin Shen, Lei Wang, Zhen Wang, and Peng Liu. Styleau: Stylegan based facial action unit

manipulation for expression editing. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[83] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar): 1157–1182, 2003.

[84] Helene Haker, Maya Schneebeli, and Klaas Enno Stephan. Can bayesian theories of autism spectrum disorder help improve clinical practice? *Frontiers in psychiatry*, 7:185543, 2016.

[85] Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. Are you on my wavelength? interpersonal coordination in dyadic conversations. *Journal of nonverbal behavior*, 44(1):63–83, 2020.

[86] J Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450(7169):557–559, 2007.

[87] Todd C Handy. *Event-related potentials: A methods handbook*. MIT press, 2005.

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

[89] Bettina Heinz. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics*, 35(7):1113–1142, 2003.

[90] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[91] Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. Backchannel relevance spaces. In *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*, pages 137–146. Peter Lang Publishing Group, 2013.

[92] Bertrand Hermann, Gwen Goudard, Karine Courcoux, Mélanie Valente, Sebastien Labat, Lucienne Despois, Julie Bourmaleau, Louise Richard-Gilis, Frédéric Faugeras, Sophie Demeret, et al. "docfeeling": a new behavioural tool to help diagnose the minimally conscious state. *bioRxiv*, page 370775, 2018.

[93] Karlijn SFM Hermans, Olivia J Kirtley, Zuzana Kasanova, Robin Achterhof, Noëmi Hagemann, Anu P Hiekkaranta, Aleksandra Lecei, Leonardo Zapata-Fonseca, Ginette Lafit, Ruben Fossion, et al. Capacity for social contingency detection continues to develop across adolescence. *Social Development*, 31(3):530–548, 2022.

[94] Eckhard H Hess and Slobodan B Petrovich. Pupillary behavior in communication. In *Nonverbal behavior and communication*, pages 327–348. Psychology Press, 2014.

[95] Ursula Hess and Patrick Bourgeois. You smile–i smile: Emotion expression in social interaction. *Biological psychology*, 84(3):514–520, 2010.

[96] Judith Holler and Stephen C Levinson. Multimodal language processing in human communication. *Trends in cognitive sciences*, 23(8): 639–652, 2019.

[97] Judith Holler and Katie Wilkin. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of nonverbal behavior*, 35(2):133–153, 2011.

[98] Paul Hömke, Judith Holler, and Stephen C Levinson. Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, 50(1):54–70, 2017.

[99] Paul Hömke, Judith Holler, and Stephen C Levinson. Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS one*, 13(12):e0208030, 2018.

[100] Richard Huskey, Amelia Couture Bue, Allison Eden, Clare Grall, Dar Meshi, Kelsey Prena, Ralf Schmälzle, Christin Scholz, Benjamin O Turner, and Shelby Wilcox. Marr's tri-level framework integrates biological explanation across communication subfields. *Journal of Communication*, 70(3):356–378, 2020.

[101] Peter Indefrey and Willem JM Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144, 2004.

[102] Rachael E Jack and Philippe G Schyns. The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634, 2015.

[103] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19): 7241–7244, 2012.

[104] Rachael E Jack, Oliver GB Garrod, and Philippe G Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192, 2014.

[105] Mikael Jensen. Smile as feedback expressions in interpersonal interaction. *International Journal of Psychological Studies*, 7(4):95–105, 2015.

[106] Peiqing Jin, Jiajie Zou, Tao Zhou, and Nai Ding. Eye activity tracks task-relevant structures during speech and auditory sequence perception. *Nature communications*, 9(1):5374, 2018.

[107] Eshin Jolly. Pymer4: Connecting r and python for linear mixed modeling. *Journal of Open Source Software*, 3(31):862, 2018.

[108] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Hind Almerekhi, and Bernard 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.

[109] Karel J Keesman. *System identification: an introduction*. Springer Science & Business Media, 2011.

[110] Adam Kendon. An agenda for gesture studies. *Semiotic review of books*, 7(3):8–12, 1996.

[111] David A Kenny, Deborah A Kashy, and William L Cook. *Dyadic data analysis*. Guilford Publications, 2020.

[112] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[113] Ami Klin, David J Lin, Phillip Gorrindo, Gordon Ramsay, and Warren Jones. Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, 459(7244):257–261, 2009.

[114] Birgit Knudsen, Ava Creemers, and Antje S Meyer. Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, 11:593671, 2020.

[115] S Oliver Kobald, Edmund Wascher, Holger Heppner, and Stephan Getzmann. Eye blinks are related to auditory information processing: Evidence from a complex speech perception task. *Psychological Research*, 83(6):1281–1291, 2019.

[116] Jorie Koster-Hale and Rebecca Saxe. Theory of mind: a neural prediction problem. *Neuron*, 79(5):836–848, 2013.

[117] Ágnes Melinda Kovács, Ernő Téglás, and Ansgar Denis Endress. The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012):1830–1834, 2010.

[118] John W Krakauer, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017.

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

[120] Eva G Krumhuber, Lina I Skora, Harold CH Hill, and Karen Lander. The role of facial movements in emotion recognition. *Nature Reviews Psychology*, 2(5):283–296, 2023.

[121] Joshua P Kulasingham and Jonathan Z Simon. Algorithms for estimating time-locked neural response components in cortical processing of continuous speech. *IEEE Transactions on Biomedical Engineering*, 70(1):88–96, 2022.

[122] Theresa Küntzler, T Tim A Höfling, and Georg W Alpers. Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Frontiers in psychology*, 12: 627561, 2021.

[123] Edmund Lalor and Aaron Nidiffer. On the generative mechanisms underlying the cortical tracking of natural speech: a position paper. 2025.

[124] Edmund C Lalor and John J Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European journal of neuroscience*, 31(1):189–193, 2010.

[125] Edmund C Lalor, Sherlyn Yeap, Richard B Reilly, Barak A Pearlmutter, and John J Foxe. Dissecting the cellular contributions to early visual sensory processing deficits in schizophrenia using the vespa evoked response. *Schizophrenia research*, 98(1-3):256–264, 2008.

[126] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, 2016.

[127] Han Z Li, Yanping Cui, and Zhizhang 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.

[128] Yuezun Li, Ming-Ching Chang, and Siwei 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.

[129] Elsa Lindboom, Aaron Nidiffer, Laurel H Carney, and Edmund C Lalor. Incorporating models of subcortical processing improves the ability to predict eeg responses to natural speech. *Hearing research*, 433:108767, 2023.

[130] Lennart Ljung. *System identification toolbox: User's guide*. MathWorks Incorporated Natick, MA, USA, 1995.

[131] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[132] Lennart Ljung et al. Theory for the user. *System identification*, 1987.

[133] Valeria Lucarini, Martine Grice, Simon Wehrle, Francesco Cangemi, Francesca Giustozzi, Stefano Amorosi, Francesco Rasmi, Nikolas Fascendini, Francesca Magnani, Carlo Marchesi, et al. Language in interaction: turn-taking patterns in conversations involving individuals with schizophrenia. *Psychiatry Research*, 339:116102, 2024.

[134] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang,

Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[135] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

[136] Sibo Ma, Alejandro Salinas, Julian Nyarko, and Peter Henderson. Breaking down bias: On the limits of generalizable pruning strategies. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2437–2450, 2025.

[137] Lara Maister, Sophie De Beukelaer, Matthew R Longo, and Manos Tsakiris. The self in the mind's eye: Revealing how we truly see ourselves through reverse correlation. *Psychological Science*, 32(12): 1965–1978, 2021.

[138] Puspita Majumdar, Surbhi Mittal, Richa Singh, and Mayank Vatsa. Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3786–3795, 2021.

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

[140] Valeria Manera, Cristina Becchio, Ben Schouten, Bruno G Bara, and Karl Verfaillie. Communicative interactions improve visual detection of biological motion. *PloS one*, 6(1):e14594, 2011.

[141] Valeria Manera, Marco Del Giudice, Bruno G Bara, Karl Verfaillie, and Cristina Becchio. The second-agent effect: communicative gestures increase the likelihood of perceiving a second agent. *PLoS One*, 6(7):e22650, 2011.

[142] Valeria Manera, Ben Schouten, Karl Verfaillie, and Cristina Becchio. Time will show: real time predictions during interpersonal action perception. *PloS one*, 8(1):e54949, 2013.

[143] Panos Z Marmarelis and Vasilis Z Marmarelis. The white-noise method in system identification. In *Analysis of physiological systems: the white-noise approach*, pages 131–180. Springer, 1978.

[144] Panos Z Marmarelis and Ken-Ichi Naka. White-noise analysis of a neuron chain: an application of the wiener theory. *Science*, 175 (4027):1276–1278, 1972.

[145] Vasilis Z Marmarelis. *Nonlinear dynamic modeling of physiological systems*. John Wiley & Sons, 2004.

[146] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[147] Ann S Masten, Glenn I Roisman, Jeffrey D Long, Keith B Burt, Jelena Obradović, Jennifer R Riley, Kristen Boelcke-Stennes, and Auke Tellegen. Developmental cascades: linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental psychology*, 41(5):733, 2005.

[148] Magdalena Matyjek, Isabel Dziobek, Antonia Hamilton, and Thalia Wheatley. Social interaction style in autism: A critical review of social behaviours and outcomes in autistic and neurotypical interactions. 2025.

[149] Oded Mayo and Simone Shamay-Tsoory. Dynamic mutual predictions during social learning: A computational and interbrain model. *Neuroscience & Biobehavioral Reviews*, 157:105513, 2024.

[150] Monica Mazza, Melania Mariano, Sara Peretti, Francesco Masedu, Maria Chiara Pino, and Marco Valenti. The role of theory of mind on social information processing in children with autism spectrum disorders: A mediation analysis. *Journal of autism and developmental disorders*, 47(5):1369–1379, 2017.

[151] Evelyn Z McClave. Linguistic functions of head movements in the context of speech. *Journal of pragmatics*, 32(7):855–878, 2000.

[152] Judith McLean, S Raab, and LA Palmer. Contribution of linear mechanisms to the specification of local motion by simple cells in areas 17 and 18 of the cat. *Visual neuroscience*, 11(2):271–294, 1994.

[153] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[154] Andrew N Meltzoff, Anna Waismeyer, and Alison Gopnik. Learning about causes from people: observational causal learning in 24-month-old infants. *Developmental psychology*, 48(5):1215, 2012.

[155] Beren Millidge, Anil Seth, and Christopher L Buckley. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*, 2021.

[156] Kibum Moon, SoJeong Kim, Jinwon Kim, Hackjin Kim, and Young-gun Ko. The mirror of mind: Visualizing mental representations of self through reverse correlation. *Frontiers in Psychology*, 11:1149, 2020.

[157] Nikki Moran, Lauren V Hadley, Maria Bader, and Peter E Keller. Perception of 'back-channeling'nonverbal feedback in musical duo improvisation. *PLoS One*, 10(6):e0130070, 2015.

[158] Louis-Philippe Morency, Iwan De Kok, and Jonathan Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *International workshop on intelligent virtual agents*, pages 176–190. Springer, 2008.

[159] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas 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.

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

[161] Lynne Murray, Adriane Arteche, Pasco Fearon, Sarah Halligan, Ian Goodyer, and Peter Cooper. Maternal postnatal depression and the development of depression in offspring up to 16 years of age. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50 (5):460–470, 2011.

[162] Richard F Murray. Classification images: A review. *Journal of vision*, 11(5):2–2, 2011.

[163] Evelien Nackaerts, Johan Wagemans, Werner Helsen, Stephan P Swinnen, Nicole Wenderoth, and Kaat Alaerts. Recognizing biological motion and emotions from point-light displays in autism spectrum disorders. 2012.

[164] Tamami Nakano and Shigeru Kitazawa. Eyeblink entrainment at breakpoints of speech. *Experimental brain research*, 205(4):577–581, 2010.

[165] Peter Neri. How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review*, 17(6):802–808, 2010.

[166] Peter Neri, M Concetta Morrone, and David C Burr. Seeing biological motion. *Nature*, 395(6705):894–896, 1998.

[167] Peter Neri, Jennifer Y Luu, and Dennis M Levi. Meaningful interactions can enhance visual discrimination of human agents. *Nature neuroscience*, 9(9):1186–1192, 2006.

[168] John G Neuhoff. Perceptual bias for rising tones. *Nature*, 395(6698): 123–124, 1998.

[169] Christine Nussbaum, Annett Schirmer, and Stefan R Schweinberger. Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. *Social Cognitive and Affective Neuroscience*, 17(12):1145–1154, 2022.

[170] Łukasz Okruszek, Aleksandra Piejka, Adam Wysokiński, Ewa Szczepocka, and Valeria Manera. Biological motion sensitivity, but not interpersonal predictive coding is impaired in schizophrenia. *Journal of Abnormal Psychology*, 127(3):305, 2018.

[171] Łukasz Okruszek, Aleksandra Piejka, Adam Wysokiński, Ewa Szczepocka, and Valeria Manera. The second agent effect: Interpersonal predictive coding in people with schizophrenia. *Social Neuroscience*, 14(2):208–213, 2019.

[172] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25(7):1697–1706, 2015.

[173] Eleanor R Palser, Clare E Palmer, Alejandro Galvez-Pol, Ricci Hannah, Aikaterini Fotopoulou, and James M Kilner. Alexithymia mediates the relationship between interoceptive sensibility and anxiety. *PloS one*, 13(9):e0203212, 2018.

[174] Marina A Pavlova, Jonas Moosavi, Claus-Christian Carbon, Andreas J Fallgatter, and Alexander N Sokolov. Emotions behind a mask: the value of disgust. *Schizophrenia*, 9(1):58, 2023.

[175] Luiz Pessoa. The entangled brain. *Journal of cognitive neuroscience*, 35(3):349–360, 2023.

[176] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[177] Jean Piaget. *Language and Thought of the Child: Selected Works vol 5*. Routledge, 2005.

[178] Elise A Piazza, Marius Cătălin Iordan, and Casey Lew-Williams. Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology*, 27(20):3162–3167, 2017.

[179] Gianluigi Pillonetto, Francesco Dinuzzo, Tianshi Chen, Giuseppe De Nicolao, and Lennart Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

[180] Isabella Poggi, Francesca D'Errico, Laura Vincze, et al. Types of nods. the polysemy of a social signal. In *LREC*, 2010.

[181] Emmanuel Ponsot, Patrick Susini, and Sabine Meunier. A robust asymmetry in loudness between rising-and falling-intensity tones. *Attention, Perception, & Psychophysics*, 77(3):907–920, 2015.

[182] Emmanuel Ponsot, Pablo Arias, and Jean-Julien Aucouturier. Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America*, 143(1): EL19–EL24, 2018.

[183] Emmanuel Ponsot, Juan José Burred, Pascal Belin, and Jean-Julien Aucouturier. Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, 115 (15):3972–3977, 2018.

[184] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. Backchannel strategies for artificial listeners. In *International conference on intelligent virtual agents*, pages 146–158. Springer, 2010.

[185] Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

[186] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[187] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[188] Elizabeth Redcay and Leonhard Schilbach. Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature reviews neuroscience*, 20(8):495–505, 2019.

[189] R Clay Reid and Robert M Shapley. Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus. *Nature*, 356 (6371):716–718, 1992.

[190] Olivia M Rifai, Sue Fletcher-Watson, Lorena Jiménez-Sánchez, and Catherine J Crompton. Investigating markers of rapport in autistic and nonautistic interactions. *Autism in Adulthood*, 4(1):3–11, 2022.

[191] Dario L Ringach, Guillermo Sapiro, and Robert Shapley. A subspace reverse-correlation technique for the study of visual neurons. *Vision research*, 37(17):2455–2464, 1997.

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

[193] Etienne B Roesch, Lucas Tamarit, Lionel Reveret, Didier Grandjean, David Sander, and Klaus R Scherer. Facsgen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal of Nonverbal Behavior*, 35(1):1–16, 2011.

[194] Lawrence D Rosenblum, Claudia Carello, and Richard E Pastore. Relative effectiveness of three stimulus variables for locating a moving sound source. *Perception*, 16(2):175–186, 1987.

[195] Nikolai F Rulkov, Mikhail M Sushchik, Lev S Tsimring, and Henry DI Abarbanel. Generalized synchronization of chaos in directionally coupled chaotic systems. *Physical Review E*, 51(2):980, 1995.

[196] David E Rumelhart, Geoffrey E Hinton, James L McClelland, et al. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1 (45-76):26, 1986.

[197] Moses Rupenga and Hima 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.

[198] Dana Samson, Ian A Apperly, Jason J Braithwaite, Benjamin J Andrews, and Sarah E Bodley Scott. Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of experimental psychology: human perception and performance*, 36 (5):1255, 2010.

[199] Wataru Sato, Sylwia Hyniewska, Kazusa Minemoto, and Sakiko Yoshikawa. Facial expressions of basic emotions in japanese laypeople. *Frontiers in psychology*, 10:259, 2019.

[200] Scott Schaefer, Travis McPhail, and Joe Warren. Image deformation using moving least squares. In *ACM SIGGRAPH 2006 Papers*, pages 533–540. 2006.

[201] Martin Schetzen. *The Volterra and Wiener theories of nonlinear systems*. Krieger Publishing Co., Inc., 2006.

[202] William Schiff and Rivka Oldak. Accuracy of judging time to arrival: effects of modality, trajectory, and gender. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):303, 1990.

[203] Leonhard Schilbach, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley. Toward a second-person neuroscience1. *Behavioral and brain sciences*, 36(4): 393–414, 2013.

[204] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171, 2020.

[205] Erich Seifritz, John G Neuhoff, Deniz Bilecen, Klaus Scheffler, Henrietta Mustovic, Hartmut Schächinger, Raffaele Elefante, and

Francesco Di Salle. Neural processing of auditory looming in the human brain. *Current Biology*, 12(24):2147–2151, 2002.

[206] Tricia Seow and Stephen M Fleming. Perceptual sensitivity is modulated by what others can see. *Attention, Perception, & Psychophysics*, 81(6):1979–1990, 2019.

[207] Ignacio Serna, Alejandro Pena, Aythami Morales, and Julian Fierrez. Insidebias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3720–3727. IEEE, 2021.

[208] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[209] Siv Skotheim, Hanne Cecilie Braarud, Kjartan Høie, Maria Wik Markhus, Marian Kjellevold Malde, Ingvild Eide Graff, Jan Øystein Berle, and Kjell Morten Stormark. Subclinical levels of maternal depression and infant sensitivity to social contingency. *Infant Behavior and Development*, 36(3):419–426, 2013.

[210] Nicholas A Smith, Valerie F McDaniel, Jean M Ispa, and Bob McMurray. Maternal depression and the timing of mother–child dialogue. *Infant and child development*, 32(1):e2389, 2023.

[211] Ryan Smith, Paul Badcock, and Karl J Friston. Recent advances in the application of predictive coding and active inference models within clinical neuroscience. *Psychiatry and Clinical Neurosciences*, 75(1):3–13, 2021.

[212] Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C Samson. Facial expression analysis with affdex and facet: A validation study. *Behavior research methods*, 50(4):1446–1460, 2018.

[213] Tricia Striano and Philippe Rochat. Developmental link between dyadic and triadic social competence in infancy. *British Journal of Developmental Psychology*, 17(4):551–562, 1999.

[214] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of neuroscience methods*, 274:141–145, 2016.

[215] Yukari Takarae, Michael K McBeath, and R Chandler Krynen. Perception of dynamic point light facial expression. *The American Journal of Psychology*, 134(4):373–384, 2021.

[216] Evan Thompson. *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press, 2010.

[217] Etienne Thoret, Thomas Andrillon, Damien Léger, and Daniel Pressnitzer. Probing machine-learning classifiers using noise, bubbles, and reverse correlation. *Journal of neuroscience methods*, 362: 109297, 2021.

[218] Xue Tian, Yiying Song, and Jia Liu. Decoding face identity: A reverse-correlation approach using deep learning. *Cognition*, 254: 106008, 2025.

[219] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[220] Edward Tronick, Heidelise Als, Lauren Adamson, Susan Wise, and T Berry Brazelton. The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child psychiatry*, 17(1):1–13, 1978.

[221] Khiet Phuong Truong, Ronald Walter Poppe, IA de Kok, and Dirk KJ Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*, pages 2973–2976. International Speech Communication Association, 2011.

[222] Laura C Trutoiu, Jessica K Hodgins, and Jeffrey 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.

[223] Kristian Tylén, Micah Allen, Bjørk K Hunter, and Andreas Roepstorff. Interaction vs. observation: distinctive modes of social cognition in human brain and behavior? a combined fmri and eye-tracking study. *Frontiers in human neuroscience*, 6:331, 2012.

[224] Sander Van de Cruys, Kris Evers, Ruth Van der Hallen, Lien Van Eylen, Bart Boets, Lee De-Wit, and Johan Wagemans. Precise

minds in uncertain worlds: predictive coding in autism. *Psychological review*, 121(4):649, 2014.

[225] Francisco J Varela. Invitation aux sciences cognitives. *(No Title)*, 1996.

[226] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[227] Tabea Von Der Lühe, Valeria Manera, Iva Barisic, Cristina Becchio, Kai Vogeley, and Leonhard Schilbach. Interpersonal predictive coding, not action perception, is impaired in autism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693): 20150373, 2016.

[228] Basil Wahn, Laura Schmitz, Alan Kingstone, and Anne Böckler-Raettig. When eyes beat lips: speaker gaze affects audiovisual integration in the mcgurk illusion. *Psychological Research*, 86(6):1930–1943, 2022.

[229] Anna Waismeyer and Andrew N Meltzoff. Learning to make things happen: Infants' observational learning of social and physical causal events. *Journal of experimental child psychology*, 162:58–71, 2017.

[230] Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207, 2000.

[231] Simon Wehrle, Kai Vogeley, and Martine Grice. Backchannels in conversations between autistic adults are less frequent and less diverse prosodically and lexically. *Language and cognition*, 16(1):108–133, 2024.

[232] Katharina Weitz, Teena Hassan, Ute Schmid, and Jens-Uwe Garbas. Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable ai methods. *tm-Technisches Messen*, 86(7-8):404–412, 2019.

[233] Ron White. Back channelling, repair, pausing, and private speech. *Applied linguistics*, 18(3):314–344, 1997.

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

[235] Nobert Wiener and PESI Masani. The prediction theory of multivariate stochastic processes, ii: The linear predictor. *Acta Mathematica*, 99(1):93–137, 1958.

[236] Marcus Wilms, Leonhard Schilbach, Ulrich Pfeiffer, Gary Bente, Gereon R Fink, and Kai Vogeley. It's in your eyes—using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social cognitive and affective neuroscience*, 5(1):98–107, 2010.

[237] Tian Xu, Jiayu Zhan, Oliver GB Garrod, Philip HS Torr, Song-Chun Zhu, Robin AA Ince, and Philippe G Schyns. Deeper interpretability of deep networks. *arXiv preprint arXiv:1811.07807*, 2018.

[238] Victor H Yngve. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466, 1960.

[239] Hui Yu, Oliver GB Garrod, and Philippe G Schyns. Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3):152–162, 2012.

[240] Sarra Zaied, Catherine Soladié, and JJ Aucouturier. The psychophysics of empathy: Using reverse-correlation to quantify the overlap between self & other representations of emotional expressions. *PsyArXiv preprint rdmve*, 2023.

[241] Aynaz ADL ZARRABI. L'université marie et louis pasteur reverse-correlation modeling of deficits of prosody perception in right-hemisphere stroke. 2025.

[242] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[243] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10): 1084–1102, 2018.

[244] Yong Zhao, Le Yang, Ercheng Pei, Meshia Cedric Oveneke, Mitchel Alioscha-Perez, Longfei Li, Dongmei Jiang, and Hichem Sahli. Action unit driven facial expression synthesis from a single image with patch attentive gan. In *Computer Graphics Forum*, volume 40, pages 47–61. Wiley Online Library, 2021.

[245] Ruicong Zhi, Caixia Zhou, Tingting Li, Shuai Liu, and Yi Jin. Action unit analysis enhanced facial expression recognition by deep neural network evolution. *Neurocomputing*, 425:135–148, 2021.

[246] Muhammad SA Zilany, Ian C Bruce, and Laurel 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.