Deep construction of an affective latent space via multimodal enactment

Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro Damelio, Giuliano Grossi, Raffaella Lanzarotti

This is the author's version of an article that will be published in this journal. Changes will be made to this version by the publisher prior to publication.
The final version of record will be available at DOI: 10.1109/TCDS.2017.2788820

Deep construction of an affective latent space via multimodal enactment

Giuseppe Boccignone, *Member, IEEE,* Donatello Conte, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, and Raffaella Lanzarotti

***Abstract*—We draw on a simulationist approach to the analysis of facially displayed emotions - e.g., in the course of a face-to-face interaction between an expresser and an observer. At the heart of such perspective lies the enactment of the perceived emotion in the observer. We propose a novel probabilistic framework based on a deep latent representation of a continuous affect space, which can be exploited for both the estimation and the enactment of affective states in a multimodal space (visible facial expressions and physiological signals). The rationale behind the approach lies in the large body of evidence from affective neuroscience showing that when we observe emotional facial expressions, we react with congruent facial mimicry. Further, in more complex situations, affect understanding is likely to rely on a comprehensive representation grounding the reconstruction of the state of the body associated with the displayed emotion. We show that our approach can address such problems in a unified and principled perspective, thus avoiding *ad hoc* heuristics while minimising learning efforts.**

***Index Terms*—Emotion, human-agent interaction, deep learning, simulation, Bayesian models**

I. Introduction

SEAMLESSLY, in the course of our entanglements and conflicts, dealings and struggles, we "perceive" the social signals brought on by others, and we recognise and understand their meaning. Yet, gazing at a gesture, glimpsing a smile or hearing a laugh involves a kind of perception which is different from the appraisal of the lifeless world.

A large body of evidence [1], [2] shows that alongside the sensory information concerning others' social stimuli - actions, in a wide sense -, one's own motor and visceromotor representations of those stimuli are enacted. Humans mirror gestures, postures, emotions, speech of other perceived humans, at least neurally, and sometimes bodily and behaviourally. Mirroring grounds the capability of own reproduction of the action in question "as if" a similar action were performed or a similar emotion experienced. Such simulation-based mechanism is likely to play a crucial role in individual cognition and social interaction [1], [2].

The rationale behind this study is thus straightforward and stems from the attempt at answering a deceptively simple, albeit overlooked question: can we exploit such primitive and fundamental simulation-based mechanism for designing artificial agents?

G. Boccignone, V. Cuculo, A. D'Amelio, G. Grossi and R. Lanzarotti are with PHuSeLab - Department of Computer Science, University of Milan, Italy. E-mail: see http://phuselab.di.unimi.it.

D. Conte is with the Computer Science Laboratory (EA 6300) at University of Tours, France. E-mail: donatello.conte@univ-tours.fr.

Manuscript received XXX; revised XXX.

Answering this question is important as witnessed by the growing interest for the computational modelling of emotion, as an attempt to develop and validate computational models of human emotion mechanisms [3], and it is crucial in the realm of cognitive and social robotics [4]–[7] (but see Section VI, for a wider discussion).

Challenges: In this paper we shall focus on the case of the perception of emotional facial expressions, but in a simulation-based context also involving the observer's physiological reactions (see [8] and [9] for a review). In particular we address the issue of facial expression mirroring and mimicry, which is at the heart of simulationist accounts.

Face perception is likely to be the most developed visual perceptual skill in humans and, cogently, most face viewing occurs in the context of social interactions [9]. Undeniably, part of the ability to extract affective information from faces can be attributed to visual expertise.

Yet, facial expressions are facial actions; as such, their perception is likely to draw on simulation mechanisms underlying action perception in general [8], [9]. Beyond visual expertise, it is increasingly apparent that visuomotor simulation activates autonomic activities [9], [10]. These participate in building a deep understanding of the perceived affective expression [8].

Our approach: We propose a novel, probabilistic computational model for dynamic affective facial expression perception relying on a mirroring mechanism. The latter involves both facial gesture and physiological simulation. In brief, in the course of a dyadic engagement, the observer's visual system, while perceiving expresser's facial display, interacts with an extended system, which takes in the emotion system. Interaction is regulated by the mediation of a visuomotor component for somatic action perception, which transforms the sensory information of observed facial actions into the observer's own motor representation. In turn, a continuous core affect space [11] characterised by the valence (pleasure-displeasure) and arousal (sleepy-activated) state variables, triggers autonomic, visceromotor processes. The simulation-based dynamics involving both the visuomotor and visceromotor routes can generate observer's actual responses, namely facial mimicry and physiological responses.

The overall goal of the approach is to allow the modelled observer to reach a core affect state similar to that of the expresser. Indeed, meeting such condition is preliminary, in the embodied perspective, to ground subsequent processing for affect understanding, e.g. the retrieval of conceptual knowledge about the emotion signalled by the expresser [8], [9].

Technically, when the observer's model is put into work, at the learning stage, inputs are (i) a video clip of facial

expressions displayed by a subject engaged in spontaneous interactions along with (ii) subject's physiological recordings (cardiac and electro-dermal activities), (iii) the annotated continuous values of valence (V) and arousal (A); at the testing stage, only the facial expression clips of new subjects are provided to the observer's model (see Fig. 1). Measurable outputs are the observer's facial and autonomic mimicry, together with the trajectories of the observer's internal V/A core affect dynamics. For our experiments we exploit data of spontaneous and natural emotions collected in the RECOLA dataset [12], a multimodal corpus designed to monitor subjects as they worked in pairs remotely to complete a collaborative task. The corpus includes audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA) modalities; the data are manually annotated with the continuous dimensional labels of arousal and valence (cfr. Fig. 1).

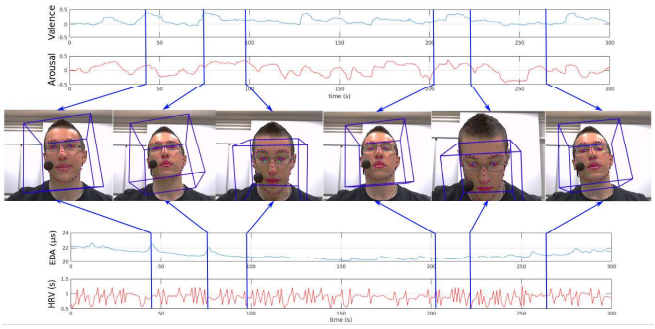


Fig. 1. Inputs available to our model as provided by the RECOLA Dataset. The centre panel shows an excerpt (few video frames) of an expresser's facial action; for completeness, the facial landmarks and head pose (box), as computed by the model along the perceptual stage, are also overlaid on the detected face. The bottom panel displays the time course of the expresser's physiological signals (HRV and EDA). Continuous V/A annotations (the attributed core affect state) are shown in the top panel. Arrows indicate the HRV, EDA and V/A values corresponding to the shown frames.

Methodology: We follow a multilevel approach to set up the computational model, in order to fully exploit the cross-level link between the psychological and neurophysiological levels of explanation [13], [14]. Marr proposed three distinct levels of description/explanation in the cognitive sciences [13]: the *what/why* level (computational theory, i.e. the individuation of a computable function as a model of a given behavioural phenomenon), the *how* level (algorithm) and the *realisation* level (implementation). Here, we comply with Chater *et al.* proposal [14] arguing that, in the light of the growing exploitation of Bayesian methods, Marr's three-fold hierarchy should be reorganised into two levels: the computational theory level, formalised in terms of Bayesian theory, and the implementation theory level, embedding both Marr's algorithmic and realisation levels. Henceforth, we will call the formal instantiation of these two levels the *theoretical model* and the *implementation model*, respectively.

The rationales behind the model are discussed in Section II. The theoretical model is presented in Section III, while the implementation model is outlined in Section IV. In Section V we show that this model-driven approach, beyond its theoretical appeal, can address flexible representation and analysis of facial expressions, that lends itself well to the task of learning from few as possible examples, while predicting some general results obtained at the level of psychological explanation.

To the best of our knowledge, not much effort hitherto has been spent in the course along which we are moving; thus, a general discussion of the approach and related work is postponed and carried out in-depth in the concluding Section VI.

II. Background and Rationales

How does affective neuroscience spell out the understanding of facial expression of emotions? In a nutshell, according to Adolph's model [8], upon the onset of an emotionally meaningful stimulus, observer's response undergoes the following stages: 1) fast early perceptual processing of highly salient stimuli (120 ms); 2) detailed perception and emotional reaction involving the body (170 ms); 3) retrieval of conceptual knowledge about the emotion signalled by the expresser's face (> 300 ms).

Stage 2 and the onset of stage 3 together pave the way to the understanding of affective expressions. Adolph's scheme points at the orbitofrontal cortex (OFC), the amygdala and the insula, three highly interconnected areas, as the key structures for *central affective control*, regulating the development of an emotional episode (cfr. Fig. 2 for the discussion that follows). Cogently, the amygdala and the OFC generate an emotional response in the observer, via connections to the motor structures, hypothalamus (HYP), and brainstem nuclei, where components of an emotional response to the facial expression can be activated. This mechanism contributes to the elicitation of knowledge about the expresser's emotional state, via the process of internal simulation, and would draw on the insula and somatosensory related cortices for representing the emotional changes in the observer. The insula cortex integrates visceral, pain, and temperature sensations and also provides visceromotor control of both the sympathetic and parasympathetic outputs, [15]. Functional imaging studies suggest that the posterior insula (PIns) receives topographically organised interoceptive inputs via the thalamus, and projects to the anterior insula (AIns). The latter integrates those inputs with inputs from cortical areas involved in perceptive, emotional, and cognitive processing [15]. It has been argued [16] that functional interactions between the amygdala and the OFC form a potential neural substrate for the encoding of the psychological core affect dimensions [11] of valence and arousal (V/A). A similar role is also played by the insula [15]. The V/A dimensions, can be thought of as "emotion primitives" supporting at the neurobiological level a central continuous emotion space [17].

Adolph's account does not provide further details about the explicit involvement of motor mechanisms in the process, albeit these being acknowledged [8]. This is not a minor issue. While investigating whether emotion facial expressions modulate the functional connectivity of the amygdala with the rest of the brain, it has been shown [18] that all queried emotions enhanced functional integration with premotor cortices (e.g., ventral premotor cortex, vPMC). Also, the amygdala forms a closed processing loop with cortical motor areas M1, M3,

M4 and supplementary motor areas (SMA) targeting the facial nucleus in the brain stem, which hosts the motor neurons that synapse on the muscles of facial expressions [19]

Facial expressions are facial actions and it has been posited that a "resonance behaviour" is one such mechanism where an individual repeats overtly a movement made by another individual. A striking example is provided by either human or rhesus monkey newborns imitating adult facial gestures, [20]. At the neurobiological level, the resonance or mirroring behaviour has been initially explained in terms of mirror neuron (MN) activity. The critical feature of MNs is the functional matching between a motor response and a perceptual one. MNs have been first localised in monkeys' ventral premotor area F5 and then in the inferior parietal lobule (IPL). Subsequently, other areas have been endowed with mirroring capabilities, so that it is currently more appropriate to address the properties of a MN system (MNS) or network [20].

Crucially, evidence has highlighted that also the human brain is provided with mirroring capabilities, and internal simulation has been related to coding the intentions of actions performed by others [1]. Mukamel *et al.* [21] have performed direct recordings of MNs activity in human patients while executing or observing facial emotional expressions.

The initial studies on the neurobiology of imitation in humans suggested a core imitation circuitry composed of three major neural systems [20], [22]: the posterior part of the superior temporal sulcus (pSTS), the rostral part of the inferior parietal lobule (rIPL), and the posterior part of the inferior frontal gyrus and adjacent ventral premotor cortex (pIFG/vPMC complex). The information processing flow is likely to occur as follows [23]: the pSTS provides a higher order visual processing of the observed action by coding an early visual description of the action; this information is sent to the rIPL and pIFG/vPMC complex that form a parieto-frontal mirroring (both motor and visual) system; the posterior parietal cortex codes the precise kinesthetic aspect of the movement and sends this information to inferior frontal mirror neurons in the pIFG; efferent copies of motor plans are sent from parietal and frontal mirror areas back to the STS; here there would be a matching process between the visual description of the observed action and the anticipated outcome of the planned imitative action. In this perspective Chakrabarti *et al.* [24] have proposed that the MNS acts as an intermediate module for facial action perception. Such scheme represents a starting point to identify a *visuomotor route* for the perception and mirroring of dynamic facial expressions of emotion. The STS traditionally accounts for changeable visual aspects of faces and contributes to a core visual system for face perception [25] by interacting with the occipital face area (OFA, for the perception of face parts) and the fusiform face area (FFA, invariant aspects of faces, identity); OFA and FFA receive inputs from early visual cortices. Crucially, in a mirroring framework, the STS enables the *visual route* to the MNs so to allow matching between sensory predictions of imitative motor plans and a visual description of observed actions [26].

Recent findings have prompted the idea that a mirror mechanism is also present in the cortical areas involved in coding emotions [2], and markedly along affective facial expression processing [20]. Namely, motor knowledge required is grounded in visceromotor actions, that is motor command sequences directed to visceral organs [20], establishing a *visceromotor route* to affective expression perception. In both monkeys and humans, a region much involved in this kind of output are the insular cortices [20]. Interoceptive experience may largely reflect limbic predictions about the expected state of the body that are constrained by ascending visceral sensations (see, [27], [28]. The anterior insula informs the rest of the brain (markedly, the amygdala and the OFC) of interoceptive changes by sending predictions based on anticipated visceromotor consequences and information forwarded by the posterior insula; meanwhile, the latter propagates prediction-error signals back to visceromotor regions to modify predictions. At a lower level, as viscerosensory signals undergo substantial signal conditioning in the brainstem, it is likely that brainstem and subcortical structures contribute directly to active inference, for instance, by computing themselves a prediction error [27], [28].

All the issues touched in the above discussions can be synthetically subsumed under the architecture of the distributed neural system for perception of dynamic facial expressions of emotion outlined in Fig. 2. The scheme shows at a glance the three main routes contributing to affective facial expression processing: the *visual route*, the *visuomotor route* and the *visceromotor route*. The affective core is provided by the close interactions occuring among the amygdala, the insula (AIns) and the OFC.

Modelling assumptions. To sum up, at the neurobiological level, core affect dynamics is consequent on the activity of a complex, open system. Such an open system is more suitably conceived as subject to stochastic variability resulting from the entanglement of many internal (and external) activities that influence it [29]. Cogently, at the psychologic/behavioural description level, a person always has a core affect and, moment-by-moment, a person's emotional state can be described in terms of how pleasant or unpleasant (valence) and how activated (arousal) the person is [17], [30]. Kuppens *et al.* [29] have remarkably shown that, across time, the core affect unfolding can be represented as a trajectory, i.e. a realisation of a stochastic process reflecting the typical pattern of affective changes and fluctuations that V/A levels undergo across time and that characterise an individual.

Neurobiological findings summarised above suggest that at the heart of the entanglement is the "core affect $\rightarrow$ action $\rightarrow$ motor" hierarchy replicated in the visuomotor and visceromotor routes. From a modelling standpoint, the different components involved in such hierarchy can be conceived as dynamic input-state-output model. The input or control is provided top-down (TD) by an upper level component and the output is an emission to a lower-level component or, equivalently, a bottom-up (BU) observation of the lower-level state. Under Markov assumption, we describe each component as a discrete-time (nonlinear) dynamical system

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t-1), \mathbf{c}(t), \boldsymbol{\epsilon}_{\mathbf{x}}(t)), \tag{1}$$

$$\mathbf{x}_{\mathrm{BU}}(t) = \mathbf{g}(\mathbf{x}(t), \mathbf{c}(t), \boldsymbol{\epsilon}_{\mathbf{x}_{\mathrm{BU}}}(t)), \tag{2}$$

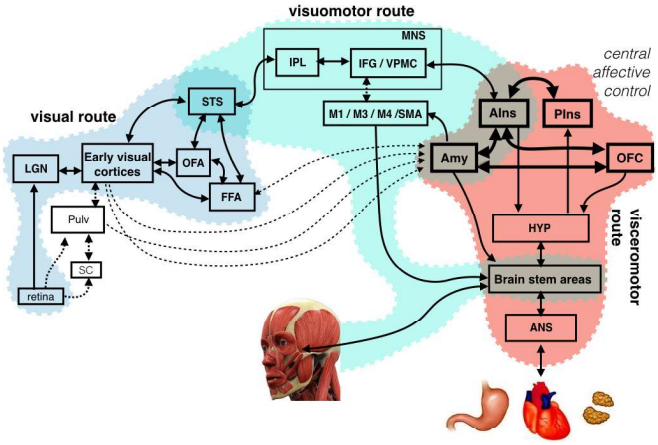


Fig. 2. Architecture of a distributed neural system for the perception of dynamic facial expressions of emotion (observer's side). Two, reciprocal, heavy arrowheads indicate "forward" and "backward" projections between areas (boxes). Only the main areas of interest have been included. The architecture incorporates a module for "action perception" based on the human MNS, (IPL, IFG/VPMC complex), which mediates between the external stimuli (expresser's facial action) as processed along the visual route (retina, lateral geniculate nuclei,LGN, early Visual cortices, OFA, FFA, STS), and the internal motor/action representation, provided by the MNS via the STS interface. The MNS feeds the necessary input to activate the core affect system, represented by the amygdala (Amy), the anterior insula (AIns) and the OFC. This system coordinates the dynamics of the activities occurring along the *visuomotor* (STS, IPL, IFG/VPMC, cortical motor areas M1/M3/M4, supplementary motor area, SMA, and subcortical motor nuclei in the brainstem), and *visceromotor* routes (PIns, HYP, brainstem visceromotor nuclei and the autonomic nervous system ANS) either by modulating perceptual representations via feedback, and by generating an emotional response in the subject, via connections to motor structures, hypothalamus (HYP), and brainstem nuclei, where components of an emotional response to the facial expression can be activated. Light dotted projections indicate the subcortical dual route from superior colliculus, SC, and pulvinar to limbic areas (not included in the current model).

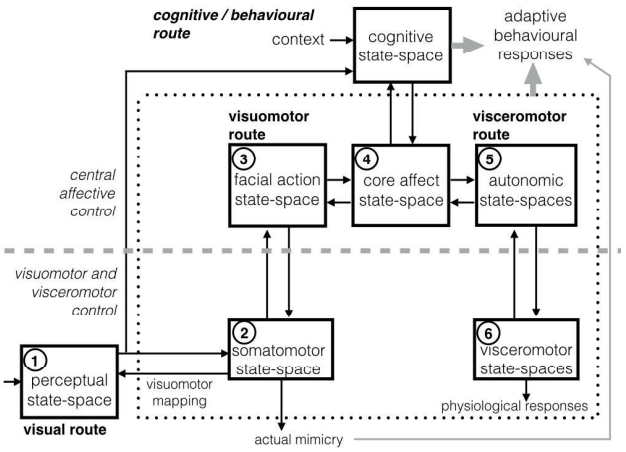


Fig. 3. The functional architecture for face-based emotion understanding. It provides an high-level decomposition of the neural architecture outlined in Fig. 2 into major components together with a characterisation of the interaction of the components. Numbered components are those considered in this study and numbering follows their presentation in text. To keep to the neural architecture, ①→②→③→④ and ⑥→⑤→④ one-head arrows indicate "forward", bottom-up projections; ①←②←③←④ and ⑥←⑤←④ denote "backward", top-down projections. The visual system for dynamic facial expression perception interacts with an extended system, which involves the emotion system (dotted box) and high level cognitive/conceptual processes. Interaction is regulated by the visuomotor mediation of a component for action perception. The latter transforms the sensory information of observed facial actions into the observer's own somatomotor representations. The activation of the visuomotor route in turn triggers visceromotor reactions through the mediation of the core affect state-space. From there the loop of simulation-based dynamics involving all components unfolds to support the whole process. The dashed grey line distinguishes between the hierarchical levels of control.

where $\mathbf{x}(t)$ is the hidden state, $\mathbf{c}(t)$ the input signal, $\mathbf{x}_{BU}(t)$ the observation, $\boldsymbol{\epsilon}_{\mathbf{x}}(t)$ and $\boldsymbol{\epsilon}_{\mathbf{x_{BU}}}(t)$ the system and observation noise at time t, while $\mathbf{f}$ and $\mathbf{g}$ are the transition and observation models, respectively. Input signal $\mathbf{c}(t)$ represents, in general, the system control vector, which can be shaped in many ways; for example, as a function of either TD or BU signals (e.g. to introduce feedback). A TD signal can also represent an exogenous input, e.g., labelling sequence provided along supervised learning. Equations 1 and 2 allow to recursively estimate the hidden state $x(t)$ at any level of the hierarchy and to convert the inferred hidden state into predictions about future observations $\mathbf{x_{BU}}(t)$.

The dynamics of each component defines a trajectory over time, say $\{\mathbf{x}(t), 1 \leq t \leq T\}$, within a manifold or state-space. Taking stock of the above discussion and by abstracting from neurobiological details, the proposed functional model relies on the following state-space representation (numeration is in accordance with Fig. 3): a *perceptual state-space* (1) resulting from visual facial cue processing (functionally accounting for the joint activity of early visual cortices, OFA, FFA, and STS); (2) a *somatomotor state-space* of the internal motor representation of facial dynamics (STS, IPL, and motor areas); (3) a *facial action state-space*, where trajectories encode facial actions (IFG /VPMC); (4) a *core affect state-space*, embedding internal V/A trajectories (Amy, AIns, OFC); (5) an *autonomic state-space* (PIns, AIns), encoding visceromotor actions; (6) a *visceromotor state-space*, representation of physiological response generation (HYP, brain stem nuclei, ANS).

At the onset of the process (presentation of stimuli), a visuomotor mapping links perceived facial cue dynamics to the internal motor dynamics. The latter is controlled by facial motor parameters, a parameter trajectory representing a facial action. Facial actions activate the core affect, which in turn triggers autonomic actions and eventually visceromotor responses. Thereafter, resonance between the expresser and the observer is established, and simulation-based dynamics unfolds to support the process, jointly involving all the introduced components.

III. Theoretical Model

To set up the model in a Bayesian framework, it is convenient to note that the dynamic stochastic process defined by Eqs. 1 and 2 can be mapped to the probabilistic generative model described via the state/observation sampling $\widetilde{\mathbf{x}}(t) \sim P(\mathbf{x}(t) \mid \mathbf{x}(t-1), \mathbf{c}(t))$ and $\widetilde{\mathbf{x}}_{BU}(t) \sim P(\mathbf{x}_{BU}(t) \mid \widetilde{\mathbf{x}}(t), \mathbf{c}(t))$, where P is a distribution associated to the probability measure over latent trajectories [31]. State variables of interest can be devised as follows.

Assume a face-to-face interaction between two agents, an *expresser* ($\mathcal{E}$) and an *observer* ($\mathcal{O}$), that share the common

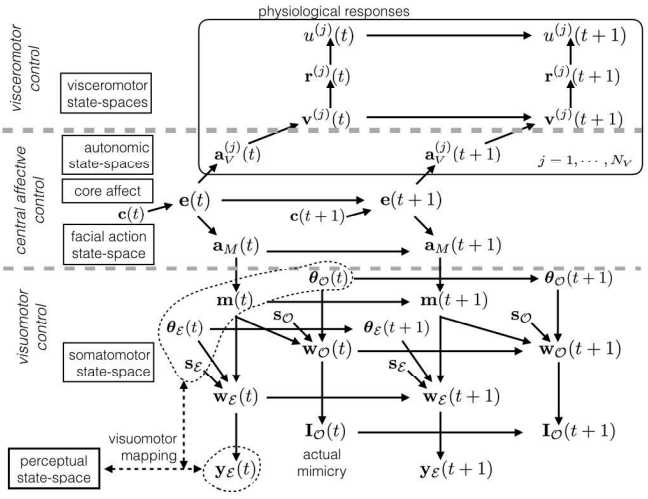


Fig. 4. The theoretical model at a glance. The conditional dependencies (arrows) among the core components (cfr. Fig. 3) and their generative dynamics represented as a dynamic PGM over the time slice $(t, t+1)$. Dashed grey lines emphasise the hierarchical levels of control of the dynamics.

model underlying the state-spaces and related dynamics introduced in Section II (Fig. 3). The inputs to the model are provided by the facial display of the expresser, in the form of a time-varying random variable (RV) $\mathbf{I}_{\mathcal{E}}(t)$, and a set of suitable control variables $\mathbf{c}(t)$, which will be used in the learning stage to account for ground-truth valence/arousal pairs. The outputs are the facial display or actual mimicry $\mathbf{I}_{\mathcal{O}}(t)$ of the observer and a set of physiological responses $\mathcal{U}(t)_{\mathcal{O}} = \{u^{(j)}(t)\}_{j=1}^{N_V}$, namely a number N_V of measurable physiological signals $u^{(j)}(t)$ originated along the observer's autonomic activity.

Then, use the following RVs related to the different state spaces. Core affect and action state-spaces:

- $\mathbf{e}(t)$ spans the latent core affect state-space, at time t;
- $\mathbf{a}_M(t)$ denotes the somatomotor action;
- $\mathbf{a}_V^{(j)}(t)$ represents the j-th visceromotor action.

Somatomotor state-space:

- $\mathbf{m}(t)$ captures the facial deformation due to muscle action "shared" between the two agents;
- $\boldsymbol{\theta}(t)$ represents the head pose parameters; in principle they need not to be shared between the agents, thus $\boldsymbol{\theta}(t) = \{\boldsymbol{\theta}_{\mathcal{O}}(t), \boldsymbol{\theta}_{\mathcal{E}}(t)\}$, though here in practice we will use the same set of pose parameters for both;
- $\mathbf{w}(t)$ accounts for somatomotor state-space dynamics;
- $\mathbf{s}_{\mathcal{I}}$ is a fixed set of static parameters encoding the biometric characteristics of each individual $\mathcal{I} \in \{\mathcal{E}, \mathcal{O}\}$, namely $\{\mathbf{s}_{\mathcal{E}}, \mathbf{s}_{\mathcal{O}}\}$; expresser's $\mathbf{s}_{\mathcal{E}}$ are inferred by the observer at the onset of the interaction, while observer's parameters are given;
- $\mathbf{y}_{\mathcal{E}}(t)$ predicts the visual facial cues of the expresser.

Visceromotor state-space:

- $\mathbf{v}^{(j)}(t)$ spans the j-th visceromotor state space;
- $\mathbf{r}^{(j)}(t)$ represents the internal observation, or feature vector, of the j-th physiological response $u^{(j)}(t)$.

The above RVs specify the main components of the functional model presented in Fig. 3. Their probabilistic conditional dependencies can be formalised in the directed Probabilistic Graphical Model (PGM, [32]), say $\mathcal{G}$, presented in Fig. 4.

If we denote $\mathcal{M}(t), \mathcal{S}(t)$ the time-varying state ensembles of the visuomotor and visceromotor routes, these are subgraphs of model $\mathcal{G}$. Then, given a core affect state $\mathbf{e}(t+1) = \mathbf{e}^*(t+1)$, the dynamics of the joint distribution represented by $\mathcal{G}$ can be factorised as $P(\mathcal{M}(t+1), \mathcal{S}(t+1) \mid \mathcal{M}(t), \mathcal{S}(t), \mathbf{e}^*(t+1)) = P(\mathcal{M}(t+1) \mid \mathcal{M}(t), \mathbf{e}^*(t+1)) \times P(\mathcal{S}(t+1) \mid \mathcal{S}(t), \mathbf{e}^*(t+1))$, that is $\mathcal{M}(t)$ and $\mathcal{S}(t)$ are conditionally independent given the current core affect state, i.e., $(\mathcal{M}(t) \perp \mathcal{S}(t) \mid \mathbf{e}(t))$ (cfr., Koller [32], Theorem 3.1), while the other RVs are marginally independent, i.e., $\mathcal{M}(t+1) \perp \mathcal{S}(t)$ and $\mathcal{S}(t+1) \perp \mathcal{M}(t)$. Since each hidden, latent state-space variable partitions the graph into independent subgraphs, such evolution can be recursively applied at any level.

The dynamics builds upon the backward/forward hierarchical information exchange between levels outlined in Fig. 3 and detailed as follows.

A. Dynamics of affect enactment

The dynamics relies on a nested, double simulation loop. The outer loop is a perception-action cycle based on the current observation of expresser's facial display. The generative properties of the model are exploited to hierarchically predict core affect states and in turn visuomotor and visceromotor states that will eventually determine facial mimicry and physiological responses. To such end, the inner loop of measurements and predictions within the central affect state-spaces implements a kind of "as if" internal simulation [10] to jointly optimise variables $\widetilde{\mathbf{m}}, \widetilde{\mathbf{v}}$. At the end of the inner loop, optimal $\widetilde{\mathbf{m}}^*, \widetilde{\mathbf{v}}^*$ are provided as top-down controls to motor and visceromotor state-spaces in the outer loop. Such dynamics, relying on prediction and measurement steps, is outlined in Algorithm 1.

Evolution of central affective states. The construction of the latent affect space model grounds in the probabilistic dependencies that relate visuomotor and visceromotor components to the core affect, namely $\mathbf{e} \rightarrow \mathbf{a}_M \rightarrow \mathbf{m}$ and $\mathbf{e} \rightarrow \mathbf{a}_V^{(j)} \rightarrow \mathbf{v}^{(j)}$, respectively. The dynamics can be summarised as sampling a time dependent affect state from the latent core affect space

$$\widetilde{\mathbf{e}}(t+1) \sim P(\mathbf{e}(t+1) \mid \mathbf{e}(t), \mathbf{c}(t+1)); \tag{3}$$

then, due to local independency, sampling in parallel somatic and visceromotor actions

$$\widetilde{\mathbf{a}}_M(t+1) \sim P(\mathbf{a}_M(t+1) \mid \mathbf{a}_M(t), \widetilde{\mathbf{e}}(t+1)), \tag{4}$$

$$\widetilde{\mathbf{a}}_V^{(j)}(t+1) \sim P(\mathbf{a}_V^{(j)}(t+1) \mid \mathbf{a}_V^{(j)}(t), \widetilde{\mathbf{e}}(t+1)), \tag{5}$$

where $j = 1, \ldots, N_V$. This way, a core affect trajectory $\{\mathbf{e}(t)\}_{t=1}^T$ generates specific action trajectories $\{\mathbf{a}_M(t)\}_{t=1}^T$ and $\{\mathbf{a}_V^{(j)}(t)\}_{t=1}^T$ to be taken in the somatomotor and visceromotor routes. Note that the control variable $\mathbf{c}(t)$ in Eq. 3 may represent either exogenous inputs, when given (e.g. V/A pairs at the learning stage) or bottom-up feedbacks $\widehat{\mathbf{a}}_M(t), \{\widehat{\mathbf{a}}_V^{(j)}(t)\}$, that can be inferred by using posterior distributions $P(\mathbf{a}_M(t) \mid \widehat{\mathbf{m}}(t))$ and $P(\{\mathbf{a}_V^{(j)}(t)\} \mid \{\widehat{\mathbf{v}}^{(j)}(t)\})$, respectively.

Algorithm 1 Simulation-based dynamics

Input: - Dynamic sequence of expresser's facial display $\mathbf{l}_{\mathcal{E}}(t)$ at times denoted by $t = 1, 2, 3, \ldots$ and corresponding to multiple of frame interval Δt
- suitable initialisation of state-space parameters

Output: Predictions $\widetilde{\mathbf{I}}_{\mathcal{O}}(t')|_{t'>t}$ and $\{\widetilde{u}^{(j)}(t')|_{t'>t}\}$ (Eq. 16)

$t \leftarrow 1$
while in interaction **do**
  {*Visuomotor mapping*}
  Given $\mathbf{I}_{\mathcal{E}}(t)$ measure $\widehat{\mathbf{l}}_{\mathcal{E}}(t), \widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t)$ (Eqs. 12, 13)
  if $t = 1$ **then**
    Measure $\mathbf{s}_{\mathcal{E}}$ and scale parameters
  end if
  $k \leftarrow 0$; $\tau_k \leftarrow t$
  {*Internal "as if" simulation*}
  repeat
    {*Forward/bottom-up measure step*}
    $\widehat{\mathbf{a}}_M(\tau_k) \sim P(\mathbf{a}_M(\tau_k) \mid \widehat{\mathbf{m}}(\tau_k))$
    $\{\widehat{\mathbf{a}}_V^{(j)}(\tau_k)\} \sim P(\{\mathbf{a}_V^{(j)}(\tau_k)\} \mid \{\widehat{\mathbf{v}}^{(j)}(\tau_k)\})$;
    $\widehat{\mathbf{e}}(\tau_k) \sim P(\mathbf{e}(\tau_k) \mid \widehat{\mathbf{a}}_M(\tau_k), \{\widehat{\mathbf{a}}_V^{(j)}(\tau)\})$
    {*Backward/top-down core affect step*}
    $\mathbf{c}(\tau_k) \leftarrow \widehat{\mathbf{e}}(\tau_k)$
    Predict $\widetilde{\mathbf{e}}(\tau_{k+1})$ (Eq. 3)
    {*Backward/top-down visuomotor step*}
    Predict $\widetilde{\mathbf{a}}_M(\tau_{k+1})$, then $\widetilde{\mathbf{m}}^*(\tau_{k+1})$ (Eqs. 4, 6);
    {*Backward/top-down visceromotor step*}
    Predict $\widetilde{\mathbf{a}}_V(\tau_{k+1})$, then $\{\widetilde{\mathbf{v}}^{(j),*}(\tau_{k+1})\}$ (Eqs. 5, 14)
    Save pred. $\widetilde{\mathbf{m}}^*, \widetilde{\mathbf{v}}^*$ as the current observed $\widehat{\mathbf{m}}, \widehat{\mathbf{v}}$
    $k \leftarrow k + 1$; $\tau_k \leftarrow t + k\delta t$
  until $\tau_k - t \leq \Delta t$
  Use predicted states as the current ones
  {*Backward/top-down visuomotor step*}
  Predict $\widetilde{\boldsymbol{\theta}}(t+1)$
  Predict $\widetilde{\mathbf{w}}(t+1)$ (Eq. 8), then $\widetilde{\mathbf{y}}_{\mathcal{E}}(t+1)$ (Eq. 9);
  {*Backward/top-down visceromotor step*}
  Use predicted $\{\widetilde{\mathbf{v}}^{(j),*}\}$ as control parameters and predict actual $\{\widetilde{\mathbf{v}}^{(j)}(t+1)\}$ (Eq. 14)
  Predict $\{\widetilde{\mathbf{r}}^{(j)}(t+1)\}$ (Eq. 15)
  {*Mimicry and physiological responses*};
  Predict $\widetilde{\mathbf{I}}_{\mathcal{O}}(t+1)$
  Predict $\{\widetilde{u}^{(j)}(t+1)\}$ (Eq. 16)
  $t \leftarrow t + 1$;
end while

The somatic visuomotor route. A trajectory $\{\mathbf{a}_M(t)\}_{t=1}^T$ in the latent space of facial actions is used to sample a sequence of facial motor control parameters $\mathbf{m}(t)$. These tune the facial action unfolding in the motor state-space spanned by $\mathbf{w}(t)$, namely the observer's internal representation of the face. To such end, we assume a parametric representation $\mathbf{w}(t) = \mathbf{w}(\mathbf{s}_{\mathcal{I}}, \mathbf{m}(t), \boldsymbol{\theta}(t))$. Parameters $\mathbf{s}_{\mathcal{I}}$ encode the invariant facial biometric traits of the agent $\mathcal{I} = \{\mathcal{O}, \mathcal{E}\}$: thus, $\mathbf{s}_{\mathcal{O}}$ are used for building the observer's inner representation of his own face, whilst $\mathbf{s}_{\mathcal{E}}$ are adopted for predicting expresser's facial action. The motor parameters $\mathbf{m}(t)$ and $\boldsymbol{\theta}(t)$ control the facial deformation due to muscle action and the head pose, respectively, and are "shared" between the two agents. They are inferred/perceived by $\mathcal{O}$ looking at $\mathcal{E}$ and used as observer's own parameters, a process which we address as *visuomotor mapping*. Assume that an estimate of face deformation and global head motion parameters, $\widehat{\mathbf{m}}(t)$ and $\widehat{\boldsymbol{\theta}}(t)$ respectively, is available at time t after the perceptual stage. The somatomotor space will be characterised by the following dynamics. First, sample facial action control parameters:

$$\widetilde{\mathbf{m}}(t+1) \sim P(\mathbf{m}(t+1) \mid \widehat{\mathbf{m}}(t), \widetilde{\mathbf{a}}_M(t+1)), \tag{6}$$

$$\widetilde{\boldsymbol{\theta}}(t+1) \sim P(\boldsymbol{\theta}(t+1) \mid \hat{\boldsymbol{\theta}}(t)). \tag{7}$$

By using sampled control parameters, set $\mathbf{w}(t+1) = \mathbf{w}(\widetilde{\mathbf{m}}(t+1), \widetilde{\boldsymbol{\theta}}(t+1), \mathbf{s}_{\mathcal{E}})$, predict the facial configuration of $\mathcal{E}$,

$$\widetilde{\mathbf{w}}(t+1) \sim P(\mathbf{w}(t+1) \mid \mathbf{w}(t), \widetilde{\mathbf{m}}(t+1), \widetilde{\boldsymbol{\theta}}(t+1)), \tag{8}$$

and sample a predicted observation of $\mathcal{E}$'s facial cues (landmarks)

$$\widetilde{\mathbf{y}}_{\mathcal{E}}(t+1) \sim P(\mathbf{y}_{\mathcal{E}}(t+1) \mid \mathcal{T}(\widetilde{\mathbf{w}}(t+1))), \tag{9}$$

where $\mathcal{T}(\cdot)$ is a projection of the internal face model in the $2D$ visual space where (retinal) sensing of the expresser occurs. Facial mimicry is obtained by setting $\mathbf{w}(t+1) = \mathbf{w}(\widetilde{\mathbf{m}}(t+1), \widetilde{\boldsymbol{\theta}}(t+1), \mathbf{s}_{\mathcal{O}})$, using Eq. 8, and generating $\mathcal{O}$'s facial expression:

$$\widetilde{\mathbf{I}}_{\mathcal{O}}(t+1) \sim P(\mathbf{I}_{\mathcal{O}}(t+1) \mid \mathbf{w}(t+1), \mathbf{I}_{\mathcal{O}}(t)) \tag{10}$$

The observer's perception of the expresser. The goal for $\mathcal{O}$ is to estimate: i) $\mathcal{E}$'s actual facial landmarks $\mathbf{l}_{\mathcal{E}}(t)$, conditioned on the set of facial patch feature responses $\mathbf{X}_{\mathcal{E}}(t)$ computed on frame $\mathbf{I}_{\mathcal{E}}(t)$, and on the currently predicted facial shape state $\widetilde{\mathbf{w}}_{\mathcal{E}}(t)$; ii) the hidden motor control parameters $\boldsymbol{\theta}(t)$ (facial pose) and $\mathbf{m}(t)$ (facial deformation) that most likely modulate the visible facial configuration of the expresser. Inference relies on the joint posterior

$$P(\mathbf{l}_{\mathcal{E}}(t), \boldsymbol{\theta}(t), \mathbf{m}(t) \mid \widetilde{\mathbf{y}}_{\mathcal{E}}(t), \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)) = P(\boldsymbol{\theta}(t), \mathbf{m}(t) \mid \mathbf{l}_{\mathcal{E}}(t), \widetilde{\mathbf{y}}_{\mathcal{E}}(t)) \times P(\mathbf{l}_{\mathcal{E}}(t) \mid \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)). \tag{11}$$

The first factor on the r.h.s substantiates the visuomotor mapping; the second factor supports the visual processing of facial landmarks. Hence, the perception stage boils down to the following.

1) Compute the most likely configuration of actual landmarks:

$$\widehat{\mathbf{l}}_{\mathcal{E}}(t) = \arg\max P(\mathbf{l}_{\mathcal{E}}(t) \mid \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)); \tag{12}$$

2) Back-project into the expresser's image space the current predicted observer's facial state $\widetilde{\mathbf{y}}_{\mathcal{E}}(t)) = \mathbf{w}(\widetilde{\mathbf{m}}(t), \widetilde{\boldsymbol{\theta}}(t), \mathbf{s}_{\mathcal{E}})$ for estimating control parameters that best explain observed landmarks $\widehat{\mathbf{l}}_{\mathcal{E}}(t)$:

$$(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t)) = \arg\max_{\mathbf{m}, \boldsymbol{\theta}} P(\widehat{\mathbf{l}}_{\mathcal{E}}(t) \mid \widetilde{\mathbf{y}}_{\mathcal{E}}(t)). \tag{13}$$

The visceromotor route. In this case, we deal with a number N_V of measurable physiological signals $\mathcal{U}(t)_{\mathcal{O}} = \{u^{(j)}(t)\}_{j=1}^{N_V}$. Typically, $u^{(j)}(\cdot)$ are 1-D time-series, that can be assumed to be the realisation of N_V independent stochastic

processes. Prediction and observation steps can be performed for the $j = 1, \cdots, N_V$ spaces as follows:

- predict the autonomic visceromotor state conditioned on the current action $\widetilde{\mathbf{a}}_V^{(j)}$

$$\widetilde{\mathbf{v}}^{(j)}(t+1) \sim P(\mathbf{v}^{(j)}(t+1) \mid \mathbf{v}^{(j)}(t), \widetilde{\mathbf{a}}_V^{(j)}(t+1)), \quad (14)$$

- predict the observation of physiological features

$$\widetilde{\mathbf{r}}^{(j)}(t+1) \sim P(\mathbf{r}^{(j)}(t+1) \mid \mathbf{v}^{(j)}(t+1)), \quad (15)$$

- predict a physiological response

$$\widetilde{u}^{(j)}(t+1) \sim P(u^{(j)}(t+1) \mid \widetilde{\mathbf{r}}^{(j)}(t+1)). \quad (16)$$

IV. Implementation model

Exact Bayesian inference of states and parameters on complex models such as the one we are dealing with is computationally intractable [33] (but see [34] for a general discussion). We thus exploit the PGM compositionality to devise apt approximations for the subgraphs of interest, which correspond to the main components of the underlying functional architecture. Meanwhile, we give precise form to the assumption that the involved state-spaces are associated to input-output state-space models.

Multimodal action and core affect latent spaces. The requirements here are: i) to devise an effective and efficient nonlinear mapping such that trajectories of similar control parameters and, in turn, of actions, are placed nearby in the core affect space whilst dissimilar trajectories are far away; ii) to functionally account for the entanglement, at the neural level, of somotamotor and visceromotor components. To meet such requirement, we use a Deep Gaussian Process (GP) approach [35].

A Deep GP is a deep directed graphical model that consists of multiple layers of latent variables and employs GPs to govern the mapping between consecutive layers. Precisely, a single layer of the deep GP is a Gaussian process latent variable model (GP-LVM). Denote: $\{\mathbf{Z}^{(h)}\}_{h=2}^{H}$ the layers of latent, hidden variables where $\mathbf{Z}^{(h)} \in \mathbb{R}^{N \times Q^{(h)}}$, $Q^{(h)}$ being the dimension of the layer at level h and N the sample size; $\mathbf{Y} \in \mathbb{R}^{N \times D}$ the down-most layer, being D the input space dimension. Each hidden layer can be modelled as a GP-LVM employing a product of $Q^{(h)}$ independent GPs as prior for the latent mapping $\mathbf{F}^{(h)} = \{f_q^{(h)}\}_{q=1}^{Q^{(h)}}$, which component $f_{q,n}^{(h)} = f_q^{(h)}(\mathbf{z}_n^{(h)})$ represents the n-th sampled value ($1 \leq n \leq N$). Thus, $f_q^{(h)} \sim \mathcal{GP}(0, k^{(h)}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(h)}))$, where $k^{(h)}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(h)})$ is the kernel function at level h. For the down-most layer, $\mathbf{F}^{(1)} = \{f_d^{(1)}\}_{d=1}^{D}$, and $f_{d,n}^{(1)} = f_d^{(1)}(\mathbf{z}_n^{(1)})$. The generative process from the upper-most to the down-most layer is given by the following state-space model:

$$z_{q,n}^{(h-1)} = f_{q,n}^{(h)} + \epsilon_{q,n}^{(h)}, \quad q = 1, \ldots, Q^{(h)}, \quad (17)$$

$$y_{d,n} = f_{d,n}^{(1)} + \epsilon_{d,n}^{(1)}, \quad d = 1, \ldots, D. \quad (18)$$

where $\epsilon_{q,n}^{(h)} \sim \mathcal{N}(0, (\sigma_{q,n}^{(h)})^2)$ and $\epsilon_{d,n}^{(1)} \sim \mathcal{N}(0, (\sigma_{d,n}^{(1)})^2)$. Clearly, the size of each latent layer is crucial but does not need to be a priori defined. It has been shown [35] that it is possible to define automatic relevance determination (ARD) covariance functions for the GPs, $k^{(h)}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(h)}) = \sigma_{\text{ARD}}^2 \exp\{-\frac{1}{2}\sum_{q=1}^{Q^{(h)}} w_q^{(h)}(z_{q,i}^{(h)} - z_{q,j}^{(h)})^2\}$, such that a different weight $w_q^{(h)}$ is assumed for each latent dimension. This can be exploited at the training stage in order to prune irrelevant dimensions by driving their corresponding weight to zero. Nonlinearities introduced by such a covariance function are treated via non-standard variational inference methods that allow to define analytically an approximate Bayesian training procedure (see [35] for details).

Crucially, such deep structure can be naturally extended "horizontally" by segmenting each layer into different partitions. Thus, the latent space at level h is partitioned into $\pi^{(h)}$ conditionally independent subsets, matching exactly the conditional independence statements assumed from the beginning to design the PGM representation provided in Fig. 4. Eventually, it allows to handle in a principled and efficient way the multimodal nature of visual cues and of the different physiological signals. This can be achieved by defining the down-most layer of the deep GP as a $N \times D$ matrix $\mathbf{Y} = [\mathbf{m} \mid \mathbf{v}^{(1)} \mid \mathbf{v}^{(2)} \mid \cdots]$, D being the sum of the dimensions of state/control parameters of each modality.

It is worth noting that the deep GP has the expressive power that indeed we need to map the trajectories taking place in the core affect state-space, onto trajectories at the motor state-space level. Because of the recursive warping of latent variables through the core affect $\rightarrow$ action $\rightarrow$ motor hierarchy, deep GP allows for modeling non-stationarities and cumbersome non-parametric functional properties [35].

In the supervised learning scenario, which is the one addressed here, the inputs of the top hidden layer $\mathbf{Z}^{(H)}$ is observed, namely is the valence/arousal time sequence or trajectory $\mathbf{e}(t)$ provided at the learning stage. When the latent space is set up, then new estimated controls $\widehat{\mathbf{m}}, \{\widehat{\mathbf{v}}^{(j)}\}$ can be stochastically backprojected through the latent space layers up to the core affect state-space, namely, $\widehat{\mathbf{a}}_M(t) \sim P(\mathbf{a}_M(t) \mid \widehat{\mathbf{m}}(t))$, $\{\widehat{\mathbf{a}}_V^{(j)}(t)\} \sim P(\{\mathbf{a}_V^{(j)}(t)\} \mid \{\widehat{\mathbf{v}}^{(j)}(t)\})$, allowing the estimate $\widehat{\mathbf{e}}(t) \sim P(\mathbf{e}(t) \mid \widehat{\mathbf{a}}_M(t), \{\widehat{\mathbf{a}}_V^{(j)}(t)\})$. This is achieved by using variational posteriors available from the model learning stage [36], but, different from [36], by using a bottom up sampling-like approach, in order not to disregard the uncertainty predicted at each time. As to pose evolution formalised in Eq. 7, we simply put $P(\boldsymbol{\theta}(t+1) \mid \hat{\boldsymbol{\theta}}(t)) = \delta(\boldsymbol{\theta}(t+1), \hat{\boldsymbol{\theta}}(t))$, hence we straightforwardly exploit the inferred $\hat{\boldsymbol{\theta}}(t)$.

Somatic motor space and visuomotor mapping. We instantiate $\mathbf{w}(t)$ as a 3D deformable shape model. Thus, $\mathbf{w}(t)$ is a vector of vertices such that the evolution of the face model at time t is represented by the ensemble of vertex state vectors $\mathbf{w}_i(t) = [X_i(t), Y_i(t), Z_i(t)]^T$. To give a precise form to the parametric face model $\mathbf{w}(t) = \mathbf{w}(\boldsymbol{\Theta}(t))$, being $\boldsymbol{\Theta}(t)$ the vector of all involved parameters, we exploit the 3D face model Candide-3 [37]. This is a 3D deformable wireframe model consisting of approximately 113 vertices $\mathbf{w}_i$ and 184 triangles (cfr. Fig. 5). The face shape $\mathbf{w}_{\mathcal{I}}$ can be generated from the standard mean shape $\overline{\mathbf{w}}$, which is deformed by both individual biometric characteristics and the facial action

(expression) performed at time t.

Denote $\mathbf{dW}_i^S$ and $\mathbf{dW}_i^M$, biometric and facial action-based deformations. Precisely, $\mathbf{dW}_i^S = \left[\mathbf{dw}_{i,1}^S, \cdots, \mathbf{dw}_{i,N_s}^S\right]$ and $\mathbf{dW}_i^M = \left[\mathbf{dw}_{i,1}^M, \cdots, \mathbf{dw}_{i,N_m}^M\right]$ are constant $3 \times N_s$ and $3 \times N_m$ matrices, respectively, where each 3×1 vector $\mathbf{dw}_{i,j}^S$ and $\mathbf{dw}_{i,k}^M$ represents the single Shape Unit (SU) and Action Unit (AU) deformation at vertex i, respectively. Here, $N_s = 14$ and $N_m = 11$. The columns of $\mathbf{dW}_i^S$ are vectors of control point displacements due to biometric traits of the individual (mouth width, eye distance, etc.). The columns of $\mathbf{dW}_i^M$ encode vectors of point displacements, each vector corresponding to AUs related to Ekman's FACS (Facial Action Coding System [38]); these describe the change in face geometry when the corresponding AU is enabled due to the motor activation of facial muscles. The effect of the SUs and AUs is controlled via the shape and motor/action parameter vectors $\mathbf{s}_\mathcal{I} = [s_1, \cdots, s_{N_s}]^T, \mathbf{m}_\mathcal{I} = [m_1, \cdots, m_{N_m}]^T$.

Under the face-to-face interaction assumption, the shape model dynamics is that of a deformable (i.e., not rigid) body and by assuming small rotations, it can be shown that at any time t, facial movement will locally move a 3D vertex $\mathbf{w}_i$ to position $\mathbf{w}_i(t + \delta t) = \mathbf{w}_i(t) + \mathbf{dw}_i$ (for unitary time step $\delta t = 1$, without loss of generality) according to the law:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \mathbf{R}(t)\mathbf{w}_i(t) + \mathbf{dW}_i^S\mathbf{s}_\mathcal{I} + \mathbf{dW}_i^M\mathbf{m}(t) + \mathbf{t}(t), \quad (19)$$

where $\mathbf{R}$ and $\mathbf{t}$ represent the rotation matrix $\mathbf{R} = \mathbf{R}(\boldsymbol{\omega})$, $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$, and the translation vector, respectively, that is the global rigid motion constrained by cranial pose dynamics. Eq. 19, applied to all vertices i, represents the state equation of the 3D face model evolving in time, i.e. the forward model, which is used in the action stage (Eq. 8). Its dynamic control parameters are the pose parameters $\boldsymbol{\theta}(t) = (\mathbf{R}(t), \mathbf{t}(t))$ and deformation parameters $\mathbf{m}(t)$. Individual biometric control parameters $\mathbf{s}_\mathcal{I}$ are considered fixed along the interaction.

Recall that, in order to estimate parameters $(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t))$, given the computed landmarks $\widehat{\mathbf{y}}_\mathcal{E}(t)$ (visuomotor mapping), the second step of the perceptual stage (Eq. 13) relies on projecting into the image space the predicted facial configuration of the expresser, namely $\widehat{\mathbf{w}}_\mathcal{E}$. The latter is obtained by sampling in the action stage the current state of the face model $\widetilde{\mathbf{w}}(\mathbf{m}, \boldsymbol{\theta})$, and assigning expresser's identity parameters, i.e. $\widetilde{\mathbf{w}}_\mathcal{E}(t)) = \widetilde{\mathbf{w}}(\mathbf{m}(t), \boldsymbol{\theta}(t), \mathbf{s}_\mathcal{E})$. Then, as to projection $\mathcal{T}$ of the 3D vertices on the 2D image coordinate system, a weak perspective projection can be adopted given the small depth of the face (see, e.g., [39] for details), thus $\widetilde{\mathbf{y}}_{\mathcal{E},l} = \mathcal{T}_s(\widetilde{\mathbf{w}}_{\mathcal{E},l})$, s being the weak-perspective scale parameter, and $l \in (1, ..., L)$ the index of the extracted facial landmarks. Denote $\boldsymbol{\Theta} = s[1, \boldsymbol{\omega}^T, \mathbf{s}^T, \mathbf{m}^T, \mathbf{t}^T]^T$ the full parameter vector. Under Gaussian noise assumption, the observation equation is

$$\widehat{\mathbf{l}}_{\mathcal{E},l} = \widetilde{\mathbf{y}}_{\mathcal{E},l} + \boldsymbol{\epsilon}_{\widetilde{\mathbf{y}}}, \quad (20)$$

and parameter estimation via Eq. 13 boils down to the negative log-likelihood minimisation problem, $(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t)) = \arg\min_{\boldsymbol{\Theta}} 1/(2\sigma_{\mathbf{y}_\mathcal{E}}^2) \sum_{l=1}^L \|\widehat{\mathbf{l}}_{\mathcal{E},l} - \widetilde{\mathbf{y}}_{\mathcal{E},l}\|^2 + L\log(2\pi\sigma_{\mathbf{y}_\mathcal{E}}^2)$, which can be easily solved in closed matrix form. Note that the full parameter vector $\boldsymbol{\Theta}$ needs to be estimated only at the onset of the interaction; in subsequent steps only $(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t))$ are needed.

Perceptual state-space. We adopt the Constrained Local Neural Field (CLNF) undirected graphical model (see [40], for details). The model uses the multivariate normal distribution

$$P(\mathbf{l}_\mathcal{E}(t) \mid \mathbf{X}_\mathcal{E}(t), \mathbf{I}_\mathcal{E}(t)) = \mathcal{N}(\mathbf{l}_\mathcal{E}(t); \boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}) \quad (21)$$

to specify the landmark prediction probability at locations $\mathbf{l}_\mathcal{E}(t) = [\ell_1, \cdots \ell_L]$ given patches $\mathbf{X}_\mathcal{E}(t) = [\mathbf{x}_1(t), \cdots, \mathbf{x}_L(t)]$ selected at frame $\mathbf{I}_\mathcal{E}(t)$. The mean vector $\boldsymbol{\mu}_\mathbf{X}$ captures the feature extractor responses on patches, after preliminary face detection [40].

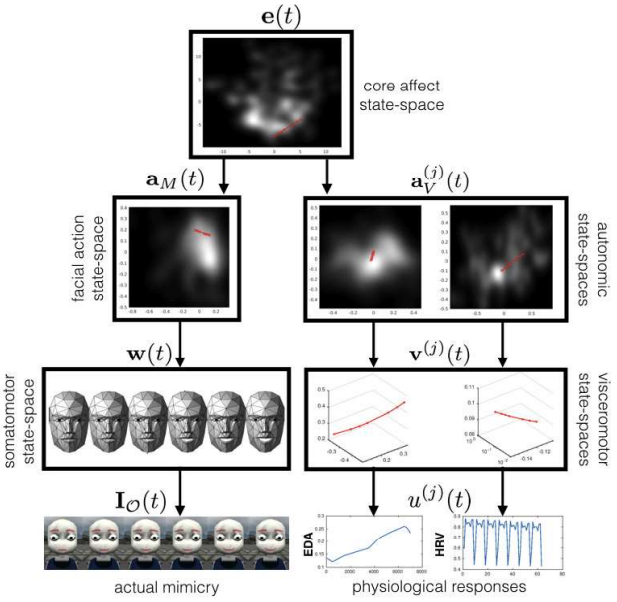


Fig. 5. Example of backward / top-down generative sampling of trajectories in the relevant state-spaces of the implementation model. Generation is induced by an initial tiny trajectory in the core affect and propagated down to produce actual facial mimicry (smile) and physiological responses (EDA and HRV). For visualisation purposes, the core affect, facial action and autonomic state-spaces are displayed as grey level images, higher brightness indicating higher probability (trajectories are derived as posterior means); somatomotor evolution $\mathbf{w}(t)$ is shown in frontal pose and before assigning biometric parameters; also, only 3 of the k dimensions of the visceromotor state-spaces spanned by $\mathbf{v}^{(j)}$ are shown. Actual facial mimicry has been obtained by exploiting the publicly available simulator of the iCub humanoid robot.

Visceromotor state-space For what concerns physiological signals $\mathcal{U}(t)_\mathcal{O} = \{u^{(j)}(t)\}_{j=1}^{N_V}$, in order to implement the predict and update step of Eq. 14, 15, we use an input driven linear dynamical system model [41]

$$\mathbf{v}^{(j)}(t+1) = \mathbf{A}\mathbf{v}^{(j)}(t) + \mathbf{B}\mathbf{c}^{(j)}(t+1) + \boldsymbol{\epsilon}_\mathbf{v}^{(j)}(t+1), \quad (22)$$

$$\mathbf{r}^{(j)}(t+1) = \mathbf{C}\mathbf{v}^{(j)}(t+1) + \boldsymbol{\epsilon}_\mathbf{r}^{(j)}(t+1), \quad (23)$$

with $\boldsymbol{\epsilon}_\mathbf{v}^{(j)}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{v}^{(j)})$, $\boldsymbol{\epsilon}_\mathbf{r}^{(j)}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{r}^{(j)})$, and, assuming $\mathbf{v}$ is a k-dimensional state vector, $\mathbf{A}$ is the $(k \times k)$ state dynamics matrix, $\mathbf{B}$ is the $(k \times d)$ input-to-state matrix, and $\mathbf{C}$ is the $(p \times k)$ observation matrix. In current implementation, the p-dimensional feature vector $\mathbf{r}^{(j)}(t)$ is obtained via the wavelet transform of the physiological response $u^{(j)}(t)$ (further details in Experimental setting, Section V).

The input driven model is able to incorporate a displacement for the hidden state dynamics $\mathbf{c}^{(j)}(t+1) = (\mathbf{A}\mathbf{v}^{(j)}(t+1) - \widetilde{\mathbf{v}}^{(j)}(t+1))$, where $\widetilde{\mathbf{v}}^{(j)}(t+1) \sim P(\mathbf{v}^{(j)}(t+1) \mid \widetilde{\mathbf{a}}_V^{(j)}(t+1))$ is the current emission predicted by the sampled visceromotor action; $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{\Sigma}_{\mathbf{v}}^{(j)}, \mathbf{\Sigma}_{\mathbf{r}}^{(j)})$ is performed via the variational Bayesian EM algorithm (see [41], for details).

At any time t the estimates of the mean $\boldsymbol{\mu}_{\mathbf{v}}^{(j)}(t)$ and covariance $\mathbf{\Sigma}_{\mathbf{v}}^{(j)}(t)$ of the hidden state $\mathbf{v}^{(j)}$ can be obtained via forward recursion (filtering), $P(\mathbf{v}^{(j)}(t) \mid \mathbf{r}_{1:t}^{(j)}, \mathbf{c}^{(j)}(t)) \propto P(\mathbf{r}^{(j)}(t) \mid \mathbf{v}^{(j)}(t)) \times \int d\mathbf{v}^{(j)}(t-1) \times P(\mathbf{v}^{(j)}(t-1) \mid \mathbf{r}_{1:t-1}^{(j)}, \mathbf{c}^{(j)}(t-1)) \times P(\mathbf{v}^{(j)}(t) \mid \mathbf{v}^{(j)}(t-1), \mathbf{c}^{(j)}(t))$. The mean $\mu_{\mathbf{v}}^{(j)}(t)$ is provided at inference time in input to the j-th autonomic state-space as the observation of the current visceromotor state $\mathbf{v}^{(j)}(t)$.

Facial mimicry and physiological signal generation The quality of facial mimicry animation is not a crucial concern in our current research work, the model being agnostic to the adopted output. Here, for visualisation purposes, Eq. 10 might be exploited to drive either a synthetic avatar animation or an actual agent (Fig. 5 shows one example using the simulator of the iCub humanoid robot).

As to physiological sampling, by inverting the wavelet transform we reconstruct the physiological responses. In this work we do not focus on the use of such signals, so we simply plot their dynamics (cfr. Figs. 6a, 6b).

V. Experiments

The focus here is on assessing the hypotheses that (i) the model suitably supports observer's affective mirroring of the expresser's affective state and that (ii) the simulation-based mechanism together with the extra autonomic activity information available during learning can improve the analysis of facial expressions when only visual information is available. Three experiments are taken into account: I) the generation of both physiological signals and visible cues given a learnt core affect state space; II) the assessment of observer's capability to reach a core affect state similar to that of the expresser, measured in terms of "internal" V/A values on the basis of his autonomic activity information and only relying on the visible facial cues of a novel expresser; III) the predictive capability of the model with respect to the results obtained by Kuppens [29], in terms of "external" V/A values, as provided by data annotation.

The observer's training stage exploits both visible facial expressions displayed by one subject serving as the teaching expresser along with subject's physiological recordings; at the testing stage, only the expresser's facial actions are provided to the observer. We recall that training entails parameter learning to set up the latent core affect space, which is performed via Variational Bayes optimisation of the Deep GP (embedding core affect, facial action and autonomic state-spaces) and the variational Kalman filter (visceromotor state-space). The somatomotor state-space relies upon an online filtering procedure based on M-L estimation (accounting for prediction error $\widehat{\mathbf{l}}_{\mathcal{E},l} - \widetilde{\mathbf{y}}_{\mathcal{E},l}$), which does not require specific learning. The early visual stage processing (landmark extraction) is trained offline since independent of the nature of facial actions performed.

Dataset. Experiments have been conducted on the public available dataset RECOLA [12] which is a multimodal corpus of spontaneous collaborative and affective interactions in French. Aiming at studying the impact of emotional feedback on teamwork quality and efficiency, 46 participants took part in the test where several multimodal data, i.e., audio, video and physiological signals (specifically, electrocardiogram ECG, and electrodermal activity EDA) were recorded continuously and synchronously. In addition to these biosignals, 6 annotators measured emotion continuously on the two dimensions of arousal and valence.

Experimental setting. Given the dataset at hand, we consider as physiological signals $u^{(j)}(t)$ the EDA, and the heart rate variability (HRV), derived from the ECG, being a good indicator of the autonomic nervous system. HRV is obtained by measuring the variation in the beat-to-beat interval in ECG. Since these kind of signals are dynamic and exhibit time-varying statistics in both the time and frequency domain, in all experiments we extract the features $\mathbf{r}^{(j)}(t)$ (Eq. 23) using discrete wavelet transform (DWT). This approach allows for the analysis of non-stationary signals at multiple scales making use of an analysis window to extract signal segments.

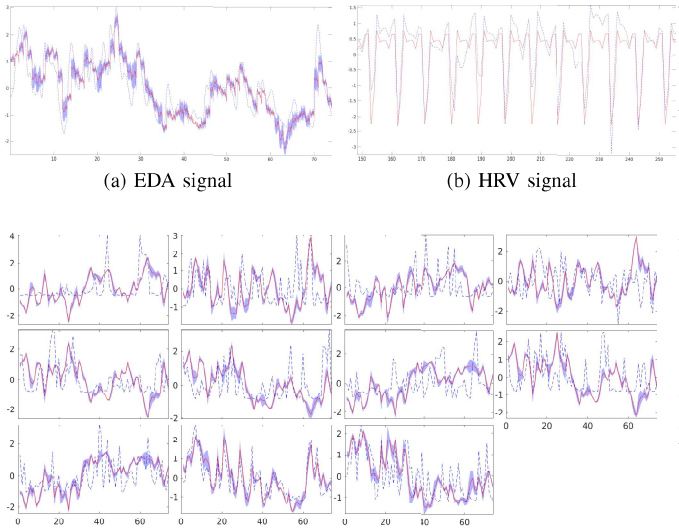
Procedurally, for EDA we perform a standard continuous decomposition analysis aiming at unbiased scores of phasic and tonic activity, thus retaining only the phasic data. In regards to the HRV signal, once denoised the ECG signal, we achieve the beat-to-beat fluctuations as RR-interval time series from ECG using standard techniques [42]. After preprocessing stage, including signal segmentation which divides the signal into 4-sec overlapped windows, we select empirically a suitable level of Daubechies 3 (db3), following the rule $L_{max} = (\log_2 N) - 1$, where N is the signal length, and we retain only the approximate coefficients as feature vector both for EDA and HRV, respectively.

During the learning stage, the resulting facial deformations $\mathbf{m}(t)$ and autonomic states $\mathbf{v}^{(j)}(t)$ are placed in the downmost layer ($h = 1$) of a 3-layer deep GP model (cfr. Eqs. 17, 18), treated as different "modalities". Such modalities share the same latent space while keeping private some of their dimensions, resulting at the $h = 2$ level in the $\mathbf{a}_M(t)$ and $\mathbf{a}_V^{(j)}(t)$ state-spaces, respectively. In this setting, the V/A annotations, obtained as the result of the Evaluator Weighted Estimator [43], perform the role of control variables $\mathbf{c}(t)$ placed as inputs of the top layer ($h = 3$). The dimensionality of such architectural setting is the following: $\mathbf{r}^{(j)}$, $p = 14$; $\mathbf{v}^{(j)}$, $k = d = 7$; $\mathbf{m}$, $N_{\mathbf{m}} = 11$; $\mathbf{m}$, $\mathbf{v}^{(EDA)}$ and $\mathbf{v}^{(HRV)}$ are preprocessed through a standard PPCA stage [44] so that the input Deep GP layer is partitioned in three subspaces, each having dimension $D = 4$; action state-spaces $\mathbf{a}_M(t)$ and $\mathbf{a}_V^{(j)}(t)$ have dimension $Q^{(2)} = 2 \times 3$, mirroring the $Q^{(3)} = 2$ dimensional core affect state-space. The latter is chosen akin to Russell's core affect [11].

Experiment I. The aim is to assess the generative capability of the system. To such end, in Fig. 6 we report the comparison between the original multimodal data and their generation via the learnt model on one subject's session randomly chosen. The experiment starts from a known V/A sequence in the

top layer, which is generatively propagated to the bottom ones. For the sake of comparison, all the generated values $\mathbf{v}^{(j)}(t)$ and $\mathbf{m}(t)$ are brought to their original 1-dimensional representation. In the former case, the feature vector $\mathbf{r}^{(j)}(t)$ is obtained through Eq. 23 and, thanks to the orthogonal property of the considered wavelet transformation, we are able to generate back the physiological responses $u^{(j)}(t)$ via inverse discrete wavelet transform (IDWT). In the latter case, generated facial deformation controls $\mathbf{m}(t)$ are simply reshaped according to the AU cardinality, $N_m = 11$, and plotted against time as the corresponding ground truth.

To give a quantitative evaluation of the system generative capability, we compute the mean square error (μ_{MSE}) and mean Pearson's correlation coefficient (μ_r), at the 0.05 significance level, between the ground truth and the predicted sequences obtained as the result of 10 sampling processes. In particular, for the EDA signal (Fig. 6a) we obtained a $\mu_{MSE} = 0.2341$ and a $\mu_r = 0.8829$ ($p < 0.001$). A similar result is achieved also for the HRV (Fig. 6b), where $\mu_{MSE} = 0.2755$ and $\mu_r = 0.8618$ ($p < 0.001$). In both cases the correlation is statistically significant. Finally, for the action units activation values (Fig. 6c), we obtained $\mu_{MSE} = 1.1011$ and $\mu_r = 0.4419$ as a result of evaluation over the 11 considered AUs, where AU_k , $k = \{2, 4, 5, 7, 9, 10, 15, 20, 23, 26, 45\}$. The correlation coefficients were respectively $r_k = \{0.43, 0.24, 0.23, 0.07, 0.34, 0.38, 0.17, 0.24, 0.49, 0.87, 0.63\}$. In all cases the p-value was under significance level (0.05), apart for AU7 and AU15 where $p = 0.57, 0.15$, respectively.



(a) EDA signal (b) HRV signal

(c) Action Unit activation values. From top-left to bottom-right: AU_k , $k = \{2, 4, 5, 7, 9, 10, 15, 20, 23, 26, 45\}$

Fig. 6. Generation of a session of physiological signals $u^{(j)}(t)$ (a), (b) and facial actions $\mathbf{m}(t)$ (c) (red) compared to the ground truth (dashed blue). In shaded light blue the 95% prediction confidence interval.

Experiment II. The goal here is to evaluate the aptness of the multimodal simulation-based mechanism to provide the observer an internal core affect dynamics, which mirrors that of novel expressers. In this case, the observer can only rely on the visual information displayed by the expresser,

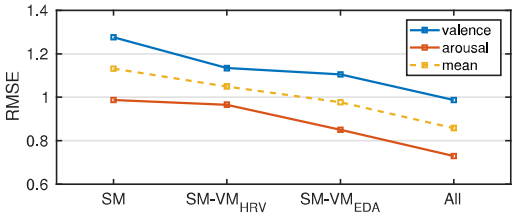


Fig. 7. Results of Experiment II, for each of the considered settings: only somatomotor route (SM), somatomotor and electrodermal routes (SM-VM$_{EDA}$), somatomotor and heart-rate routes (SM-VM$_{HR}$) and all somato- and viscero- motor routes.

whilst autonomic activity is at this point "innate", i.e. learnt previously from an expresser different from the new ones.

This experiment has been conducted adopting a one vs. all strategy. In order to assess the observer/expresser congruence of the internal core affect state, at the learning stage, a model for each subject of the RECOLA dataset is learnt; at the testing stage, one learnt model is used as the observer, while all the others are exploited as expressers interacting with the chosen observer; this procedure is repeated for all subjects in the dataset. The use of internal core affect values allows a fair comparison, since, as made clear from the beginning, the current model does not account for the cognitive level. In fact, it is worth remarking, that "external" V/A values are usually derived in psychological experiments through attribution. This can be accomplished by experts (annotators) as in RECOLA [12] or by participants themselves, who self-report felt valence and arousal according to an established protocol (e.g., [29]). Either way, affect attribution entails a cognitive step, which is in principle out of the scope of the model proposed here.

The interaction process is driven on the expresser's side by random sampling of internal core affect trajectories from which via top-down forward sampling, visible and physiological cues are generated (cfr. Fig. 5). The observer can only rely upon expresser's visible cues inferred from the facial expression. Along the interaction, observer's affective dynamics unfolds as described in Algorithm 1.

In order to assess the effectiveness of the different visible and hidden cues in determining an observer core affect state (that is predictive of that of the expresser), four different settings were adopted. In particular, we simulated the prediction process by relying 1) only on the observer's somatomotor route (SM), 2) combining SM and the electrodermal route (VM$_{EDA}$), 3) combining SM and heart-rate route (VM$_{HR}$), 4) considering SM and all available visceromotor routes VM$_{EDA}$, VM$_{HR}$. The results, shown in Fig. 7, provide evidence of the importance of physiological internal cues in the prediction of other's internal core affect state. In particular it is shown that for arousal, the root mean square error (RMSE) value between expresser's and observer's trajectories improves from 0.987 of the first setting to 0.73 of the last one. A similar behaviour can be noticed also for the valence, from 1.276 for the first setting to 0.987 for the "complete" setting.

Experiment III. We started out on the assumption that core affect dynamics is consequent on the activity of a complex,

open system subject to stochastic variability resulting from the many internal and external events, that influence it. Kuppens *et al.* [29] have remarkably shown that "external", actually observable V/A trajectories from a single subject, say $\mathbf{e}^{ext}(t)$, can be modelled with an Ornstein-Uhlenbeck (OU, [45]) state-space process. The OU process can be written in discrete time, via the Euler-Maruyama representation of the continuous time stochastic differential equations [31], as

$$\mathbf{e}(t) = \mathbf{B}\left(\mathbf{m} - \mathbf{e}(t-1)\right) + \mathbf{D}^{1/2}\boldsymbol{\epsilon}_{\mathbf{e}}(t-1), \tag{24}$$

$$\mathbf{e}^{ext}(t) = \mathbf{e}(t) + \boldsymbol{\epsilon}_{\mathbf{e}^{ext}}(t), \tag{25}$$

where $\mathbf{m}$ is a vector with two components and $\mathbf{B}$ and $\mathbf{D}$ are positive-definite 2×2 matrices. The measurement error in Eq. 25 is represented by $\boldsymbol{\epsilon}_{\mathbf{e}^{ext}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{e}^{ext}})$, i.e. random draw from a bivariate normal distribution; the state error $\boldsymbol{\epsilon}_{\mathbf{e}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ derives from the standard bivariate Wiener process. Equation 24 shows that the state dynamics is a mean-reverting process with $\mathbf{m}$ playing the role of steady state or attractor (the affective "home base" of the individual, [29]).

For the OU process we know the exact solution for the stationary covariance kernel of Eq. 24, which is the exponential kernel induced by the corresponding prior Gaussian Markov process [31], [46]. We can thus perform standard GP regression to learn kernel parameters and compute the exact posterior process and make prediction on unseen data.

Under such circumstances, we can proceed analogously to Experiment I. In this case at the learning stage we also learn the $\mathbf{e} \mapsto \mathbf{e}^{ext}$ mapping (i.e., Eq. 24 parameters). Then, at the testing stage, the observer's internal core affect trajectories are mapped into the external V/A trajectories and the latter are directly compared with the corresponding dataset annotated V/A sequences. In simple terms, we use Equation 24 as a proxy for bridging the gap between the internal core affect values and the cognitively attributed V/A values. Remarkably, the generative OU process driven by the observer's internal core affect predicts the external V/A behaviour of the expresser. The mean Pearson's correlation coefficients, over all subjects, obtained for V and A are $\mu_r^V = 0.77$ and $\mu_r^A = 0.78$, respectively (0.05 significance level, $p < 0.001$). One typical example of the model fitting expresser's external V/A trajectory is shown in Fig. 8.

VI. Discussion

Results of the first two experiments show that the modelled observer after learning, is endowed with a core affect representation that allows enactment, i.e. the activation within the observer of the emotional state underlying the facial action of the expresser, or a relevantly similar state. In turn, the generative capability of the model yields to mimicry; also, in the case of overt facial mimicry, the continuous representation entailed by the model, does not limit its expressive capability to the six typical basic emotions, the output only being constrained by the "faceware" of the addressed agent. Experiments bear witness to the fact that exploiting a multimodal representation, which is sophisticated to model the intricate nonlinear time-varied relationships between the different modalities, provides more accurate results than unimodal counterparts. This is consistent with results reported in the affective computing literature [47]. More important, it substantiates empirical findings reported in the psychological literature, that impairments in motor or limbic areas hamper the recognition of affective expressions [10]. On the other hand, the fact that inducing a concurrent motor load reduces expression understanding [48], may have concrete relevance in designing actual agents, e.g. robots. The third experiment indicates that the model can predict some remarkable result obtained, at the psychological level, by [29], namely, that empirical core affect trajectories in the arousal/valence dimensions can be captured by an OU process, reflecting the typical pattern of affective changes and fluctuations that characterises an individual [29]. This achievement should anyway be handled with caution. On the one hand, we have bridged the gap between the core affect and the cognitive levels by a straightforward regression procedure, which is a proxy for the complex processes occurring across levels. On the other hand, and more subtly, V/A values at the behavioural level, used as a ground truth, are always the empirical result of a human-based attribution process. The latter, though performed with established protocols and tools, is not granted in general to determine the very emotion felt by others [49]. Such issue is an open problem and is largely overlooked in the literature, and benchmarking procedures, e.g. those borrowed from the computer vision community [50], may turn to be inappropriate in these circumstances.

Mimicry (either covert or overt) that we have mainly addressed here is assumed to be a privileged route to affective empathy [6], [51]. However, the link to cognitive empathy, which requires higher levels of cognition [52] is left open. When coming to concrete fields of application such as robotics, dealing with such issue can be crucial [6]. In general, the integration of cognitive components compels attention since the meaning of complex expressions may be context-dependent. Yet, extending towards higher levels of cognition can be done in a principled way relying on the probabilistic framework we have set up (as discussed below). In brief, the proposed model makes a step forward in the emergent direction of going beyond detecting affect so to dynamically responding to the sensed affect, thereby closing the affective loop [47].

While addressing a flexible, embodied representation of perceived facial actions, the model lends itself well to the task of learning from few as possible examples; actually, in the experiments reported here just one expresser example, much like as in a mother/infant interaction. This can be particularly advantageous in terms of learning performance. Realms where the model could be applied (e.g. robotics, virtual agent design) it is unlikely to deal with a huge amount of training data. In Experiment II we have adopted a one vs. all strategy to marginalise possible biases (either positive or negative) depending on the chosen expresser. On the other hand in more practical exploitations, such issue should be taken into account when training for instance a model-based agent.

More generally, our work builds upon integrative knowledge from various fields, combining current insights provided by social and affective neuroscience [53] to the psychology of emotion [9]. The premise of our model, i.e., that people engage

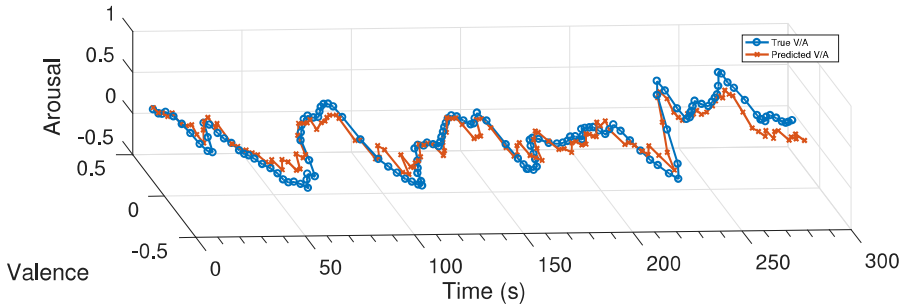


Fig. 8. V/A mapping via the OU process. The expresser's external V/A behaviour (dotted, blue) predicted by the OU process driven by the observer's core affect (cross,red)

in motor simulation of each other's emotional expressions (a.k.a., the Simulation-Theory, ST, account, but see Goldman et al. [10]) has received considerable support [9], [54]. This assumption is not without challenges [55]; however, findings at the neurobiological level are likely to reconcile the initial dispute between the ST and the Theory-Theory account (TT or naïve theory of mind) where the mindreader selects a mental state for attribution to a target based on inference; e.g., for facial expressions by *direct mapping* the representations of particular facial configurations to names for emotion states [10]. Indeed, a simulation-based mechanism (mainly concerning social detection) may be necessary in development as part of the input to later theory-building relying upon observer's beliefs, judgment and reasoning (social evaluation system) [52], [55]. What is clear is that perceiving an emotional response of another person elicits the same emotional response in ourselves, and cogently overt or covert mimicry facilitates physiological and motor feedback inducing emotion in the observer. Emotional contagion allows us to share the emotion of the person we are observing, a prerequisite of empathy [2], [51], [56], [57].

The construction of the theoretical model, in particular the PGM structure, hinges on the distributed neural architecture (Fig. 2) devised in Section II, which calls for symmetrical and intertwined activities of somatomotor and visceromotor components and shares many features with other contemporary neural models of emotional unfolding, e.g. [28], [51]. As to the implementation model, inference on states and parameters mostly relies on the Variational Bayes optimisation procedure that at heart depends on free energy maximisation [33]. A key observation here is that the optimisation of the free energy in stochastic input-output state-space dynamical systems involves the optimisation of the prediction error on states and parameters, and this also holds when such systems are layered in a hierarchical structure [58]. In such terms, the implementation model can be conceived as realising a form of hierarchical predictive coding [58], [59]. An alternative option could have been an implementation of Bayesian inference relying upon sampling, Monte Carlo based approximation [34]. The neural, realisation level plausibility [13] of either approach is currently matter of debate [34], [58].

Relations of our proposal with more technically oriented computational models can be devised by considering the areas of artificial intelligence (AI) oriented models, machine learning-based models and robotics. As to AI oriented models, work in this field does not specifically address the issue of affective facial expressions; however, by and large (see the in-depth review by Reisenzein *et al.* [3]) they do address the theoretical level, by formalising emotion theories in an implementation-independent formal language (e.g., [60]). The implementation level concerns the instantiation of emotion theories in general-purpose cognitive or agent architectures (e.g., belief-desire-intention or BDI architecture). At the heart of these proposals lies the TT approach; by and large the issue of embodiment and the neurobiological basis of affect construction. However, in a two-stage or hybrid perspective [52], [55], a viable path to address the cognitive level could be provided by probabilistic approaches [61], [62]. The latter share with us the effort in devising a theoretical model, which is well grounded in probabilistic structures at least informed by psychological theoretical constraints.

Machine learning based analysis of facial affect has fostered a wealth of approaches (for in-depth reviews, see [50] and [47]). Overall, these approaches share the assumption that understanding affective states from facial expressions can be accomplished through direct mapping, in the form of a computer vision and pattern recognition "pipeline" [50]: namely, visual feature extraction/reduction followed by classification (discrete emotion recognition) or regression (continuous affect detection). For these approaches affect detection basically boils down to a pattern recognition problem [47]. Remarkably, facial expression analysis recognition plays an important role [47], [50], [63], [64]. However, computational modelling as an attempt to develop and validate computational models of human emotion mechanisms [3] is, in general, not at stake neither at the computational theory level nor at the implementation modelling level; ST approaches are by and large overlooked. It is reasonable that, for specific applications (e.g., facial expression recognition on the Web [65] social behavioural biometrics [66], etc.), systems that can conceivably perform the task of direct mapping input data (image/video, physiological signals, etc) to affective states (discrete or continuous via classification/regression), need not to be biologically inspired, just technologically capable of producing the desired output. Yet, even in such cases this approach is not without challenges [67], whilst in other contexts the assumption that understanding affective states from facial expressions can be accomplished through the classic "pipeline" [50] is at best questionable.

A different state of affairs is tangible in robotics, markedly in social robotics [6], [7]. Clearly, since part of the ability to extract affective information from faces can be attributed to visual expertise, current computer vision techniques can provide flexibility and adaptivity to robotics systems [68]. Yet, roboticists do address embodiment and enactment, though the "bodyware" puts severe constraints (degree of freedom, real-time, etc.) and limitations [7]. Asada argued that the design of artificial empathy is one of the most essential issues in social robotics [6]. Motor mimicry, mirroring mechanisms, accounting for the functional roles of the amygdala and the insula are among the essential requirements [6]. The effort of embodiment has fostered valuable studies on the physical grounding of emotion display and perception. Indeed, there is a tradition of robotic research that utilised neuroscience studies as a starting point, e.g. [69]–[71]. These approaches led to accurate models of muscular-skeletal systems, and in particular of facial features appearance [72], [73]. Early work, e.g., by Ogata and Sugano [74], Breazeal and Scassellati [75], has attempted to ground affective behaviour in robot's drives. By reason of "bodyware" being a major concern, implementation models and architectures play prominent role; the computational theory level is seldom addressed. However, proposals that, *prima facie*, might seem very different at the implementation level, can be reconciled at the theoretical model level. For instance, a cross-modal, long-term associative memory for encoding feelings has been proposed by Lim and Okuno [76] in the shape of a Gaussian Mixture Model (GMM), whose parameters are learned via the EM algorithm, whilst [77] address a similar issue by resorting to a Self-Organizing Map (SOM). Though apparently different at the implementation level (SOMs being nothing but regularised GMM (as formally shown in [78]), they basically involve the same model when considered at the computational theory level: emotion encoding as a discrete latent variable model representation solving an (unsupervised) clustering problem. These two significant works are thus related to work described here (projecting into a latent space), albeit our core affect state-space is a continuous one, and builds upon a more complex, nonlinear implementation model (deep GP). In other cases [77], [79], much like in machine learning approaches, a direct mapping via deep neural networks is pursued, from the latent space of affective expressions learnt in a bottom-up, feed-forward sweep to facial gestures, synthetic speech etc. Indeed the use of deep architectures, can lead to efficient implementation models capable to handle the multimodal nature of emotion [80]. Though, for the reasons we have discussed, these can be hardly assumed as models *tout court*, and deep networks *per se* should not be viewed as an implementation model accounting for brain mechanisms [81].

An interesting departure from such trend has been put forward by Horii *et al.* [49]. Their work builds on [80] and exploits a Restricted Boltzmann Machine (RBM) generative architecture to actualise mental simulation for inferring emotion from multimodal signals. This approach is the most related to ours, at least at the implementation modelling level (deep construction and simulation), whilst their theoretical model can be assumed to coincide with the generative RBM description. Multimodal signals are those arising from facial expressions, hand movements, and speech, but rely on an acted dataset (IEMOCAP) where facial expressions have been recorded with a motion-capture system, thus avoiding cumbersome issues related to actual processing of visual cues. Also, physiological signals, a fundamental aspect of emotional unfolding, are not taken into account. In other proposals, such as that by [82] there is an attempt to address the computational theory level so to frame a ST approach, but constraints from neuroscience are overlooked. Motor representation is not explicitly addressed and the latent space of actions is assumed as the affective space *tout court*; only static images are considered and visuomotor mapping is instantiated as a projection to a GP-LVM latent space, which is achieved through a simple variant of the PCA.

A remarkable effort to bridge the different levels of explanations is made in very recent work by Ahmadi and Tani [83]. Though not directly addressing the issue of emotion, cogently, they aim at framing in a general Bayesian setting at the computational theory level, previous work (at the implementation modelling level) on sensorimotor learning via multiple timescale Recurrent Neural Network for robotic imitative interaction with human subjects [84], [85]. Interestingly, bridging across levels brings in the idea, that has recently gained currency, of dealing with complex Bayesian inferential steps via deterministic approximations. Exact inference on directed nonlinear probabilistic models is typically intractable due to the required marginalisation of the latent component [33]. These circumstances are leading to the development of probabilistic generative models relying on mainstream deep neural networks, e.g. [86]–[90]. Such a strategy could also be pursued to cope with a current limitation of our implementation model concerning scalability issues in the adopted Deep GP architecture [89]. This is indeed related to backward inference and should be taken into account when dealing with larger datasets.

Another limitation of the model, referring to Fig. 2, is that we are not currently considering the subcortical, dual visual route from SC/Pulvinar to emotion-related structures (e.g. the amygdala, cfr. light dotted projections). These shortcut pathways are generally deemed to be important to rapidly trigger emotional response before full-fledged processing of the visual stimuli, although this assumption is not without challenge [91], [92]. In our context, such information could be used as an empirical Bayesian prior on the core affect space [93] an aspect which has not been integrated in the model.

Beyond current limitations, the probabilistic model presented here is an attempt to account for multimodal mirroring and enactment mechanisms in a novel and unified way. These are likely to be at the heart of face-based emotion understanding and, more generally of affective interactions [8], [9], and to gain currency in the design of artificial agents [6], [7].

ACKNOWLEDGMENT

This research was part of the project "Interpreting emotions: a computational tool integrating facial expressions and biosignals based shape analysis and Bayesian networks", supported by the Italian Government-MIUR, *Future in Research* Fund.

REFERENCES

[1] G. Rizzolatti and C. Sinigaglia, "The mirror mechanism: a basic principle of brain function," *Nat. Rev. Neurosci.*, vol. 17, no. 12, pp. 757–765, 2016.

[2] V. Gallese, C. Keysers, and G. Rizzolatti, "A unifying view of the basis of social cognition," *Trends Cogn. Sci.*, vol. 8, no. 9, pp. 396–403, 2004.

[3] R. Reisenzein, E. Hudlicka, M. Dastani, J. Gratch, K. Hindriks, E. Lorini, and J.-J. C. Meyer, "Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange," *IEEE Trans. Affect. Comput.*, vol. 4, no. 3, pp. 246–266, 2013.

[4] C. Breazeal, "Emotion and sociable humanoid robots," *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1, pp. 119–155, 2003.

[5] T. Ziemke and R. Lowe, "On the role of emotion in embodied cognitive architectures: From organisms to robots," *Cogn. Comput.*, vol. 1, no. 1, pp. 104–117, 2009.

[6] M. Asada, "Towards artificial empathy," *Int. J. Soc. Robot.*, vol. 7, no. 1, pp. 19–33, 2015.

[7] E. Wiese, G. Metta, and A. Wykowska, "Robots as intentional agents: Using neuroscientific methods to make robots appear more social," *Front. Psychol.*, vol. 8, p. 1663, 2017.

[8] R. Adolphs, "Recognizing emotion from facial expressions: Psychological and neurological mechanisms," *Behav. Cogn. Neurosci. Rev.*, vol. 1, no. 1, pp. 21–62, 2002.

[9] A. Wood, M. Rychlowska, S. Korb, and P. Niedenthal, "Fashioning the face: sensorimotor simulation contributes to facial expression recognition," *Trends Cognit. Sci.*, vol. 20, no. 3, pp. 227–240, 2016.

[10] A. I. Goldman and C. S. Sripada, "Simulationist models of face-based emotion recognition," *Cognition*, vol. 94, no. 3, pp. 193–213, 2005.

[11] J. A. Russell, "A circumplex model of affect." *J. Pers. Soc. Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[12] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. Int. Conf. and Workshops Automatic Face and Gesture Recognition*. IEEE, 2013, pp. 1–8.

[13] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman, 1982.

[14] N. Chater, J. Tenenbaum, and A. Yuille, "Probabilistic models of cognition: Conceptual foundations," *Trends Cognit. Sci.*, vol. 10, no. 7, pp. 287–291, 2006.

[15] A. Craig, "Interoception: the sense of the physiological condition of the body," *Curr. Opin. Neurobiol.*, vol. 13, no. 4, pp. 500–505, 2003.

[16] C. D. Salzman and S. Fusi, "Emotion, cognition, and mental state representation in amygdala and prefrontal cortex," *Annu. Rev. Neurosci.*, vol. 33, pp. 173–202, 2010.

[17] D. J. Anderson and R. Adolphs, "A framework for studying emotions across species," *Cell*, vol. 157, no. 1, pp. 187–200, 2014.

[18] M. Diano, M. Tamietto, A. Celeghin, L. Weiskrantz, M.-K. Tatu, A. Bagnis, S. Duca, G. Geminiani, F. Cauda, and T. Costa, "Dynamic changes in amygdala psychophysiological connectivity reveal distinct neural networks for facial expressions of basic emotions," *Sci. Rep.*, vol. 7, p. 45260, 2017.

[19] K. M. Gothard, "The amygdalo-motor pathways and the control of facial expressions," *Front. Neurosci.*, vol. 8, 2014.

[20] A. Tramacere and P. F. Ferrari, "Faces in the mirror, from the neuroscience of mimicry to the emergence of mentalizing," *J. Anthropol. Sci.*, vol. 94, pp. 113–126, 2016.

[21] R. Mukamel, A. D. Ekstrom, J. Kaplan, M. Iacoboni, and I. Fried, "Single-neuron responses in humans during execution and observation of actions," *Curr. Biol.*, vol. 20, no. 8, pp. 750–756, 2010.

[22] M. Iacoboni, "Neurobiology of imitation," *Curr. Opin. Neurobiol.*, vol. 19, no. 6, pp. 661–665, 2009.

[23] L. Carr, M. Iacoboni, M.-C. Dubeau, J. C. Mazziotta, and G. L. Lenzi, "Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 9, pp. 5497–5502, 2003.

[24] B. Chakrabarti, E. Bullmore, and S. Baron-Cohen, "Empathizing with basic emotions: common and discrete neural substrates," *Soc. Neurosci.*, vol. 1, no. 3-4, pp. 364–384, 2006.

[25] D. Pitcher, V. Walsh, and B. Duchaine, "The role of the occipital face area in the cortical face perception network," *Exp. Brain Res.*, vol. 209, no. 4, pp. 481–493, 2011.

[26] P. Molenberghs, C. Brander, J. B. Mattingley, and R. Cunnington, "The role of the superior temporal sulcus and the mirror neuron system in imitation," *Hum. Brain. Mapp.*, vol. 31, no. 9, pp. 1316–1326, 2010.

[27] L. F. Barrett and W. K. Simmons, "Interoceptive predictions in the brain," *Nat. Rev. Neurosci.*, vol. 16, no. 7, p. 419, 2015.

[28] A. K. Seth, "Interoceptive inference, emotion, and the embodied self," *Trends Cognit. Sci.*, vol. 17, no. 11, pp. 565–573, 2013.

[29] P. Kuppens, Z. Oravecz, and F. Tuerlinckx, "Feelings change: accounting for individual differences in the temporal dynamics of affect," *J. Pers. Soc. Psychol.*, vol. 99, no. 6, p. 1042, 2010.

[30] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychol Rev.*, vol. 110, no. 1, p. 145, 2003.

[31] C. Archambeau and M. Opper, "Approximate inference for continuous-time markov processes," in *Bayesian Time Series Models*, D. Barber, A. T. Cemgil, and S. Chiappa, Eds. Cambridge University Press, 2011, pp. 125–140.

[32] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. Cambridge, MA: MIT press, 2009.

[33] M. Beal and Z. Ghahramani, "The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Stat.*, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Eds., vol. 7. Oxford University Press, 2003, pp. 453–464.

[34] A. N. Sanborn and N. Chater, "Bayesian brains without probabilities," *Trends Cognit. Sci.*, vol. 20, no. 12, pp. 883–893, 2016.

[35] A. Damianou and N. Lawrence, "Deep gaussian processes," in *Proc. Int. Conf. Artificial Intelligence and Statistics*, ser. JMLR Workshop and Conference Proceedings, vol. 31. Scottsdale, AZ, USA: JMLR, 2013, pp. 207–215.

[36] A. C. Damianou, M. K. Titsias, and N. D. Lawrence, "Variational inference for latent variables and uncertain inputs in gaussian processes," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1425–1486, 2016.

[37] J. Ahlberg, "CANDIDE-3 An updated parameterized face," Linköping University, Department of Electrical Engineering, Linköping, Sweden, Tech. Rep. LiTH-ISY-R-2326, 2010.

[38] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[39] J. Orozco, O. Rudovic, J. Gonzlez, and M. Pantic, "Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises," *Image Vision Comput.*, vol. 31, no. 4, pp. 322 – 340, 2013.

[40] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proc. Int. Conf. Comput. Vision Workshops*, 2013, pp. 354–361.

[41] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild, "A bayesian approach to reconstructing genetic regulatory networks with hidden factors," *Bioinformatics*, vol. 21, no. 3, pp. 349–356, 2004.

[42] G. Grossi, R. Lanzarotti, and J. Lin, "High-rate compression of ecg signals by an accuracy-driven sparsity model relying on natural basis," *Digit. Signal Process.*, vol. 45, pp. 96–106, 2015.

[43] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. Automat. Speech Recognition and Understanding Workshop*. IEEE, 2005, pp. 381–385.

[44] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. R. Statist. Soc. B*, vol. 61, no. 3, pp. 611–622, 1999.

[45] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the brownian motion," *Phys. Rev.*, vol. 36, no. 5, p. 823, 1930.

[46] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. Cambridge, MA, USA: The MIT Press, 2006.

[47] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, p. 43, 2015.

[48] A. Ipser and R. Cook, "Inducing a concurrent motor load reduces categorization precision for facial expressions." *J. Exp. Psych.*, vol. 42, no. 5, p. 706, 2016.

[49] T. Horii, Y. Nagai, and M. Asada, "Imitation of human expressions based on emotion estimation by mental simulation," *Paladyn, J. Behav. Robot.*, vol. 7, no. 1, 2016.

[50] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, 2015.

[51] E. Prochazkova and M. E. Kret, "Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion," *Neurosci. Biobehav. Rev.*, vol. 80, no. Supplement C, pp. 99 – 114, 2017.

[52] K. Vogeley, "Two social brains: neural mechanisms of intersubjectivity," *Phil. Trans. R. Soc. B*, vol. 372, no. 1727, p. 20160245, 2017.

[53] R. Adolphs, "The social brain: neural basis of social knowledge," *Ann. Rev. Psych.*, vol. 60, p. 693, 2009.

[54] A. P. Atkinson and R. Adolphs, "The neuropsychology of face perception: beyond simple dissociations and functional selectivity," *Philos. Trans. R. Soc., B*, vol. 366, no. 1571, pp. 1726–1738, 2011.
[55] R. Saxe, "Against simulation: the argument from error," *Trends Cognit. Sci.*, vol. 9, no. 4, pp. 174–179, 2005.
[56] C. D. Frith and U. Frith, "Mechanisms of social cognition," *Annu. Rev. Psychol.*, vol. 63, pp. 287–313, 2012.
[57] V. Gallese, "The 'shared manifold' hypothesis. From mirror neurons to empathy," *J. Conscious. Stud.*, vol. 8, no. 5-7, pp. 33–50, 2001.
[58] K. Friston, "Hierarchical models in the brain," *PLoS Comput. Biol.*, vol. 4, no. 11, p. e1000211, 2008.
[59] D. M. Wolpert, K. Doya, and M. Kawato, "A unifying computational framework for motor control and social interaction," *Philos. Trans. R. Soc., B*, vol. 358, no. 1431, pp. 593–602, 2003.
[60] J. Broekens, D. Degroot, and W. A. Kosters, "Formal models of appraisal: Theory, specification, and computational model," *Cogn. Syst. Res.*, vol. 9, no. 3, pp. 173–197, 2008.
[61] C. Conati, "Probabilistic assessment of user's emotions in educational games," *Appl. Artif. Intell.*, vol. 16, no. 7-8, pp. 555–575, 2002.
[62] J. Hoey, T. Schroder, and A. Alhothali, "Bayesian affect control theory," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* IEEE, 2013, pp. 166–172.
[63] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Autom. Control*, vol. 3, no. 1, pp. 69–87, 2012.
[64] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010.
[65] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 58–65.
[66] M. Sultana, P. P. Paul, and M. Gavrilova, "Social behavioral biometrics: an emerging trend," *Int. J. Patt. Recogn. Artif. Intell.*, vol. 29, no. 08, p. 1556013, 2015.
[67] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? the rise of the new kids on the block," *Image Vis. Comput.*, vol. 55, no. Part 1, pp. 6 – 8, 2016.
[68] S. R. Fanello, C. Ciliberto, N. Noceti, G. Metta, and F. Odone, "Visual recognition for humanoid robots," *Robotics Auton. Syst.*, vol. 91, pp. 151–168, 2017.
[69] M. Kawato, "Internal models for motor control and trajectory planning," *Curr. Opin. Neurobiol.*, vol. 9, no. 6, pp. 718–727, 1999.
[70] B. Scassellati, "Theory of mind for a humanoid robot," *Auton. Robots*, vol. 12, no. 1, pp. 13–24, 2002.
[71] Y. Demiris, L. Aziz-Zadeh, and J. Bonaiuto, "Information processing in the mirror neuron system in primates and machines," *Springer Ser. Bio-/Neuroinf.*, vol. 12, no. 1, pp. 63–91, 2014.
[72] J.-H. Oh, D. Hanson, W.-S. Kim, Y. Han, J.-Y. Kim, and I.-W. Park, "Design of android type humanoid robot albert hubo," in *Proc. Int. Conf. Intelligent Robots and Systems*. IEEE, 2006, pp. 1428–1433.
[73] C. Becker-Asano, K. Ogawa, S. Nishio, and H. Ishiguro, "Exploring the uncanny valley with geminoid hi-1 in a real-world application," in *Proc. Int. Conf. Interfaces and Human Computer Interaction*, 2010, pp. 121–128.
[74] T. Ogata and S. Sugano, "Emotional communication between humans and the autonomous robot which has the emotion model," in *Proc. Int. Conf. Robotics and Automation*, vol. 4. IEEE, 1999, pp. 3177–3182.
[75] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *Proc. Int. Conf. Intelligent Robots and Systems*, vol. 2. IEEE, 1999, pp. 858–863.
[76] A. Lim and H. G. Okuno, "A recipe for empathy," *Int. J. Soc. Robotics*, vol. 7, no. 1, pp. 35–49, 2015.
[77] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adapt. Behav.*, vol. 24, no. 5, pp. 373–396, 2016.
[78] T. Heskes, "Self-organizing maps, vector quantization, and mixture modeling," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1299–1305, 2001.
[79] N. Churamani, M. Kerzel, E. Strahl, P. Barros, and S. Wermter, "Teaching emotion expressions to a human companion robot using deep neural architectures," in *Proc. Int. Joint Conf. Neural Networks*. IEEE, 2017, pp. 627–634.
[80] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3687–3691.
[81] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[82] J. Vitale, M.-A. Williams, B. Johnston, and G. Boccignone, "Affective facial expression processing via simulation: A probabilistic model," *Biol. Inspir. Cogn. Arch.*, vol. 10, pp. 30–41, 2014.
[83] A. Ahmadi and J. Tani, "Bridging the gap between probabilistic and deterministic models: A simulation study on a variational bayes predictive coding recurrent neural network model," *arXiv preprint arXiv:1706.10240*, 2017.
[84] ——, "How can a recurrent neurodynamic predictive coding model cope with fluctuation in temporal patterns? robotic experiments on imitative interaction," *Neural Netw.*, 2017.
[85] S. Murata, Y. Yamashita, H. Arie, T. Ogata, S. Sugano, and J. Tani, "Learning to perceive the world as probabilistic or deterministic via interaction with others: a neuro-robotics experiment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 830–848, 2017.
[86] D. Tran, M. D. Hoffman, R. A. Saurous, E. Brevdo, K. Murphy, and D. M. Blei, "Deep probabilistic programming," in *Proc. Int. Conf. Learning Representations*, 2017.
[87] M. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta, "Composing graphical models with neural networks for structured representations and fast inference," in *Adv. Neural. Inf. Process. Syst.*, 2016, pp. 2946–2954.
[88] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 1050–1059.
[89] Z. Dai, A. Damianou, J. González, and N. Lawrence, "Variational auto-encoded deep gaussian processes," in *Proc. Int. Conf. Learning Representations*, San Juan, Puerto Rico, 2016.
[90] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski, "Deep generative stochastic networks trainable by backprop," in *Proc. Int. Conf. Machine Learning*, 2014, pp. 226–234.
[91] L. Pessoa and R. Adolphs, "Emotion processing and the amygdala: from a 'low road' to'many roads' of evaluating biological significance," *Nat. Rev. Neurosci.*, vol. 11, no. 11, pp. 773–783, 2010.
[92] D. Rudrauf, O. David, J.-P. Lachaux, C. K. Kovach, J. Martinerie, B. Renault, and A. Damasio, "Rapid interactions between the ventral visual stream and emotion-related structures rely on a two-pathway architecture," *J. Neurosci.*, vol. 28, no. 11, pp. 2793–2803, 2008.
[93] C. Ceruti, V. Cuculo, A. D'Amelio, G. Grossi, and R. Lanzarotti, "Taking the hidden route: deep mapping of affect via 3D neural networks," in *Proc. Automatic Affect Analysis and Synthesis Workshop*, ser. LNCS, 2017.

Giuseppe Boccignone received the Laurea degree in theoretical physics from the University of Turin (Italy) in 1985. In 1986, he joined Olivetti Corporate Research, Ivrea, Italy. From 1990 to 1992, he served as a Chief Researcher of the Computer Vision Lab at CRIAI, Naples, Italy. From 1992 to 1994, he held a Research Consultant position at Research Labs of Bull HN, Milan, Italy, leading projects on biomedical imaging. In 1994, he joined as an Assistant Professor the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Italy. In 2008 he joined the Dipartimento di Informatica, University of Milan, Italy, where he currently is a Full Professor of Perception Models, Natural Interaction, and Affective Computing. His research interests include affective computing, visual attention, Bayesian models and stochastic processes for vision and the cognitive sciences.

Donatello Conte received a Ph.D. degree in Information Engineering from the University of Salerno, Fisciano, Italy in 2006. From 2006 to 2013 he has been Assistant Professor at University of Salerno. He is currently Associate Professor of Computer Science with the University of Tours in France and at the LIFAT Laboratory. His current research interests include structural pattern recognition, real-time video analysis, affective computing and document images processing.

Vittorio Cuculo received his B.Sc. in Computer Science from the Univ. of Pisa (Italy) in 2011. In 2013 he starts a collaboration as Senior Developer with the interaction design studio Dotdotdot, Milan (Italy) applying his programming skills to a creative environment. He received his M.S. (2013) from the Univ. of Milan (Italy) and in 2014 started a Ph.D. in Mathematical Sciences. He is currently a member of PHuSe Lab research group, Milan (Italy) and his research interests concern with affective computing and signal processing.

Alessandro D'Amelio received the M.Sc. (Hons.) degree in Computer Science from the University of Milan (Italy) in 2017. He is now pursuing the Ph.D degree in Computer Science at the University of Milan. His research interests include deep learning and affective computing.

Giuliano Grossi received the Ph.D. degree in 2000, from the University of Milan. Since 2001 it has been Assistant Professor at the University of Milan, Dept. of Computer Science. His research interests are in affective computing and sparse representation in signal and image processing, solving combinatorial optimization problems with meta-heuristics based on neural and genetic models.

Raffaella Lanzarotti received her Ph.D degree in computer science from the University of Milan in 2003. Since 2004 she has been Assistant Professor at the Department of Computer Science at the University of Milan. Her research interests concerns the image and signal processing and affective computing, deepening issues concerning face images, such as face recognition and facial expression analysis, and physiological signal processing such as ECG and EMG.