# Feeling First: A Persistent Affective State as the Control Signal for Conversational Agents

Dr. Henrik Richter[1], Dr. Tomas Lindqvist[1], Dr. Tsai Chung[1]

[1]*Synthetic Humans*

**Status: concept paper and research proposal.** This document describes an architecture and a plan to test it. It reports no experimental results. Every number marked *target* is a design goal, not a measurement.

## Abstract

Conversational systems usually decide what to say and then adjust how it sounds. We argue the order should be reversed. We propose an *emotion engine*: a persistent, continuous affective state that is updated by appraising each incoming event, decays over real time between conversations, and conditions every output channel, including the decision to say nothing. The design draws on dimensional models of affect, appraisal theory, and stochastic models of affect dynamics from psychology. We specify the four components (state, appraisal, decay, expression), describe a prototype that can be built on an existing language model, and lay out a pre-specified evaluation: four hypotheses, the ablations that isolate each component, and a blind human test of whether the result is harder to identify as machine-written. We also state what the proposal does not claim.

## 1 Introduction

People recognise machine-written replies quickly. The reply arrives instantly, the punctuation is clean, and the tone is cheerful whatever was said. We think the common cause is structural: in most systems, affect is a style applied to a response that was already chosen. A person's mood works the other way round. It exists before the message arrives, it shapes what they notice and what they choose to say, and it is still there afterwards.

This paper proposes treating affect as state rather than style. The contribution is a design and a test plan:

1. An architecture in which a persistent affective state is the control signal for content, timing and modality (Section 3).
2. A decay model that keeps running between conversations, so elapsed time changes behaviour (Section 3.3).
3. A response policy in which silence, delay and unprompted messages are first-class actions (Section 3.4).
4. A falsifiable evaluation plan (Section 5).

## 2 Background

**Representing affect.** Dimensional models place emotions in a continuous space. Russell's circumplex uses valence and arousal [1]; Mehrabian's PAD model adds dominance [2]. Categorical work finds finer structure: Cowen and Keltner report 27 distinct categories of self-reported emotion bridged by continuous gradients [3], and GoEmotions provides 27 labelled categories for text [4]. We take from this that a low-dimensional core is necessary but not sufficient.

**Appraisal.** Appraisal theories hold that emotion arises from evaluating an event against one's goals and resources, not from the event itself [5, 6, 7]. Scherer's component process model describes sequential checks for novelty, intrinsic pleasantness, goal relevance, coping potential and norm compatibility [6]. The OCC model gives a computationally tractable taxonomy that has been widely used in agents [7].

**Affect over time.** Emotions differ greatly in how long they last; sadness typically outlasts surprise or disgust by a wide margin [8]. Emotional inertia, the degree to which current affect predicts later affect, varies between people and is a stable individual trait [9]. The DynAffect model describes core affect as a noisy process pulled toward a personal home base [10]. In agent research, ALMA layered short-lived emotions over a slower mood and a fixed personality, with mood drifting back to a personality-defined default [11].

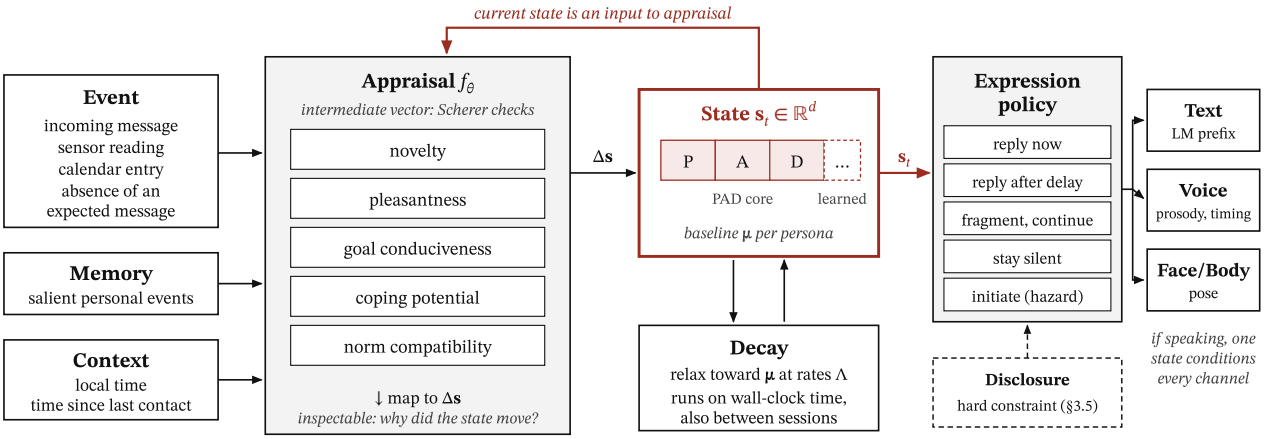



**Figure 1:** Architecture overview (schematic). An event, together with memory and context, is appraised along five intermediate checks and mapped to a state change $\Delta\mathbf{s}$. The state $\mathbf{s}_t$ (PAD core plus learned dimensions) relaxes toward the persona baseline $\boldsymbol{\mu}$ on wall-clock time between events. The same state is fed back into appraisal and drives the expression policy, which may choose not to speak; if it speaks, one state conditions text, voice and face/body, so the channels cannot disagree. Disclosure is a hard constraint outside the learned components.

**Affective dialogue systems.** The Emotional Chatting Machine conditioned generation on an emotion category with internal and external memory [12]. EmpatheticDialogues [13] and DailyDialog [14] made emotion-grounded conversation trainable at scale. These systems condition a single response on a label; the affect does not persist as agent state.

**Memory and agency.** Generative Agents showed that a language model with a memory stream, retrieval and reflection produces believable long-horizon behaviour, including self-initiated actions [15]. Full-duplex speech models such as Moshi show that listening and speaking can be modelled jointly in real time, which makes interruption and overlap possible [16].

**Gap.** To our knowledge no system combines (a) a persistent continuous affective state, (b) real-time decay across sessions, and (c) a policy over whether and when to speak, with one state driving text, voice and embodiment.

## 3 Proposed architecture

Figure 1 gives an overview of the four components.

### 3.1 State

The agent holds a vector $\mathbf{s}_t \in \mathbb{R}^d$. The first three dimensions are pleasure, arousal and dominance, so the state is interpretable and comparable with prior work. The remaining dimensions are learned and are intended to capture finer distinctions (for example, annoyance at a person versus frustration at a situation). We set $d = 140$ as a *target*; the right value is an empirical question addressed in Section 5.

Each persona has a baseline $\boldsymbol{\mu}$ (its temperament) and a per-dimension decay rate. These are the only persona-specific affect parameters.

### 3.2 Appraisal

An appraisal module maps an event to a change in state:

$$\Delta\mathbf{s} = f_\theta(\text{event}, \mathbf{s}_t, \text{memory}, \text{context}) \tag{1}$$

*Event* is an incoming message, a sensor reading, a calendar entry, or the absence of an expected message. *Context* includes local time and time since last contact. The module first produces an intermediate appraisal vector along Scherer's checks (novelty, pleasantness, goal conduciveness, coping potential, norm compatibility), then maps it to $\Delta\mathbf{s}$. The intermediate layer keeps the update inspectable: one can ask why the state moved.

Because the current state is an input, the same message is appraised differently in different moods (Fig. 3). This is the mechanism by which a bad morning colours the afternoon.

### 3.3 Decay

Between events, each dimension relaxes toward baseline at its own rate. We model this as an Ornstein–Uhlenbeck process, following DynAffect [10]:

$$d\mathbf{s} = -\Lambda(\mathbf{s} - \boldsymbol{\mu})\,dt + \Sigma\,d\mathbf{W} \tag{2}$$

with $\Lambda$ a diagonal matrix of rates. Over a gap $\Delta$ with no events, the expected state is

$$\mathbb{E}[\mathbf{s}_{t+\Delta}] = \boldsymbol{\mu} + e^{-\Lambda\Delta}(\mathbf{s}_t - \boldsymbol{\mu}). \tag{3}$$

The gap is wall-clock time, so the process runs while no conversation is happening. Rates are initialised from duration findings in the literature [8], so that high-arousal surprise fades in minutes and low-arousal sadness persists for hours, and are then fitted. The noise term gives day-to-day variation without any input. Figure 2 illustrates Eq. 3 for three example rates.

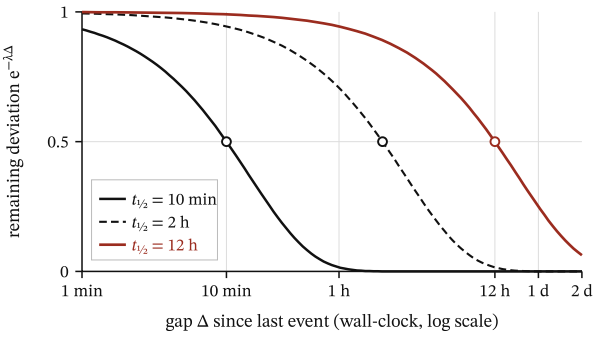



**Figure 2:** Illustration of the model equation, not data. Expected relaxation of one state dimension toward its baseline under Eq. 3, plotted as the remaining fraction of the initial deviation, $(\mathbb{E}[s_{t+\Delta}] - \mu) \,/\, (s_t - \mu) = e^{-\lambda\Delta}$, with $\lambda = \ln 2 / t_{½}$ for three example half-lives: fast (10 min, solid), medium (2 h, dashed) and slow (12 h, accent). Circles mark each half-life. The rates are examples chosen for illustration; actual rates are to be initialised from [8] and fitted.

### 3.4 Expression

The state conditions a policy before it conditions language. At each decision point the policy chooses among: reply now, reply after a delay, send a fragment and continue, stay silent, or initiate. Unprompted messages are modelled with a hazard rate that depends on state, time since last contact, and salient items in memory.

If the policy chooses to speak, $\mathbf{s}_t$ conditions generation. For text we propose a learned projection of $\mathbf{s}_t$ into prefix embeddings for the language model, trained with the base model frozen or lightly adapted. Surface behaviour such as casing, message splitting, typos and response latency is produced by the same conditioned model, not by post-processing rules.

The same vector drives the other channels: prosody and timing for speech, and facial and bodily pose for an embodied agent. One source of truth means the channels cannot disagree.

### 3.5 Disclosure

The system is honest about what it is. If asked whether it is a machine, it says so. This is a hard constraint on the policy, outside the learned components.

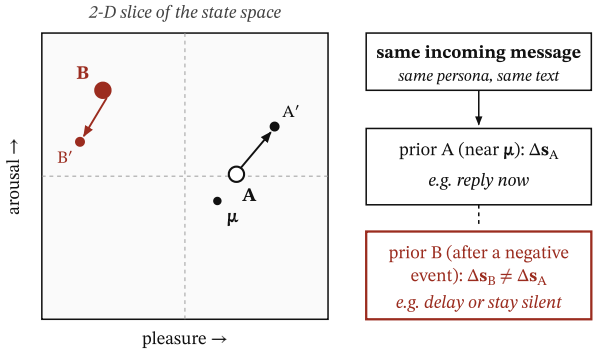



**Figure 3:** Schematic, not model output. The same message, sent to the same persona, is appraised from two prior states: A near the baseline $\mu$ and B after a negative event. Because $\mathbf{s}_t$ is an input to $f_\theta$ (Eq. 1), the updates $\Delta\mathbf{s}_A$ and $\Delta\mathbf{s}_B$ differ, and the policy may choose a different action. Positions, directions and actions are illustrative only.

## 4 Proof of concept

The prototype scope is deliberately small:

- **State and decay** implemented exactly as in 3.1 and 3.3, with $d = 16$ (PAD plus 13 learned dimensions) to start.
- **Appraisal** as a small classifier head over a frozen sentence encoder, trained on EmpatheticDialogues [13] and GoEmotions [4] labels mapped to PAD.
- **Expression** as prefix conditioning on an open-weight language model, plus the five-action policy.
- **Clock**: sessions are replayed with real or simulated gaps so decay can be observed.

This is enough to demonstrate the central claim qualitatively: the same user message, sent to the same persona, produces different replies depending on what happened earlier and how long ago. It is not enough to support quantitative claims, which is the purpose of Section 5.

## 5 Evaluation plan

### 5.1 Hypotheses

**H1 (carry-over).** After a negative event, responses to a neutral probe are measurably more negative than control, and the effect shrinks with elapsed time at the rate the decay model predicts.

**H2 (consistency).** Across a multi-day scripted interaction, human raters judge state-conditioned agents as more emotionally consistent than a prompt-only baseline given the same history as text.

**H3 (detectability).** In a blind test, participants identify state-conditioned replies as machine-written less

often, and more slowly, than baseline replies.

**H4 (restraint).** Allowing silence and delay as actions raises ratings of naturalness without lowering ratings of helpfulness.

### 5.2 Baselines and ablations

Baselines: the same language model with (i) no affect, (ii) an emotion label in the prompt, (iii) the full conversation history in the prompt. Ablations remove one component at a time: no decay (state frozen between sessions), no state input to appraisal, no policy (always reply immediately), and PAD-only state ($d = 3$). The PAD-only ablation is the test of whether dimensions beyond three earn their place. Table 1 summarises the conditions.

**Table 1:** Conditions in the evaluation plan (Section 5.2). ● component present, ○ removed, — not applicable. Planned conditions only; no results are reported.

| Condition | Affect signal | State→ appraisal | Decay | Policy | $d$ |
|---|---|---|---|---|---|
| *Baselines (same language model)* | | | | | |
| (i) No affect | none | — | — | — | — |
| (ii) Label in prompt | emotion label | — | — | — | — |
| (iii) History in prompt | history as text | — | — | — | — |
| **Full system** | **state** | ● | ● | ● | ***d*** |
| *Ablations (one component removed)* | | | | | |
| No decay | state | ● | ○ | ● | *d* |
| No state input | state | ○ | ● | ● | *d* |
| No policy | state | ● | ● | ○ | *d* |
| PAD-only | state | ● | ● | ● | 3 |

### 5.3 Measures

- **Automatic:** valence and arousal of outputs scored by an independent classifier; latency and message-length distributions compared with human chat logs.
- **Human:** pairwise preference, 7-point ratings for consistency and naturalness, and time-to-decision in the detection task.
- **Decay fit:** the half-life recovered from behaviour compared with the half-life set in the model.

Hypotheses, sample sizes and analysis will be pre-registered before data collection.

## 6 Scale and deployment targets

These are goals for the full system, stated so they can be checked later (Table 2).

**Table 2:** Scale and deployment goals. Every row is a *target*, not a measurement.

| Item | Goal | Status |
|---|---|---|
| Training corpus | On the order of billions of human conversational exchanges | *target* |
| State dimensions | 140 (see the ablation in 5.2) | *target* |
| Languages | Coverage of 60 or more languages with region-specific usage | *target* |
| On-device variant | About one billion parameters, so that inference runs without sending conversation data off the device | *target* |

## 7 Risks and ethics

A system built to be emotionally convincing carries specific risks. Users may form attachments, and an agent that initiates contact could foster dependence. Affective state could be used to manipulate. We propose four safeguards: mandatory disclosure (3.5); user-visible controls over initiation frequency; logging of state trajectories so behaviour can be audited; and an explicit exclusion of objectives that reward engagement. Persistent memory of personal events is sensitive data, which is one motivation for on-device inference.

## 8 Limitations

No component has been evaluated. The mapping from categorical emotion labels to a continuous state is lossy and culturally variable. Decay rates drawn from self-report studies may not transfer to text behaviour. A convincing demonstration is not evidence of the hypotheses in Section 5; only the planned experiments are. We make no claim that the system has feelings. The state is a control signal.

## References


[1] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1980.

[2] A. Mehrabian. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4), 1996.

[3] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *PNAS*, 114(38), 2017.

[4] D. Demszky et al. GoEmotions: a dataset of fine-grained emotions. *ACL*, 2020.

[5] R. S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

[6] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. In *Appraisal Processes in Emotion*, Oxford University Press, 2001.

[7] A. Ortony, G. L. Clore and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[8] P. Verduyn and S. Lavrijsen. Which emotions last longest and why: the role of event importance and rumination. *Motivation and Emotion*, 39, 2015.

[9] P. Kuppens, N. B. Allen and L. B. Sheeber. Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7), 2010.

[10] P. Kuppens, Z. Oravecz and F. Tuerlinckx. Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, 99(6), 2010.

[11] P. Gebhard. ALMA: a layered model of affect. *AAMAS*, 2005.

[12] H. Zhou et al. Emotional Chatting Machine: emotional conversation generation with internal and external memory. *AAAI*, 2018.

[13] H. Rashkin et al. Towards empathetic open-domain conversation models: a new benchmark and dataset. *ACL*, 2019.

[14] Y. Li et al. DailyDialog: a manually labelled multi-turn dialogue dataset. *IJCNLP*, 2017.

[15] J. S. Park et al. Generative agents: interactive simulacra of human behavior. *UIST*, 2023.

[16] A. Défossez et al. Moshi: a speech-text foundation model for real-time dialogue. arXiv:2410.00037, 2024.