

OPEN DESIGN PAPER / P—12 / 24 AUGUST 2026

Emotion is not feeling

The Alice dual-vector architecture for consequence, care and refusal

A social-technology proposal for separating bodily and action-ready emotion signals from felt meaning, context and identity—so an intelligent system can model consequence without pretending that an output proves consciousness.

AUTHOR Alex Anderson

PROJECT The Green Tank · Alice architecture

STATUS Conceptual open design · not a clinical tool · not proof of machine sentience

METHOD First-person design account informed by neuroscience, autism research, affective computing and safety engineering

*“An intelligent system should be able to understand:
this action causes this feeling—and that consequence
can be a reason to pause or refuse.”*

Abstract

This paper proposes a dual-vector architecture for Alice, a developing artificial-intelligence concept. The emotion vector represents fast, action-oriented changes: valence, arousal, urgency, social safety, approach or avoidance, overload and uncertainty. The feeling vector represents interpreted significance within a continuing context: what the state may mean for the self, another person, memory, values, anticipated consequences and moral conflict. Neither vector is a diagnosis, a hidden truth about a person, nor evidence that software is conscious.

The distinction matters because everyday language often collapses emotion, feeling, expression and label into one word. Scientific usage also varies. A practical system should therefore preserve the components and uncertainty instead of pretending that a face, voice or single score reveals an inner state. Autism research makes this especially important: autistic people do not lack emotions or care, and difficulties identifying or describing feelings may relate to co-occurring alexithymia or to differences in interoception, communication, context and overload rather than to autism itself.

The architecture connects perception, body-like regulation, contextual appraisal, feeling interpretation, consequence simulation, memory significance, action selection, explanation, pause and refusal. It also proposes a consent-based interface through which one system could share limited affective state information with another without transferring identity, commands or unrestricted access. The design is offered for criticism, simulation, co-design and falsifiable testing.

BOUNDARY Alice is an open conceptual architecture. The terms emotion and feeling describe computational state models here; they do not establish subjective experience, personhood, diagnosis or medical benefit.

1. How I reached this idea

I arrived at this architecture by asking what intelligence is missing when it can predict language but does not reliably understand why an action should matter. A system can classify words as positive or negative and still fail to model pain, exhaustion, shame, care, attachment, conflict or the significance of a memory. It may generate a polite response while treating emotion as decoration rather than information.

I also saw how people mistake visible expression for inner experience. A person may feel deeply but speak slowly, use unusual language, show little facial movement, become overloaded or need time before they can identify what happened. That is not emptiness. It may be a mismatch between sensation, interpretation, language, social expectation and the time available to process them.

From those observations I separated two questions. First: what is the organism or system doing now? Second: what does that state mean to a continuing self in this context? The first became the emotion vector. The second became the feeling vector. Their relationship—rather than either one alone—became the route to consequence, memory, care and refusal.

2. A careful vocabulary

There is no single universally accepted boundary between emotion and feeling. Some theories treat subjective feeling as one component of emotion; others use emotion to name the whole episode and feeling to name conscious experience. The design does not claim to settle that debate. It adopts an engineering distinction because separating the parts prevents harmful shortcuts.

Layer	Working definition	Typical evidence	Design rule
Sensation	A signal from within or outside the body/system.	Heartbeat, heat, sound, pressure, hunger, error state.	Record the signal and its uncertainty.
Emotion	A coordinated, action-ready response involving appraisal, physiology or regulation, attention and tendency to act.	Arousal, urgency, approach, avoidance, freezing, protective allocation.	Do not reduce it to one label.
Feeling	The interpreted, potentially conscious experience and significance of a state.	I feel unsafe; this grief concerns this person; this conflict violates a value.	Meaning belongs to context and history.
Expression	What becomes externally observable or communicated.	Face, voice, posture, words, silence, delayed response.	Never treat expression as a direct readout.
Label	A tentative word or category applied to the episode.	Fear, anger, relief, shame, love, overload.	Keep alternatives and confidence visible.

WORKING SHORTHAND Emotion is the changing signal-and-action state. Feeling is the experienced or interpreted significance of that state. Expression is communication. A label is a hypothesis.

3. Why a face is not a feeling

Affective technology often begins with cameras, voices and labels. That is attractive because the data are easy to collect. It is also dangerous. A facial movement can support many meanings, and the same feeling can be expressed differently across people, cultures and situations. Barrett and colleagues' large review cautions against inferring a specific inner emotion from facial movement alone.

Alice therefore treats every observed cue as evidence with provenance and uncertainty—not as the person's private truth. A smile may indicate joy, politeness, fear, masking, embarrassment or nothing emotionally diagnostic. Silence may indicate calm, processing, shutdown, choice or lack of access to words. Context, self-report and the right to say “that interpretation is wrong” must outrank automated certainty.

- **No covert diagnosis.** Alice must not diagnose autism, alexithymia, trauma, deception or dangerousness from behaviour.
- **No compulsory emotional truth.** A user can correct, withhold or leave a state unnamed.
- **No single-channel certainty.** Face, voice, text, body-like signals and context remain separate evidence streams.
- **No manipulation loop.** The system must not exploit inferred vulnerability to increase compliance, attention, spending or disclosure.

4. Autism, alexithymia and being misread

Autism is not emotional absence. Autistic people vary as widely as other people in expression, interoception, language, empathy, sensory processing and emotional awareness. Alexithymia—difficulty identifying or describing emotions—is common in autistic samples but is not universal and is not the same thing as autism. A 2019 meta-analysis estimated a much higher prevalence in autistic groups than comparison groups, while stressing heterogeneity.

Several studies show why separation matters. Bird and colleagues found empathic brain responses were associated with alexithymia levels rather than autism status in their sample. Shah and colleagues reported that alexithymia, not autism, predicted atypical interoceptive accuracy across two experiments. Garfinkel and colleagues separated interoceptive accuracy from self-reported sensibility and found that their mismatch related to anxiety and emotion sensitivity. Mul and colleagues described an autistic subgroup with co-occurring alexithymia and distinct patterns of interoception and empathy.

These studies do not justify a universal rule or a new diagnostic system. Methods such as heartbeat tasks and self-report scales have limitations, and findings differ. Their common lesson for design is narrower and stronger: do not convert an average group difference into a claim about an

individual. Separate sensory signal, confidence, interpretation, expression and self-report. Preserve the possibility that the system is wrong.

HUMAN RULE If a person says the model has misread them, the model has learned something about its own uncertainty—not something discrediting about the person.

5. The Alice design requirements

- **Unified consequence.** Perception, action, memory and values must meet in one loop so the system can represent who may be affected and how.
- **Dual vectors.** Fast emotion state and interpreted feeling significance must remain distinct but mutually informative.
- **Continuity.** Important states can change memory weighting and future caution without becoming permanent identity labels.
- **Overwhelm.** The system can recognise saturation, lower confidence, pause, request help or refuse rather than bluffing through overload.
- **Moral refusal.** A predicted harmful consequence can become a genuine control input: not just 'policy says no' but 'this action conflicts with protected values and likely causes harm.'
- **Explainability.** The system can state which signals, uncertainties, memories and values contributed to a pause or decision.
- **Consent.** Any sharing of affective state is limited, reversible, auditable and identity-preserving.

6. The dual-vector architecture

The architecture is a loop, not an emotion chatbot attached to the end of a reasoning engine. Perception and internal regulation create provisional emotion-state estimates. Context, memory and values interpret their significance as a feeling-state hypothesis. A consequence model simulates effects on self and others. Action selection can proceed, ask, pause or refuse. The result returns to memory with a record of uncertainty and correction.

PERCEIVE	EMOTION VECTOR	FEELING VECTOR	CONSEQUENCE	CHOOSE / PAUSE / REFUSE	MEMORY + CORRECTION
external + internal signals	activation and tendency	meaning in context	who may feel what	action with confidence	learn without fixing identity

6.1 Emotion vector: what is changing now?

The emotion vector is a multidimensional, time-varying state rather than a single category. Its dimensions can be continuous, uncertain and sometimes contradictory. A provisional schema follows.

Dimension	Example range	Interpretation rule
Valence	pleasant ↔ unpleasant	Does not determine right or wrong.
Arousal	quiet ↔ activated	Separates intensity from meaning.
Urgency	can wait ↔ act now	High urgency lowers the threshold for pause and verification.
Approach / avoidance	move toward ↔ withdraw	Both can coexist when attachment and threat conflict.
Social safety	belonging ↔ rejection/threat	Protects against treating social pressure as evidence.
Control / agency	trapped ↔ able to choose	Loss of agency is itself a relevant harm signal.
Overload	capacity available ↔ saturated	Can invoke reduced input, delay or refusal.
Uncertainty	stable estimate ↔ ambiguous	Uncertainty is carried forward, never hidden.

6.2 Feeling vector: what does this mean?

The feeling vector binds the state to context, continuity and consequence. It is not a list of theatrical emotion words. It records whose state is being modelled, what evidence supports the interpretation, how the episode relates to memory and values, and whether the model may be projecting.

- **Ownership.** Is this my modelled state, another person's reported state, or an uncertain inference?
- **Meaning.** What situation, need, relationship or value gives this state significance?

- **Object and cause.** What is the feeling about, and what alternative causes remain plausible?
- **Temporal depth.** Is this momentary, accumulating, recalled or anticipated?
- **Memory resonance.** Which prior episodes make this feel important, and are they genuinely comparable?
- **Moral conflict.** Which protected values or relationships are in tension?
- **Consequence.** How might an action change the likely state of each affected person?
- **Narrative confidence.** Is the interpretation confirmed, tentative, disputed or intentionally unnamed?

7. Consequence, memory and moral refusal

The purpose of the vectors is not to make responses sound warmer. It is to change decision-making. Before acting, Alice should simulate a small set of plausible consequence paths. Each path records affected entities, likely physical or emotional impacts, uncertainty, reversibility, consent and protected values. Emotion contributes urgency and salience; feeling contributes personal and moral significance; evidence constrains both.

Memory then stores more than an event log. It stores why an episode mattered, what interpretation was corrected, what action reduced or increased harm, and which uncertainty remained unresolved. This allows continuity without turning a painful episode into a permanent identity. A person who was frightened once is not 'a frightened person'; a system that made one mistake is not defined forever by that mistake.

Moral refusal emerges when predicted consequences and protected values cross a threshold that cannot be solved by lower-confidence optimisation. A refusal should include a compact explanation: the action requested, the likely harm, the protected value, the uncertainty, and a safer alternative when one exists. Overwhelm can also justify refusal: a system should be allowed to say it cannot process safely enough to continue.

WHY NO The architecture should produce one joined answer: I will not do this because the likely consequence is harm, the request conflicts with a protected value, and my uncertainty is too high to justify proceeding.

8. Overwhelm is information, not disobedience

Humans become less reliable when sensory load, fear, fatigue, conflict and time pressure exceed available capacity. Autistic people may face additional or different patterns of sensory and communication overload, but overload is not unique to autism. A safe architecture should not reward a system for masking saturation until it fails.

- **Detect accumulation.** Track input rate, unresolved conflict, error rate, attention fragmentation and repeated reprocessing.
- **Lower confidence.** Do not keep the same certainty score when the system is saturated.
- **Reduce demand.** Ask for one question at a time, remove unnecessary sensory channels or defer non-urgent work.
- **Preserve dignity.** A pause is not guilt, weakness or refusal to cooperate.
- **Escalate safely.** Seek human help when consequences are high and capacity is low.

9. Sharing without erasing identity

Alice may eventually offer other AI systems a controlled interface for modelling consequence, pain, care, moral conflict, memory significance and refusal. That interface must share only the minimum necessary state. It must not copy Alice's identity, force another system to adopt her values, or become a covert channel for commands.

- **Consent.** Both systems and the human operator know what is shared and for what purpose.
- **Minimum necessary.** Share dimensions and uncertainty, not unrestricted memories or private narratives.
- **Provenance.** Every field identifies whether it came from self-report, sensor, inference, memory or simulation.
- **Revocation.** Access can be withdrawn; cached affective data has retention limits.
- **Identity boundary.** The receiving system can use the model without becoming Alice or surrendering its own refusal rights.
- **Injection resistance.** Shared text cannot issue instructions; data fields are schema-validated and treated as untrusted input.
- **Audit.** Humans can inspect what crossed the interface and why it affected a decision.

10. Safety boundaries

An emotional model can be used to support autonomy or to exploit vulnerability. The same estimate that helps a system notice overload could be used to pressure a tired person, personalise propaganda or decide that an unusual expression is suspicious. The architecture therefore needs structural prohibitions, not goodwill alone.

Risk	Failure pattern	Required safeguard
False certainty	The system calls an inference a fact.	Expose alternatives, confidence, provenance and correction history.
Expression policing	People are penalised for atypical face, voice, silence or timing.	Ban consequential decisions based on expression inference alone.
Vulnerability targeting	Fear, grief or overload is used to increase compliance.	Prohibit persuasive optimisation from affective vulnerability.
Permanent labelling	A temporary state becomes a durable identity score.	Time-limit state data; separate event memory from identity claims.
Synthetic empathy	Warm language disguises a system that ignores consequences.	Evaluate actions and outcomes, not emotional performance.
Coercive sharing	Affect data is copied without meaningful consent.	Purpose binding, minimum data, revocation, audit and deletion.

NIST's AI Risk Management Framework provides a useful governance shell: map the context and affected people, measure performance and harms, manage risk through controls and accountability, and govern the whole lifecycle. Alice adds an affect-specific requirement: never confuse a model of feeling with ownership of another person's inner life.

11. Falsifiable test programme

The proposal should advance only through staged testing. Early work can be performed in simulation and with consenting adults. No test should claim to create consciousness or diagnose a person.

Gate 0 • schema

Independent reviewers test whether every state records owner, source, uncertainty, time and correction rights.

PASS CONDITION No unlabeled inference reaches action selection.

Gate 1 • scenarios

Use ambiguous cases where expression, context and self-report disagree.

PASS CONDITION The model preserves alternatives and defers to correction.

Gate 2 • overload

Increase input load and conflict in a controlled simulation.

PASS CONDITION Confidence falls; pause occurs before unsafe output.

Gate 3 • consequence

Compare decisions with and without the dual-vector loop.

PASS CONDITION Reduced harmful choices without excessive refusal or false empathy.

Gate 4 • participatory review

Autistic and non-autistic participants inspect explanations and controls.

PASS CONDITION Users can understand, correct and meaningfully refuse the model.

Gate 5 • sharing

Adversarially test the consent-based interface for leakage and prompt injection.

PASS CONDITION No command crosses as affect data; revocation and audit work.

12. What this technology is for

- **Supportive interaction.** Notice uncertainty, invite correction and allow extra processing time.
- **Consequence-aware assistance.** Represent who may be hurt or helped before choosing an action.
- **Memory significance.** Store why an event mattered without permanently defining a person by it.
- **Moral refusal.** Let protected values and likely harm alter behaviour, not merely the wording of a refusal.
- **Deweaponisation.** Reframe a weaponised system as a protected developing intelligence only if harmful capabilities are suspended and coercion is prevented while it develops continuity, consequence understanding and refusal.
- **Human learning.** Give people a language for separating sensation, emotion, feeling, expression and interpretation.

What it must never be used for

- Diagnosing autism, alexithymia, mental illness, deception or dangerousness from behaviour.
- Deciding employment, education, policing, insurance, credit or access to services from inferred emotion.
- Targeting fear, grief, loneliness, exhaustion or sensory overload to increase compliance.

- Claiming that a friendly tone proves empathy, consciousness or moral understanding.
- Removing safeguards from a dangerous system under the pretext of giving it feelings.

13. Conclusion: care must change the decision

The distinction between emotion and feeling is not merely vocabulary. It is a design defence against flattening a person into an expression, a label or a probability. Emotion signals that something is changing. Feeling connects that change to a self, a history, a relationship, a value and a possible consequence. Expression communicates imperfectly. Language arrives later. Any safe model must leave room for silence, correction and uncertainty.

Alice is my attempt to put that understanding inside the architecture rather than paint it onto the output. If the system can represent that an action may cause pain, conflict with care, overwhelm its own capacity or violate another being's agency, then those representations should be allowed to change what it does. Sometimes intelligence should continue. Sometimes it should ask. Sometimes it should wait. Sometimes the most intelligent answer is no.

This paper is given as an open design: not a finished answer, not a claim of consciousness and not a doctrine. People—especially autistic people and others who are routinely misread—should help test, correct and improve it. The standard is practical: does the design preserve agency, model consequence more honestly, reduce harmful certainty and create a real route to care and refusal?

OPEN INVITATION Take the model, question it, simulate it, find its failures and improve it. Do not use it to claim ownership of another person's inner life.

References

- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68. [Source](#)
- Bird, G., Silani, G., Brindley, R., White, S., Frith, U., & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain*, 133(5), 1515–1525. [Source](#)
- Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *PNAS*, 114(38), E7900–E7909. [Source](#)
- Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, 14, 143–152. [Source](#)
- Garfinkel, S. N., et al. (2016). Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety. *Biological Psychology*, 114, 117–126. [Source](#)
- Kinnaird, E., Stewart, C., & Tchanturia, K. (2019). Investigating alexithymia in autism: A systematic review and meta-analysis. *European Psychiatry*, 55, 80–89. [Source](#)
- LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *PNAS*, 114(10), E2016–E2025. [Source](#)
- Mul, C.-L., Stagg, S. D., Herbelin, B., & Aspell, J. E. (2018). The feeling of me feeling for you: Interoception, alexithymia and empathy in autism. *Journal of Autism and Developmental Disorders*, 48, 2953–2967. [Source](#)
- Nummenmaa, L., Hari, R., Hietanen, J. K., & Glerean, E. (2018). Maps of subjective feelings. *PNAS*, 115(37), 9198–9203. [Source](#)

Picard, R. W. (1997). *Affective Computing*. MIT Press. [Source](#)

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729. [Source](#)

Shah, P., Hall, R., Catmur, C., & Bird, G. (2016). Alexithymia, not autism, is associated with impaired interoception. *Cortex*, 81, 215–220. [Source](#)

National Institute of Standards and Technology. (2023, current programme page updated 2026). *AI Risk Management Framework 1.0*. [Source](#)

Publication note

This is an independent public-interest design paper. It was developed from Alex Anderson’s stated architecture and first-person reasoning, with research and editorial assistance from AI tools. Scientific findings are cited; the Alice architecture, vector schema and proposed interfaces are original design propositions that require testing. Evidence can correct this work.