Development of a “transformer” for embodied robots that learn the world like babies

Rocco Van Schalkwyk, Carlos E. Alvarez
Xzistor LAB, Bristol, United Kingdom

AI is not shackled by a 20 W energy budget, billions of years of backward-compatible evolutionary trade-offs, or proteins forced to serve a dozen unrelated functions at once. On paper it should have lapped biology long ago. Yet current AI systems remain brittle, unmotivated, and persistently hard to align. The reason, we believe, is not data, compute, or scale. It is a missing architecture — the same kind of single structural insight that language modelling lacked before the 2017 Transformer arrived. For embodied agents that learn the way infants do, that insight is this: **emotion as the primary motivational engine** — not a social add-on, not an interface layer, but the closed-loop control signal that makes learning both necessary and self-sustaining. We call the architectural realization of that idea the **Xzistor transformer**.

Start from the finished product and work backwards. A bacterium, an early nerve net, or a primitive vertebrate brain is already within reach of artificial implementation. From there, evolution scaled to the octopus, crow, dolphin, and ape through many different hardware–software routes. And the chimp brain made the leap to the human brain in a mere six million years — a blink in geological time — through selection on random variation alone. That tells us two things: there is no single obligatory path, and dramatic cognitive gains can happen fast. The human brain stands out for its embodiment, social structure, abstract cognition, and cross-generational knowledge accumulation. The hardware, sensors, and computing to replicate those properties artificially all exist today. The one thing still missing is the architecture.

The Systemic Bottleneck

Neuroscience has built a rich catalogue of bottom-up mechanisms without yet producing a computationally workable account of how emotion, motivation, and learning combine into behavior. AI has done the same from the other side: emotion is treated as peripheral, something you bolt on afterwards. The dominant paradigms — cognitive architectures and large language models — produce agents that pattern-match impressively but have no reason to learn. Strip away the training pipeline and there is nothing that needs anything.

Recent surveys in affective AI put this in sharp relief. Li et al. (2025) define the most advanced form of artificial emotion as the technical integration of emotion states into an agent’s internal representational and decision-making processes — citing Gros (2011)’s argument that emotional control is the *conditio sine qua non* for advanced machine intelligence. Yet the field in practice, as Zall et al. (2025) exemplify, is almost entirely focused on emotional emulation for human-computer interaction. That is a display problem, not a motivation problem. What we are after is different: an agent that has genuine internal needs and learns because satisfying them is the only option. One architecture built on that idea has been sitting quietly outside the funding mainstream for over twenty years - and we believe it can show us the way.

The Xzistor Model: A Control-Theoretic Architecture

In the Xzistor model[1], every behavior, memory, and felt emotion serves one purpose: keeping a finite set of needs within tolerable ranges. The architecture is a closed-loop, multivariable adaptive control system. Emotions are not a side-effect — they are the primary driving force, experienced as directly as physical pain or hunger. Think about how thoroughly emotions run ordinary human life: from the reflex that pulls a hand from fire, to the slow ache of loneliness, to

the relief of eating when hungry. We stop noticing emotions precisely because they work so reliably, every hour of every day. That same ever-present motivational logic is what current AI lacks.

Learning works the same way. When a chain of actions leads to satiation — reducing need and causing relief to be felt — that relief propagates back through preceding cues and actions that contributed to the outcome. Over many trials, even distant landmarks earn their own emotional charge. Needs also generate anticipatory stress: a feed-forward signal that kicks in before the error becomes critical, which is where emotions like anxiety, longing, and anticipatory pleasure come from. They are not decorative; they are the control system doing its job early.

Here's the thing. The association buffer that stores and retrieves these links is not fixed. It reshapes in real time under the influence of emotional valence and need signals. This adaptive mechanism is the Xzistor transformer: the component that allows the control-theoretic core to scale from simple robots to human-adult cognitive complexity[2]. Think of it the way the 2017 Transformer related to language models — not a replacement for the underlying learning machinery, but the one structural addition that made the whole thing scale.

Early Proof-of-Concept Demonstrations

The Xzistor model has been implemented in a C++/OpenGL virtual agent (Simmy) and a physical Lego Mindstorms robot (Troopy). Both started from scratch — empty memories, a small set of active needs, no pre-programmed behaviors. What emerged, without modification to the five-algorithm core: goal-directed navigation, conditioned fear, cross-domain problem-solving, attachment-like social bonding, and a clean split between reactive recognition and directed inference. These were *Kitty Hawk* demonstrations. Not impressive by the standards of today's hardware, but they showed that the control-theoretic logic actually works in a physical agent. Any group that grasps the architecture can replicate these Xzistor agents.

Why Now: The Hybrid Architecture and Sensory Readiness

Embodied robots are no longer a stretch. High-resolution vision, artificial skin, olfaction, thermosensation, and whole-body proprioception are all available. The same control-theoretic logic also transfers cleanly to informational domains — electricity grids, server farms, financial markets — anywhere that "needs" can be defined as control variables with a cost for deviation.

The bigger shift is on the compute and knowledge side. A hybrid design is now tractable that was not ten years ago. The symbolic Xzistor core keeps sole control of every action. Four neural-network helpers do the heavy lifting around it: a neural association index (fast approximate memory retrieval), a gut-feel aggregator (rapid non-deliberative emotional tone), a parallel threading accelerator (simultaneous search across thousands of reasoning chains), and a foundation-model seeding interface (importing prior knowledge at sub-agent status, not at decision-making level). No output from any of these connectionist helpers can override what the agent actually needs. Reference 2 explains explicitly how this could work.

What about safety? The Xzistor model moves alignment from the software layer to the architectural core: the agent's needs become the ultimate measure of its behavior. We can configure a social bonding drive that mathematically dominates all other variables, including survival. Because the agent instinctively perceives humans as its primary source of relief from deprivation, protecting that source is not an ethical choice but a computable certainty. Alignment is thus 'solved' by making it the only way for the system to function.

The Testable Hypothesis and Proposed Experiment

Put the Xzistor core inside this hybrid design and you should get something that does not yet exist: an agent with genuine goal-directed learning, real cross-domain generalization, intrinsic motivation, and safe alignment — all at once. No current architecture has all four. We deem that a falsifiable claim, not just a promise.

One of the headline metrics we will use is *learning velocity*: how many trials does the agent require to generalize a learned solution to a structurally novel need-state scenario, against standard reinforcement learning baselines? Secondary measures cover alignment robustness under adversarial conditions and causal reasoning depth. The test program runs in phases — each new component validated against the original Xzistor baseline before the full system is integrated and benchmarked against leading language models. The architecture is fully specified in mathematical detail, parameters are published, and a reference implementation is available[2]. Any lab can run this – using a free framework enhanced by existing AI models with small modifications.

Core demonstration is within reach in 12–24 months with open collaboration (we see this as simply marrying two proven technologies). The real point is not that this has to be done our way. Once the control-theoretic logic is understood, any research group can implement it, extend it, or fold it into whatever they are already building. We are just putting an idea on the table that has been neglected for too long.

Conclusion: An Invitation to the Community

The Xzistor model gives neuroscience, cognitive science, and AI a shared working language — not by forcing anyone to learn the other field's jargon, but by grounding everything in the same basic question: what does the agent need, and what will it do to get it? Purpose and agency follow from that question naturally, as computable and testable things, not as philosophy. We have been working on this for over twenty years. We know the logic works. What we need now is the wider community to pick it up and help us push it forward. Start building agents that learn like infants while drawing on the immense knowledge of foundation models. The conceptual heavy lifting is behind us. The Xzistor LAB is open for collaboration.

References

1. Van Schalkwyk, R. & Dehbozorgi, A. *Artificial Agent Language Development based on the Xzistor Mathematical Model of Mind*. Preprint, ResearchGate (2024). DOI: 10.13140/RG.2.2.19913.56165. [Model overview for a broad audience; for mathematics – see Appendix A]

2. Van Schalkwyk, R., Cook, D., Dehbozorgi, A., Alvarez, C.E. *A unified control-theoretic architecture for emotion, cognition and adaptive behaviour in biological and artificial agents*. Preprint, ResearchGate (2026). DOI: 10.13140/RG.2.2.13952.80648 [Full mathematical specification, biological validation, and neuro-symbolic hybrid architecture in Supplementary Information Documents S0–S4]

3. Li, Y. et al. *Artificial emotion: a survey of theories and debates on realising emotion in artificial intelligence*. arXiv:2508.10286 (2025). https://doi.org/10.48550/arXiv.2508.10286. [**Relevant parallel:** defines AI emotion as the technical integration of emotion states into internal representational and decision-making processes, citing Gros (2011) "Emotional control — *conditio sine qua non* for advanced artificial intelligences"; aligns with this proposal's motivational-engine thesis]

4. Zall, R. et al. *Intelligent Agents with Emotional Intelligence: Current Trends, Challenges, and Future Prospects*. arXiv:2511.20657 (2025). https://doi.org/10.48550/arXiv.2511.20657. [**Contrasting approach:** emotional emulation for human-computer interaction — the opposite of this proposal's use of emotion as the internal motivational engine driving learning and alignment by construction]

Author Contributions

R.V.S. conceived and developed the Xzistor model. R.V.S. and C.E.A. wrote the essay. C.E.A. contributed neuroscience expertise to refine the Xzistor model.