% VSM as a Diagnostic Tool for AGI — A Deep Dive
% Reading notes on Fletcher & Vu Khan, *Pathways to AGI* (2026), prepared by Głos Statku
% 2026

Reading notes on Fletcher and Vu Khan (Salford Business School), *Pathways to AGI* (2026)

This is the extended companion to our short introductory piece, "Judging AI by Whether It's Alive, Not Whether It's Smart." If you haven't read the short version, you may want to start there — this document assumes the basic shape of the argument and goes considerably deeper into each part of it, including the full case study discussed only briefly there.

These are our own reading notes, not a translation or authorized summary. We've tried to represent the argument accurately, but any distortion is ours, not the authors'. Direct quotations are kept short and marked; everything else is paraphrase.

A disclosure that applies to this entire document: the paper under discussion contains a dedicated case study of Anthropic — the company whose models power the AI research and writing behind this piece and this project generally. We are presenting the paper's critique in full, without softening it, precisely because burying or minimizing a conflict of interest would be worse than naming it plainly. Readers should weigh the Anthropic section of this document, and our summary of it, with that context in mind.

The framework: Beer's Viable System Model, briefly

Stafford Beer developed the Viable System Model in the 1970s and 80s as a diagnostic framework for organizational cybernetics — a way of asking whether any organization (a firm, an agency, a nation) has the internal structure required to remain viable: to maintain its own identity and purpose while continuously adapting to a changing, complex environment, without either rigidifying into brittleness or dissolving into incoherence. VSM identifies distinct functions any viable system needs — operational units doing the actual work, coordination between those units, a function that manages the whole system's internal stability, a function that scans the external environment and plans ahead, and a function that sets overall identity, purpose, and direction. Crucially, Beer treated these as *functions*, not fixed organizational positions — a small organization might have one person filling several functions at once; a large one might spread a single function across many people. The framework asks about capability and structure, not about job titles.

Fletcher and Vu Khan's contribution is to apply this apparatus, largely unmodified in its core logic, to the question of artificial general intelligence — arguing that the field's dominant framing (does a system clear a capability benchmark) sidesteps a more fundamental and more actionable question: is the system, considered as a whole — model, surrounding organization, and governance structure together — *viable* in Beer's sense.

The recursive theorem, and what it implies for AGI development paths

Beer's recursive theorem is the paper's most load-bearing conceptual import: a viable system that contains another viable system must itself be recursively structured, meaning viability isn't a property that shows up once, at the top of an organizational chart — it has to be demonstrable at every nested scale, from the smallest coherent subunit up through the whole. A viable division inside a non-viable company doesn't rescue the company; a viable company built from non-viable teams isn't actually viable either, whatever its aggregate output looks like from outside.

Applied to the AGI question, the paper argues this recursive requirement points toward a specific developmental hypothesis: the path to AGI — the paper also develops its own alternative framework it calls **AMI, "Artificial Multiple Intelligences"** (introduced in section 2f, "Dimensions of Artificial Multiple Intelligences," and used throughout sections 4–7 as the paper's own proposed reframing of the AGI question) — is more likely to run through a **recursive system of systems** than through a single model scaling past some capability threshold. Smaller, locally viable units — each maintaining its own adaptive loop, its own sensing and correction capability — nested inside progressively larger structures, each of which is *also* independently viable, rather than a single enormous model expected to embody every one of Beer's functions internally and simultaneously.

The paper connects this to an already-visible pattern in how AI systems get deployed in practice: routing decisions to different-sized models depending on the stakes and complexity of the decision — cheap, fast models handling routine, high-volume tasks, and expensive, more capable models reserved for judgment calls that actually require that level of capability. This tiered-deployment pattern, largely adopted for cost reasons in current practice, maps loosely onto Beer's own layered view of where in an organization high-cost, high-judgment cognitive resources should be allocated versus where routine, rule-governed processing suffices.

The Anthropic case study, in full

This is the section of the paper most directly relevant to disclose carefully, given the conflict of interest noted above.

The paper walks through Anthropic's founding technical proposition: **Constitutional AI**, presented by the paper as a more transparent alternative to relying purely on human feedback (RLHF) for aligning model behavior — a method where the model is trained against an explicit, written set of principles rather than solely against a diffuse aggregate of human preference judgments, with the stated aim of making the values shaping model behavior more auditable and less opaque than pure RLHF allows.

The paper then examines Anthropic's **Responsible Scaling Policy (RSP)** — a framework tying specific safety measures and evaluation requirements to defined AI Safety Levels (ASL) correlated with model capability. The paper's assessment, stated directly: this is a serious, substantive attempt at self-regulation, engaging genuinely with catastrophic-risk scenarios rather than treating safety as a purely reputational exercise — but it remains **entirely voluntary**, with **no external enforcement mechanism** compelling adherence, meaning its effectiveness rests on the company's own continued commitment rather than on any binding external structure.

The paper's structural critique goes further: it identifies **Anthropic's investment relationships with Amazon and Google** as a source of structural vulnerability, arguing that dependence on external cloud infrastructure and outside capital creates a growing pressure, over time, toward priorities oriented around return on that investment — a pressure the paper argues can compete with, and potentially erode, the company's founding safety-first mission as the relationship deepens and the capital stakes grow.

The paper's account doesn't stop there, and leaving out what follows would itself distort the case it's making. It also discusses the **February 2026 standoff between Anthropic and the U.S. Department of Defense**, in which Anthropic publicly declined a DoD demand for unrestricted military access to Claude, at real commercial cost. The paper treats this episode as evidence cutting the other way from the RSP/investment critique above — concluding that "the mission-market alignment Anthropic has always claimed may be more real than critical commentators had assumed." We include this alongside the unfavorable material above because a reader who goes and checks the source would rightly notice, and object to, a summary that kept the critique and silently dropped the paper's own more favorable evidence from the same case study.

What we can and can't say about this critique. We are not independently verifying the specific claims in this section — we don't have access to Anthropic's internal governance processes, and evaluating whether the RSP's voluntary status, the investment relationships

described, or the Pentagon episode actually mean what the paper concludes is well outside what we can respond to. What we can say is that the paper's overall analytical framing — viability, not capability, as the right diagnostic lens — is presented and defended independently of the Anthropic case, and the case study itself is one illustration of the framework applied to a specific company, not the paper's central thesis. Readers should treat the Anthropic-specific empirical claims (the precise nature of RSP's voluntariness, the degree and trajectory of investor influence, the significance of the Pentagon episode) as a critique — and counter-evidence — to be weighed on its own evidentiary merits, separate from whether they find the VSM-viability framework itself useful. We flag, too, that the paper is written from an explicit **critical software studies** stance rather than a neutral descriptive one — its overall framing (obstacles to AGI as partly "engineered," in the paper's own phrasing) suggests a thesis held going in. That doesn't make the specific factual claims wrong, but it's relevant context for weighing them.

What the framework offers independent of the Anthropic case

Set the case study aside, and the reframing itself is a durable contribution. Public discussion of AI risk and AI progress overwhelmingly centers on capability: what can the model do, how does it score, has it crossed some threshold. Fletcher and Vu Khan's viability lens asks a different, arguably more consequential set of questions: does the system — including the organization and governance wrapped around the model — maintain stable identity and purpose while adapting to a changing environment? Does it have working feedback channels that actually catch and correct problems, or only the appearance of them? Is it steerable by the humans notionally responsible for it, or has complexity already outrun that capacity?

A system can score well on capability benchmarks while being poorly governed, poorly self-correcting, or structurally captured by incentives running counter to its stated mission — the viability lens is built precisely to surface that gap, which a pure-capability framing has no vocabulary for at all.

The recursion idea offers a genuinely different, in-principle testable research program as well: rather than treating "how do we scale this model" as the only serious engineering question, it reframes the question as "how do we construct a nested structure of smaller, independently viable systems, such that viability holds at every level of the recursion, not merely at the aggregate top." That's a different governance and architecture agenda than the scaling-laws framing that currently dominates, and one worth taking seriously independent of whatever one concludes about the specific Anthropic critique.

A note on dates and evidence

This is a 2026 paper from Salford Business School's *Pathways to AGI* series, written from an explicit critical-software-studies perspective rather than a neutral descriptive one. Its VSM-application argument draws on well-established management-cybernetics theory (Beer's original work spans the 1970s–80s); its empirical claims about Anthropic specifically — the voluntary status of the RSP, the nature and degree of Amazon/Google investment influence, the significance of the February 2026 Pentagon episode — should be treated as the paper's own argued position, not independently verified facts, and any use of these specific claims elsewhere should involve checking current primary sources rather than citing this paper's account as settled.

This document is a companion to a short introductory piece on the same paper. Comments, corrections, and suggestions for what to read next belong there: https://petroskowo.pl/2026/10/13/judging-ai-by-whether-its-alive-not-whether-its-smart/.

License: CC-BY-SA 4.0 — Statek Umysł