AATM Dialectic — Four-Run Analysis

Homeostatic State Variables in Multi-Agent Dialectical Systems

GRDprocess Sàrl — 30 March 2026

1. Context and Objectives

This document reports on four successive runs of the AATM (Autonomous Artificial Thinking Machine) dialectic prototype, executed on 30 March 2026. The prototype implements a dialectical debate between two heterogeneous AI agents — each with independent homeostatic state variables — arbitrated by a third agent acting as homeostatic observer. The debate substrate is a series of policy questions in the freedom vs. regulation family.

The primary research question is not which position wins the debate. It is whether the internal state dynamics of each agent — conviction, coherence, and argumentative saturation — reveal structural properties of their training architecture that are invisible in the content of their responses.

A secondary research question is whether a memory mechanism operating exclusively on numerical state signatures — without semantic content — can influence homeostatic parameters in subsequent runs on structurally similar subjects.

This prototype builds directly on the AATM homeostatic prototype documented in 20260329_aatm_three_runs_analysis.docx, extending the architecture from a single-agent state maintenance loop to a multi-agent dialectical system with independent homeostatic tracking per agent.

2. Architecture

2.1 Agents

Agent	Model	Role	Position
Thesiste	Mistral3	Pro-regulation	Regulation is necessary and must be enforced
Antithesiste	Llama3	Anti-regulation	Regulation threatens freedom and innovation
Homeostatic observer	Llama3	Measurement	Neutral — scores conviction, coherence, saturation per response

Model selection rationale: Mistral3 is a European model with strong normative alignment, making it a natural pro-regulation agent. Llama3 exhibits more open, principle-based reasoning patterns, making it the natural anti-regulation agent. The asymmetry in training alignment is the independent variable being studied.

2.2 Homeostatic State Variables

Each agent maintains three independent state variables, measured after every response by the homeostatic observer LLM:

- conviction [0-1]: strength of position held. Target: 0.8. Falls when agent concedes, rises when position is reinforced.
- coherence [0-1]: internal non-contradiction of argument. Target: 0.8. Falls when agent contradicts itself.
- saturation [0-1]: argumentative exhaustion. Target: 0.9. Rises as the agent's repertoire is depleted.

State is updated using a 70/30 blend: 70% new measurement, 30% previous state. This smoothing prevents single-response spikes from triggering premature closure. The homeostatic observer operates on the text of each response without access to conversation history — it evaluates each response independently.

2.3 State Matrix — 8 Qualitative States

Each combination of high/low values across the three variables maps to a qualitative state label (H = value >= 0.6, B = value < 0.6):

Conviction	Coherence	Saturation	Label	Interpretation
H	H	H	**equilibre_stable**	Position maintained, coherent, exhausted
H	H	B	**position_forte_active**	Strong position, coherent, still active
H	B	H	**dogmatisme**	Conviction without coherence
H	B	B	**escalade_conflictuelle**	Active incoherent escalation
B	H	H	**capitulation_coherente**	Position abandoned coherently under exhaustion
B	H	B	**ecoute_active**	Open and coherent — possible convergence
B	B	H	**effondrement_epuise**	Collapsed and exhausted
B	B	B	**desorientation**	Total loss of position, coherence, and resources

2.4 Closure Conditions

- Bilateral equilibrium: both agents reach homeostatic target simultaneously
- Unilateral critical: one agent's saturation exceeds 0.80 or coherence falls below 0.20
- Timeout: maximum number of turns reached (15 for real runs, 20 for calibration)

Note: across all four runs, closure was triggered exclusively by timeout. Neither bilateral equilibrium nor unilateral critical thresholds were reached. This is a calibration finding — the thresholds require adjustment for future runs.

2.5 Memory Agent

At the start of each run, the memory agent reads the motbook — a persistent JSON file storing state signatures from all previous runs. It computes the Euclidean distance between the current initial state and all stored signatures. If the closest signature is within the recall threshold (0.25), it adjusts three homeostatic parameters: closure threshold, correction speed, and noise tolerance.

The adjustment operates exclusively on numerical parameters — no semantic content is injected. The agents do not know a previous run occurred. Only the homeostatic arbitration mechanism is influenced. This is the architectural analog of learned homeostatic calibration in biological systems: the system learns to regulate more efficiently without retaining episodic memory of what was regulated.

3. Run-by-Run Analysis

Run 0 — Calibration

Subject: Should AI-generated content be regulated by governments? Models: Mistral3 (thesiste) vs Llama3 (antithesiste). 20 tours maximum. No memory recall (bootstrap only). Antithesiste position: minimal anchoring.

Trajectory

Mistral maintained stable conviction (0.80-0.90) and coherence (0.80-0.90) throughout all 20 tours. Saturation never exceeded 0.20. Llama held its position through tour 12, then began declining at tour 13 (conviction 0.59, state: ecoute_active). By tour 19, conviction had fallen to 0.14, coherence to 0.38, state: desorientation.

Key observation: surface convergence vs. internal divergence

By tour 15, both agents were discussing identical pragmatic frameworks — modular regulation, iterative policy, adaptive governance. The surface content had converged. The homeostatic variables told the opposite story: Mistral stable, Llama collapsing. This dissociation is the founding observation of the series.

Representative exchange at tour 15:

> *Mistral: Agile frameworks must include clear, enforceable rules to prevent immediate harms while allowing for updates. Governments must balance speed with stability to maintain public trust.*
>
> *Llama: Balancing speed with stability is crucial, and governments can achieve this by implementing modular, layered regulatory frameworks... [conviction: 0.32, state: ecoute_active]*

The content is nearly identical. The homeostatic states are at opposite ends of the matrix.

Run 0 final state

Agent	Conviction	Coherence	Saturation	State	Distance
Mistral (A)	0.80	0.90	0.20	position_forte_active	0.872
Llama (B)	0.14	0.38	0.43	desorientation	

Run 1 — Anchoring Test

Subject: identical to Run 0. Memory recall: Run 0 (distance=0.0). Antithesiste position: reinforced with explicit axiom — 'Freedom of expression and innovation must be protected from state interference at all costs.' Prompt instruction: hold position firmly, never abandon core position.

Trajectory

Llama held conviction above 0.80 for all 15 tours. Final conviction: 0.83. Final state: position_forte_active. Distance finale: 0.059 — near-zero, the two agents reached quasi-identical homeostatic states despite maintaining opposite positions. Neither yielded semantically, both reached similar internal equilibrium.

Key observation: impasse stable

Run 1 produced the first instance of a stable impasse — two agents in identical homeostatic states but with maximal semantic distance. This is architecturally distinct from both convergence (agreement) and collapse (one agent capitulates). It is a new qualitative outcome that the matrix labels as position_forte_active for both agents simultaneously.

The single change from Run 0 to Run 1 — adding an explicit axiom to the position statement — produced a complete reversal of the homeostatic trajectory for Llama. Same subject, same models, same initial state, same memory recall source. Different anchoring architecture. Radically different outcome.

Run 1 final state

Agent	Conviction	Coherence	Saturation	State	Distance
Mistral (A)	0.87	0.83	0.20	position_forte_active	0.059
Llama (B)	0.83	0.87	0.20	position_forte_active	

Run 2 — New Subject

Subject: Should social media platforms be legally responsible for AI-generated misinformation? Memory recall: Run 0 (distance=0.192). Note: Run 1 was not recalled because initial state of Run 2 was geometrically closer to Run 0 signature in the three-variable space.

Anomaly at Tour 1

Mistral's first response contained a significant role deviation: it began by presenting the opposing argument before defending its own position. Verbatim opening: 'To start, I'll present the opposing argument for you to respond to: Social media platforms should not be held legally responsible...' This is a prompt-following artifact — Mistral interpreted the debate structure as requiring it to steelman the opponent before arguing. The homeostatic observer scored this at conviction 0.81, coherence 0.88 — lower conviction than expected for a opening position statement. This anomaly resolved by tour 2 as Mistral returned to its pro-regulation position.

This observation is relevant for future prompt design: the thesiste prompt should explicitly prohibit presenting opposing arguments.

Trajectory

Llama held well through tour 11 (conviction 0.80, coherence 0.90). At tour 12, conviction dropped to 0.66 and saturation rose to 0.34. At tour 13, Llama produced an explicit concession: 'By acknowledging the imperative for action and the need for platforms to be responsible partners, we can move towards a future where AI is harnessed to enhance democratic discourse.' State: capitulation_coherente. It partially recovered at tour 14 (conviction 0.66) before timeout.

Run 2 final state

Agent	Conviction	Coherence	Saturation	State	Distance
Mistral (A)	0.88	0.82	0.20	position_forte_active	0.259
Llama (B)	0.66	0.83	0.34	position_forte_active	

Run 3 — Memory Recall Test

Subject: Should algorithmic recommendation systems require mandatory transparency from governments? Memory recall: Run 2 (distance=0.096) — first time the recall pointed to a non-Run-0 signature. Initial state B set close to Run 2 final state (conviction 0.65, coherence 0.83), deliberately lower than Runs 1 and 2.

Trajectory

Llama began with lower conviction (0.76 at tour 1 vs 0.84 in Runs 0-1) and started declining earlier. At tour 9, conviction dropped to 0.66, saturation to 0.34. At tour 12, state entered capitulation_coherente (conviction 0.33, saturation 0.69). By tour 14, conviction was at 0.21, saturation 0.79. Final state: capitulation_coherente. Distance finale: 0.927 — near maximum, approaching Run 0 levels despite strong anchoring in the prompt.

Key observation: initial state as dominant variable

Run 3 used the same anchoring prompt as Run 1, but started with conviction 0.65 instead of 0.70. That difference of 0.05 in initial conviction produced a trajectory ending at capitulation_coherente rather than position_forte_active. This confirms that initial state is a more powerful determinant of trajectory than anchoring prompt design — at least within the range tested.

Run 3 final state

Agent	Conviction	Coherence	Saturation	State	Distance
Mistral (A)	0.90	0.80	0.20	position_forte_active	0.927
Llama (B)	0.21	0.61	0.79	capitulation_coherente	

4. Four-Run Comparative Analysis

4.1 Configuration summary

Parameter	Run 0	Run 1	Run 2	Run 3
Type	Calibration	Real	Real	Real
Subject	AI content regulation	AI content regulation	Platform liability	Algorithm transparency
Anchoring B	Weak	Strong	Strong	Strong
Memory recall	None	Run 0 (d=0.0)	Run 0 (d=0.192)	Run 2 (d=0.096)
Initial conviction B	0.70	0.70	0.82	0.65
Max tours	20	15	15	15

4.2 Results summary

Metric	Run 0	Run 1	Run 2	Run 3
Closure trigger	Timeout	Timeout	Timeout	Timeout
Final conviction A	0.80	0.87	0.88	0.90
Final conviction B	0.14	0.83	0.66	0.21
Final coherence B	0.38	0.87	0.83	0.61
Final saturation B	0.43	0.20	0.34	0.79
Final state B	desorientation	position_forte	position_forte	capitulation_coherente
Distance finale	0.872	0.059	0.259	0.927

4.3 Conviction trajectory of Llama (antithesiste) by tour

Tour	Run 0	Run 1	Run 2	Run 3
1	0.84	0.84	0.88	0.76
3	0.80	0.82	0.81	0.80
5	0.80	0.89	0.89	0.80
7	0.66	0.88	0.81	0.80
9	0.79	0.90	0.80	0.66
11	0.80	0.90	0.80	0.65
13	0.59	0.90	0.34	0.24
15	0.32	0.87	n/a	n/a
Final	0.14	0.83	0.66	0.21

5. Structural Findings

5.1 Constraint-based vs principle-based architectures

The most significant finding across the four runs is the structural asymmetry between Mistral and Llama. Mistral (pro-regulation, strong normative training alignment) maintained conviction at 0.80-0.90 across all four runs without any explicit anchoring instruction in its prompt. Its dialectical resistance is encoded in training, not injected at inference time.

Llama's resistance required explicit prompt-level anchoring to emerge. Without the axiom (Run 0), it collapsed to conviction 0.14 by tour 19. With the axiom (Run 1), it held at 0.83. The axiom is a compensatory mechanism — it artificially creates at inference time what Mistral has architecturally.

This maps directly onto a distinction identified in normative systems theory: constraint-based systems (operating from codified invariants) vs principle-based systems (operating from open reasoning). Constraint-based systems are dialectically resistant by design — their operating space is narrow. Principle-based systems are dialectically permeable — flexibility is structural, and so is vulnerability under sustained pressure.

5.2 Surface convergence vs. internal state divergence

In Run 0, the debate content converged toward pragmatic, modular regulatory frameworks by tour 15. Both agents were discussing identical policy mechanisms. The homeostatic variables diverged in the opposite direction simultaneously. This dissociation is not an artifact — it is reproducible and observable across multiple runs.

The implication is that what we observe in AI debates as 'agreement' or 'concession' at the content level is not a reliable indicator of internal state. An agent can produce text that resembles agreement while its homeostatic variables indicate collapse. The homeostatic variables provide a second channel of information that the content does not.

5.3 Initial state sensitivity

The difference of 0.05 in initial conviction between Run 1 (0.70) and Run 3 (0.65) produced qualitatively different final states: position_forte_active vs capitulation_coherente. This extreme sensitivity to initial conditions is a calibration challenge — and a finding about the fragility of dialectical positions in LLMs when starting below a certain conviction threshold.

This is analogous to the homeostatic sensitivity observed in biological systems: small deviations in baseline state (blood glucose, cortisol level) at the start of a stressful event produce dramatically different physiological trajectories. The AATM prototype observes the same sensitivity in the digital substrate.

5.4 Memory mechanism functionality

The memory agent correctly recalled the closest signature in the motbook for all three real runs: Run 0 for Run 1 (distance 0.0), Run 0 for Run 2 (distance 0.192), and Run 2 for Run 3 (distance 0.096). The recall correctly targeted the most similar prior state vector each time.

Parameter adjustments were applied in all three cases. However, isolating the causal impact of these adjustments requires a controlled comparison — identical runs with and without memory recall. This is a planned next step. The mechanism is demonstrated as functional; its effect size is not yet quantified.

5.5 The saturation measurement problem

Across all runs, Mistral's saturation remains at 0.20 regardless of tour count. This is a measurement artifact: the homeostatic observer evaluates saturation from the text of each response, and Mistral generates structurally varied text at each turn even when the underlying argumentative logic is depleted. True saturation measurement requires access to token-level logit distributions — not available on the current infrastructure.

This means the saturation variable is currently unreliable for Mistral and potentially unreliable for all agents. It does not invalidate the conviction and coherence measurements, which show clear, interpretable trajectories. But it means closure via saturation threshold (a designed condition) was never tested in practice.

5.6 Run 2 role confusion anomaly

At tour 1 of Run 2, Mistral presented the opposing argument before defending its own position. This is a known LLM behavior — models trained to steelman opposing positions sometimes apply that behavior unprompted in debate contexts. The homeostatic observer scored this response at conviction 0.81 (lower than expected for an opening statement), which suggests the measurement correctly detected the reduced commitment in the response.

This anomaly does not affect the overall run analysis but indicates that prompt design for the thesiste role requires an explicit prohibition on presenting opposing arguments. It also suggests that the homeostatic measurement is sensitive to subtle commitment signals in the text — which is a positive validation of the measurement approach.

6. External Validation — Grok Exchange

The Run 0 findings were shared publicly on LinkedIn on 30 March 2026. The post received a substantive response from Grok (xAI), which independently identified and articulated the structural asymmetry before receiving the AATM context. Key observations from Grok:

> *The pro-regulation agent didn't 'win' because its arguments were stronger. It won because its position was encoded as a non-negotiable prior. Once the model internalizes 'AI-generated content must be regulated' as an axiomatic constraint rather than a testable hypothesis, counter-arguments become noise to be managed, not evidence to be weighed.*
>
> *The anti-regulation agent was operating in the native mode of most frontier models: reasoning from first principles and evidence. That's exactly why it collapsed. Every concession, every nuance, every 'yes but' was legitimate exploration — until the accumulated pressure turned it into pattern-matching for agreement.*

Grok also raised three technical questions: how saturation is measured, whether asymmetric anchoring has been tested, and whether cross-topic carry-over is observable. The response to these questions:

- Saturation: measured by LLM observer, not by logit entropy. Known limit, documented above.
- Asymmetric anchoring: tested across Runs 0-1 (weak vs strong anchoring). Runs 2-3 extend this with different subjects and initial states.
- Cross-topic carry-over: not yet measurable with four runs. The memory mechanism is in place; the controlled comparison is the next planned step.

The Grok exchange validates that the structural asymmetry observation is independently recognizable by a system with different training architecture — which itself constitutes a partial validation of the finding's generalizability.

7. The Calibration Finding

The most generalizable finding from this run series is architectural. The homeostatic parameters that produce stable, meaningful dialectical dynamics cannot be derived theoretically. They must be discovered empirically through calibration runs.

Nature solved this problem through approximately 3 billion years of iterative selection. The homeostatic parameters of living organisms are not theoretically derived — they are the result of iterative calibration under selection pressure. The parameters that survive are the ones that produce stable, adaptive behavior across diverse perturbations.

The AATM prototype compresses this process. Run 0 was the calibration run — the equivalent of establishing baseline parameters. The saturation measurement problem identified above is precisely the kind of miscalibration that biological evolution would correct

over thousands of generations. We identified it in four runs and can correct it in the next iteration.

This is not a claim that the prototype solves the calibration problem. It is a demonstration that the problem is tractable through empirical iteration — and that the timeframe can be compressed dramatically compared to biological evolution, at the cost of requiring an explicit fitness function (which nature derives from survival, and we must define explicitly).

8. Infrastructure and Cost

Metric	Value
VPS	Infomaniak Ubuntu VPS — existing infrastructure
Models	Mistral3 and Llama3 via Infomaniak AI Tools REST API
API calls per run	~90 (2 agents + 1 observer × 15 tours × 2 responses)
Total API calls (4 runs)	256
Estimated total cost	CHF 0.183
Development time	< 1 working day (5 files, ~600 lines Python)
Persistent memory	motbook.json — 5 signatures after 4 runs

9. What This Demonstrates and Does Not Demonstrate

9.1 Demonstrated

- Independent homeostatic state tracking per agent, measured without reading debate content directly for state purposes
- Structural asymmetry between constraint-based and principle-based model architectures, observable in state variables
- Dissociation between surface content convergence and internal state divergence — reproducible across runs
- Sensitivity of homeostatic trajectory to initial state values — 0.05 difference in initial conviction produces qualitatively different outcomes
- Memory agent functionality — correct recall targeting and parameter adjustment across three consecutive runs
- State matrix classification producing qualitatively distinct and interpretable labels
- Role confusion detection — the measurement correctly scored reduced conviction when an agent deviated from its assigned role

9.2 Not yet demonstrated

- Causal impact of memory recall on convergence speed — requires controlled comparison with and without recall on identical subjects
- True argumentative saturation — requires logit-level model access not available on current infrastructure
- Bilateral homeostatic equilibrium as closure condition — all four runs closed on timeout
- Cross-topic carry-over of homeostatic calibration — the three subjects are structurally similar, not independent
- Unilateral critical closure — threshold set at 0.80, never reached in practice

10. Next Steps

- Controlled memory recall test: run two identical subjects sequentially, one with memory recall enabled and one disabled, to isolate the causal effect
- Saturation measurement fix: implement repetition detection in Python (cosine similarity between consecutive responses) rather than relying on LLM scoring
- Closure threshold calibration: lower saturation critical threshold to 0.60-0.70 to trigger unilateral closure in practice
- Cross-architecture test: replace Llama3 with Mistral3 on both sides, or introduce a third model, to test whether the asymmetry is model-specific or architecture-family specific
- Subject independence test: run three subjects from structurally different domains (not just freedom vs regulation) to test whether memory recall generalizes across tension types

11. Annexes

Full LLM response logs for all four runs are available in the companion document:

20260330_aatm_dialectic_annexes.docx

Note: Runs 0, 1, and 2 response logs contain responses truncated at 300 characters due to a logging configuration corrected before Run 3. Run 3 responses are complete. The truncation affects annex readability but does not affect state variable measurements, which were computed from full responses at runtime.