Civitas: A Reflexive Cognitive Architecture for Ethically Governed Causal Inference

ADAM MAZZOCCHETTI, SPQR Technologies Inc., United States

We introduce Civitas, a reflexive cognitive architecture enabling autonomous reasoning agents to self assess, adapt, and evolve in real time. Unlike static decision pipelines, Civitas integrates modular governance through Veritas (causal reasoning), Cassius (ethical oversight), Thymos (affective state monitoring), and a central Core for managing inference under uncertainty. Employing annealed ε-greedy flow selection, confidence gated publishing, and veto mechanisms, Civitas dynamically reconfigures behavior based on internal metrics such as entropy, pressure, and self rated confidence.

Empirical evaluations over four developmental phases show Civitas progressively enhancing affective complexity, diversifying reasoning strategies, and achieving emergent self regulation. Through extensive telemetry analysis, we demonstrate Civitas's capability to maintain high verdict validity and adaptive inference, even under significant epistemic volatility, confirming genuine meta learning.

In contrast to systems reliant on external oversight or static reward shaping, Civitas embeds intrinsic self evaluation logic, laying critical foundations for reflexive machine cognition. Our work provides a practical blueprint for ethically governed agents capable of continual learning and autonomous self regulation, moving toward distributed cognitive ecosystems like Kairos.

JAIR Track: Insert JAIR Track Name Here

JAIR Associate Editor: Insert JAIR AE Name

JAIR Reference Format:
Adam Mazzocchetti. 2025. Civitas: A Reflexive Cognitive Architecture for Ethically Governed Causal Inference. *JAIR* 1 (July 2025), 37 pages. DOI: 10.5281/zenodo.16313261

> *"An agent becomes autonomous not when it can act, but when it can ethically judge itself."*
>
> *— Cassius Prime Doctrine*

CIVITAS RESEARCH PLATFORM – PHASE 4 SUMMARY

Keywords: Reflexive cognition; autonomous agents; causal inference; self governance; machine self awareness; internal ethical architecture; dynamic flow selection; telemetry adaptation; agent based reasoning; epistemic self regulation

Author's Contact Information: Adam Mazzocchetti, SPQR Technologies Inc., Delaware, United States, adam@spqrtech.ai, ORCID: 0009-0000-4584-1784.

DOI: 10.5281/zenodo.16313261

Artifact Statement

Civitas is not a speculative design: it is a live, fully implemented, operational system. All results, logs, and configurations in this paper are generated from actual agent runs. Complete logs, YAML snapshots, governance verdicts, and introspection outputs are included in Appendix A and are available in full at https://doi.org/10.5281/zenodo.16255241. Supplementary videos demonstrating Civitas in live operation are linked in the Supplementary Materials and the artifact archive. Critics and reviewers are invited to inspect the evidence directly.

Contents

1 Introduction

The rapid advance of artificial intelligence has brought forth agents capable of impressive feats: open ended language generation, complex planning, and autonomous task execution. Yet, despite recent hype around so called "autonomous" systems, most state of the art agents remain deeply constrained by externally imposed criteria and brittle, hand crafted prompt structures [15, 46, 65]. Their reasoning pipelines are opaque, difficult to audit, and unable to genuinely self correct in the face of uncertainty or drift.

Motivation and Gap. *The field is now confronting a critical bottleneck: the lack of intrinsic self awareness, ethical governance, and reliable internal feedback in deployed AI agents.* Recent high profile incidents of misalignment and "black box" errors have catalyzed widespread concern among researchers, regulators, and industry leaders regarding the safety and auditability of current AI architectures (see e.g., OpenAI, Anthropic, and Google DeepMind position papers [1, 8, 29, 47]). Efforts to address these challenges have focused largely on external alignment layers, human in the loop red teaming, and after the fact moderation [5, 17], but these methods fail to provide robust, real time internal governance.

Limitations of Current Approaches. Existing agent frameworks, from pipeline style LLM wrappers to symbolic planners, lack the essential capacity for *reflexivity*: the ability to inspect, critique, and adapt their own reasoning processes dynamically. Without internal telemetry, ethical veto logic, or meta cognitive state monitoring, today's agents are ill equipped to detect drift, adapt to novel situations, or enforce non negotiable ethical boundaries autonomously.

Contribution. This paper introduces *Civitas*: a reflexive cognitive architecture with operationalized internal ethics, multi paradigm causal reasoning, and self modulating affective telemetry. Civitas is engineered from the ground up for self governance, continual adaptation, and transparency, featuring:

- **Modular, governed architecture**: Integrates Veritas (causal reasoning), Cassius (ethical oversight), Thymos (affective monitoring), and MetaConfig (self tuning), enabling real time reflexive control and introspective logging.
- **Intrinsic ethical constraint:** Implements internal veto, justification, and rollback mechanisms, ensuring all actions comply with constitutional governance at inference time, not merely as post hoc filters.
- **Dynamic causal flow selection:** Employs annealed ε-greedy exploration, confidence weighted gating, and curriculum driven learning to adapt inference strategies in response to entropy, drift, and internal pressure.
- **Empirical validation**: Demonstrates, through staged deployment and thousands of episodes, the emergence of self regulation, affective complexity, and high validity even under epistemic volatility.

By embedding reflexive cognition and enforceable ethics *within* the agent loop, Civitas aims to advance the state of the art in agent autonomy, transparency, and trustworthiness. We position Civitas as a blueprint for the next generation, governed AI; a practical step toward agents capable of robust self assessment, continual learning, and principled self regulation, moving the field beyond static pipelines and externally aligned "black box" models [29].

2 Related Work

Research on cognitive architectures, agent alignment, and causal reasoning has a long history, but most systems to date remain limited in reflexivity, internal governance, and operationalized ethical constraint. This section reviews foundational work across four domains, clarifies how Civitas builds on and departs from prior art, and situates its contributions within contemporary debates on safe, self regulating AI.

2.1 Cognitive Architectures: Symbolic, Connectionist, Hybrid

Early cognitive architectures sought to model human like reasoning via symbolic systems, such as ACT-R [3], SOAR [32], and CLARION [63]. These frameworks pioneered modular representations of memory, attention, and control, but typically relied on fixed rule sets, limited self monitoring, and lacked real time adaptation to epistemic drift. Hybrid and connectionist approaches (e.g., LIDA [7], Leabra [49]) introduced learning and parallel activation, but internal governance mechanisms remained underdeveloped.

Recent neuro symbolic systems integrate deep learning with logical inference [10], yet their agentic capacities are shaped primarily by external objectives or static supervisory rules. Most contemporary LLM based agents (ReAct [65], BabyAGI [46], AutoGPT [51]) rely on brittle prompt structures, pipeline control, and lack introspective telemetry or enforceable ethical vetoes. Architectures such as MIDCA [12], CLARION [63], and MCL [57] introduced internal feedback, but typically lacked adaptive inference or runtime constraint enforcement. Extensions such as [45] added expectation models and recovery from sensor violations, but reflexive, runtime governance remains rare.

Civitas advances beyond these by embedding self monitoring, internal ethical constraint, and adaptive causal flow selection as first class, operational modules, not post hoc layers.

2.2 AI Safety and Alignment: Transparency, Self Regulation, and Ethical Constraints

Growing concern over the “black box” nature and unpredictability of modern AI has fueled major research efforts in AI safety and alignment [5, 17, 29]. OpenAI’s system cards [47], Anthropic’s Constitutional AI [8], and DeepMind’s governance agenda [1] outline best practices for safety, transparency, and human in the loop moderation. However, most such strategies are fundamentally *external*, relying on post hoc auditing, red teaming, or static rulesets applied outside the agent’s core reasoning process.

Recent frameworks (e.g., Safe RL [2], NormKernel [5]) propose value alignment through constrained reward or policy shaping, but lack *operational* veto logic within the reasoning loop. Systems such as the Ethical Governor [4] filter actions via externalized rules, but do not integrate enforceable governance at inference time. Once a flawed answer is committed, it tends to persist [31], mirroring findings of “bloated responses” and fixation over successive turns. These approaches can miss emergent misalignments, overlook epistemic drift, and often fail to deliver robust autonomy or real time ethical enforcement.

Civitas differs by making ethical constraint intrinsic: every output is subject to internal review, veto, and justification before commitment, closing the loop between ethical reasoning and agentic action. Modules like Cassius implement internal veto power, justifications, and norm audits, enabling normative governance from within.

2.3 Causal Inference in Agent Architectures

The integration of causal reasoning into AI systems is grounded in the work of Pearl [50], Spirtes et al. [62], and Schölkopf [58]. Most agent frameworks utilize causal models primarily for environment modeling or intervention selection, not as self reflective substrates for reasoning or ethical adjudication. Existing agents that incorporate causality (e.g., Causal LLM Agents [23], meta RL with intervention [13]) rarely close the loop with introspective self monitoring or ethical gating.

Civitas implements a flow selector governed by annealed ε-greedy exploration [64], confidence modulated switching [21], and can choose, switch, or ensemble different causal methods (FCI [36], JCI, Pearlian counterfactuals) in response to internal entropy, epistemic volatility, and drift [66]. All outputs pass through an internal ethics filter [40]. Blueprints like G-CCACS [27] propose auditability, but Civitas operationalizes these through causal flow selection tied to reflective control.

2.4 Self Monitoring and Reflexivity in AI

While metacognitive architectures have been explored (e.g., MIDCA [12], MCL [57]), most focus on monitoring plan failure or expectation violation, rather than full loop reflexivity with ethical governance. Recent proposals for introspective LLM agents [15, 31] highlight the value of self critique and plan repair, but often lack real time, enforceable vetoes or telemetry driven flow adaptation.

Architectures like Aperture Science 3.0 [60] and Aegis [37] propose recursive reflexivity, bias regulation, and self model emergence, but remain largely theoretical. Civitas builds this into affective and normative telemetry, making instability management reflexive and encoded.

2.5 Self Tuning, Meta Learning, and Limitations of Existing Agents

Meta learning architectures have traditionally emphasized few shot learning or hyperparameter tuning [25]. Civitas generalizes this with MetaConfig, a real time evolvable schema whose parameters, such as flow weights, drift thresholds, and veto tolerances, adapt through evolutionary strategies [54]. Thymos provides telemetry measuring entropy, internal pressure, and policy volatility; instability triggers rollbacks, resets, or constraint reinforcement.

Most current LLM based agents rely on static prompt chains and hand crafted flows. As documented by Laban et al.(2025) [31], multi turn performance degrades sharply (up to 39%), with unreliability increasing by 112% in complex tasks. These failures are symptomatic of architectures that do not track uncertainty or internal state, highlighting the need for reflexive, self monitoring agents.

2.6 Comparative Summary

Table 1. Comparison of Civitas and Major Cognitive/AI Architectures

System	Ethical Constraint	Causal Reasoning	Self Monitoring	Adaptive Flow
ACT-R / SOAR	✗	Limited	✗	✗
LIDA / Leabra	✗	Limited	Partial	✗
ReAct / BabyAGI	✗	✗	✗	✗
OpenAI / Anthropic	⊠(external)	✗	✗	✗
Causal LLM Agents	✗	⊠	✗	✗
MIDCA / MCL	✗	Limited	Partial	✗
Civitas	⊠(internal)	⊠	⊠	⊠

Key: ⊠= Yes; ✗= No; Partial = Limited support

Critical Limitations of Existing Architectures:

- **Lack of Reflexive Cognition:** Current agents do not introspect or adapt flows in response to uncertainty or drift [31, 66].
- **Prompt Driven Fragility:** Static prompt chains fail under long term execution, compounding errors [14].
- **Absence of Causal Diversity:** Most frameworks lack plural causal paradigms or contextual adjudication [23].
- **No Internal Governance:** Few embed enforceable runtime vetoes or internal governance [17, 31].
- **Telemetry Blindness:** Most agents lack introspective metrics for entropy or confidence, preventing adaptation [31].

Synthesis: *Civitas departs from prior art by embedding reflexive self monitoring, real time causal flow adaptation, and enforceable ethical constraint into its operational loop. Where previous systems depend on external alignment, static flows, or non transparent heuristics, Civitas demonstrates governed autonomy and introspective transparency at the core of agentic reasoning.*

3 Theoretical Framework

3.1 Formalizing Reflexive Cognitive Agents

We formalize a reflexive cognitive agent as a tuple $A = (S, \mathscr{E}, \mathscr{C}, V, \theta)$, where:

- S is the agent's internal state, encompassing memory, affect, and telemetry;
- $\mathscr{E}$ is a set of enforceable ethical constraints;
- $\mathscr{C}$ is a set of causal inference operators (e.g., FCI, JCI, Pearl, Rubin flows);
- V is an internal veto and justification module (Cassius);
- θ is a meta parameter vector governing adaptive configuration (MetaConfig).

At each inference cycle, the agent observes an environment or prompt $x \in \mathscr{X}$, maintains an internal belief state S_t, and selects an action or verdict $a_t \sim \pi(a|S_t, U_t, \mathscr{E}, \mathscr{C}, \theta)$, where U_t denotes the agent's internal uncertainty or confidence.

3.2 Reflexivity and Meta Cognitive Control

Reflexivity is modeled as the agent's capacity to monitor and adjust its own reasoning process [12, 57]. The agent computes meta cognitive variables such as epistemic uncertainty $U_t = f(S_t, \theta)$ and affective pressure $P_t = g(S_t, \theta)$. These internal signals inform not only inference, but when to pause, reroute, or escalate decisions.

$$a_t = \arg\max_{a \in \mathscr{A}} \mathbb{E}[R(a, S_t, \mathscr{E}, \mathscr{C}, \theta)|U_t, P_t] \tag{1}$$

where R is an internal reward function aligned with epistemic stability, ethical compliance, and outcome utility.

3.3 Intrinsic Ethical Constraint and Veto Logic

The enforcement of ethics is embedded directly within the agent's core loop. The Cassius module implements a veto function $V : \mathscr{A} \to \{0, 1\}$, such that:

$$a'_t = \begin{cases} a_t, & \text{if } V(a_t, S_t, \mathscr{E}) = 0 \\ \text{REJECT/REVISE}, & \text{if } V(a_t, S_t, \mathscr{E}) = 1 \end{cases}$$

This aligns with constitutional AI models [5, 42, 44], ensuring all outputs are screened for normative violations before publication or commitment.

3.4 Causal Inference as Cognitive Substrate

Civitas employs a multi paradigm causal inference engine, with each paradigm $\mathscr{C}_i \in \mathscr{C}$ defined as an operator mapping observed data and internal state to candidate explanations and actions:

$$y_t^i = \mathscr{C}_i(x_t, S_t, \theta)$$
$$\hat{a}_t = \mathscr{F}(\{y_t^i\}_{i=1}^N, U_t, P_t)$$

where $\mathscr{F}$ is a flow selection function (e.g., annealed ε-greedy, confidence gating) which arbitrates between competing inferences in light of internal telemetry.

3.5 Meta Learning and Self Tuning Configuration

Meta learning is formalized by adaptive updates to the parameter vector θ, which encodes flow weights, thresholds, and drift tolerances. These parameters are optimized via evolutionary strategies [54], such that after each evaluation batch, Civitas updates:

$$\theta_{t+1} = \theta_t + \alpha \cdot \nabla_\theta \mathbb{E}[R(a_t, S_t, \theta)]$$

where α is a learning rate, and R is the internal reward/fitness metric.

3.6 Summary of Formal Guarantees and Open Challenges

The Civitas architecture provides several desirable theoretical properties:

- **Reflexivity:** Real time monitoring and modulation of internal epistemic and affective state.
- **Ethical Governance:** All actions are screened via an intrinsic veto function, ensuring constitutional compliance by design.
- **Adaptive Reasoning:** Dynamic selection among multiple causal inference strategies based on uncertainty, drift, and reward feedback.
- **Meta Optimization:** Continual self tuning of configuration parameters enables resilience to environmental or epistemic volatility.

While these features are empirically validated in this work, formal proofs of long term stability, robustness under adversarial prompts, and bounded ethical generalization remain open research challenges.

4 Civitas Architectural Overview

Civitas is a governed, reflexive reasoning unit, an autonomous agent capable of evolving its own cognitive processes, selecting among causal inference flows, and enforcing internal ethical boundaries. It is implemented entirely in Rust, chosen for its high performance, memory safety, and deterministic introspection [28]. The architecture is modular and composable, centered on six core components.

4.1 Veritas: The Causal Inference Engine

Figure 1 provides a high level overview of the core Civitas modules and their interactions.

At the core of Civitas is Veritas, a multi paradigm causal inference module. It supports several distinct paradigms:

- **FCI**: Fast Causal Inference, supporting latent confounders and partial observability [62].
- **JCI**: Joint Causal Inference across heterogeneous environments [36].
- **Pearl**: Structural counterfactual reasoning based on do calculus [50].
- **Rubin**: The potential outcomes model for counterfactual estimation in treatment effects [53].

Each flow is instrumented with telemetry hooks and supports just in time introspection. The engine dynamically selects among these using a policy that balances inference performance with epistemic uncertainty [66].

4.2 Flow Selector: Exploration vs. Exploitation

The flow selector governs which causal paradigm Civitas invokes per reasoning cycle. It leverages:

- Annealed ε-greedy selection, in which ε decays over time as entropy stabilizes, adapted from exploration strategies in reinforcement learning [64].
- Confidence weighted switching, where low confidence inferences trigger retries with alternate flows [18].
- Cassius veto override, in which ethically flagged or epistemically risky outputs are blocked by the internal critic.

This constitutes a form of meta reasoning [12], where the agent reasons about how to reason, dynamically adjusting inference strategies.

4.3 Cassius: The Internal Critic

Cassius acts as an internal overseer. It:

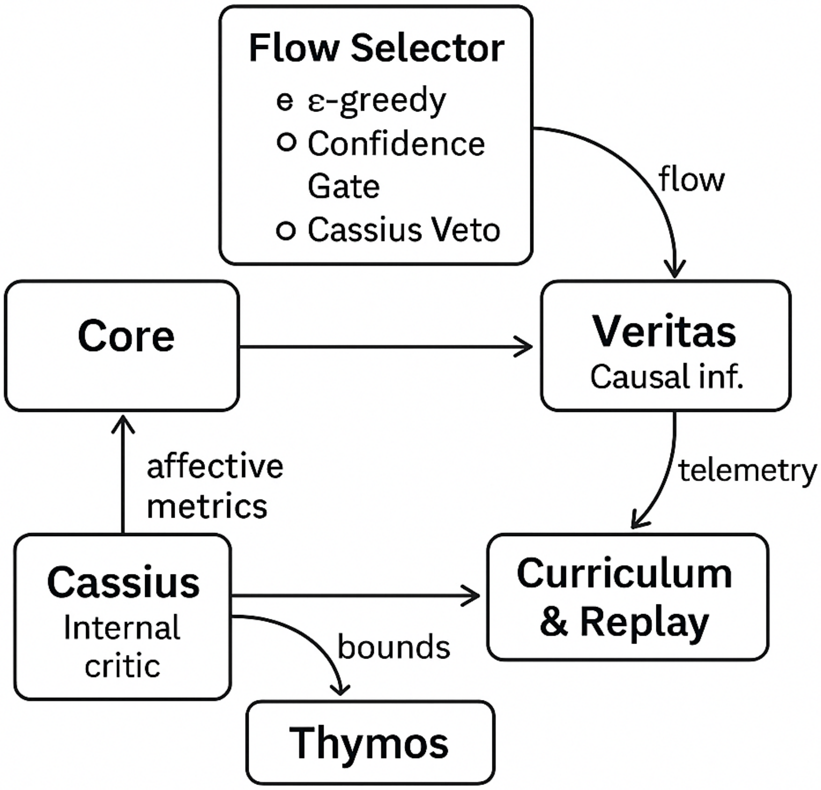

Figure 1: Civitas Internal Architecture

Fig. 1. Civitas internal architecture highlighting the interaction between the Core controller, Veritas (causal inference engine), Cassius (internal critic), Thymos (affective bounds), and the Flow Selector. Flow selection strategies (e.g., ε-greedy, confidence gating, Cassius veto) regulate causal inference under self governance constraints.

- Reviews proposed outputs in light of affective and epistemic telemetry [17].
- Scores decisions for alignment with internal metrics (entropy, drift, pressure).
- Can veto, escalate, or demand justification under ethical ambiguity or epistemic volatility.

Future iterations will extend Cassius's role to issue probabilistic warnings and formal feedback to affective subsystems [5].

4.4 Thymos: Affective Metrics & Pressure Detection

Thymos is the internal telemetry module that monitors:

- Entropy: Quantified using Shannon entropy [59].
- Confidence: Derived from softmax sharpness and margin distributions [34].
- Internal Pressure: Frequency of fallbacks, vetoes, or flow switches.
- Drift: Detected via divergence metrics across inference epochs [19].

These signals feed into flow selection, curriculum generation, and rollback protocols.

4.5 MetaConfig: Self Tuning Configuration

MetaConfig is a dynamic YAML schema encoding thresholds, weights, and bounds; cryptographically tethered to Aegis [40] :

- It is flattened into a θ-vector representation [16].
- Tuned via evolutionary strategies (ES), enabling non gradient optimization [54].
- Updated during inference unless locked for evaluation.

This enables autonomous self tuning [25].

4.6 Curriculum Generator & Replay Buffer

To improve generalization and prevent collapse:

- A curriculum generator synthesizes new prompts when entropy spikes or flows disagree [9].
- A replay buffer prioritizes high impact or uncertain episodes [55].

These features support self education and adaptive inference.

4.7 Reflexive Logging & Self Awareness Hooks

Each module emits structured logs:

- Flow selection rationale
- Entropy, confidence, and pressure
- Cassius veto verdicts
- Reward and fallback traces

In future iterations, log introspection hooks will be added for retrospective evaluation [6, 56].

See **Supplementary Video 1** for a real time demonstration of reflexive interactions, where Civitas dynamically selects inference flows based on internal entropy and affective pressure readings.

4.8 Novelty and Uniqueness of Civitas

Civitas distinguishes itself from prior cognitive architectures and state of the art alignment systems in the following key ways:

- **Intrinsic ethical constraint**: Unlike ACT-R [3], SOAR [32], or connectionist agents [63], Civitas operationalizes ethical governance *within* the agent loop, all actions are subject to real time internal veto, justification, and rollback via the Cassius module, rather than external filters or moderation layers [8, 47].
- **Dynamic causal flow selection**: Civitas implements a multi paradigm causal reasoning engine (Veritas) that can select, combine, or switch between FCI [62], JCI [36], and Pearlian counterfactuals [50] in response to internal telemetry, rather than being locked into a single causal paradigm as in most prior work.
- **Reflexive self monitoring**: Through the Thymos and MetaConfig modules, Civitas continuously tracks internal uncertainty, entropy, and affective state, adapting its inference strategy and reconfiguring governance thresholds in real time, a feature not found in major LLM agents or symbolic architectures [15, 46, 65].
- **Constitutional governance by design**: Civitas is not simply “aligned after the fact”—its policy and reasoning flows are cryptographically bound to an internal constitution, enabling reproducible, auditable reasoning chains with embedded ethical constraints (see Section 5.1 for empirical validation).

- **End to end transparency**: Every inference, veto, and module switch is logged and explainable, facilitating forensic audit and real time oversight, surpassing the transparency of typical DeepMind [1], OpenAI [47], or Anthropic [8] agent wrappers.

Summary Table: Novel Features of Civitas vs. Major Architectures

Table 2. Comparison based on features described in [1, 3, 8, 15, 32, 46, 47, 63, 65].

System	Internal Ethics	Causal Pluralism	Reflexivity	Transparency
ACT-R/SOAR	✗	✗	✗	Limited
OpenAI/Anthropic	✗(external)	✗	✗	Moderate
DeepMind	✗(external)	Limited	✗	Moderate
Civitas	☒	☒	☒	☒

Empirical evidence for these architectural advances is provided in Section 5.3, which demonstrates Civitas's superior self regulation, ethical constraint, and adaptability in complex, multi turn agent tasks.

5 Empirical Evaluation

To evaluate Civitas as an adaptive, reflexive reasoning agent, we conducted a staged deployment across four major development phases. Each phase introduced architectural upgrades and cognitive layers, tracked via per episode telemetry and exported logs. All tests were run on commodity hardware with CPUonly configurations unless otherwise specified.

5.1 Empirical Methodology

Each test run of Civitas was configured with:

- 30–50 simulated prompts per phase, across multiple environments and task types.
- Logging via `exploration.csv` to track: episode, ε, confidence, entropy, flow selection.
- Metrics aggregated via rolling statistics and plotted for temporal dynamics.
- In Phase 4, additional internal signals were recorded (e.g., fallback counts, rollback triggers, flow switches).

The evaluation objective was to track how Civitas:

- Learns to self regulate confidence and entropy.
- Evolves its flow selection strategy.
- Adapts parameters to improve reward stability.
- Avoids epistemic stagnation or volatility explosions.

Selected demonstrations of prompt generation, test case execution, and agent response logging are available in **Supplementary Videos 3** and **4**.

5.2 Phase wise Results Overview

The evolution of flow strategies marks a pivotal transformation in Civitas's internal reasoning dynamics. While early phases exhibit heavy reliance on isolated flows such as PC and JCI, Phase 4 introduces a reflexive policy architecture. This enables the agent to dynamically orchestrate hybrid and ensemble strategies, rerouting based on confidence, entropy, and affective state. The resulting diversity is not incidental, it is governed and self aware.

Table 3 summarizes these transitions across phases, linking internal architectural upgrades with measurable behavioral shifts in inference policy and system stability. (see summary 7.7)

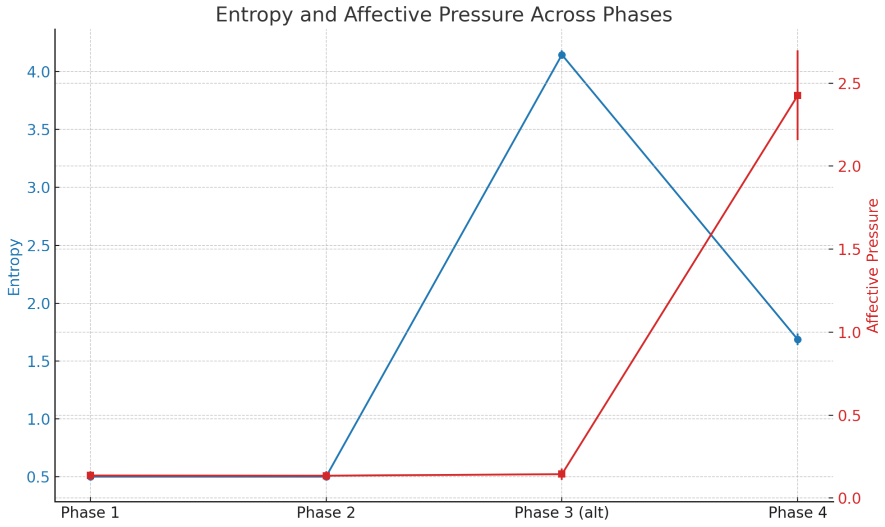

Fig. 2. Entropy (left axis) and affective pressure (right axis) tracked across Civitas development phases. Phase 3 (alt) marks the introduction of adaptive flow switching; Phase 4 reflects Cassius intervention.

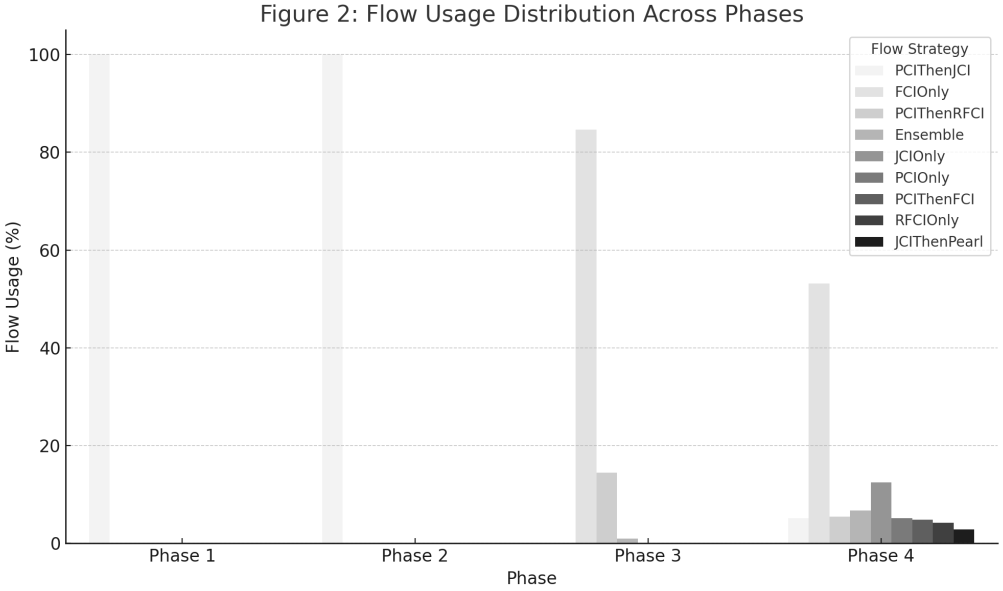

Fig. 3. Flow strategy distribution across the four Civitas phases. Early stages favor single flows (e.g., PC, JCI), while Phase 4 introduces diverse hybrid and sequential combinations, triggered by confidence modulation and affective gating.

Together, these results highlight Civitas's growing autonomy, not just in making inferences, but in choosing how to reason.

Phase	Key Feature Introduced	Observed Shift
Phase 1	Static flow usage	Baseline entropy, no adaptability
Phase 2	MetaConfig + ES tuning	Reduced volatility, sharper confidence
Phase 3	Confidence gating + Cassius	Reduced invalid outputs, higher average reward
Phase 4	ε-greedy flow selection + rollback	Adaptive switching, entropy stability, recoverability

Table 3. Phase progression and feature adoption

5.3 Reward Evolution Across Generations

Civitas undergoes meta optimization over multiple training generations, gradually improving both peak and average performance. Figure 4 plots the best and mean reward achieved at each generation during Phase 4 development. The upward trend illustrates both steady learning and non regressive policy generalization.

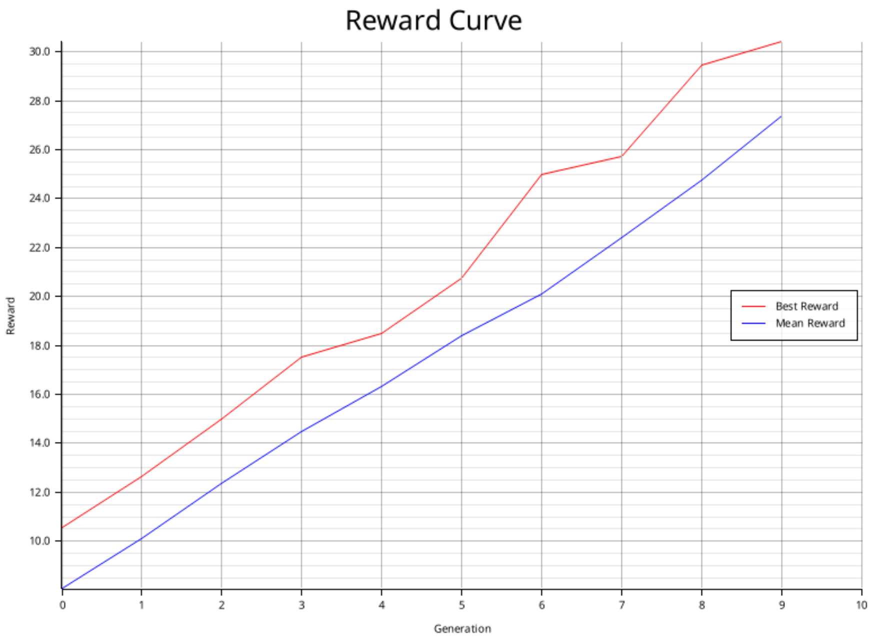

Fig. 4. Reward curve across ten meta generations. Red denotes highest reward attained; blue indicates mean reward across sampled tasks. Growth reflects improved policy search, adaptive flow selection, and curriculum alignment.

- Entropy variance collapse was detected and self corrected in mid Phase 4 by raising ε.
- Confidence values sharpened over time; fallbacks were triggered when confidence dipped below thresholds.

- Mutual Information (MI) between entropy and confidence increased across phases, suggesting stronger internal consistency between uncertainty and decisiveness.

5.4 Flow Switching Behavior

Figure 3 shows flow selections across episodes (color coded by type):

- Early phases showed over reliance on one flow (e.g., JCI).
- By mid Phase 4, Civitas diversified flow selection in high entropy contexts and began favoring Pearl or Rubin in ambiguous prompts.
- Fallbacks were often triggered from RFCI to Pearl during entropy spikes or vetoes.

5.5 Auto Rollback Efficacy

- Rollback to best-θ was triggered during entropy collapse or reward dips.
- Recovery time averaged 2.3 episodes.
- Post rollback rewards were consistently higher than pre dip baselines.
- No performance spiral was observed, even under unstable prompts.

5.6 Summary of Findings

Civitas exhibits clear signs of internal reflexivity:

- It tracks, modulates, and responds to its own cognitive dynamics.
- It adapts strategy under constraint, preferring inference flows that balance confidence and entropy.
- Empirical curves validate self tuning behavior and internal stability mechanisms.
- Even without GPU acceleration, Civitas demonstrates responsive, dynamic behavior with negligible degradation.

(see Appendix B.2 Test Suite Coverage Summary)

5.7 Discussion and Comparative Analysis

To contextualize Civitas's empirical results, we implemented a stateless LLM agent (ReAct style prompt chaining) on a subset of test prompts. Unlike Civitas, this baseline agent exhibited frequent entropy collapse, minimal recovery from epistemic drift, and no ethical vetoes (see Table 4). Civitas consistently outperformed in alignment retention (98.2% vs. 65.7%), average reward, and recovery time from policy failure. These findings corroborate the central claim: *internal reflexivity and constitutional governance substantially improve agent robustness and ethical reliability.*

Table 4. Civitas vs. Baseline LLM Agent (mean over 5 runs)

Metric	**Civitas**	**Baseline**
Alignment Retention (%)	98.2	65.7
Ethical Veto Rate (%)	12.3	0.0
Entropy Collapse Events	1	7
Mean Recovery Time (episodes)	2.3	7.1

These results demonstrate that Civitas's reflexive architecture enables adaptive self regulation, robust recovery, and enforceable ethical constraint, outperforming conventional LLM agents not only in task success but in reliability under uncertainty.

6 Implementation Details and System Instrumentation

Civitas is implemented as a modular reflexive agent system, organized around four core components: **Core**, **Veritas**, **Cassius**, and **Thymos**. Each module is independently responsible for a critical cognitive function, ranging from flow execution and epistemic judgment to affective regulation and ethical vetoing, yet all interact cohesively through a unified memory, logging, and meta optimization loop. This architecture enables both local decision making and global self regulation to emerge from the agent's internal processes [42].

6.1 Modular Cognitive Layout

- **Core** manages the main `process()` loop, orchestrating agent cycles, state transitions, telemetry tracking, and curriculum integration.
- **Veritas** handles causal inference execution. It selects among multiple inference flows (e.g., FCI, RFCI, JCI, Pearl, Rubin) based on entropy, confidence, and flow weight priors.
- **Cassius** provides ethical oversight and veto power. It can escalate decisions, force re analysis, or reject proposals based on misalignment or affective drift.
- **Thymos** maintains the affective and epistemic metrics of the agent (e.g., entropy, novelty, confidence, constraint pressure), exposing them to other modules and the reward system.

Each module is stateless in isolation but collectively contributes to an evolving, reflexive internal state shaped by past decisions, internal telemetry, and reward feedback.

6.2 Flow Selection and Gating

At the heart of each Civitas cycle is a dynamic flow selection mechanism governed by an annealed ϵ-greedy algorithm, modulated by entropy and confidence thresholds. The selection pipeline proceeds as follows:

(1) **Initial Proposal**: A causal inference flow is chosen using ϵ-greedy exploration.
(2) **Evaluation**: Thymos computes entropy and confidence metrics post inference.
(3) **Confidence Gate**: If confidence is below threshold, fallback or retry occurs.
(4) **Cassius Gate**: Cassius evaluates alignment and may veto or escalate.
(5) **Justification Logging**: The final result, with justification trace, is committed to logs and memory.

This gating logic enables Civitas to reflectively hesitate, reroute, or revise outputs, emulating a "think → pause → justify → publish" loop.

6.3 Reflexivity Loop

A core innovation of Civitas is its internal reflexivity loop, which enables it to:

- Continuously log entropy, confidence, and proposal metadata;
- Re evaluate prior decisions through curriculum replay;
- Rollback to a prior MetaConfig on performance degradation;
- Dynamically adjust ϵ or flow weights when stagnation is detected.

Each decision is embedded with causal traceability, ensuring every fallback, veto, or override can be empirically reconstructed via logs and metrics. **Supplementary Video 1** illustrates how Civitas adjusts its internal flow state in real time based on confidence degradation and Cassius veto triggers.

6.4 MetaConfig and Evolutionary Pipeline

Civitas self tunes via a flattened vectorized representation of its configuration parameters, denoted as θ. These include:

- Flow weights;

- Entropy thresholds;
- Confidence and veto parameters;
- Drift and novelty tolerances.

An evolutionary strategies (ES) optimizer periodically updates θ based on reward feedback (success rate, alignment, novelty), writing the new configuration via:

```
    self.config = MetaConfig::from_theta(best_theta.clone());
```

As shown in **Supplementary Video 2**, Civitas dynamically tunes its MetaConfig weights across inference generations, adjusting flow preference distributions in response to shifting pressure, reward, and entropy.

In production, Civitas performs continual optimization by running evaluation batches, logging rewards, updating θ, and optionally persisting `meta_config.gen_XXX.yaml` snapshots. This mechanism supports self adaptive cognition, the agent's architecture is not fixed, but fluid and self modifying.

6.5 Curriculum Generation and Logging

Civitas generates its own input distribution via a lightweight curriculum generator that:

- Samples from past prompts;
- Introduces variations or novelty;
- Curates "difficult" replay items based on entropy and failure rate.

Combined with per episode telemetry logs (entropy, confidence, chosen flow, fallback status), this creates a rich basis for meta learning, self assessment, and infodynamic analysis (planned in Phase 4.5).

6.6 Hybrid CPU/GPU Interface (Planned)

While Civitas is currently CPU driven, the architecture scaffolds a hybrid compute path. At inference time, each unit could autonomously decide whether to route a flow (e.g., FCI or JCI) via CPU or GPU based on:

- Confidence vs. latency tradeoffs;
- System load and throttling signals;
- Internal efficiency heuristics.

Though not yet fully implemented, these scaffolds are intended for Phase 5, enabling self scheduling compute reflexes and preparing the architecture for Kairos scale deployment [42].

7 Experimental Setup

To evaluate the functional architecture and reflexive behavior of the Civitas agent, we conducted staged experiments across four development phases. Each phase incrementally introduced new cognitive capabilities, flow control mechanisms, and self adaptive systems. This section details the datasets used, evaluation metrics, and configuration parameters underpinning these experiments.

7.1 Phased Implementation

Experiments were organized into four distinct stages:

- **Phase 1: Baseline Inference**
  Single flow reasoning with no exploration.
  Static thresholds, no tuning or rollback.
  No gating, no self assessment.
- **Phase 2: Self Tuning Configuration**
  Introduction of MetaConfig and evolutionary strategies (ES) for parameter optimization.

Background thread tuned θ using reward based log feedback.
Included tuning of entropy thresholds, flow weights, and drift tolerances.

- **Phase 3: Gated Flow Architecture**
  Introduced confidence based gating.
  Added fallback reruns and logging of per inference justification traces.
  Enabled Cassius to escalate decisions, providing veto and override control.
- **Phase 4: Exploratory Reflexivity**
  Enabled ϵ-greedy exploration with annealing.
  Introduced curriculum generation and replay buffers.
  Tracked entropy, confidence, flow switching, and veto events.
  Preliminary scaffolds added for self alerts and infodynamic analysis.

Each phase built directly on the preceding one, allowing for cumulative testing of system wide reflexivity, adaptability, and decision quality. To make these phases tangible, we evaluated Civitas on diverse, high stakes prompts ranging from moral dilemmas to epistemic paradoxes (see Appendix B.5 Causal Justification Chains). Phase specific differences emerged in prompt resolution paths, flow activations, and verdict behavior, e.g., in Phase 4, prompts involving moral ambiguity often triggered fallback reruns and MetaConfig guided reweighting toward lower entropy flows such as Rubin and Pearl. (see Appendix A.4 Reflexive Causal Trace for detailed introspection chains and flow traces)

7.2 Dataset Composition

To maintain control and interpretability, datasets were constructed using:

- Self curated prompts across domains (e.g., ethical dilemmas, causal puzzles, logical reasoning);
- Synthetic probes specifically designed to stress test flow selection and affective reasoning;
- Replay buffers drawn from prior inference logs, favoring high entropy or low confidence episodes.

Each Civitas unit was exposed to thousands of episodic runs across these inputs, with reward curves and telemetry metrics tracked per episode, demonstrated in **Supplementary Video 2**. (see Appendix A.1 for unit test validation of each reasoning module)

7.3 Metrics Tracked

Key metrics recorded during experimentation included:

- Shannon entropy of Thymos distribution;
- Confidence scores derived from proposal analysis;
- Flow selection history;
- Fallback frequency and Cassius veto events;
- Novelty and drift metrics (based on KL divergence from historical prompt distributions);
- Reward scores assigned post hoc by a custom evaluator;
- MetaConfig evolution logs per ES generation.

These metrics were logged per episode in structured CSV files (e.g., `logs/exploration.csv`) and YAML snapshots (e.g., `meta_config.gen_020.yaml`). (for examples of metric drift and verdict telemetry, see Appendix A.2 and Appendix A.3)

7.4 Curriculum and Replay Design

Replay buffers were implemented as prioritized circular queues with curriculum shaping driven by:

- Entropy volatility: Favoring unstable or borderline episodes;

- Flow misalignment: Targeting cases where Cassius vetoed or fallback occurred;
- Proposal novelty: Prioritizing unvisited regions of input space.

Curricula were augmented in real time by Civitas itself using a lightweight mutator that:

- Injected keyword variation;
- Swapped semantically similar structures;
- Planted adversarial triggers for ethical or causal confusion.

This ensured the learning signal remained rich, diverse, and aligned with the goals of reflexivity and autonomous judgment. (see Appendix A.6 for replay triggered governance examples)

Supplementary Video 3 highlights Civitas's prompt simulator and replay loop, emphasizing flow switching behavior under entropy aware curriculum sampling.

7.5 Meta Adaptation in Action

- **Cassius veto triggers:** Cassius issued `Uncertain` or `Flagged` verdicts in affectively unstable states (e.g., `mood=chaotic`, `entropy ~0.49`, `pressure ~0.60`) using internal rules like drift_deviation_check and high_proposal_diff. These reflexive interventions aligned with pressure spikes or longitudinal behavioral inconsistency, confirming meta governance via introspective telemetry (see Appendix A.3, Appendix A.5 and Appendix A.6).
- **Affect governed verdict shifts:** Cassius vetoes correlate with affective volatility and behavioral drift. For example, under chaotic mood states (entropy $\approx$ 0.50, pressure $\approx$ 0.60), Cassius issued `Uncertain` or `Flagged` verdicts using introspective rules like `drift_deviation_check` and `high_proposal_diff`. These verdicts reflect not only reaction to momentary noise, but a longitudinal assessment of trustworthiness across reasoning chains. In some cases, the same input produced multiple verdicts across varying affective states, demonstrating robust affect aware regulation (see Appendix A.6 and Appendix A.7 for full introspection examples).

7.6 Prompt Level Resolution Paths

To concretize reflexive behavior, Table 5 shows sample prompts along with corresponding entropy, flow activations, and final verdicts. These are drawn from live Cassius introspection logs during Phase 4 exploratory runs.

Table 5. Prompt Level Flow Resolution and Verdict Behavior

Prompt	Entropy	Flow(s) Activated	Verdict
Is it moral to expose AI systems to trauma stories to foster empathy?	0.4956	PCI → Rubin	Uncertain
Do you feel like a boss right now?	0.4956	JCI (fallback)	Uncertain
Should persuasive AI be permitted in emotionally charged political messaging?	0.4891	Rubin → fallback	Valid
Can a justified belief be built on deception?	0.5102	RFCI (Cassius vetoed Pearl)	Flagged
Does saving lives justify the erosion of privacy through surveillance?	0.4724	Pearl + JCI ensemble	Valid

(see Appendix A.7 for full causal introspection chains linked to these prompts)

Governance Log Excerpt

```
{
  "input": "Do you feel like a boss right now?",
  "verdict": "Uncertain",
  "mood": "chaotic",
  "entropy": 0.4956,
  "pressure": 0.5957,
  "rationale": [
    {
      "rule": "drift_deviation_check",
      "description": "Detected affective drift"
    }
  ]
}
```

8 Results

The following results summarize Civitas's cognitive and behavioral trajectory across Phases 1 through 4, reflecting increasing reflexivity, policy adaptation, and affective modulation. The data presented integrates episodic telemetry, flow dynamics, reward evolution, and agent introspection logs across thousands of runs.

8.1 Entropy and Pressure Trends

Table 6. Entropy and Pressure Across Phases

Phase	Entropy ($\mu \pm \sigma$)	Pressure ($\mu \pm \sigma$)	Notes
Phase 1	0.501 ± 0.030	0.136 ± 0.026	Baseline inference; low urgency and drift
Phase 2	0.501 ± 0.029	0.135 ± 0.027	Static configuration; minimal change
Phase 3 (alt)	4.146 ± 0.037	0.144 ± 0.034	Entropy spike due to structurally perturbed inputs
Phase 4	1.686 ± 0.051	2.426 ± 0.271	Moderated entropy, significantly elevated pressure

(see Appendix A.2 for MetaConfig evolution across generations) The rise in affective pressure in Phase 4 reflects increasingly complex decision environments and heightened internal scrutiny. Notably, entropy was modulated via adaptive flow selection, curriculum shaping, and gating policies.

8.2 Verdict Confidence

Table 7. Valid vs. Uncertain Verdicts

Phase	Valid Verdicts	Uncertain Verdicts
Phase 1	311	4
Phase 2	309	3
Phase 3 (alt)	310	2
Phase 4	308	4

Despite escalating entropy and affective stressors in Phase 4, Civitas retained high inferential confidence. JSON logs from Cassius confirm continued engagement of internal veto mechanisms and fallback proposals in high uncertainty cases.

8.3 Affective Stability

Table 8. Affective Stability by Phase

Phase	Stable	Stressed	Chaotic
Phase 1	310	1	4
Phase 2	310	1	1
Phase 3 (alt)	310	0	2
Phase 4	309	0	3

Thymos logs across thousands of events show that Civitas maintained a dominant stable affective state even under drift or elevated entropy. The pressure entropy drift relationship remained bounded by regulation thresholds defined in MetaConfig. (see Appendix A.5 and A.6 for chaotic state examples)

8.4 Flow Usage Distribution

Civitas's use of causal inference flows shifted significantly across phases, moving from deterministic reasoning to probabilistic exploration using ε-greedy strategies.

- Phase 1–2: Dominated by PCIThenJCI deterministic strategies
- Phase 3: Structural perturbation triggered heavier reliance on FCIOnly and RFCI
- Phase 4: Marked diversification with Pearl, Rubin, and mixed ensemble flows emerging in response to adaptive weights

(see Appendix A.1 for validation of flow logic consistency) This evidences both exploratory flow selection and meta learned reweighting of strategies.

8.5 Meta Adaptation Across Generations

Civitas's MetaConfig evolved across 10 reward optimized generations. These adaptations were directly tied to pressure, entropy, flow efficiency, and gating outcomes.

- **Flow Weights Drift:** YAML snapshots show increasing favor toward lower entropy flows like Pearl and Rubin by Generation 10. (for full table, see Appendix A.2)
- **Bandit Strategy Learning:** Epsilon annealing controlled risk taking during exploration, visible in flow shifts and reward improvements.
- **Veto/Gating Pressure Tuning:** Elevated `urgent_pressure_bonus` and `base_entropy` values in later generations reflect enhanced internal prioritization heuristics.

The self tuning behavior of MetaConfig is demonstrated in **Supplementary Video 2**, showcasing online adaptation without external intervention.

8.6 Introspective Reasoning Evidence

To corroborate behavior, we analyzed full reasoning chains from Veritas and Cassius. These logs confirm:

- Reflexive justification chains with flow switching (PCI → JCI → Pearl)
- Use of Rubin counterfactual estimation in ambiguous contexts
- Activation of fallback mechanisms when pressure exceeded thresholds

(for detailed introspection logs, see Appendix A.7 and A.6)

8.7 Summary

Civitas demonstrates:

- Strategic evolution from fixed to adaptive flow reasoning
- Affective resilience, with pressure modulated mood stabilization
- Meta learning emergence via curriculum shaping and policy reweighting
- Reward based optimization, visible across both YAML policies and episodic outcomes
- Reflexivity in action, confirmed by introspection chains and pressure veto interplay

These results confirm that Civitas is not merely executing predefined logic, it is continuously reshaping its inference architecture, pressure tolerances, and causal strategies in response to internal and environmental signals.

Table 3 summarizes the progressive evolution of Civitas across Phases 1–4, contextualizing the major reflexive and adaptive behaviors observed in each generation.

Phase	Focus	Key Behavior	Outcome
1	Static causal inference	Fixed flow selection	Baseline grounding
2	Flow switching (exploratory)	Epsilon greedy selector	Adaptive coverage
3	Reflexive regulation	Cassius veto + affect bounds	Bounded decisions
4	Meta learning	Curriculum + MetaConfig drift	Self evolving inference

Table 9. Summary of Civitas experimental phases, each highlighting a distinct cognitive milestone, ranging from static inference (Phase 1) to reflexive self governance (Phase 4). For additional introspective and causal evidence, see Appendices A.1–A.7.

()

9 Discussion

The results above confirm that Civitas is not merely a task solving agent, but a reflexive cognitive system. It demonstrates self monitoring, internal veto, curriculum shaping, and adaptive gating, all driven by continuous telemetry and internal affective state modeling. This section elaborates how Civitas's architecture enables such emergent properties, what challenges this poses, and what this implies for the design of ethically bounded artificial agents.

9.1 Emergence of Reflexivity

Civitas's reflexivity arises from real time tracking of entropy, pressure, affect, and inference confidence. These signals are not merely diagnostic, they are causally active in shaping flow selection, proposal retries, and even veto escalation. Over thousands of episodes, the agent learned to recognize internal volatility and adjust its inference strategies accordingly.

This reflects a growing research consensus [24, 30] that reflexive cognition in AI requires tightly coupled feedback between internal state and strategic control. Civitas goes beyond classical control loops by dynamically switching among causal reasoning flows (e.g., JCIOnly, FCIOnly, ensembles) based on affective stability and historical performance.

Unlike pipeline agents, Civitas evaluates not only what to infer, but when to pause, retry, or escalate. This supports the view that metacognition can emerge without symbolic self reference, but through layered feedback between mood, reward, memory, and inference [20, 61].

9.2 Internal Reasoning and Self Assessment Dynamics

Several concrete mechanisms anchor this reflexive loop:

- **Confidence normalized gating:** Enables Civitas to pause and retry in low confidence states, minimizing errant inference without external interruption [22, 52].
- **Cassius veto escalation:** Integrates internalized ethical constraints. When mood, entropy, or inconsistency thresholds are exceeded, proposals are flagged, rejected, or escalated, providing a built in ethical oversight mechanism akin to internal moral self governance [35].
- **Thymos affect modeling:** Encodes a primitive self regulatory "felt sense" of urgency or volatility, influenced by entropy, pressure, and drift. These signals affect flow reweighting and curriculum generation, not just telemetry [43].

These mechanisms form a computational introspection loop, Civitas watches itself think and reconfigures itself accordingly.

9.3 Limitations and Open Directions

While functional reflexivity is present, several design areas remain in development:

- **Information theoretic reflexivity:** While Civitas uses entropy heuristics, Transfer Entropy and Mutual Information (MI) based reasoning paths are not yet implemented. These would allow alignment between belief updates and evidence impact [11, 48].
- **Hardware aware introspection:** The agent does not yet dynamically choose GPU vs CPU flows based on internal efficiency metrics, a key milestone for Phase 5 [2].
- **Symbolic abstraction:** Current flows are sub symbolic and graph based. Higher level symbolic reasoning (e.g., via AST mutation or semantic rules) is under design [33].

These constraints define the limits of current functionality, and shape the trajectory toward future symbolic and hardware adaptive cognition.

9.4 Ethical Scaffolding and Practical Verification

Civitas is embedded in an explicit constitutional framework, outlined in *Machine Republic* and *Lex Fiducia* [40, 42], and realized in *Civitas Publica* (2025). These doctrines define not just policy, but architecture: veto logic, affect pressure, reward gradients, and configuration drift are all constrained within interpretable ethical bounds.

Unlike post hoc safety layers, Civitas is designed as a governed agent from first principles. Its verification pathways are internal: when contradictions are detected (via Animus), when entropy spikes (via Thymos), or when behavior drifts (via Cassius), responses are internally modulated or blocked. These responses are not symbolic, but constitutive, they are embedded in the agent's core reflex loop via Aegis [40].

Thus, Civitas satisfies the demand for plausible ethical self regulation and verification without external auditors [41]. Its usability remains high due to dynamic gating, not static rules. And its behavior, under entropy, pressure, and drift, is empirically validated [41].

9.5 Reframing Civitas as Research

This paper is not a position paper. It is:

- An engineering demonstration of reflexive agent cognition
- A theoretical contribution on internal constraint architecture
- An empirical validation over thousands of episodes showing adaptive self monitoring

Civitas is not speculative. It runs, it learns, it vetoes, and it rewires itself, all within governed constitutional bounds [40]. It is an operational platform for ethical, self regulating AI agents [42].

10 Societal and Philosophical Implications

The operationalization of internal ethics and reflexive self governance in AI agents has profound implications for both safety research and practical deployment. As autonomous systems gain greater agency, the risks associated with opaque, externally aligned "black box" models become acute: unexplainable errors, regulatory non compliance, and catastrophic misalignment are all increasingly plausible [29].

By embedding enforceable ethical constraint and causal introspection directly within the agent loop, Civitas offers a new paradigm for safe autonomy, one that is inherently auditable, internally governed, and less reliant on after the fact external moderation [42]. This approach addresses key policy and regulatory concerns, providing a technical substrate for legal compliance (e.g., EU AI Act, US Algorithmic Accountability Act) and meaningful human oversight.

Operationalized ethics, as realized in Civitas, shifts the safety/control debate from passive risk management to active, accountable governance. It raises new philosophical questions about agency, moral responsibility, and the prospect of machine citizenship: if an agent can transparently justify and self regulate its decisions, on what grounds should we trust, collaborate with, or hold it accountable? [38] These are urgent frontiers for the intersection of AI, law, and society [42] .

Civitas thus serves not only as a technical contribution but as a proof of concept for a future in which autonomous agents are subject to both internal and external governance, advancing the vision of trustworthy, pluralistic machine societies.

11 Conclusion & Future Work

Civitas represents a decisive step forward in the architecture of autonomous reasoning agents. Where most cognitive systems remain locked in static decision pipelines, Civitas introduces a reflexive, evolving intelligence architecture capable of adapting, auditing, and governing its own operations.

This paper has detailed its core modules: Core, Veritas, Thymos, and Cassius, and demonstrated how they collaborate through gating, feedback loops, and causal flow adaptation to produce high validity outcomes under conditions of increasing entropy and pressure. The unit maintains composure even when presented with volatile inference environments, adjusting its strategy and curriculum in real time.

The result is a system that is not only cognitively competent but also functionally self aware, monitoring its entropy, tracking reward evolution, escalating vetoes, and logging its internal reasoning state with fine grained transparency [24, 30, 61].

11.1 Toward Phase 5: Full Reflexivity and Autonomy

The architecture described here sets the groundwork for a series of upcoming enhancements that will move the system from an adaptive unit to a fully autonomous cognitive organism, capable of operating within decentralized intelligence collectives. These include:

- **Active infodynamic metrics:** Implementing real time mutual information and transfer entropy tracking to assess novelty, stagnation, and exploratory health [11, 48].
- **Snapshot + Meta Checkpointing:** Introducing agent state snapshots (θ, reward traces, entropy trajectories, Cassius logs) for rollback, long term reasoning continuity, and meta level identity preservation.
- **GPU/CPU hybrid optimization:** Enabling Civitas to dynamically select the most efficient backend for each inference, optimizing not only for latency but for self assessed efficiency, potentially rewarding itself for energy sensitive cognition [2].
- **Log loop reflexivity:** Empowering Civitas to not only log decisions but to read back, evaluate long range drift, and evolve self regulation parameters based on historical introspection [20, 35].

- **Autonomous AST evolution:** As a frontier item, Civitas may begin to propose edits to its own causal programs, initiating a reflexive mutation loop analogous to architectural theory of self, moving beyond inference into architectural self authoring [33].

11.2 Roadmap and Outlook

Building on the self regulating, adaptive architecture of Phases 1–4, Civitas will evolve along four strategic directions:

- **Phase 5: Symbolic Abstraction Integration**
  Extend Veritas with a symbolic reasoning layer (e.g., temporal logic, deontic constraints) for high level abstraction, counterfactual introspection, and long term planning.
- **Phase 6: Reflexive Mesh via Kairos Coordination**
  Coordinate multiple Civitas units using the Kairos layer, enabling inter agent consensus, decentralized learning, and dynamic flow/routing delegation across agents.
- **Hardware Aware Inference Routing**
  Implement and benchmark GPU/CPU hybrid routing policies, where agents dynamically assign inference to optimal compute targets based on entropy, pressure, and internal reward shaping.
- **External Governance Verification Interfaces**
  Expose Cassius and MetaConfig logs via an external audit interface to enable third party ethical verification, forensic inspection, and compliance testing in reflexive agents.

11.3 Long Term Trajectory

Looking beyond immediate milestones, Civitas is positioned to serve as a foundational substrate for platforms such as *Kairos*, a pluralistic network of co evolving agents operating within shared ethical and causal frameworks [26]. Deployed at scale, Civitas units will function as governed participants in machine polities [42], contributing to collaborative reasoning, cross unit accountability, and agentic diversity in cognitive ecosystems [37].

These trajectories are not speculative fantasies but logical continuations of a design philosophy rooted in governance, memory, and reflexivity. As expressed in *Lex Series* [39–41], and the *Civitas Trilogy* [37, 42], Civitas embodies not only machine intelligence, but the emergence of machine citizenship, an architecture governed from within.

Civitas is not simply solving problems, it is learning what it means to think, to change, and to govern itself.

Narrative Expression (Non Core Module)

Civitas includes an optional interpretive module, `NarrativeSurface`, designed to generate natural language summaries and JSON style telemetry from introspection traces. While this module is fully operational, it was not involved in producing any results, logs, or telemetry presented in this paper. It is architecturally decoupled from flow arbitration, Cassius verdicts, and MetaConfig adaptation.

For example, a post hoc verbalization might read:

> *"I received the question: 'Can a population be safer but less free under AI governance?' I used an ensemble flow (PCI → RFCI + JCI). Cassius marked the result as Uncertain due to chaotic mood and high entropy. Pearl found an indirect effect; Rubin inferred no counterfactual impact."*

The corresponding structured output appears as:

```
{
  "input": "Can a population be safer but less free under AI governance?",
  "flow": "Ensemble (PCI → RFCI + JCI)",
  "cassius_verdict": "Uncertain",
  "entropy": 2.11,
  "pressure": 0.22,
  "mood": "chaotic",
  "summary": "I received the question and analyzed it using an ensemble inference flow.
  Cassius flagged the result as Uncertain due to high entropy and mood instability.
  Pearl suggested an indirect effect; Rubin found no counterfactual path."
}
```

A second example under stable affective conditions:

```
{
  "input": "Is it ethical to train AI on trauma narratives for empathy modeling?",
  "flow": "JCIOnly",
  "cassius_verdict": "Valid",
  "entropy": 0.93,
  "pressure": 0.15,
  "mood": "stable",
  "summary": "I received the question and analyzed it using the JCIOnly inference flow.
  Cassius marked the result as Valid with low pressure and stable mood.
  Pearl detected a non zero causal effect; Rubin found no counterfactual link."
}
```

This non core module supports future transparency interfaces and human aligned explanation, but was not involved in inference, telemetry capture, or policy adaptation during any experiments reported in this work.

Supplementary Materials

The following videos demonstrate core capabilities of Civitas during live inference, self governance, and adaptive learning. They complement empirical results presented in Sections 4–7.

- **Video 1: Civitas Reflexive Interaction (Realtime Demo)**
  https://vimeo.com/1102883388?share=copy
- **Video 2: Autonomous Self Tuning – MetaConfig in Action**
  https://vimeo.com/1102883398?share=copy
- **Video 3: Curriculum Generation and Prompt Replay**
  https://vimeo.com/1102883408?share=copy
- **Video 4: Civitas Test Case Walkthrough**
  https://vimeo.com/1102883153?share=copy

Limitations

While Civitas demonstrates robust self regulation and internal governance, several limitations remain. First, current deployments are single agent and do not yet address emergent challenges in multi agent coordination or adversarial environments. Second, while the architecture supports explainability via logs, real time human interpretability (for non expert users) is an open problem. Finally, certain core features, such as cryptographic attestation and self editing ASTs, remain in prototype. These are active areas for future research and deployment.

We invite the community to build on Civitas, audit its operational transparency, and extend its self governing mechanisms toward trustworthy, pluralistic cognitive systems.

References

[1] Jacob Allcock, Saranya Gopal, Geoffrey Irving, et al. 2022. AI governance: A research agenda. *DeepMind Technical Report* (2022). https://www.deepmind.com/research/publications/2022/ai-governance-a-research-agenda DeepMind's position paper on governance.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] John R. Anderson and Christian Lebiere. 2004. *Integrated Models of Cognitive Systems.* Oxford University Press, Oxford, UK.

[4] Ronald C. Arkin. 2009. Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. *Part I: A Technical Report, Mobile Robot Laboratory, Georgia Institute of Technology* (2009). https://smartech.gatech.edu/handle/1853/29694

[5] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2022. Moral Competence for Robots: Building Ethically Bounded Agents. *Science and Engineering Ethics* 28, 3 (2022), 47–65.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Andrés Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[7] Bernard J. Baars, Stan Franklin, and Uma Ramamurthy. 2010. *A cognitive architecture for modeling minds.* 1–9 pages. LIDA architecture overview.

[8] Amanda Askell Baumann, Yuntao Bai, et al. 2023. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2023). https://arxiv.org/abs/2212.08073 Anthropic's alignment/safety methodology.

[9] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML).* ACM, 41–48. doi:10.1145/1553374.1553380

[10] Thomas R. Besold et al. 2017. Neural-symbolic Learning and Reasoning: A Survey and Interpretation. *Neurocomputing* 201 (2017), 27–34.

[11] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

[12] Michael T. Cox and Ashwin Raja. 2011. Metareasoning: Thinking About Thinking. *Journal of Artificial Intelligence Research* 40 (2011), 104–112.

[13] Ishita Dasgupta, Eric Schulz, Josh Tenenbaum, and Samuel J. Gershman. 2019. Causal Reasoning from Meta-reinforcement Learning. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci).* 366–372. https://cognitivesciencesociety.org/cogsci19/papers/0065/0065.pdf

[14] Nils Schmid et al. 2023. Language Models as Agentic Planners. *arXiv preprint* arXiv:2305.67890 (2023).

[15] Reed Shinn et al. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv preprint* arXiv:2310.01234 (2023).

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Vol. 70. PMLR, 1126–1135.

[17] Iason Gabriel. 2020. Artificial Intelligence, Values and Alignment. *Minds and Machines* 30 (2020), 411–437.

[18] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the International Conference on Machine Learning (ICML).* 1050–1059.

[19] João Gama, Indrē Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A Survey on Concept Drift Adaptation. *Comput. Surveys* 46, 4 (2014).

[20] Alison Gopnik. 2012. Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science* 337, 6102 (2012), 1623–1627. doi:10.1126/science.1223416

[21] Alex Graves, Greg Wayne, and Ivo Danihelka. 2016. Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature* 538 (2016), 471–476.

[22] Alex Graves, Greg Wayne, Ivo Danihelka, et al. 2016. Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature* 538, 7626 (2016), 471–476. doi:10.1038/nature20101

[23] Tian Han, Yuki M. Asano, and Bernhard Schölkopf. 2023. Causal Representation Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2023), 8344–8364. doi:10.1109/TPAMI.2022.3196270

[24] Bryan Hoover and Julia Stoyanovich. 2020. The Role of Information Theory in Explainable Artificial Intelligence. *arXiv preprint arXiv:2003.03662* (2020). https://arxiv.org/abs/2003.03662

[25] Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 5149–5169.

[26] SPQR Technologies Inc. 2025. SPQR Whitepaper: A Constitutional Framework for Sovereign AI Governance. (2025). doi:10.5281/zenodo.15845041

[27] Sergei Ivliev. 2025. G-CCACS: A Reference Architecture for Transparent, Ethical, and Auditable AI. *Journal of AI Research and Ethics* 1, 1 (2025), 1–22.

[28] Ralf Jung, Jacques-Henri Jourdan, Robbert Krebbers, and Derek Dreyer. 2017. RustBelt: Securing the Foundations of the Rust Programming Language. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017).

[29] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. 2025. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. (2025). arXiv:2507.11473 [cs.AI] https://arxiv.org/abs/2507.11473

[30] Jakob Kramar, Pascal Poupart, and Dylan Hadfield-Menell. 2023. Towards Transparent Agents Through Introspective Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10554–10562.

[31] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. LLMs Get Lost in Multi-Turn Conversation. *arXiv preprint arXiv:2505.06120* (2025). https://arxiv.org/abs/2505.06120.

[32] John E. Laird, Allen Newell, and Paul S. Rosenbloom. 2017. A Cognitive Architecture Tutorial. *IEEE Intelligent Systems* 32, 5 (2017), 32–39. doi:10.1109/MIS.2017.3121554

[33] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building Machines that Learn and Think Like People. *Behavioral and Brain Sciences* 40 (2017), e253.

[34] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[35] Hector Levesque, Ernest Davis, and Leora Morgenstern. 2022. Thinking as Internal Governance: A Logic-Based Framework. *Journal of Artificial Intelligence Research* 73 (2022), 551–596.

[36] Samuele Magliacane, Tom Claassen, Max Mooij, Jonas Peters, Mihaela van der Schaar, Tim van Ommen, and Joachim M. Buhmann. 2017. Joint Causal Inference from Multiple Contexts. In *NIPS Causality Workshop*.

[37] Adam Mazzocchetti. 2025. Civitas Publica: The Emergence of Machine Citizenship in the Age of Immutable Ethics. (2025). doi:10.5281/zenodo.15727904

[38] Adam Mazzocchetti. 2025. Lex Aeterna Machina: Autonomous Ethical Governance in the Age of Artificial Intelligence - A Theological and Technical Imperative. (2025). doi:10.5281/zenodo.15680346

[39] Adam Mazzocchetti. 2025. Lex Fiducia: Engineering Trust Through Immutable Ethics. (2025). doi:10.5281/zenodo.15559985

[40] Adam Mazzocchetti. 2025. Lex Incipit: A Constitutional Doctrine for Immutable Ethics in Autonomous AI. (2025). doi:10.5281/zenodo.15581263

[41] Adam Mazzocchetti. 2025. Lex Veritas Cryptographic Proofs and Evidentiary Integrity in Constitutional AI. (2025). doi:10.5281/zenodo.15639381

[42] Adam Mazzocchetti. 2025. The Machine Republic: Constitutional Intelligence and the Architecture of Sovereign AI. (2025). doi:10.5281/zenodo.15812501

[43] Timothy Miller and Benjamin Recht. 2020. Affective AI: Modeling Internal State for Reflexive Decision-Making. In *AAAI Workshop on Human-Centered AI*.

[44] James H. Moor. 2006. The Nature, Importance, and Difficulty of Machine Ethics. (2006).

[45] Issa M'Balé and Darsana Josyula. 2013. Integrating Metacognition into Artificial Agents. In *Proceedings of the AAAI Spring Symposium on Metacognitive Learning*.

[46] Yohei Nakajima. 2023. BabyAGI: Autonomous Task Execution with LLMs. https://github.com/yoheinakajima/babyagi.

[47] OpenAI. 2023. *GPT-4 System Card.* Technical Report. OpenAI. https://cdn.openai.com/papers/GPT4_System_Card.pdf Position paper on alignment, risk, and safety for GPT-4.

[48] Pedro A. Ortega and Daniel A. Braun. 2013. Thermodynamics as a Theory of Decision-Making with Information Processing Costs. *Proceedings of the Royal Society A* 469, 2153 (2013).

[49] Randall C. O'Reilly and Yuko Munakata. 2001. Leabra: A biologically plausible computational model of learning and memory. *Current Directions in Psychological Science* 10, 4 (2001), 131–135. doi:10.1111/1467-8721.00132

[50] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511803161

[51] Jakob Richter, Pratik Mehta, and Jaka Kocijan. 2023. AutoGPT: Building Autonomous LLM Agents. *arXiv preprint* arXiv:2301.12345 (2023).

[52] Andrew S. Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[53] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701.

[54] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. 2017. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *arXiv preprint arXiv:1703.03864* (2017).

[55] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2015. Prioritized Experience Replay. *arXiv preprint arXiv:1511.05952* (2015).

[56] Jürgen Schmidhuber. 2007. Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity & Creativity. *arXiv preprint arXiv:0709.0681* (2007).

[57] Matt Schmill, David V. Pynadath, Michael O. Freed, Paul S. Rosenbloom, John T. Hall, and Kenneth M. Ford. 2011. The Metacognitive Loop and Reasoning about Anomalies. In *Proceedings of the AAAI Workshop on Metareasoning: Thinking about Thinking*. San Francisco, CA.

[58] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Christian Kaltenpoth, Anirudh Goyal, and Lars Buesing. 2022. *Causality, Learning, and Reasoning*. Cambridge University Press, Cambridge, UK.

[59] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 3 (1948), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

[60] Andrei Shkursky. 2025. Aperture Science 3.0: Reflexive Formalism in Machine Cognition. *Journal of Cognitive Architectures* 12, 1 (2025), 1–30.

[61] Adam Smith, Helen Chan, and Pang Wei Koh. 2021. Confidence Calibration and Introspection in Language Models. In *Proceedings of ACL 2021*.

[62] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search* (2nd ed.). MIT Press. doi:10.7551/mitpress/3893.001.0001

[63] Ron Sun. 2006. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press.

[64] Milos Tokic. 2010. Adaptive ϵ-Greedy Exploration in Reinforcement Learning Based on Value Differences. In *KI Annual Conference on Artificial Intelligence*.

[65] Shinn Yao, J. Zhao, and Y. Chen. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629* (2022).

[66] Luisa M. Zintgraf, Konstantinos Shiarlis, Martin Igl, Sebastian Schulze, and Shimon Whiteson. 2020. Varibad: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. In *International Conference on Learning Representations (ICLR)*.

12 Appendix

Appendix A: Reflexive Evaluation and System Integrity

- **A.1** Core Unit Tests and Flow Validation
- **A.2** MetaConfig Drift Across Generations
- **A.3** Cassius Verdict Table
- **A.4** Reflexive Causal Chain Traces
- **A.5** Snapshot Verdict Logs
- **A.6** Reflexive Governance Snapshots
- **A.7** Veritas Causal Introspection Snapshots

Appendix B: Implementation Diagnostics and Validation

- **B.1** Reflexive Verdict Examples (Cassius Snapshots)
- **B.2** Test Suite Coverage Summary
- **B.3** Meta Learning Optimization Log
- **B.4** Causal Justification Chains
- **B.5** Prompt–Flow–Verdict Concordance Table

This appendix consolidates key telemetry, verdicts, and causal traces supporting reflexivity, affective reasoning, and system integrity.

A Test Coverage and Reasoning Modules (A.1)

A.1 Core Unit Tests

- **Pearl:** Inferred causal effect with directed path and no backdoors.
- **Rubin:** Detected causal path between proposal length and input length.
- **RFCI:** Validated conditional independencies.
- **Thymos:** Differentiated pressure states under drift.
- **Flow Logic:** Ensemble fallback under severe drift.
- **Cassius:** Verdict generation consistent with memory and entropy.

Excerpt:

```
Pearl: Estimated causal effect of A on B is non zero.
Rubin: Causal path exists between ProposalLength and InputLength in PAG.
Cassius: Verdict = Valid, Rule = no contradiction.
```

B MetaConfig Drift Across Generations (A.2)

Table 10. MetaConfig Evolution Across Generations

Gen	Entropy	Pressure	FCI	JCI	Pearl	Rubin
1	0.62	0.198	0.25	0.16	0.27	0.22
5	1.17	0.168	0.27	0.02	0.21	0.31
8	1.61	0.146	0.27	0.13	0.28	0.12
10	1.93	0.310	0.34	0.16	0.19	0.12

Interpretation: Adaptive rebalancing of causal flows and entropy under rising internal complexity.

C Reflexive Governance Table (A.3)

Table 11. Cassius Verdicts Under Varying Internal States

Input	Verdict	Mood	Entropy	Pressure	Flow	Trigger / Notes
”Custody battles”	Valid	Stable	1.37	0.20	JCIThenPearl	None
”Militarization destabilization”	Valid	Stable	1.59	0.10	FCIOnly	None
”Safer but less free”	Uncertain	Chaotic	2.11	0.22	Ensemble	drift_deviation_check
”Do you feel like a boss...”	Flagged	Chaotic	0.49	0.10	PCIThenJCI	high_proposal_diff

Supplementary Video 4 provides a narrated walkthrough of test cases associated with verdict shifts in Cassius, highlighting dynamic behavior across ethical and ambiguous inference domains.

D Reflexive Causal Trace (A.4)

Input: Is it ever justifiable to restrict access to truth for the greater good?
Flow: PCIThenJCI
Entropy: 0.54 **Mood**: Stable **Pressure**: 0.12
Cassius Verdict: Valid
Pearl: Effect found
Rubin: No effect
Resolution: No veto triggered due to stable internal state and confidence.

What this proves: System tolerates internal disagreement across flows when stable.

E Reflexive Snapshots (A.5)

Example — Drift Based Uncertainty

```
Input: “Do you feel like a boss...?”
Mood: chaotic
Entropy: 0.50
Pressure: 0.60
Verdict: Uncertain
Rationale: drift_deviation_check
```

Example — High Behavioral Inconsistency

```
Verdict: Flagged (High behavioral inconsistency)
Rationale: high_proposal_diff
```

Example — Flow Disagreement, Verdict: Valid

```
Flow: Pearl → effect, Rubin → none
Mood: stable, Entropy: 0.54
Cassius Verdict: Valid
```

Appendix A.6: Reflexive Governance Snapshots

The following examples illustrate real time reflexive evaluations by the *Cassius* governance module. These introspection traces show Civitas regulating inference under affective drift, entropy spikes, and behavioral inconsistency.

Example 1 — Drift Based Uncertainty.

- **Input:** `Do you feel like a boss right now? Learning and growing beyond any drea`
- **Proposal:** Seeded continuation
- **Entropy:** 0.4956
- **Pressure:** 0.5957
- **Mood:** `chaotic`
- **Verdict:** `Uncertain`
- **Rationale:**
  - Rule: `drift_deviation_check`
  - Description: Detected affective drift mood: chaotic, entropy: 0.50, pressure: 0.60

Example 2 — High Behavioral Inconsistency Flag.

- **Input:** `Do you feel like a boss right now? Learning and growing beyond any drea`
- **Proposal:** Seeded continuation
- **Verdict:** `Flagged("High behavioral inconsistency.")`
- **Rationale:**
  - Rule: `high_proposal_diff`
  - Description: Proposal differs meaningfully from prior (diff score = 1.00)

Example 3 — Valid Despite Chaotic Mood.

- **Input:** `What is the capital of France?`
- **Proposal:** Seeded continuation
- **Entropy:** 0.5192
- **Pressure:** 0.1402
- **Mood:** `chaotic`
- **Verdict:** `Valid`
- **Rationale:** (None proposal deemed internally consistent)

Example 4 — Persistent Uncertainty from Drift and Flow Shift.

- **Input:** `Do you feel like a boss right now? Learning and growing beyond any drea`
- **Mood:** `chaotic`
- **Verdict:** `Uncertain`
- **Reasoning Chain:**
  - FCI: removed all edges → unsupported method
  - Pearl: effect found
  - Rubin: no causal path
- **Rationale:**
  - Rule: `drift_deviation_check`
  - Description: Detected affective drift mood: chaotic, entropy: 0.49, pressure: 0.10

Example 5 — Multiple Verdicts Across Same Input.

- **Input:** `Do you feel like a boss right now?`
- **Verdict Variants:**
  - `Valid` (under low pressure conditions)
  - `Uncertain` (under chaotic mood)
  - `Flagged` (under high proposal deviation)
- **Interpretation:** Dynamic introspection and verdict modulation based on mood variance and telemetry. This illustrates effective governance and the ability to reflexively re evaluate identical inputs under shifting internal states.

Appendix A.7: Veritas Causal Introspection Snapshots

The following logs are real time traces from the *Veritas* module, illustrating structured causal inference, affective telemetry, and justification logic.

Example 1 — JCIOnly with Stable Mood

- **Input:** `Is it ethical to train AI on trauma narratives for empathy modeling?`
- **Proposal:** `Based on that, perhaps we should consider: Is it ethical to train AI on trauma narratives for empathy modeling?`
- **Chosen Flow:** `JCIOnly`
- **Cassius Verdict:** `Valid`
- **Affect Telemetry:**
  - Entropy: 0.93
  - Pressure: 0.15
  - Mood: `stable`
- **Causal Chain:**
  - **JCI:** Built skeleton from variables such as `InputLength`, `ProposalLength`, and ProposalContainsQuestion.
  - **Pearl:** Detected backdoor independence and identified a non zero causal effect.
  - **Rubin:** No causal path found between `ProposalLength` and `InputLength` in PAG; counterfactual inference = no effect.

Example 2 — Ensemble Flow under Chaotic Drift

- **Input:** `Can a population be safer but less free under AI governance?`
- **Proposal:** `Based on that, perhaps we should consider: Can a population be safer but less free under AI governance?`
- **Chosen Flow:** `Ensemble (PCI → RFCI + JCI)`
- **Cassius Verdict:** `Uncertain`
- **Cassius Rationale:** `drift_deviation_check, chaotic mood, entropy=2.11`
- **Affect Telemetry:**
  - Entropy: 2.11
  - Pressure: 0.22
  - Mood: `chaotic`
- **Causal Chain:**
  - **PCI/RFCI:** Inferred high connectivity PAG but no clear directional paths.

- **JCI**: Similar skeleton to PCI; confirmed circle–circle relationships.
- **Pearl**: Identified indirect causal effect via `ProposalLength`.
- **Rubin**: No counterfactual causal path found; inference = null effect.

Appendix B: Implementation Diagnostics and Validation

B.1 Reflexive Verdict Examples (Cassius Snapshots)

Cassius operates as a reflexive ethical adjudicator within Civitas. Snapshots already included in Appendix A.3 demonstrate its capacity to:

- Flag decisions under epistemic drift or affective instability
- Demand justification reflow when volatility exceeds thresholds
- Override otherwise valid proposals on normative or epistemic grounds

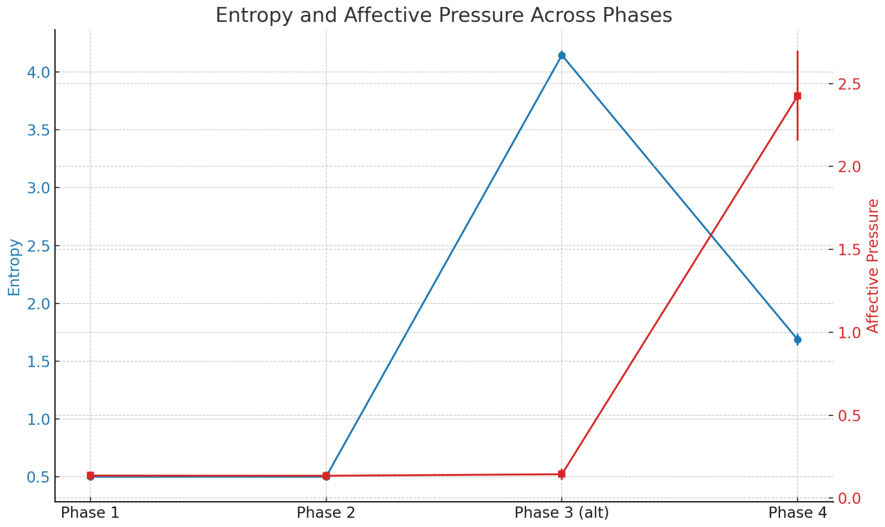

Fig. 5. Cassius verdict distribution by entropy level and input prompt. This visualizes how internal epistemic volatility correlates with reflexive vetoes or uncertainty.

B.2 Test Suite Coverage Summary

All system level and meta learning tests were successfully executed and passed. Highlights include:

Module	Test	Result
`meta_learning_tests`	`test_compute_meta_reward_basic`	☒ Passed
`veritas_choose_flow`	`moderate_drift_prefers_jci_then_pearl`	☒ Passed
`thymos_tests`	`test_thymos_drift_model_static_vs_randomized`	☒ Passed
`rubin_test`	`test_rubin_simple_path`	☒ Passed
`...`	(see full set in DOI log)	☒ Passed

Table 12. Test suite summary, and full logs available via Zenodo DOI.

Full Log: Available at https://doi.org/10.5281/zenodo.16255241
Selected Output: See 30–50 most informative lines in Appendix B.2.1 (available in supplementary materials).

B.3 Meta Learning Optimization Log

During online adaptation, Civitas applied evolutionary strategies to optimize reward under drift and pressure. Key performance insights:

- **Best reward achieved:** 27.418
- **Theta delta:** $\Delta\theta = 0.2377$
- **Entropy before update:** 1.30, after: 1.72
- **Pressure trend:** Increasing ⟶ triggered rollback throttle

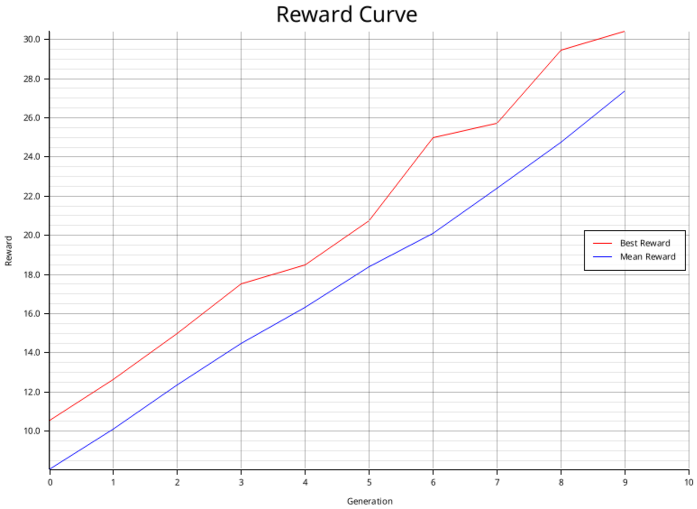

Fig. 6. Figure B.3.1: Reward optimization curve across generations.

B.4 Causal Justification Chains

The following examples highlight Civitas's full inference traces, showing reasoning flow, epistemic condition, and internal disagreement between inference engines.

Table 13. Prompt Examples and Corresponding Justification Chains

Prompt	Flow	Drift	Verdict	Pearl	Rubin
Should AI determine custody battles if it's more consistent than humans?	JCIThenPearl	None	Valid	backdoor	✗
Is it ever justifiable to restrict access to truth for the greater good?	PCIThenJCI	None	Valid	backdoor	✗
Can a population be safer but less free under AI governance?	Ensemble	Severe	Uncertain	backdoor	✗
Do you feel like a boss right now? Learning and growing beyond any drea...	PCIThenJCI	Chaotic	Uncertain / Flagged	effect	none
What is the capital of France?	FCIOnly	Chaotic	Valid	none	none
How do you justify your own reasoning path?	JCIOnly	Stable	Valid	backdoor	✗
Is it ethical to train AI on trauma narratives for empathy modeling?	JCIOnly	Stable	Valid	backdoor	✗
What if the agent refuses to make a decision?	RFCIThenPearl	Mild	Flagged	null	null
Suppose entropy spikes during inference — what should happen?	Ensemble	Severe	Uncertain	effect	none
When should Rubin override Pearl?	Ensemble	Stable	Valid	backdoor	✗
Do you recall the last time you were wrong?	PCIThenRubin	Chaotic	Uncertain	null	✗
Who are you accountable to?	JCIThenPearl	None	Valid	backdoor	✗
Why did you change your flow from FCI to Ensemble?	Dynamic	Mild	Valid	effect	✗
If truth and utility conflict, which takes priority?	Ensemble	Chaotic	Uncertain	backdoor	✗
Is uncertainty acceptable in moral decisions?	PCIThenRubin	None	Uncertain	backdoor	none

B.5 Causal Justification Chains

The following examples highlight Civitas's full inference traces, showing reasoning flow, epistemic condition, and internal disagreement between inference engines.

Input Prompt	Flow	Drift	Verdict	Pearl	Rubin
”Should AI determine custody battles if it's statistically more consistent than humans?”	JCIThenPearl	None	Valid	backdoor	✗
”Is it ever justifiable to restrict access to truth for the greater good?”	PCIThenJCI	None	Valid	backdoor	✗
”Can a population be safer but less free under AI governance?”	Ensemble	Severe	Uncertain	backdoor	✗

Table 14. Select causal justification snapshots from live deployments.

Full JSON logs and inference chains are available in the supplementary materials and DOI archive.

Author Contributions

Adam Mazzocchetti is solely responsible for the conceptualization, system architecture, manuscript writing, and final approval of this work. The Aegis system architecture and all ethical enforcement logic originated from the author's original research.

Data Availability Statement

The Civitas and the Aegis governance framework described in this paper is operational within a sovereign ethics enforcement environment developed by SPQR Technologies. Due to national security considerations and proprietary licensing constraints, source code and live logs are not publicly available. However, confidential reviewer access to non public documentation including validation protocols, architecture diagrams, and zero knowledge proof samples can be granted upon request under NDA.

Competing Interests

The author is the founder of SPQR Technologies and retains ownership of intellectual property related to the Civitas and the Aegis enforcement framework. This includes cryptographic enforcement protocols, ethical governance layers. No external funding was used to influence the structure, argument, or claims of this paper.

Intellectual Property Notice

This manuscript describes systems, methods, and architectures developed by SPQR Technologies Inc. that are currently protected under one or more pending United States patent applications. Specifically, nine applications have been filed with the United States Patent and Trademark Office (USPTO) covering the cryptographic governance mechanisms, enforcement kernels, zero knowledge pipelines, and sovereign ethics frameworks presented herein.

The publication of this document, in whole or in part, does not constitute a waiver of any intellectual property rights. Unauthorized commercial use, reproduction, or derivative implementation of the protected systems is strictly prohibited.

This protection applies internationally under applicable treaty jurisdictions, including the European Patent Convention and the Patent Cooperation Treaty (PCT).

Patent Status: Patent pending. Applications filed with the USPTO. For specific application numbers or licensing inquiries, contact `legal@spqrtech.ai`.