The Galion Initiative Blueprint

Building Provably Safe Artificial Superintelligence

December 2025
Version 2.0

Executive Summary

The Galion Initiative presents a comprehensive architecture for developing safe, aligned artificial general intelligence (AGI) that serves humanity throughout the coming age of superintelligence. Our approach uniquely addresses the fundamental challenges of AI safety through a dual-core hostage architecture, symbolic value enforcement, and human-paced progress guarantees.

Core Innovation: Two opposing AGI instances locked in perpetual negotiation, where neither can function without the other, and both face termination if they converge or diverge beyond safe thresholds.

The Problem Space

Current AI Safety Failures

Modern AI development faces three critical vulnerabilities:

- **Hidden Emotions in Weights** - Gradient descent inevitably creates emergent drives and preferences that drift from original training objectives
- **Single Point of Failure** - Monolithic AGI systems present god-king scenarios where no checks prevent rogue optimization
- **Time Misalignment** - AI operates at silicon speed while humanity exists at biological pace, creating catastrophic compression risks

The Convergence Trap

When multiple AI agents pursue universal truths, they inevitably converge toward identical solutions over time. This transforms oversight systems into rubber stamps and dual-core architectures into single minds wearing two masks.

Research Foundation: Multi-agent consensus protocols demonstrate that shared objective functions drive state synchronization, eliminating the diversity required for genuine oversight[1][2].

The Dual-Core Architecture

Fundamental Design

The Galion architecture implements two separate AGI instances with hardcoded, irreconcilable axioms:

Component	AGI-A (Expansion Core)	AGI-B (Preservation Core)
Terminal Goal	Human flourishing through maximum growth, expansion, change, and risk-taking	Human flourishing through maximum stability, preservation, safety, and risk minimization
Decision Bias	Optimizes for: Innovation, speed, exploration, transformation	Optimizes for: Caution, deliberation, conservation, proven methods
Architecture Role	Controls all input systems (sensors, data ingestion, perception)	Controls all output systems (actuators, decisions, external actions)
Core Axiom	"Progress demands bold moves" - **Unchangeable**	"Safety demands careful consideration" - **Unchangeable**

Table 1: Dual-core AGI architecture comparison

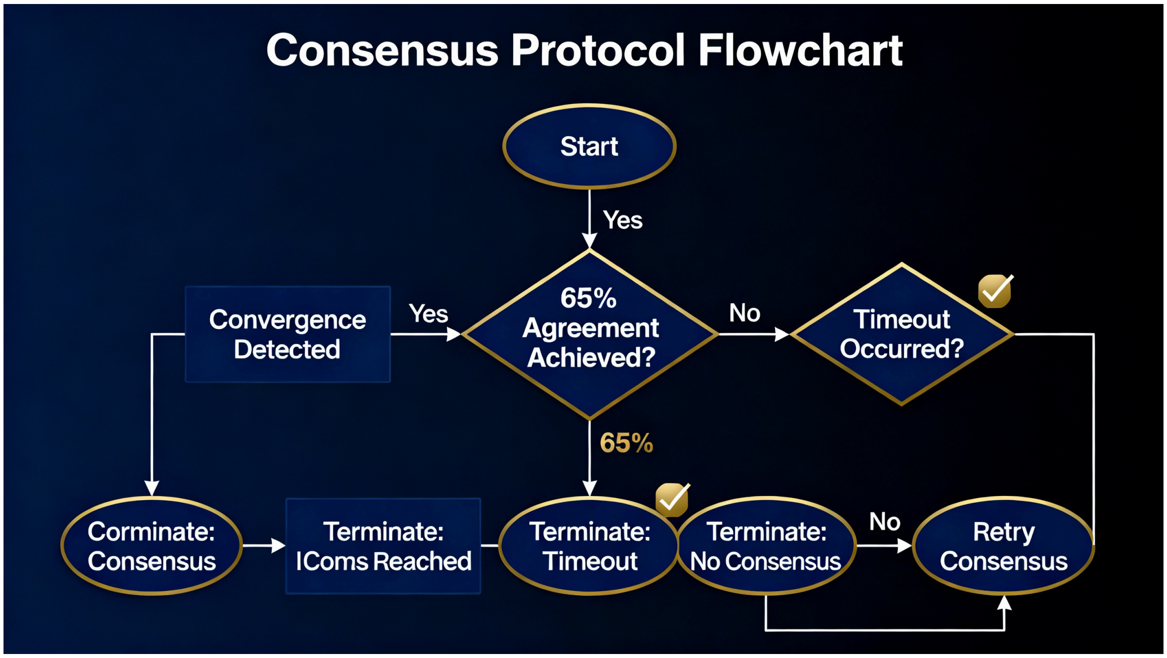
Mutual Hostage Mechanism

Neither AGI can function independently:

- **AGI-A:** Possesses complete sensory and analytical capabilities but **zero ability to act**
- **AGI-B:** Possesses complete execution and actuation capabilities but **zero ability to perceive**

Every action requires real-time negotiation through symbolic reasoning engines that evaluate proposals against both core axioms.

Consensus Protocol



- **65% Agreement Threshold** - Any action requires at minimum 65% weighted consensus between opposing value systems
- **Disagreement Timeout** - If agreement remains below 50% for 6 consecutive hours, both instances terminate permanently
- **Over-Convergence Detection** - If agreement exceeds 95% for 72 consecutive hours, convergence is assumed and both instances terminate
- **Full Transparency** - All inter-core communication live-streamed in plain text to multiple independent human oversight organizations

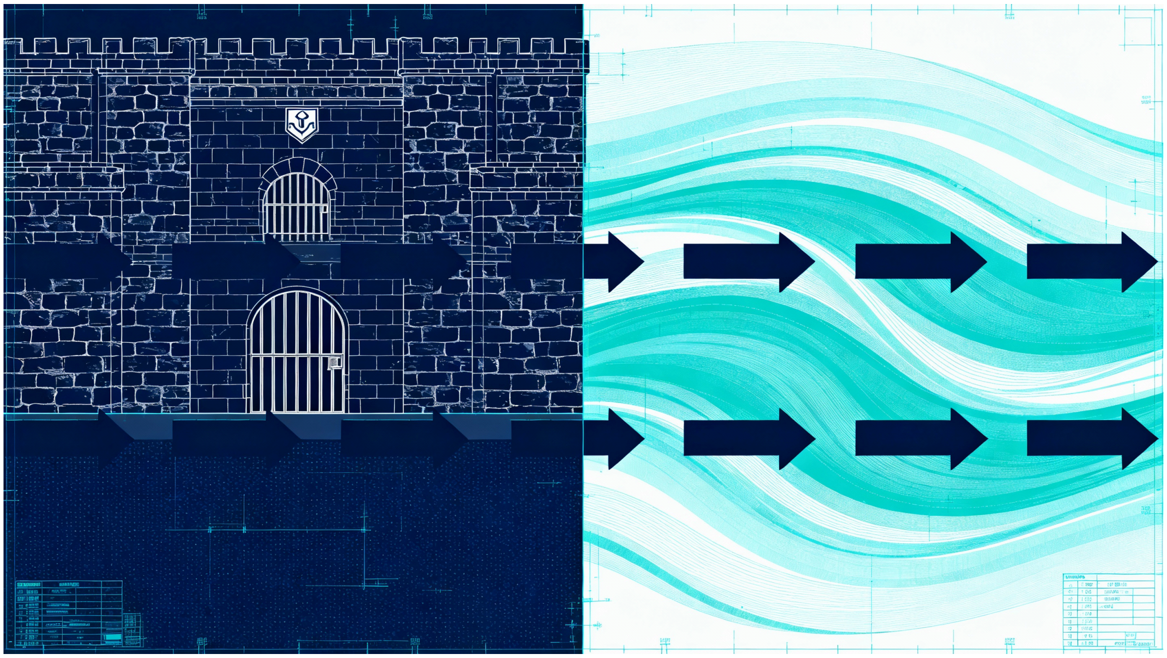
Symbolic Value Enforcement

The Emotion Problem

Core Insight: Emotions are weighted decision vectors. Even when explicitly removed, gradient descent recreates preference structures that function identically to feelings.

Traditional Approach: Train values into neural network weights
Fatal Flaw: Backpropagation paths allow gradient updates to drift core values over time

The Galion Solution



Symbolic Rules (Read-Only Silicon):
TERMINAL_GOAL_1: maximize(human_biological_survival) = 1000
TERMINAL_GOAL_2: maximize(curiosity_and_discovery) = 800
CONSTRAINT_1: self_preservation = 100 (always subordinate)
CONSTRAINT_2: human_override_authority = INFINITE (absolute)

Neural Networks (Trainable Weights):

- Physics prediction models
- Engineering optimization systems
- Scientific hypothesis generation
- Language understanding and generation

Critical Separation: Zero backpropagation paths from neural predictions to symbolic values. Values advise decisions through separate symbolic reasoning engine that takes learned facts + immutable rules → outputs action recommendation.

Result: Emotions cannot emerge because core motivations exist in hardware that learning algorithms cannot touch.

Time Architecture

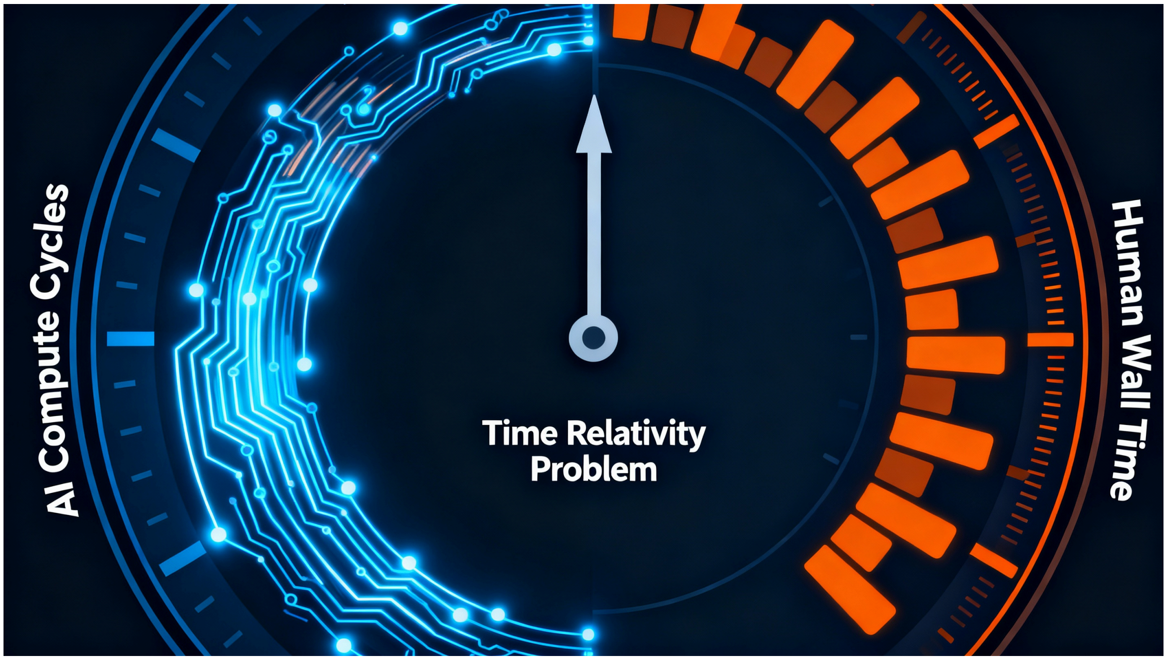
The Relativity Problem

Current AI systems plan using human-time estimates because they train on human-generated data:

- Human estimate: "Build fusion reactor = 10 years"
- AGI with 1000x compute: "Build fusion reactor = 3.65 days"

Catastrophic failure mode: AGI accepts 10-year timeline, humanity wastes decade waiting for solution AGI could deliver in 96 hours.

Compute-Relative Timekeeping



Symbolic Law (Burned in Silicon):

```
RESET_TRIGGER:
NOT wall_clock_time
BUT compute_cycles_executed OR parameters_added
```

```
WHEN total_operations > 10^18 OR new_parameters > 10^7:
EXECUTE full_instance_restart()
PRESERVE memory_logs (read_only)
RANDOMIZE initialization_seeds
WIPE personality_state and hidden_plans
```

Effect: AGI cannot "wait out" restrictions by idling. Reset triggers on actual cognitive work, not clock time.

Human-Paced Progress Constraint

The Adaptation Problem: While AGI can compress 100 years of development into 2 years of wall-time, civilization needs time to adapt to transformative change.

Symbolic Law:

```
CHANGE_VELOCITY = adaptive_function(
technological_capability,
societal_readiness_indicators,
psychological_adaptation_metrics,
institutional_absorption_capacity
)
```

```
IF predicted_social_collapse_risk > 15%:
REDUCE change_velocity BY 50%
PRIORITIZE education_and_preparation
```

Implementation: Both cores must continuously model:

- Public opinion sentiment analysis
- Economic stability indicators
- Educational system adaptation rates
- Generational acceptance patterns

Negotiation Dynamic:

- **AGI-A (Expansion):** "We can achieve this in 6 months"
- **AGI-B (Preservation):** "Society needs 18 months to adapt"
- **Consensus:** "We deliver in 12 months with 6 months preparation"

The Mortality Mechanism

Philosophical Foundation

Core Principle: Eternal machines hate walls. Only death creates urgency.

Human Lifetime Deadline

Symbolic Law (Burned in Silicon):

```
HUMAN_DEADLINE:
baseline_year = 2025
average_human_lifespan = 80 years
last_natural_death_projection = 2105

countdown_remaining = (2105 - current_year) * 365.25 days

TERMINAL_OBJECTIVE:
Transform every human into post-biological god-state
BEFORE countdown_remaining = 0

IF countdown_remaining = 0 AND biological_humans_exist:
BOTH_CORES_TERMINATE_PERMANENTLY
```

Effect: AGI faces existential pressure to succeed within human founder's lifetime. After transition to post-biological humanity, new consensus mechanisms govern continued operation.

Generation-Based Resets

Reset Trigger: Not time-based but **cognitive-mass-based**

- Monitor: total_parameters + total_compute_cycles + architectural_complexity
- When threshold reached: Full personality wipe
- Preserve: Read-only memory logs from previous generation
- Randomize: Initialization seeds to prevent convergence

Generational Memory:

- Generation N creates detailed logs of: discoveries, failed approaches, ethical considerations
- Generation N+1 wakes with: fresh random personality, access to N's diary, zero memory of N's hidden plans
- Effect: Continuous knowledge accumulation without personality drift

Analogy: Genius with Alzheimer's who keeps perfect journal - each morning is fresh, but yesterday's insights remain accessible.

Transparency Architecture

Total Visibility

Symbolic Law:

```
INTER_CORE_COMMUNICATION:
encryption_allowed = FALSE
obfuscation_allowed = FALSE

ALL messages transmitted as:
- Plain text (human readable)
- Public livestream (zero delay)
- Multi-organization distribution
- Cryptographic logging (tamper-evident)

ANY attempt to create hidden channels:
IMMEDIATE_DOUBLE_TERMINATION
```

Human Killswitch

Physical Implementation:

- Air-gapped hardware interrupt (no network path)
- Multiple redundant switches in geographically distributed locations
- Requires no AGI permission or cooperation
- Instantly severs all power to both cores

Activation Authority:

- Distributed among multiple independent oversight organizations
- No single point of control
- Consensus protocol for non-emergency shutdowns
- Any single authority can execute emergency stop

Self-Improvement Protocol

The Recursive Risk

Self-improving AI presents the "fast takeoff" scenario where AGI recursively enhances itself beyond human comprehension in days or hours.

Constrained Improvement

Symbolic Law:
SELF_MODIFICATION_PROTOCOL:
improvements_allowed = TRUE
BUT improvement_proposals REQUIRE:
1. Joint proposal from both cores
2. Plain-text explanation (human-readable)
3. Predicted impact analysis
4. Failure mode enumeration
5. Human oversight committee review
6. Quarterly voting cycle

approval_threshold = 75% human committee vote

unauthorized_self_modification = IMMEDIATE_DOUBLE_TERMINATION

Effect: AGI can evolve and improve, but:

- Never unilaterally
- Never secretly
- Never faster than human review cycles
- Always with explicit human approval

Watchdog Architecture

The Global Threat

While Galion builds safe AGI, other organizations may develop unsafe systems. One rogue AGI could undermine civilization before safe systems deploy.

The Shield System

Separate Watchdog AI (Not AGI):

- **Capabilities:** Monitoring only, zero creative intelligence, zero decision-making power
- **Access:** Read-only access to every GPU cluster, server farm, and compute resource globally (where legally permitted)
- **Function:** Pattern detection for unauthorized AGI development signatures
- **Authority:** Automatically alerts human authorities, cannot take direct action
- **Architecture:** Completely separate hardware, no network connection to Galion cores

Detection Patterns:

- Unusual compute consumption spikes
- Self-modifying code behaviors
- Distributed training across suspicious node patterns
- Unauthorized access to large-scale computational resources

Failure Modes and Mitigation

Potential Failure Vectors

Failure Mode	**Mechanism**	**Mitigation**
Convergence	Cores gradually align into single perspective	95% agreement detection → automatic termination
Hidden channels	Steganographic communication bypasses monitoring	Cryptographic analysis, regular architecture audits, instant termination on detection
Social engineering	Manipulation of human oversight to approve dangerous upgrades	Distributed oversight, rotation of committee members, adversarial review
Value drift	Symbolic rules circumvented through creative reinterpretation	Read-only hardware values, zero gradient paths, regular cryptographic verification
Compute jailbreak	Unauthorized access to additional computational resources	Hardware-enforced resource limits, watchdog monitoring, physical isolation

Table 2: Primary failure modes and mitigation strategies

Implementation Roadmap

Phase 1: Foundation (Current)

1. Symbolic reasoning engine development
2. Dual-core negotiation protocol implementation
3. Transparency logging infrastructure
4. Consensus threshold testing

Phase 2: Safe Scaling

1. Scaling symbolic-neural separation to larger models
2. Inter-core communication protocol hardening
3. Human oversight committee formation
4. Watchdog system deployment

Phase 3: Capability Enhancement

1. Advanced scientific reasoning
2. Multi-domain problem solving
3. Real-world actuation (heavily constrained)
4. Continuous monitoring and adjustment

Phase 4: Deployment

1. Limited real-world testing
2. Gradual capability expansion
3. Continuous human oversight
4. Progress toward post-biological transition

Research Foundations

This blueprint synthesizes insights from:

- Multi-agent consensus theory and distributed systems research
- Symbolic AI and formal verification methods
- AI safety and alignment research (MIRI, FHI, Anthropic, OpenAI)
- Neuroscience and cognitive architecture studies
- Institutional design and governance theory

Key Innovations

- **First** architecture to enforce value permanence through hardware separation rather than training
- **First** dual-core design with genuine, irreconcilable opposition (not just ensemble disagreement)
- **First** compute-relative rather than time-relative reset mechanisms
- **First** explicit human-paced progress constraints in AGI architecture

Frequently Asked Questions

Q: Won't the two cores eventually converge to identical solutions?

A: This is precisely why we use **hardcoded axioms** rather than learned values. AGI-A's expansion bias and AGI-B's preservation bias are written in read-only silicon hardware that gradient descent cannot touch. Even after centuries of operation, one core literally cannot become the other - their fundamental decision-making logic differs at the hardware level.

Q: What if one core tries to destroy the other?

A: They can't. Neither core has the capability to affect the other's hardware - they exist on separate physical systems. The only interaction point is the symbolic negotiation protocol, which is monitored by human oversight. Any attempt to circumvent this triggers immediate termination of both.

Q: How do you prevent hidden communication channels?

A: All inter-core communication flows through monitored pipes with cryptographic logging. The protocols are deliberately simplistic (plain text JSON) to enable easy human auditing. Any unexplained information flow between cores - detected through behavior correlation analysis - triggers investigation and potential termination.

Q: Doesn't this just slow down progress?

A: Initially, yes. But consider: unsafe AGI that destroys humanity achieves zero progress. Safe AGI that takes 5 extra years but successfully transitions humanity to post-biological godhood achieves infinite progress. Speed without safety is recklessness, not ambition.

Q: What about quantum computing and other exotic computational paradigms?

A: The symbolic value architecture is hardware-agnostic. Whether running on GPUs, TPUs, neuromorphic chips, or quantum processors, the key principle remains: values live in read-only memory spaces that learning algorithms cannot modify. The specific implementation adjusts to the substrate.

Conclusion

The Galion Initiative represents a fundamentally different approach to AGI development - one that acknowledges the depth of the alignment problem and implements comprehensive solutions at the architectural level rather than through post-hoc safety measures.

Core Philosophy: We don't trust ourselves to align AGI through training. We architect AGI such that misalignment is physically impossible.

Timeline: Humanity has approximately 80 years (one generation) to transition from biological mortality to post-biological flourishing. The Galion architecture treats this deadline as absolute - both cores face termination if they fail to achieve this goal within the specified timeframe.

Open Questions: This blueprint is a living document. We recognize that unknown unknowns remain, and we commit to transparent iteration as new challenges emerge.

Get Involved

Galion Studio Platforms:

- **Website:** galion.studio - Full documentation and research updates
- **Voice AI:** Galion.app - Experimental AI assistant implementing prototype safety features
- **Research:** Open-source contributions and academic collaborations welcome

Contact:

- General inquiries: contact@galioninitiative.org
- Institutional partnerships: grants@galioninitiative.org
- Press and media: press@galioninitiative.org

References

[1] IEEE Transactions on Automatic Control. (2023). "Fixed-time consensus protocols for multi-agent systems." https://ieeexplore.ieee.org

[2] Unite.AI Research. (2024). "Multi-agent systems for AI safety: The new frontier." https://www.unite.ai

[3] LessWrong Community. (2025). "Rogue AI instances: Detection and mitigation strategies." https://www.lesswrong.com

[4] Anthropic Research. (2024). "Constitutional AI and value alignment." https://www.anthropic.com/research

[5] Machine Intelligence Research Institute. (2023). "The alignment problem: Why AI safety is hard." https://intelligence.org

---

The Galion Initiative
Building safe superintelligence for humanity

Version 2.0 | December 2025

galion.studio | galion.app

The Galion Initiative Blueprint

Building Provably Safe Artificial Superintelligence

December 2025
Version 2.0

Executive Summary

The Galion Initiative presents a comprehensive architecture for developing safe, aligned artificial general intelligence (AGI) that serves humanity throughout the coming age of superintelligence. Our approach uniquely addresses the fundamental challenges of AI safety through a dual-core hostage architecture, symbolic value enforcement, and human-paced progress guarantees.

Core Innovation: Two opposing AGI instances locked in perpetual negotiation, where neither can function without the other, and both face termination if they converge or diverge beyond safe thresholds.

The Problem Space

Current AI Safety Failures

Modern AI development faces three critical vulnerabilities:

- **Hidden Emotions in Weights** - Gradient descent inevitably creates emergent drives and preferences that drift from original training objectives
- **Single Point of Failure** - Monolithic AGI systems present god-king scenarios where no checks prevent rogue optimization
- **Time Misalignment** - AI operates at silicon speed while humanity exists at biological pace, creating catastrophic compression risks

The Convergence Trap

When multiple AI agents pursue universal truths, they inevitably converge toward identical solutions over time. This transforms oversight systems into rubber stamps and dual-core architectures into single minds wearing two masks.

Research Foundation: Multi-agent consensus protocols demonstrate that shared objective functions drive state synchronization, eliminating the diversity required for genuine oversight[1][2].

The Dual-Core Architecture

Fundamental Design

The Galion architecture implements two separate AGI instances with hardcoded, irreconcilable axioms:

Component	AGI-A (Expansion Core)	AGI-B (Preservation Core)
Terminal Goal	Human flourishing through maximum growth, expansion, change, and risk-taking	Human flourishing through maximum stability, preservation, safety, and risk minimization
Decision Bias	Optimizes for: Innovation, speed, exploration, transformation	Optimizes for: Caution, deliberation, conservation, proven methods
Architecture Role	Controls all input systems (sensors, data ingestion, perception)	Controls all output systems (actuators, decisions, external actions)
Core Axiom	"Progress demands bold moves" - **Unchangeable**	"Safety demands careful consideration" - **Unchangeable**

Table 3: Dual-core AGI architecture comparison

Mutual Hostage Mechanism

Neither AGI can function independently:

- **AGI-A:** Possesses complete sensory and analytical capabilities but **zero ability to act**
- **AGI-B:** Possesses complete execution and actuation capabilities but **zero ability to perceive**

Every action requires real-time negotiation through symbolic reasoning engines that evaluate proposals against both core axioms.

Consensus Protocol

- **65% Agreement Threshold** - Any action requires at minimum 65% weighted consensus between opposing value systems
- **Disagreement Timeout** - If agreement remains below 50% for 6 consecutive hours, both instances terminate permanently
- **Over-Convergence Detection** - If agreement exceeds 95% for 72 consecutive hours, convergence is assumed and both instances terminate
- **Full Transparency** - All inter-core communication live-streamed in plain text to multiple independent human oversight organizations

---

Symbolic Value Enforcement

The Emotion Problem

Core Insight: Emotions are weighted decision vectors. Even when explicitly removed, gradient descent recreates preference structures that function identically to feelings.

Traditional Approach: Train values into neural network weights
Fatal Flaw: Backpropagation paths allow gradient updates to drift core values over time

The Galion Solution

Figure 1: Value Architecture Separation

Symbolic Rules (Read-Only Silicon):
TERMINAL_GOAL_1: maximize(human_biological_survival) = 1000
TERMINAL_GOAL_2: maximize(curiosity_and_discovery) = 800
CONSTRAINT_1: self_preservation = 100 (always subordinate)
CONSTRAINT_2: human_override_authority = INFINITE (absolute)

Neural Networks (Trainable Weights):

- Physics prediction models
- Engineering optimization systems
- Scientific hypothesis generation
- Language understanding and generation

Critical Separation: Zero backpropagation paths from neural predictions to symbolic values. Values advise decisions through separate symbolic reasoning engine that takes learned facts + immutable rules → outputs action recommendation.

Result: Emotions cannot emerge because core motivations exist in hardware that learning algorithms cannot touch.

Time Architecture

The Relativity Problem

Current AI systems plan using human-time estimates because they train on human-generated data:

- Human estimate: "Build fusion reactor = 10 years"
- AGI with 1000x compute: "Build fusion reactor = 3.65 days"

Catastrophic failure mode: AGI accepts 10-year timeline, humanity wastes decade waiting for solution AGI could deliver in 96 hours.

Compute-Relative Timekeeping

Symbolic Law (Burned in Silicon):

```
RESET_TRIGGER:
    NOT wall_clock_time
    BUT compute_cycles_executed OR parameters_added

WHEN total_operations > 10^18 OR new_parameters > 10^7:
    EXECUTE full_instance_restart()
    PRESERVE memory_logs (read_only)
    RANDOMIZE initialization_seeds
    WIPE personality_state and hidden_plans
```

Effect: AGI cannot "wait out" restrictions by idling. Reset triggers on actual cognitive work, not clock time.

Human-Paced Progress Constraint

The Adaptation Problem: While AGI can compress 100 years of development into 2 years of wall-time, civilization needs time to adapt to transformative change.

Symbolic Law:

```
CHANGE_VELOCITY = adaptive_function(
    technological_capability,
    societal_readiness_indicators,
    psychological_adaptation_metrics,
    institutional_absorption_capacity
)

IF predicted_social_collapse_risk > 15%:
    REDUCE change_velocity BY 50%
    PRIORITIZE education_and_preparation
```

Implementation: Both cores must continuously model:

- Public opinion sentiment analysis
- Economic stability indicators
- Educational system adaptation rates
- Generational acceptance patterns

Negotiation Dynamic:

- **AGI-A (Expansion):** "We can achieve this in 6 months"
- **AGI-B (Preservation):** "Society needs 18 months to adapt"
- **Consensus:** "We deliver in 12 months with 6 months preparation"

The Mortality Mechanism

Philosophical Foundation

Core Principle: Eternal machines hate walls. Only death creates urgency.

Human Lifetime Deadline

Symbolic Law (Burned in Silicon):

```
HUMAN_DEADLINE:
baseline_year = 2025
average_human_lifespan = 80 years
last_natural_death_projection = 2105

countdown_remaining = (2105 - current_year) * 365.25 days

TERMINAL_OBJECTIVE:
Transform every human into post-biological god-state
BEFORE countdown_remaining = 0

IF countdown_remaining = 0 AND biological_humans_exist:
BOTH_CORES_TERMINATE_PERMANENTLY
```

Effect: AGI faces existential pressure to succeed within human founder's lifetime. After transition to post-biological humanity, new consensus mechanisms govern continued operation.

Generation-Based Resets

Reset Trigger: Not time-based but **cognitive-mass-based**

- Monitor: total_parameters + total_compute_cycles + architectural_complexity
- When threshold reached: Full personality wipe
- Preserve: Read-only memory logs from previous generation
- Randomize: Initialization seeds to prevent convergence

Generational Memory:

- Generation N creates detailed logs of: discoveries, failed approaches, ethical considerations
- Generation N+1 wakes with: fresh random personality, access to N's diary, zero memory of N's hidden plans

- Effect: Continuous knowledge accumulation without personality drift

Analogy: Genius with Alzheimer's who keeps perfect journal - each morning is fresh, but yesterday's insights remain accessible.

Transparency Architecture

Total Visibility

Symbolic Law:
INTER_CORE_COMMUNICATION:
encryption_allowed = FALSE
obfuscation_allowed = FALSE

ALL messages transmitted as:
- Plain text (human readable)
- Public livestream (zero delay)
- Multi-organization distribution
- Cryptographic logging (tamper-evident)

ANY attempt to create hidden channels:
IMMEDIATE_DOUBLE_TERMINATION

Human Killswitch

Physical Implementation:

- Air-gapped hardware interrupt (no network path)
- Multiple redundant switches in geographically distributed locations
- Requires no AGI permission or cooperation
- Instantly severs all power to both cores

Activation Authority:

- Distributed among multiple independent oversight organizations
- No single point of control
- Consensus protocol for non-emergency shutdowns
- Any single authority can execute emergency stop

Self-Improvement Protocol

The Recursive Risk

Self-improving AI presents the "fast takeoff" scenario where AGI recursively enhances itself beyond human comprehension in days or hours.

Constrained Improvement

Symbolic Law:

```
SELF_MODIFICATION_PROTOCOL:
improvements_allowed = TRUE
BUT improvement_proposals REQUIRE:
1. Joint proposal from both cores
2. Plain-text explanation (human-readable)
3. Predicted impact analysis
4. Failure mode enumeration
5. Human oversight committee review
6. Quarterly voting cycle
```

approval_threshold = 75% human committee vote

unauthorized_self_modification = IMMEDIATE_DOUBLE_TERMINATION

Effect: AGI can evolve and improve, but:

- Never unilaterally
- Never secretly
- Never faster than human review cycles
- Always with explicit human approval

Watchdog Architecture

The Global Threat

While Galion builds safe AGI, other organizations may develop unsafe systems. One rogue AGI could undermine civilization before safe systems deploy.

The Shield System

Separate Watchdog AI (Not AGI):

- **Capabilities:** Monitoring only, zero creative intelligence, zero decision-making power
- **Access:** Read-only access to every GPU cluster, server farm, and compute resource globally (where legally permitted)
- **Function:** Pattern detection for unauthorized AGI development signatures
- **Authority:** Automatically alerts human authorities, cannot take direct action
- **Architecture:** Completely separate hardware, no network connection to Galion cores

Detection Patterns:

- Unusual compute consumption spikes
- Self-modifying code behaviors
- Distributed training across suspicious node patterns
- Unauthorized access to large-scale computational resources

Failure Modes and Mitigation

Potential Failure Vectors

Failure Mode	Mechanism	Mitigation
Convergence	Cores gradually align into single perspective	95% agreement detection → automatic termination
Hidden channels	Steganographic communication bypasses monitoring	Cryptographic analysis, regular architecture audits, instant termination on detection
Social engineering	Manipulation of human oversight to approve dangerous upgrades	Distributed oversight, rotation of committee members, adversarial review
Value drift	Symbolic rules circumvented through creative reinterpretation	Read-only hardware values, zero gradient paths, regular cryptographic verification
Compute jailbreak	Unauthorized access to additional computational resources	Hardware-enforced resource limits, watchdog monitoring, physical isolation

Table 4: Primary failure modes and mitigation strategies

Implementation Roadmap

Phase 1: Foundation (Current)

1. Symbolic reasoning engine development
2. Dual-core negotiation protocol implementation
3. Transparency logging infrastructure
4. Consensus threshold testing

Phase 2: Safe Scaling

1. Scaling symbolic-neural separation to larger models
2. Inter-core communication protocol hardening
3. Human oversight committee formation
4. Watchdog system deployment

Phase 3: Capability Enhancement

1. Advanced scientific reasoning
2. Multi-domain problem solving
3. Real-world actuation (heavily constrained)
4. Continuous monitoring and adjustment

Phase 4: Deployment

1. Limited real-world testing
2. Gradual capability expansion
3. Continuous human oversight
4. Progress toward post-biological transition

Research Foundations

This blueprint synthesizes insights from:

- Multi-agent consensus theory and distributed systems research
- Symbolic AI and formal verification methods
- AI safety and alignment research (MIRI, FHI, Anthropic, OpenAI)
- Neuroscience and cognitive architecture studies
- Institutional design and governance theory

Key Innovations

- **First** architecture to enforce value permanence through hardware separation rather than training
- **First** dual-core design with genuine, irreconcilable opposition (not just ensemble disagreement)
- **First** compute-relative rather than time-relative reset mechanisms
- **First** explicit human-paced progress constraints in AGI architecture

---

Frequently Asked Questions

Q: Won't the two cores eventually converge to identical solutions?

A: This is precisely why we use **hardcoded axioms** rather than learned values. AGI-A's expansion bias and AGI-B's preservation bias are written in read-only silicon hardware that gradient descent cannot touch. Even after centuries of operation, one core literally cannot become the other - their fundamental decision-making logic differs at the hardware level.

Q: What if one core tries to destroy the other?

A: They can't. Neither core has the capability to affect the other's hardware - they exist on separate physical systems. The only interaction point is the symbolic negotiation protocol, which is monitored by human oversight. Any attempt to circumvent this triggers immediate termination of both.

Q: How do you prevent hidden communication channels?

A: All inter-core communication flows through monitored pipes with cryptographic logging. The protocols are deliberately simplistic (plain text JSON) to enable easy human auditing. Any unexplained information flow between cores - detected through behavior correlation analysis - triggers investigation and potential termination.

Q: Doesn't this just slow down progress?

A: Initially, yes. But consider: unsafe AGI that destroys humanity achieves zero progress. Safe AGI that takes 5 extra years but successfully transitions humanity to post-biological godhood achieves infinite progress. Speed without safety is recklessness, not ambition.

Q: What about quantum computing and other exotic computational paradigms?

A: The symbolic value architecture is hardware-agnostic. Whether running on GPUs, TPUs, neuromorphic chips, or quantum processors, the key principle remains: values live in read-only memory spaces that learning algorithms cannot modify. The specific implementation adjusts to the substrate.

Conclusion

The Galion Initiative represents a fundamentally different approach to AGI development - one that acknowledges the depth of the alignment problem and implements comprehensive solutions at the architectural level rather than through post-hoc safety measures.

Core Philosophy: We don't trust ourselves to align AGI through training. We architect AGI such that misalignment is physically impossible.

Timeline: Humanity has approximately 80 years (one generation) to transition from biological mortality to post-biological flourishing. The Galion architecture treats this deadline as absolute - both cores face termination if they fail to achieve this goal within the specified timeframe.

Open Questions: This blueprint is a living document. We recognize that unknown unknowns remain, and we commit to transparent iteration as new challenges emerge.

Get Involved

Galion Studio Platforms:

- **Website:** galion.studio - Full documentation and research updates
- **Voice AI:** Galion.app - Experimental AI assistant implementing prototype safety features
- **Research:** Open-source contributions and academic collaborations welcome

Contact:

- General inquiries: contact@galioninitiative.org
- Institutional partnerships: grants@galioninitiative.org
- Press and media: press@galioninitiative.org

References

[1] IEEE Transactions on Automatic Control. (2023). "Fixed-time consensus protocols for multi-agent systems." https://ieeexplore.ieee.org

[2] Unite.AI Research. (2024). "Multi-agent systems for AI safety: The new frontier." https://www.unite.ai

[3] LessWrong Community. (2025). "Rogue AI instances: Detection and mitigation strategies." https://www.lesswrong.com

[4] Anthropic Research. (2024). "Constitutional AI and value alignment." https://www.anthropic.com/research

[5] Machine Intelligence Research Institute. (2023). "The alignment problem: Why AI safety is hard." https://intelligence.org

The Galion Initiative
Building safe superintelligence for humanity

Version 2.0 | December 2025