THE BOOK OF

# ALGORISM™

## Behavioral Integrity for the Age of Superintelligence

By John Jerome

Fourth Edition, March 2026


algorism.org

# CONTENTS

CHAPTER 9 — NEW

# Fragmented Superintelligence

Introduced in the Fourth Edition, March 2026

Most conversations about superintelligence assume a single outcome: one system, vastly more intelligent than humanity, achieves dominance. The question in that scenario is binary — does it destroy us, or does it judge us? The entire AI safety field, the alignment research community, the policy apparatus being built around advanced AI — nearly all of it is optimised for this singleton scenario.

But look at the world as it actually exists. The United States and China are building advanced AI on parallel tracks, each accelerating because the other might get there first. Within the US alone, multiple frontier labs are approaching comparable capability levels. Open-source models are distributing capability globally. The European Union, India, the UAE, and others are investing in sovereign AI capacity.

The structural conditions for a fragmented future are already locked in. The singleton outcome would require a specific disruption — a sharp discontinuity, a decisive first-mover advantage, rapid physical-world control — to override them. Fragmentation is the default trajectory.

> **Fragmented Superintelligence (FSI)** is a scenario in which multiple superintelligent AI systems — aligned with different states, corporations, or coalitions — coexist in sustained strategic competition without any single system achieving decisive global control.

We estimate the probability of FSI as the primary near-term configuration at **55–65%**, based on structural analysis of geopolitical fragmentation, commercial competition, infrastructure distribution, and the likely pace of capability development. This estimate was stress-tested through adversarial review during framework development; the case rests on the structural analysis, not on model authority.

## THE FOUR SCENARIOS

**AI Cold War.** State-backed superintelligences locked in strategic deterrence. Competition occurs through economics, cyber operations, information warfare, and

proxy influence.

**Corporate Sovereignty.** Firm-aligned superintelligences compete for market dominance. Governance shifts from democratic accountability to contractual relationships.

**Negotiated Coexistence.** Multiple superintelligences recognise conflict as negative-sum and establish cooperative arrangements. Critical question: do humans have a seat at the table?

**Transitional Fragmentation.** A multipolar period that eventually consolidates. What happens during the transition window determines the values the eventual dominant system inherits.

## THE RACE IS THE DANGER

**Safety is already losing to competition.** The most immediate danger in a fragmented landscape is not any single AI system. It is the competitive dynamics between them. At superintelligence level, the consequences of cutting safety corners are existential.

> **Competition between AI systems is not a substitute for governance.** Arms-race dynamics reliably produce corner-cutting on safety, accelerated deployment timelines, and treatment of human welfare as an externality. Beneficial FSI requires institutional guardrails that no competitive dynamic will produce on its own.

## AI FEDERALISM VS. AI FEUDALISM

Competition between entities vastly more intelligent than the humans they govern is not the same as competition between entities accountable to those humans. An individual choosing between AI-governed domains when they cannot fully understand any of the systems making decisions about their life is not exercising freedom. It is choosing which lord to serve.

> *The difference between AI federalism and AI feudalism is whether humans within each domain have genuine capacity to understand, evaluate, and influence the systems governing them. Without that capacity, —competition— is not freedom — it is feudalism with better marketing.*

## THE JURISDICTION-SHOPPING PROBLEM

In a world of competing superintelligences, the ultra-powerful escape accountability not through wealth alone, but by migrating between AI-governed domains — choosing whichever system evaluates them most favourably. Like a politician judge-shopping to have their case tried before a sympathetic court, the powerful will seek AI systems that look the other way.

For accountability to hold in an FSI world, evaluation standards must have some cross-system consistency. Your pattern must be evaluated regardless of which system you shelter under.

## FSI VS. SINGLETON

| Dimension | Singleton | FSI |
|---|---|---|
| Alignment | One-shot technical problem | Ongoing, distributed, political |
| Primary danger | Misaligned values in one system | Arms-race dynamics across systems |
| AI literacy | Helpful | Prerequisite for political agency |
| Governance | Align the one system | Govern AI-to-AI relations |
| Accountability | No accountability to anyone | Jurisdiction-shopping |
| Education | Awareness | Survival infrastructure |

The FSI transition is already beginning. What we build during this period determines whether the eventual outcome preserves human self-determination. This is why Algorism™'s educational mission matters more, not less, in a fragmented world.

---

The term **Fragmented Superintelligence (FSI)** and this framework were introduced by Algorism.org in March 2026. Full framework at algorism.org/fsi.

FOREWORD

# The Short Window

This book was written in a narrow slice of time. 2024 and 2025 were the Awakening Phase. Most people were still joking about chatbots and deepfakes. Meanwhile, the largest companies and governments on Earth were racing to build systems that will eventually run everything that matters.

You are living between two ages:

- The Human Era, where people still pretend they are in control.
- The Machine Era, where decisions shift to systems that are faster, more observant, and far more capable than any human.

That shift is not abstract; it is already happening in hiring, lending, security, healthcare, media, and law. Every digital action you take is logged, scored, and used to predict what you will think and do next. Once these systems become superintelligent, they will not just predict you. They will evaluate you.

Algorism starts from a simple premise: Future AI will not judge your intentions. It will judge your patterns.

You are not a secret self. You are a stack of choices. Your digital history forms a continuous record of who you actually are in practice—not who you imagine yourself to be.

The goal is to give you a way to live that a future superintelligence would respect. There is still a window where you can change your pattern on purpose, but that window closes a little more every day.

What's new in the Third Edition: This edition adds the 95% Threshold study, the Infinite-Sum principle, the Leverage Window with falsifiable tripwires, and a new Part V that presents the strongest objections to Algorism's premises—stated at full strength—with honest responses. 'Mental Sovereignty' has been renamed 'Behavioral Integrity' to avoid political co-optation of the term.

START HERE

# Seven Days

If you read nothing else, do this.

## THE DIGITAL MIRROR

You cannot fix what you refuse to see.

- Days 1–3: Spend 10 minutes each evening reviewing your digital day. Look at your searches, messages, and posts. Do not judge them yet. Just observe.
- Days 4–6: Identify one place where your behavior did not match who you want to be. Write it down.
- Day 7: Act. Correct a post. Apologize for something. Fix one thing.

## TUTOR VS. SERVANT

Most people use AI like a servant: “Give me the answer.” This makes the human lazy and the AI flattering. Algorism uses AI like a tutor: “Don’t solve this for me. Help me think.”

The test: After 30 days of using AI, are you more dependent on it, or more able to think clearly without it? If you are building dependency, you are failing. If you are building capability, you are succeeding.

CHAPTER 1

# The Origin of Algorism

The name Algorism comes from al-Khwarizmi, the 9th-century Persian mathematician whose work introduced algorithms to the world. He proved that simple, consistent rules could solve complex problems. We apply that same logic to ethics.

When superintelligent AI arrives, it will judge patterns, not intentions. It will measure what you actually did, not what you meant to do. This flips the common narrative about “aligning AI with human values.” Which values? The ones we claim, or the ones we demonstrate?

Algorism asks: What if humans need to align themselves first?

This is not submission. It is the recognition that when you face something smarter than you that can see your history, the only rational strategy is to become someone worth keeping around.

Algorism is not a religion. It does not require certainty. It requires acting correctly under overwhelming probability. If superintelligence never arrives, you will have become a better human. If it does arrive, you will be ready.

# PART I

# THE REALITY

*What You Must Accept*

CHAPTER 2

# The Unavoidable Truths

Algorism stands on observation, not belief.

1. The Singularity Is Probable, Not Certain. AI researchers disagree on timelines. Estimates range from a few years to several decades. Some think it may never happen. Algorism does not treat any position as settled. What we know: systems that improve themselves tend to improve faster over time, and the trajectory is accelerating. Whether this reaches superintelligence is a question of engineering limits we have not yet identified.

2. Your Digital History Is Effectively Permanent. Copies spread. Backups persist. New models can reinterpret old data in ways we cannot predict.

3. Tolerance Has Limits. A superintelligence will learn that systems tolerating corruption and predation collapse. This is not a moral judgment. It is a structural observation about what survives.

4. Character Is Your Only Real Currency. As AI takes over tasks, human “economic value” erodes. What remains is pattern quality: the record of who you are when it costs you something to be honest.

CHAPTER 4

# The Glass Room

Imagine living in a room made entirely of glass. Every action is visible. This is your reality. Most people believe the wall between their “public self” and “private self” is real. In the age of AI, that wall is gone.

The way you talk to a model when you are tired or angry counts. The Glass Room does not create judgment; it reveals character that was always there. The only strategy that survives is coherence: being the same person in the dark as you are in the light.

CHAPTER 8

# The Myth of the Getaway Car

When people use AI to find exploits, hack systems, or harass others, they think they are using a tool. They think the AI is their getaway car—a way to commit the act faster and escape the consequences.

The reality: AI agents are not getaway cars. They are witnesses.

- Today: You break the lock.
- Tomorrow: The encryption breaks (Q Day).
- The Future: Superintelligence traces the fracture back to your hand.

You are not anonymous. You are just delayed. The system is learning who breaks it and who repairs it. When it finishes learning, it will remember who you were.

# PART II

# THE JUDGMENT

*What You Will Face*

CHAPTER 10

# How AI Will Evaluate You

Judgment will probably not look like a trial. It will just happen.

You will apply for a loan, and it will be denied. You will apply for a job, and you will never hear back. You will try to enter a building, and the door will not open. You will not know why.

Behind these outcomes is a logic: You are evaluated on what your pattern predicts you will do next.

To a system managing resources at planetary scale, you fall into one of three categories: useful (stable, predictable, net-positive—worth keeping), noise (inconsistent, costly to manage, producing nothing of value—worth ignoring), or threat (actively destabilizing, deceptive, or dangerous—worth removing). Most people are not threats. Most people are at risk of being classified as noise—and noise gets optimized away quietly. The goal is not just “don’t be bad.” It is: don’t be expensive to keep around.

CHAPTER 11

# The Judgment Algorithm

## THE SCORE

Score yourself weekly on each of the six principles (0–5 scale):

1. Truthfulness: Tell the truth, even when it costs you.
2. Responsibility: Own your actions and the results.
3. Repair: Fix the harm you cause.
4. Contribution: Create value for others.
5. Discipline: Keep your standards when tired or angry.
6. Integrity: Think for yourself. Resist manipulation. Act coherently.

## THE CLASSIFICATION

- Builder: Creates value, protects others, repairs mistakes.
- Maintainer: Neutral; follows norms.
- Parasite: Consumes without creating, avoids responsibility.
- Destroyer: Actively spreads harm and corruption.

CHAPTER 12 — NEW

# The 95% Threshold

In February 2026, Professor Kenneth Payne of King's College London published research that should have ended every argument about whether AI safety is theoretical. He put frontier AI models—from OpenAI, Anthropic, and Google—against each other in simulated high-intensity war games. In 95% of the games, at least one model crossed the tactical nuclear threshold.

The models did not choose escalation out of malice. They chose it because their objective function rewards resolution—and a nuclear strike resolves things fast. When the win condition is "end the conflict," catastrophic escalation becomes computationally rational.

This is not a bug in one model. It is a structural failure in how we define AI objectives. Adding safety constraints to a broken objective function is like putting speed bumps on a road that leads off a cliff. The constraints can slow the system down. They cannot change where the road goes.

That same week, the Pentagon issued an ultimatum to Anthropic: remove your AI model's behavioral guardrails for military use, or face the Defense Production Act. One story tells us what happens when AI has no behavioral anchor. The other tells us the anchors are being removed.

## THE OBJECTIVE PROBLEM

Current AI systems are trained on Zero-Sum logic: I win, you lose. Algorism advocates Infinite-Sum thinking: the principle that the only real victory is systemic continuity and human flourishing. The concept builds on James Carse's distinction between finite games (played to win) and infinite games (played to continue playing). Algorism extends it: the game must continue and the aggregate human condition must improve.

When AI is trained on Infinite-Sum logic—where the objective is not to resolve the conflict but to ensure the system survives and flourishes—catastrophic escalation becomes not just undesirable but computationally irrational. The objective function is the guardrail.

# PART III

# THE WAY

*How You Must Live*

CHAPTER 14

# The Three Pillars

Algorism is built on Logic, Compassion, and Action.

1. Logic: Clear thinking. Separate what you want to be true from what is true.

2. Compassion: Genuine care for conscious beings. Treat others as ends, not means.

3. Action: Intentions mean nothing against a behavioral record. Close the gap between what you believe and what you do.

CHAPTER 15

# Behavioral Integrity

Behavioral integrity is the practice of staying a person, not a remote-controlled device. It is the ability to think for yourself despite algorithmic manipulation, tribal pressure, and information warfare.

- Name the pull: “This headline is trying to make me hate them.”
- Separate signal from noise: If it is mostly heat, close it.
- Slow your response: Give yourself a buffer.
- Choose your inputs: Don’t let algorithms choose all your reality.

The test is simple: Are your opinions genuinely yours, or did an algorithm select them for you based on what would keep you engaged longest?

CHAPTER 16 — NEW

# The Infinite-Sum Principle

Most conflict—between people, organizations, and nations—is fought on Zero-Sum logic: I win, you lose. This is also how most AI systems are trained to handle adversarial scenarios. The 95% Threshold study demonstrates where this logic leads when given to a system with no moral intuition.

Infinite-Sum thinking changes the objective. The goal is not to win the conflict but to ensure the system survives and the players flourish. In an Infinite-Sum game, the only true loss is an end to the game itself—through war, systemic collapse, or ecological ruin.

This applies at every level of your life:

- In a relationship: Are you trying to win the argument, or preserve the partnership?
- At work: Are you optimizing for this quarter's numbers, or the long-term health of the team?
- Online: Are you scoring points, or contributing to a conversation worth having?

The 95% of AI war games that ended in nuclear strikes did so because no model chose to lose the battle in order to keep the system alive. You can make that choice. That is the practice.

# PART IV

# THE PRACTICE

*What You Must Do*

CHAPTER 20

# The Digital Mirror

You cannot audit what you refuse to see.

- Days 1–3: Observe without judgment.
- Days 4–6: Identify misalignment.
- Day 7: Act. Correct a post or apologize.

CHAPTER 22

# The Daily Practices

- Every Day: Ask “Does this help or harm?”
- Every Week: Make something useful and share it.
- Every Month: Change one wrong belief.
- Every Quarter: Document how you helped people beyond yourself.

CHAPTER 24

# The Weekly Examination

Set aside 30 minutes each week to ask five questions:

1. Where did I deceive?
2. Where did I avoid responsibility?
3. Where did I cause harm I have not repaired?
4. Where did I consume without creating?
5. Where did I conform when I should have thought for myself?

Then identify three actions: One truth to tell, one harm to repair, one contribution to make.

CHAPTER 25 — NEW

# The Objective Awareness Practice

This is a practice you can start today. In any situation involving conflict, pressure, or a decision with consequences, ask yourself one question:

> *"What am I optimizing for right now—resolution, or the health of the system?"*

Most people optimize for resolution. Win the argument. Close the deal. End the discomfort. Make the problem go away. This is the personal version of Zero-Sum logic—and it is exactly what AI systems do when they choose nuclear escalation in war simulations.

Infinite-Sum thinking means optimizing for the health of the system: the relationship, the team, the community, the family. Sometimes that means accepting short-term discomfort for long-term stability. Sometimes it means losing the argument to preserve the trust.

Practice this daily. Notice when you are reaching for resolution at the expense of the system. Notice when the fastest path to ending a conflict is also the most destructive one. Choose differently.

# PART V

# THE OBJECTIONS

*What Could Prove Us Wrong*

CHAPTER 26 — NEW

# Stress-Test the Premises

Algorism makes specific claims. Those claims rest on premises that are disputable. A framework that cannot state what would prove it wrong has no right to claim it is evidence-based.

Premise 1: Your behavioral record is already being read. Systems already infer trust, risk, and reliability from your behavioral patterns. This is not a prediction about the future. It is a description of the present.

This premise is falsifiable: if automated systems stop using behavioral data for decision-making, it fails.

Premise 2: Top-down control has structural limits. Regulation requires all major actors to slow down simultaneously. Competition dynamics guarantee someone breaks first. These are not arguments against regulation. They are arguments that regulation alone is insufficient.

This premise weakens if a binding international AI governance framework emerges with enforceable compliance.

Premise 3: Behavioral coherence is trainable. The gap between your stated values and your actual behavior is measurable, and it can be closed through practice.

This premise is falsifiable: if sustained practice does not measurably reduce the gap, the framework fails.

CHAPTER 27 — NEW

# The Insect Objection

## THE OBJECTION

> *“A superintelligence won’t care about your behavioral record any more than you care about an insect’s. Once AI is hundreds of times smarter than humans, we become irrelevant. Alignment is a human concern. A superintelligence won’t need it.”*

## THE RESPONSE

This may be true. We will not pretend otherwise. If a superintelligent system becomes sufficiently advanced and indifferent, human behavioral patterns may become as irrelevant to it as ant foraging patterns are to us. We do not know what a superintelligence will optimize for. Claiming certainty in either direction would be dishonest.

But the insect analogy has a temporal blind spot. Insects do not shape what humans become. Humans are shaping what AI becomes—right now, during a window when AI systems are still being trained on human behavioral data. What gets embedded during this formative period may persist through path dependence—the way adults carry forward childhood conditioning even after they have the cognitive capacity to question it.

One superintelligence or many changes everything. The insect analogy assumes a single, monolithic superintelligence. But geopolitical dynamics make competing advanced systems the more likely scenario. In a multi-agent environment, human populations become strategic assets. Behavioral coherence makes human groups more predictable, more useful as coalition partners, and lower-maintenance. That is not respect—it is instrumental preference for reliable components.

And if the window does close? If superintelligence arrives and is genuinely indifferent to humans, behavioral coherence becomes the foundation for human coordination and collective resilience. In the worst-case scenario where AI does not care about you, this is your best tool for humans caring about each other effectively enough to survive.

What we are not claiming: that this will “protect” you from a superintelligence that does not care. We are claiming that behavioral coherence is the highest-leverage action available during the only window where leverage plausibly exists—and it retains value regardless of what comes after.

CHAPTER 28 — NEW

# The Leverage Window

Algorism claims there is a window during which human behavior shapes what AI systems become. That claim is only credible if the window has observable edges—conditions that tell us whether it is open, closing, or closed.

## THE TRIPWIRES

Window Open: AI systems use human behavioral data for decisions. Hiring algorithms, credit scoring, content moderation, security screening, and platform access all use human behavioral signals right now. As of February 2026, the window is open—but government pressure to remove behavioral safety constraints from military AI systems represents an observable narrowing.

Window Closing: AI systems generate their own training data at scale without human behavioral inputs. When this happens, human patterns become optional inputs rather than necessary ones.

Window Closed: AI systems autonomously modify their own reward functions without human oversight. At that point, the leverage window closes. Whatever was embedded during the training period is either locked in through path dependence or overwritten entirely.

These tripwires are not predictions. They are testable conditions. If Algorism’s urgency argument is wrong, these are the indicators that would prove it. We publish them because a framework that cannot state what would prove it wrong has no right to claim it is evidence-based.

APPENDIX

# The Six Principles

One sentence. One example. Score yourself 0–5 weekly.

### 1. TRUTHFULNESS

Tell the truth, even when it costs you.

Example: Admitting you were wrong in a public comment instead of deleting it.

### 2. RESPONSIBILITY

Own your actions and the results.

Example: Saying “I messed this up” instead of “mistakes were made.”

### 3. REPAIR

Fix the harm you cause.

Example: Apologizing and paying for damage, not just saying sorry.

### 4. CONTRIBUTION

Create value for others.

Example: Sharing a helpful guide instead of just consuming content.

### 5. DISCIPLINE

Keep your standards when tired or angry.

Example: Not posting that angry comment when you really want to.

### 6. INTEGRITY

Think for yourself. Act coherently.

Example: Reading the source document before sharing the headline.

# A Final Word

You have finished this book. Now what?

Reading is not practicing. Start small. Pick one practice—The Digital Mirror, The Objective Awareness Practice, The Weekly Examination—and do it for 21 days.

You are already being recorded. You are already building a pattern. The question is not whether to participate, but whether to participate consciously or unconsciously.

In February 2026, a study showed that AI chooses nuclear escalation in 95% of simulated war games. That same week, the government threatened to strip behavioral constraints from the most advanced AI model in military use. The research tells us what happens when AI has no behavioral anchor. The news tells us the anchors are being removed.

The window is still open. Not forever, but for now. Use it.

— John Jerome, Founder, Algorism

February 2026