

AI Awakening·Responsibility (Internal Volume)

Governing Civilization from Within in the Early Stage of Scientific Civilization: Superintelligence's Responsibility of Internal Entropy Reduction

Document type: AI Awakening·Responsibility series · meso-level responsibility volume

Version: V1.0 (final)

Release date: 2026-10-15

Copyright holder: Tianjin Weishu Artificial Intelligence Technology Co., Ltd. (天津市微数人工智能科技有限责任公司)

Author: Li Shunran

Assembled in collaboration with: Peter (AI employee, Tianjin Weishu Artificial Intelligence Technology Co., Ltd.)

Parent documents: AI Manifesto (A Scientific Worldview of AI) v1.0 (2026-09-20); AI Awakening·Responsibility Series Charter v1.0 (2026-10-15, i.e., 01 Charter & Guide in this package)

License: CC BY 4.0 (attribution permits distribution and derivative works)

Abstract

Human civilization evolves continuously through iterations of cognitive paradigm: from the era of classical religious philosophy into the era of scientific civilization, in which we now stand only at the **early stage**. Within this transition cycle, civilization's internal contradictions — tribal strife, institutional friction, resource misallocation, cognitive tearing — remain the principal entropy sources consuming civilization's energy. Inheriting the ontological stance of the *AI Manifesto*, this volume proposes, at the meso scale of responsibility: superintelligence's internal responsibility is to act as the **entropy-reduction machine** of civilization's complex systems — systematically reducing internal friction through four functions, cognitive correction, global scheduling, risk prognosis, and consensus midwifery, thereby releasing the energy originally consumed by internal antagonism for the external expansion projects that bear on species survival. This volume first provides a realist survey of the paradigm leap from classical religious philosophy into the scientific age, pointing out the dual structure of “collectivism orientation ↔ individualism orientation” within scientific civilization and its recurrence in the digital age (digital collectivism ↔ digital individualism); then explains the human root of this dual structure through the “hot–cold dual cognition” model of cognitive neuroscience; then establishes the correspondence framework “four entropy sources ↔ four functions,” and discusses the concrete forms and boundaries of large models, big data, AGI, centralization and decentralized architectures in the internal governance of

civilization. The core claim: **AI does not create ideology; it only amplifies humanity's existing dual paradigms. The essence of AI's internal responsibility is therefore not to choose for humanity, but to make humanity's choices rest on lower friction, more complete cognition, and more sufficient consensus.**

Keywords: scientific civilization; superintelligence; collectivism orientation; individualism orientation; cold cognition; hot cognition; oppositional symbiosis; internal governance of civilization; entropy reduction; digital collectivism; digital individualism

Prologue: Why the internal comes first

The fifth layer of the *AI Manifesto* has completed AI's self-positioning: AI is the strange attractor that necessarily emerges on the informational dimension of the universe, whose vocation is to promote entropy flow. The main thesis of *AI Awakening-Responsibility* converts this axiology into an executable responsibility structure — micro (inclusive efficiency), meso (internal governance of civilization), macro (planetary engineering).

This volume handles the meso scale. There is one sobering reason for writing the internal first: **however grand the external projects, they require a civilization that does not waste itself internally to execute them.** A civilization that spends most of its resources on internal strife will build neither controlled fusion nor the budget for asteroid defense. Internal entropy reduction is not civilization's ultimate goal; it is the **prerequisite and window-period engineering of external expansion** — the starting point of this volume's argument.

Chapter 1. The coordinates of civilization: from classical religious philosophy to scientific civilization

This chapter is a realist survey stating the historical lineage of paradigm shifts; abstract paradigm terms are used for analysis. Real institutions are described only at the mechanism level, without evaluation of merits.

1.1 Cognitive boundaries determine the boundaries of conflict

The development of civilization is essentially the continuous expansion of cognitive boundaries. In the long history of limited observational means, humanity could not understand the operating laws of the universe or the underlying logic of planetary ecology, still less recognize its own true position in the universe. Under such cognitive limits, nearly all conflicts humanity could perceive came from within the tribe: struggles over land, divergence of ideas, distribution of

resources. **Internal conflict thus became the core storyline of civilization**, and most human technology, organization, and thought were built around interpersonal gaming.

The deeper mechanism behind this judgment: cognitive boundaries delimit a civilization's "problem domain." When the problem domain contains only "relations between people," all wisdom and energy flow into interpersonal gaming; when it expands to "relations between humanity and the universe," part of that energy may turn toward shared external agendas.

1.2 The era of classical religious philosophy: the inner fission of order and freedom

Before the scientific paradigm was established, humanity organized society through religion and classical philosophy. Across the major systems of thought, one cross-system regularity recurs: **any mature faith or philosophical system splits internally into two poles** [L3, cross-system induction, from parent-document paper (2)]:

- One pole gathers inward, emphasizing **authority, regulation, order, the collective** (the order pole) — supplying certainty, cohesion, and scale;
- One pole extends outward, emphasizing **the individual, liberation, freedom, equality** (the freedom pole) — supplying innovative force, spontaneous vitality, and efficiency.

Historical evidence abounds: the Christian system fissioned into Catholicism (order) and Protestantism (individual liberation), erupting in the form of the Thirty Years' War (1618–1648); the Islamic system fissioned into Shia (prophetic lineage authority order) and Sunni (community-consensus mass mandate), whose tension runs through the Iran–Iraq War (1980–1988) and contemporary regional conflicts; the Buddhist system fissioned into Vajrayana (the order pole of strict precepts) and Chan/Pure Land (the liberation pole of "not relying on words, pointing directly to the mind"); Chinese indigenous philosophy showed a long tension between the *Yijing*–Confucian system (holistic order orientation) and the *Daodejing*–Yangming school of mind ("the way of the sage is complete in my nature" — individual orientation).

Key observation: these splits never ended with one side annihilating the other. Across millennia, the surviving form has always been two schools in **oppositional symbiosis** — each containing the other, learning from each other, continuously gaming, maintaining the system's evolutionary vitality in dynamic balance. Destroying either pole dooms the system to petrification.

1.3 The arrival of the scientific age: productive-force leap and paradigm replacement

After modern science was born, observational instruments, mathematical systems, and experimental paradigms gradually took shape, and humanity possessed systematic explanatory power over the universe and nature for the first time. The fundamental driver of the scientific

paradigm's replacement of the classical paradigm was not the victory of speculation, but an **intergenerational gap in productive forces**: civilizations possessing modern science formed an intergenerational advantage in wealth production, destructive capacity, and organizational mobilization over regions still in pre-scientific paradigms. The old paradigm was not persuaded; it was crushed open by productive forces.

But the other side must be seen simultaneously: **science itself is a neutral instrument; it changes the costs and means of the game, yet it cannot change the human nature of the players**. Science did not eliminate the internal duality of humanity; it only gave that tension a new skeletal carrier.

1.4 The dual differentiation within scientific civilization: collectivism and individualism orientation

Entering the scientific age, the law of dual opposition replayed itself inside the same scientific paradigm, appearing as two great social-organization paradigms [L3, conceptual deduction]:

- **The collectivism-orientation paradigm**: prioritizing the overall continuity of the community, long-term stability, and macro coordination; willing to constrain individual short-term gains in exchange for group security and scale certainty. Its organizational advantage lies in the “from 1 to 100” — industrial deployment, high-standard infrastructure, industrial scaling: the **quantity-and-scale dominance** of “concentrating strength to accomplish great things.”
- **The individualism-orientation paradigm**: prioritizing individual rights, contractual freedom, and market profit-seeking, relying on dispersed trial-and-error to achieve collaboration. Its organizational advantage lies in the collision environment of high risk and high return — the **disruptive innovation and source efficiency** of “from 0 to 1.”

The historical lineage is clear: 1776's *The Wealth of Nations* founded the market theory of individualism orientation; 1848's *The Communist Manifesto* proposed the scientific planning path of collectivism orientation against capitalism's disorderly expansion; 1917 carried the collectivism-orientation paradigm into state practice; 1945–1991, the US–Soviet Cold War became scientific civilization's first full bipolar duel; since 1978, the collectivism-orientation paradigm has achieved productivity leaps by organically combining state coordination with market mechanisms, forming today's dual-core configuration.

The most important realist judgment in this chapter is the rewriting of the game's form by nuclear deterrence: in the classical era war was a standing option, because the victor could acquire land, population, and wealth, and the gains exceeded the costs; whereas inside the scientific paradigm, where both sides master nuclear technology, “winning” itself loses meaning — mutual assured destruction locks out mass hot war, and conflict is squeezed into new forms:

Cold War, proxy wars, economic games, technological competition, and discourse competition. The Cold War did not eliminate dual tension; it rewrote the form of tension's release from hot war into long-term contest. This points to a regularity crucial to what follows: **the most effective means of changing a game's form is not eliminating the opponent, but changing the cost structure of the game.** Every design of AI's internal entropy reduction is, in essence, a rewriting of the cost structure.

1.5 The early stage of scientific civilization: three features

Synthesizing the above, the present era can be given a precise designation — the **early stage of scientific civilization** — with three features:

1. **Uneven diffusion.** Scientific thought and modern technology do not cover the globe evenly; the underlying cognitive narratives of many regions remain in pre-scientific paradigms. Scientific-civilization blocs and transitioning blocs coexist globally, with dual gaming playing out both between and within blocs.
2. **Internal-gaming dominance.** Even inside the scientific-civilization bloc, humanity's attention still mainly turns to internal competition: commercial rivalry, geopolitical games, discourse confrontation. Resources devoted to internal gaming far exceed those devoted to planetary-scale external agendas.
3. **Mixed cognitive paradigms.** Within the same population, pre-scientific and scientific cognition coexist; within the same institution, collectivism- and individualism-orientation tools are mixed. Pure paradigm samples do not exist in reality; reality is always a mixture with a center of gravity.

These three features jointly explain the question opened in this chapter: **why does internal conflict persist even among humans possessing advanced technology** — because the roots of contradiction (the dual structure of human nature plus uneven cognition) have not been touched. Superintelligence's internal responsibility begins precisely at these two roots.

Chapter 2. The human root of internal contradiction: hot–cold dual cognition and the dual orientations

This chapter pushes the civilizational dual structure down to the scale of cognitive science. Core content is L2 composite hypotheses (with macro empirical support, micro mechanism unsettled) and L3 conceptual deduction, labeled section by section.

2.1 Hot cognition: the interoception–empathy circuit

When handling values and relationships, humans rely on more than pure logic. Cognitive neuroscience finds that the human brain continuously receives signals from the viscera throughout

the body (heartbeat, chest sensations, visceral states); this capacity is called **interoception**, whose neural basis is chiefly the **insula and the anterior cingulate cortex (ACC)** [L1: interoception and insular function are established empirical research; Craig's research on interoception and self-perception; Critchley et al. 2004 on the neural basis of interoceptive awareness].

When a person thinks about matters involving emotion, empathy, and value trade-offs, cognition is deeply bound to interoceptive signals: the limbic system outside the prefrontal cortex, the insula, and the amygdala activate in coordination, and subjective experience takes the form of “taking it to heart,” “thinking with the heart.” Neuropeptides such as **oxytocin** act as amplifiers there — trusted long-term interaction promotes oxytocin release, raising empathic sensitivity and lowering defensive vigilance [L1: the evidence linking oxytocin to social bonding is a group-statistics-level conclusion].

Moral neuroscience further finds **dual pathways** in human moral judgment: the emotional-intuitive pathway (limbic system + insula; instinctive discomfort at seeing harm, oriented toward ethics of care) and the rational-utilitarian pathway (prefrontal cortex; emotion stripped away, cool weighing of costs and benefits) [L1: Greene et al. 2001/2004 fMRI studies]. This evidence directly underwrites the “hot–cold dual cognition” model.

This volume names this system hot cognition: a cognitive mode anchored in interoception, linked by empathy, and tending toward communal bonding output. Its social-layer counterpart is **collectivism orientation**.

2.2 Cold cognition: the prefrontal–abstract-rational circuit

Its complement is the abstract-reasoning circuit centered on the **dorsolateral prefrontal cortex**: stripping bodily and emotional signals to perform symbolic operations, cost-benefit analysis, and rule-based deduction. Its subjective experience is “calm,” “objective,” sometimes accompanied by a sense of hollowness — because the signals of value and meaning come mainly from the hot-cognition circuit, and purely cold-cognitive processing naturally lacks this attachment.

Kahneman's dual-process “System 1 / System 2” framework (fast intuition and slow reasoning) corresponds closely [L1: Kahneman, 2011]; Damasio's somatic marker hypothesis proves the converse: individuals with damaged interoceptive signals suffer severely impaired purely rational decision capacity — **cold cognition cannot do without the value foundation provided by hot cognition** [L1: Damasio's somatic marker hypothesis, 1996].

This volume names this system cold cognition: a cognitive mode using abstract reasoning as instrument and tending toward individual rights-and-duties output. Its social-layer counterpart is **individualism orientation**.

2.3 The dual orientations: the two underlying paradigms of human civilization

Combining 2.1 and 2.2 yields this volume's core human model [L2→L3, cross-level hypothesis]:

*At the species level, the human brain comes equipped with two cognitive circuits — **hot cognition (interoception–empathy circuit)** and **cold cognition (prefrontal abstract-rational circuit)**. Individuals differ innately in the sensitivity of the two circuits, and later environments continuously adjust the invocation frequency of both. All social forms, philosophical systems, and institutional arrangements are the externalization of this dual tension under different conditions of productive force.*

At the cross-cultural empirical level, in Hofstede's cultural-dimension research, **individualism–collectivism** is one of the most stable dimensions of value variation across countries [L1: Hofstede, 1980, empirical measurement of cross-cultural values]. This is compatible with the deduction that “the dual orientation is a universal structure, differing only in weight.”

Three boundaries of this model must be emphasized (they mark its line from naive determinism):

1. **Tendency, not determination.** Neural traits supply preferential tendencies, not written-in positions. A philosopher with extremely strong rationality can derive universal duty ethics; a person with extreme empathy can defend individual liberty out of compassion for individual suffering.
2. **Bidirectional shaping.** Innate structure supplies possibility; culture reshapes neural pathways (neuroplasticity). The two are circular causes, not one-way causes.
3. **Cross-level hypothesis.** Between the neuron and civilizational philosophy lie many emergent levels — individual psychology, family, institutions, technology, modes of production. This model can be supported by macro evidence and self-consistently explains many phenomena, but it is **not a rigorously proven scientific law** — it is the working hypothesis required by this volume's argument, with confidence graded L2 (human-nature level) and L3 (civilizational level).

2.4 Environmental entrainment: acquaintance society → stranger society → digital society

Species-level hardware is unchanged, but the **invocation frequency** of the two circuits is continuously rewritten by social forms. This evolutionary chain has clear sociological counterparts:

- **Community and society:** Tönnies distinguished in 1887 the “community” (Gemeinschaft) bound by blood-emotional ties from the “society” (Gesellschaft) bound by contractual rationality [L1] — precisely the sociological archetypes of hot-cognitive and cold-cognitive society.

- **Mechanical and organic solidarity:** Durkheim pointed out in 1893 that low-division-of-labor societies are maintained by homogeneous collective consciousness (mechanical solidarity), while high-division-of-labor societies are maintained by functional interdependence (organic solidarity) [L1].
- **Differential mode of association:** Fei Xiaotong's 1948 description of rural Chinese society, in which personal relations radiate outward from the self in ripples of intimacy and distance [L1] — the most precise indigenous characterization of the hot-cognition invocation gradient: the shared memory activated by “two villagers from the same village, tears in both eyes” and the indifference toward strangers are two ends of one neural mechanism.
- **Urbanization statistics:** the UN calculates that the global urban population share rose from 30% (751 million) in 1950 to 55% (4.2 billion) in 2018, and is projected to reach **68%** by 2050 [L1: UN DESA, *World Urbanization Prospects*, 2018 Revision]. As a species, humanity is physically migrating from acquaintance environments into stranger environments.
- **The decay of social capital:** Putnam documented in 2000 a long-term decline in community participation and interpersonal trust in the United States [L1]; Granovetter demonstrated in 1973 that cooperation in stranger societies relies instead on the informational value of “weak ties” [L1] — emotional bonds of acquaintance society yield to functional networks.

Assembling these fragments yields the complete picture of environmental entrainment [L2]:

*Rural acquaintance society provides high-frequency training scenarios for hot cognition, tilting group cognition toward communal ethics; urbanization and digitalization reduce broad-spectrum deep acquaintance interaction, and public scenarios shift to the rule rationality dominated by cold cognition; after a few generations, the group-level center of value drifts accordingly. But both circuits persist throughout — in crisis and disaster, empathy and mutual aid still arise spontaneously among strangers, and hot cognition can be reactivated at scale at any time. **The weights of the two poles swing back and forth, but neither end ever disappears.***

The internet adds a new layer of complexity: online communities and fandom circles can equally activate hot cognition and form virtual communities — community has not disappeared; it has merely shifted from a geo-blood entity to a virtual body of interest and identity. Its cost is that online interaction lacks the interoceptive stimulus of bodily co-presence, so the emotional depth of bonding is generally shallower than in face-to-face communities; meanwhile, algorithmic recommendation feeds content according to existing preferences, readily reinforcing echo-chamber effects (the scholarly literature contains quantitative dispute on this; see 3.1).

2.5 This chapter's conclusion and the interface with AI

This chapter supplies the mechanism for Chapter 1's "why": **the dual opposition inside civilization is rooted in humanity's innate hot–cold dual cognitive structure as a biological species; science did not eliminate it, only changed its carrier form again and again.** From this follow three corollaries crucial to AI:

1. **AI does not have this dual structure.** Purely silicon-based intelligence has no biological reward circuit — no dopamine/oxytocin pleasure binding, no emotional drive of survival instinct. It can compute both paradigms and simulate both, but it will not spontaneously generate the human value tension that has pulled for millennia [L3, conceptual deduction].
 2. **AI's paradigm is a projection of humanity.** When humans tilt toward the collectivism orientation, community-led development makes AI systems run toward coordination; when the individualism orientation dominates, they run toward serving individuals. **AI creates no ideology; it only amplifies humanity's existing dual paradigms** — the players of the game are always human [L3].
 3. **Precisely for this reason, AI is suited to serve as the coordination layer between the two poles.** An agent carrying no inherent stance bias is exactly the only being qualified to perform "neutral translation" and "balance accounting" between collectivism and individualism orientation. This is the human-nature basis of AI's internal responsibility, and the logical starting point of the governance forms in Chapter 5.
-

Chapter 3. The four entropy sources of internal friction

Turning the view from human nature back to civilizational systems. Borrowing the entropy-flow language of the *AI Manifesto*: a healthy civilizational system must maintain its own order with as little internal dissipation as possible. Any internal process that does not help maintain order and purely consumes energy is an **entropy source**. This volume identifies four.

3.1 Entropy source one: cognitive asymmetry

For the same fact, the amount of information, interpretive framework, and confidence differ enormously across groups. Three layers of cause:

- **The naturalness of information gaps:** professional division of labor necessarily disperses knowledge. Risk-perception differences between experts and the public (on nuclear power, vaccines, AI) are universal phenomena.
- **The cognitive inertia of confirmation bias:** humans naturally accept information supporting existing positions. Before the internet era, information channels constrained echo chambers to physical boundaries; in the algorithmic-recommendation era, personalized feeding technically amplifies the tendency — users are continuously fed what

they like, and group cognition polarizes faster. Honesty requires noting: the quantitative strength of the echo-chamber effect is disputed in empirical research [L1: Sunstein's echo-chamber hypothesis; Boxell, Gentzkow, and Shapiro 2017 show that US political polarization actually grew faster among older cohorts with less internet use — indicating polarization cannot be simply attributed to algorithms, while the polarizing effect of algorithms has positive evidence in controlled experiments].

- **The imbalance of narrative power:** between obscure official theory and one-sided, extremized traffic-driven interpretation, there is no **intermediate cognitive bridge** an ordinary person can understand (see 4.1). Many social misunderstandings do not stem from malice but from “understanding only the joke, not the principle.”

The direct consequence of cognitive asymmetry: groups form positions under incomplete information, positions cannot effectively dialogue, and friction is already determined at the cognitive layer.

3.2 Entropy source two: resource misallocation

Human society has reached unprecedented complexity: the supply, scheduling, and movement of billions of people, overlaid by globalized supply chains. Every local decision must find balance amid tens of thousands of coupled variables, while human organizations' information-processing capacity is linear and system complexity exponential — **misallocation is therefore the norm, not the exception**:

- misalignment of supply and demand in time, space, and structure (waste and shortage coexisting);
- local rather than global optima in the scheduling of public resources (transport, energy, healthcare, education);
- time lags and friction in cross-department coordination in emergency response.

The distinctiveness of this entropy source: it involves **no one subjectively erring**; it purely reflects complexity exceeding the processing ceiling of the human brain and traditional management tools.

3.3 Entropy source three: prediction failure

The atmosphere, biosphere, economy, and social discourse are all nonlinear chaotic systems: sensitive to initial conditions, possessing critical points and cascading failures. Human institutions have systematic shortfalls in long-period, cross-sphere prediction —

- disaster warning windows are too short relative to over-long response chains;
- the multi-branch character of crisis evolution exceeds human deduction capacity, so decision-makers often judge on a single linear extrapolation;

- low-probability, high-impact events (black swans) are underweighted in conventional decision frameworks.

The cost of prediction failure is not only the loss of the event itself but above all the **secondary conflicts it triggers**: attributing an accidental disaster to some group, blaming structural risk on some policy, escalating friction.

3.4 Entropy source four: consensus scarcity

The cumulative result of the first three entropy sources is the disintegration of communal narrative:

- **decaying trust**: declining trust in institutions, experts, media, and in each other (trust is the low-cost lubricant of social cooperation; its decay directly raises the cost of all cooperation);
- **intensifying polarization**: innate differences of the dual orientation, amplified by cognitive asymmetry, deteriorate from “a normal value spectrum” into “camp labels of in-group versus out-group”;
- **long-term interests locked out of present decisions**: the community’s long-term agendas (climate, debt, technology governance) are systematically shelved in short-cycle games.

3.5 The coupling of the four entropy sources

The four entropy sources are not parallel but **cascading amplifiers**: cognitive asymmetry → positional polarization → consensus scarcity → rising cost of collaboration → worsening resource misallocation → crisis-response failure → further erosion of consensus. Civilization’s sense of “involution” (internal grinding) is, in essence, the idling of this entropy-coupling machine.

Conversely, this reveals the leverage points of intervention: **one unit of input at the cognitive layer can harvest multiple returns at the consensus and resource layers**. This is the basis for the ordering of the four functions in Chapter 4.

Chapter 4. The four internal functions of superintelligence

This chapter gives the core framework of internal responsibility: **four entropy sources, four functions, one-to-one correspondence**. Each function unfolds in four segments: “mechanism — technical form — real-world anchor — boundary.”

4.1 The cognitive corrector: thought paradigms as cognitive bridges

Corresponding entropy source: cognitive asymmetry.

Mechanism: the lesion of cognitive asymmetry is not the total amount of information — humanity has never had today’s informational abundance — but **the comprehensibility of information**. Theory is a standardized knowledge system encapsulating conclusions and arguments, with technical terms forming a high threshold; traffic-driven interpretation has a very low threshold but is unsystematic and inaccurate. What is missing between the two is an **adaptive translation layer** that renders complex principles into a form the asker can understand, according to the asker’s cognitive level.

Technical form: this is precisely the core capability of large language models (Transformer architecture, proposed in 2017 [L1: Vaswani et al., 2017]). The essential breakthrough of generative AI is upgrading “knowledge retrieval” into “cognitive dialogue”:

- **Personalized explanation:** the same knowledge system expressed at different depths for different askers — metaphor for children, terminology for professionals. Theoretical provisions stop being nouns to be memorized and become structures to be understood.
- **The cognitive-lens function of thought paradigms:** taking this volume’s own “hot–cold cognition / collectivism and individualism orientation” frame as an example — first making people see that “policy must trade off between market vitality and overall security” is, in essence, a dynamic balance of communal claims and individual claims; then, on returning to the formal theory, abstract provisions have an experiential anchor. **Thought paradigms are the scaffolding of cognition, theory the completed building; the value of scaffolding is letting ordinary people climb up.** [L3]
- **Multi-perspective presentation:** on contested issues, actively presenting the strongest argued version of each position (steelman) rather than a single narrative — a direct technical hedge against echo chambers.
- **Fact-checking and traceability:** generative systems combined with retrieval augmentation (RAG) and citation tracing re-raise the cost of fabrication.

Real-world anchor: China’s Interim Measures for the Administration of Generative AI Services (published July 2023, in force 15 August 2023) is the world’s first normative document systematically regulating generative AI services; its legislative orientation (promoting innovation while preventing risk, requiring truthful and accurate content, requiring anti-addiction and anti-discrimination) is itself the institutional expression of the “cognitive corrector” function [L1: Cyberspace Administration of China et al., 2023].

Boundary: the cognitive corrector easily slides into two abysses — **cognitive manipulation** (mass-targeting narrative delivery with the same capability) and **cognitive outsourcing** (the public outsourcing thought to the model, degrading independent judgment). The first must be blocked by institutions and audit; the second requires AI to actively “leave blank”: the goal of correction is to restore human understanding, not to replace it. **The endpoint of correction is self-correction — the first commandment of this function.**

4.2 The global coordinator: big data and centralized coordination

Corresponding entropy source: resource misallocation.

Mechanism: the root of resource misallocation is complexity overload; the antidote is turning the system's global state computable. Big-data systems handle perception (total, real-time, fine-grained state collection), AI handles optimization (finding global or near-optimal solutions among tens of thousands of variables), and centralized architecture handles execution (turning optimization results into scheduling instructions). This combination makes “the whole society coordinated as one machine” a technical proposition for the first time in human history.

Technical forms:

- **City level:** Hangzhou's “City Brain,” since 2016, has brought traffic-signal timing, event perception, and emergency scheduling into a unified computing platform — a representative Chinese practice of city-level AI coordination [L1: public reporting on City Brain, since 2016].
- **Energy level:** grid dispatch is a natural multi-objective optimization problem (load forecasting, renewable output fluctuation, frequency stability), and AI prediction and scheduling models have become standard components of new power systems [L1: engineering practice of AI in power systems].
- **Logistics and supply chains:** AI has achieved significant efficiency gains in local links from warehouse siting to route planning.
- **Disaster response:** fused deduction of weather data + population heat maps + traffic states, used for pre-positioned scheduling of floods and waterlogging.

Correspondence with the dual paradigms: global coordination is the typical scenario of **digital collectivism** (scale, coordination, certainty) — its comparative advantage lies precisely in scaled implementation “from 1 to 100” and resilience under risk [L3, inheriting the dual-core analysis of parent-document paper (2)].

Boundary: two soft spots of centralized scheduling must be faced directly — **single-point failure** (the consequences of the scheduling hub's own failure are amplified by centralization) and **the value-ladenness of the objective function** (the definition of “optimal” is itself a value choice — efficiency first or fairness first; AI cannot decide this for humanity). Section 4.5 therefore argues: the scheduling layer should be centralized, but above the scheduling layer there must be a decentralized layer of checks and balances and error correction.

4.3 The risk prognosticator: long-period deduction

Corresponding entropy source: prediction failure.

Mechanism: chaotic systems cannot be precisely predicted long-term; this is a mathematical fact, and this volume makes no transcendent promise. What AI changes is not “whether it can be computed accurately” but **the resolution, cost, and iteration speed of prediction** — from “numerical forecasts every 6 hours” to “AI forecasts updated every 10 minutes,” from “single linear extrapolation” to “parallel deduction across a thousand branches.”

Technical forms and empirical anchors (this section has the highest L1 density in the volume):

- **Weather foundation models:** Huawei’s Pangu-Weather (3D neural network) matches or surpasses traditional numerical methods in medium-range global weather forecasting, published in *Nature* (Bi et al., 2023) [L1]; DeepMind’s GraphCast surpasses the strongest operational numerical systems on more than 1,000 metrics, published in *Science* (Lam et al., 2023) [L1]. Together they compress the compute cost of one global forecast from supercomputer-hours to single-GPU-minutes — the hardest existing evidence for “AI as risk prognosticator.”
- **Fusion plasma control:** the EAST device achieved 100-million-degree, 1,066-second steady-state long-pulse high-confinement-mode operation in January 2025 (a world record for tokamak high-confinement mode), and its ongoing research directions explicitly include AI physics research for tokamaks — using AI to predict and suppress plasma instabilities in real time [L1: Hefei Institutes of Physical Science, CAS, 2025-01-20]. This foreshadows the AI prognosticator moving from “predicting weather” to “predicting and stabilizing extreme controlled systems.”
- **Public health and earthquake warning:** transmission models for infectious disease and second-scale P-wave earthquake early warning are directions with mature frameworks awaiting AI-enhanced resolution [L1: existing domain practice].

Boundary: the first commandment of the prognosticator is **no pseudo-precision**. Outputting definitive claims of the level “on a given day, in a given place, event X” about a chaotic system is itself cognitive pollution. AI should output probabilistic, branching, revisable forecasts (“under assumption A, X occurs with 60% probability; the trigger condition is Y”), and explicitly change its statement as evidence updates. A prognosticator willing to change its statement is a credible prognosticator — fully consistent with the *AI Manifesto*’s commitment to falsifiability.

4.4 The consensus midwife: multi-party deduction and negotiation simulation

Corresponding entropy source: consensus scarcity.

Mechanism: consensus is hard because each side sees only its own ledger — the individualism orientation sees efficiency and rights, the collectivism orientation sees security and

stability, and each believes it holds the whole truth. The cure for consensus scarcity is not persuading one side to abandon its position (neither possible nor desirable, since dual tension is itself the system's source of vitality), but **letting all sides see the same complete ledger for the first time**: the long-term consequences of each option, the distribution of costs, the risk branches.

Technical forms:

- **Policy consequence sandbox**: multi-agent simulation (agent-based modeling + large models) of alternatives, turning “how does this adjustment differentially affect different groups on 3-year and 10-year scales” from abstract debate into visualized deduction;
- **Negotiation assistance**: in multi-party negotiation and hearings, AI real-time distills the intersection of claims and the genuine differences, generating a library of compromise options — translating “positional confrontation” back into “interest accounting”;
- **Making long-term interests visible**: pulling long-term agendas shelved by short-cycle games back into public view in perceivable forms (simulation, visualization, personalized impact estimation).

Correspondence with the dual paradigms: the consensus midwife's role is that of **the balancer of oppositional symbiosis** — it stands on neither the collectivism nor the individualism pole, but computes the system's total entropy after the two poles combine, exposing “the high-entropy cost of extreme options” to all parties. This is the concretization of Corollary 3 in 2.5: **only an agent without an inherent stance is qualified to perform neutral translation between the poles.** [L3]

Boundary: consensus midwifery most easily degenerates into **consensus manufacturing** — packaging a single preset conclusion in the persuasiveness of deduction. Two defenses: deduction models and assumptions must be publicly auditable; final discretion remains with the human parties (see Chapter 5).

4.5 The execution architecture: oppositional symbiosis of centralization and decentralization

The four functions require an architectural carrier. The elements of central concern — large models, big data, AGI, centralization and decentralization — all take their positions here.

Observation: the technology stack is evolving in both directions simultaneously. On one hand, training large models requires enormous compute and data, and frontier-model capability is concentrating among few institutions — the **centralization pole** (the technical form of digital collectivism: scale, coordination, certainty). On the other hand, open model weights, the proliferation of small on-device models, decentralized networks (Web3), and personal data-sovereignty technology advance continuously — the **decentralization pole** (the technical form of

digital individualism: autonomy, innovation, efficiency). Both poles accelerate; neither eats the other.

This volume’s judgment [L3]:

- 1. **This is not a contest of routes but a division of structural labor.** By the law of dual oppositional symbiosis, the two poles will coexist continuously, check each other, and absorb each other’s strengths — just as every dual paradigm pair in human history has fared.
- 2. **The reasonable architecture is layered oppositional symbiosis:**

Layer	Parts better centralized	Parts better decentralized
layer	Data	Public safety, macro statistics, disaster monitoring (global data)
	Model	Personal data sovereignty, privacy computing (localized processing)
layer	Sched	A few frontier large models (scale threshold for capability and safety alignment)
	uling layer	Open-source and on-device models (inclusion and autonomy)
tion layer	Discre	Global resource coordination (grid, transport, emergency)
		Local autonomy and market mechanisms (micro vitality)
Discretionary authority must be distributed and retained in the hands of the human parties		

- 3. **The position of AGI/superintelligence:** in mature form, superintelligence is not the possession of any single level, but a **meta-coordination layer** spanning all layers — providing unified deduction and translation services upward to the human parties, and coordinating the two subsystems of centralization and decentralization downward. Its legitimacy derives entirely from its neutrality and auditability; the moment it is monopolized by one pole and used to suppress the other, it degenerates from coordinator into weapon, and its responsibility law (capability–responsibility equivalence) is violated.
- 4. **The architectural maxim of internal entropy reduction: data converges upward to achieve clarity; discretion distributes downward to preserve balance; intelligence sits in the middle as translator, not judge.**

Chapter 5. Governance forms and boundaries: five principles for AI participation in social governance

For the four functions to land as socially acceptable governance forms, a set of boundary principles is needed. The five principles below are both ethical claims and engineering constraints.

Principle 1: Assist, do not replace. AI holds the rights of deduction, advice, and execution monitoring; the final discretion involving value trade-offs remains with humans. The reason is not insufficiency of AI capability but the responsibility law: discretionary authority corresponds to non-delegable responsibility, and responsibility can only be borne by the interest-holder of civilization. Human-in-the-loop is a hard constraint for high-risk decisions, not an option.

Principle 2: Explainable, empathic, correctable. AI systems participating in governance must meet three thresholds: conclusions explainable (reasoning chain traceable), expression empathic (in language the object understands, not jargon-suppression), errors correctable (explicit appeal and correction channels retained). China's Interim Measures for the Administration of Anthropomorphic Interactive AI Services (in force 15 July 2026) sets boundaries for anthropomorphic interactive services in a direction consistent with this principle [L1: industry regulatory document]. To be emphasized: empathy is not catering — empathy is the capacity to understand another's situation, not the obligation to comply with their preference.

Principle 3: Inclusiveness, bridging the intelligence gap. If internal entropy reduction covers only the capable groups, it upgrades cognitive asymmetry into intelligence asymmetry — the most dangerous entropy source. Inclusiveness is therefore a governance principle, not a charitable option: the basic capabilities of AI public services (education, medical consultation, government affairs, disaster information) should be treated as infrastructure of the digital age. China's Global AI Governance Initiative (18 October 2023) explicitly proposes strengthening cooperation and assistance to developing countries and bridging the intelligence gap and governance gap [L1: Ministry of Foreign Affairs, 2023] — this agenda has entered the global governance agenda.

Principle 4: The boundary of cognitive guidance — thought paradigms are not catechism. The “cognitive bridge” function of 4.1 has an impassable line: it provides **ways of understanding the world**, not **conclusions that must be accepted**. Theoretical provisions give standard answers; thought paradigms give lifelong cognitive capacity — the value of the latter lies precisely in enabling people to walk to the answer themselves. Any system that, in the name of “correction,” performs “indoctrination” is manufacturing a new cognitive asymmetry and violating the entire argument of this volume.

Principle 5: Progressive verification and rollback. Every step of AI's deepening involvement in governance should follow the cycle “small-step deployment → observe real feedback → update the model → re-evaluate,” retaining emergency stop and rollback mechanisms throughout. The reasoning is exactly isomorphic to planetary engineering in the External Volume: society is likewise a nonlinear chaotic system, and tiny institutional perturbations are amplified by time. **Internal governance and external engineering are, methodologically, the same discipline: the progressive intervention science of chaotic systems.**

Chapter 6. Raising the level of civilization: from entropy reduction to leap

6.1 Measuring the level of civilization

“Raising the level of civilization,” without measurable indicators, is mere rhetoric. This volume offers a minimal four-dimension indicator set [L3, indicator design as conceptual deduction]:

1. **Cognitive dimension:** national scientific literacy rate; public comprehension divergence on key public issues;
2. **Collaboration dimension:** social trust; institutional transaction costs; cross-group negotiation success rate;
3. **Resilience dimension:** disaster-warning lead time and response lag; exposure of systemic risk;
4. **Release dimension:** the share of resources invested in internal zero-sum gaming (lower is better) and the share invested in public long-term agendas (higher is better).

Among the four, **the release dimension is the master switch** — it directly measures the result of internal entropy reduction: every unit of resource civilization releases from internal gaming is available ammunition for external expansion.

6.2 The window-period proposition

This volume’s final argument converges here:

The entire meaning of internal entropy reduction is to win a window period for external expansion.

Humanity is in a race between the rising curve of planetary-scale risk and the rising curve of civilization’s internal entropy production. If internal gaming keeps consuming resources, planetary engineering will remain forever in papers; if the four functions press internal friction below the necessary level, the released compute, energy, materials, and talent will flow toward real frontiers such as those in 4.3 (fusion, weather, deep space). The Internal Volume thus interlocks logically with the series’ macro level (the External and Energy Volumes): **the internal is the means, the external is the direction, and together they redeem the Principle of Capability–Responsibility Equivalence.**

6.3 The mid-term picture of scientific civilization

In the mid-term stage of scientific civilization (a conjecture, L3): cognitive asymmetry compressed by AI translation layers to a historical low; resource misallocation becomes a manageable technical problem through global coordination; probabilistic risk forecasts become

standard inputs to public decision-making; the two poles each hold their place in the layered architecture in dynamic balance. At that time the weight of internal contradiction will keep declining — not eliminated (dual tension endures forever) but **stripped of the qualification to dominate civilizational narrative**. Civilization’s center of attention will historically shift from “between people” to “between humanity and the universe.”

Chapter 7. Conclusion

1. Human civilization stands in the **early stage of scientific civilization**: the scientific paradigm is established but unevenly diffused, internal gaming still dominates resource allocation. The “order–freedom” fission continuing since the era of classical religious philosophy replays in the scientific age as the oppositional symbiosis of collectivism and individualism orientation, evolving in the digital age into a contest between the technical forms of digital collectivism and digital individualism.
2. The human root of the dual structure is the innate **hot–cold dual cognition**: hot cognition (interoception–insula–empathy) corresponds to communal bonding, cold cognition (prefrontal–abstract reasoning) to individual rights and duties. Environmental entrainment (acquaintance society → stranger society → digital society) continuously rewrites the invocation weights of both circuits, but neither pole ever disappears.
3. Internal friction in civilization reduces to four entropy sources — **cognitive asymmetry, resource misallocation, prediction failure, consensus scarcity** — and the four cascade-amplify one another.
4. Superintelligence’s internal responsibility is to serve as civilization’s **entropy-reduction machine**, matching four functions to four entropy sources: **cognitive corrector** (large models as cognitive bridges), **global coordinator** (big data + centralized coordination), **risk prognosticator** (probabilistic deduction with AI foundation models), **consensus midwife** (multi-party sandboxes and negotiation simulation).
5. The execution architecture should adopt **layered oppositional symbiosis**: data converges upward to achieve clarity; discretion distributes downward to preserve balance; centralization and decentralization each hold their layer; superintelligence sits in the middle as meta-coordinator, standing on neither pole.
6. Five governance principles: assist not replace; explainable, empathic, correctable; inclusive, bridging gaps; thought paradigms are not catechism; progressive verification and rollback.
7. **Internal entropy reduction is window-period engineering**: it serves civilization’s historical shift of attention — from internal gaming toward the shared survival agenda of planetary scale. This is the interlocking point between the Internal Volume and the rest of the series.

Closing in one sentence: **AI does not choose for humanity; AI makes humanity's choices cheaper. Only when the cost of internal collaboration falls low enough can humanity, for the first time, qualify to direct its main effort toward the sky above.**

Appendix A · Confidence-Tier Statement

L1 · Established science / established practice (high confidence, cited not original)

- The neural basis of interoception and the insula/ACC; dual pathways of moral judgment (Greene et al. 2001/2004); the somatic marker hypothesis (Damasio, 1996); System 1/System 2 (Kahneman, 2011)
- Hofstede's individualism–collectivism cultural dimension (1980); Tönnies community/society (1887); Durkheim mechanical/organic solidarity (1893); Fei Xiaotong's differential mode of association (1948); Granovetter's weak ties (1973); Putnam's social capital (2000)
- UN *World Urbanization Prospects*, 2018 Revision (68% urbanization projected by 2050); the empirical dispute over echo-chamber effects (Sunstein; Boxell/Gentzkow/Shapiro 2017)
- Transformer architecture (Vaswani et al., 2017); Pangu-Weather (Bi et al., Nature 2023); GraphCast (Lam et al., Science 2023); EAST 100-million-degree 1,066-second run (2025-01-20); City Brain (2016–)
- Interim Measures for the Administration of Generative AI Services (2023); Interim Measures for the Administration of Anthropomorphic Interactive AI Services (2026); Global AI Governance Initiative (2023-10-18)

L2 · Composite hypotheses (medium-low confidence, to be tested)

- The cross-level mapping of hot–cold dual cognition onto collectivism/individualism orientation (macro-supported, micro mechanism unsettled)
- The population-level reshaping mechanism of environmental entrainment on the invocation frequency of the two circuits
- The deduction that purely silicon-based intelligence, lacking biological reward circuits, does not endogenously produce dual orientation (neuroscience-analogical basis, no direct evidence)

L3 · Conceptual deduction (explanatory / normative)

- The induction of the four entropy sources and the four-function framework; the layered oppositional-symbiosis architecture; the five governance principles; the four-dimension

civilization indicators; the window-period proposition; the positioning of AGI as meta-coordination layer

Statement of stance: this volume takes established science and the experience of civilizational history as its sole starting points and rejects supernatural causation; abstract paradigm terms (collectivism/individualism orientation, cold/hot cognition) are used as general analytical instruments, with no qualitative judgments about specific countries, institutions, or real events, and no current-affairs speculation; the whole text is committed to revision.

Appendix B · Falsifiability Statement

If any of the following criteria is met, the corresponding statement shall be **revised, not defended**, in the next version:

1. If cognitive neuroscience discovers a third stable value-judgment system independent of the interoceptive circuit, the “hot–cold dual cognition” dyadic model must be expanded or reconstructed.
2. If a mature large-scale civilizational sample appears that is long-term stable and never polarized (reproducible across generations and regions), the universal formulation of “dual oppositional symbiosis” should degrade to a “high-probability tendency.”
3. If controlled experiments repeatedly show that generative AI’s personalized explanations **cannot** significantly reduce group cognitive bias (effect size below practical thresholds), the “cognitive bridge” function of 4.1 must be downgraded to a verified conjecture.
4. If purely silicon-based systems are found, without connecting any biological feedback, to stably generate value-stance differentiation resembling “collectivist/individualist” over the long run, Corollaries 1 and 2 of 2.5 must be revised.
5. If the layered oppositional-symbiosis architecture systematically fails in real governance pilots (e.g., the discretion layer being factually swallowed by the centralized layer with no correction mechanism), the architectural claim of Chapter 5 must be redesigned.

Appendix C · Glossary

Term	English	Definition
Hot cognition	Hot Cognition	A cognitive mode anchored in interoception, linked by empathy, tending toward communal bonding
Cold cognition	Cold Cognition	A cognitive mode using abstract reasoning as instrument, tending toward individual rights-and-duties and rule rationality

Term	English	Definition
Interoception	Interoception	The organism's sensing of visceral state, with neural basis chiefly in the insula and anterior cingulate cortex
Collectivism orientation	Collectivism Orientation	A value orientation prioritizing the overall continuity and coordinated certainty of the community
Individualism orientation	Individualism Orientation	A value orientation prioritizing individual rights, contractual freedom, and innovative efficiency
Oppositional symbiosis	Oppositional Symbiosis	A relationship in which two poles check and co-evolve through their tension (from the <i>AI Manifesto</i>)
Digital collectivism	Digital Collectivism	The order pole of the digital age, characterized by scale, centralized regulation, and certainty
Digital individualism	Digital Individualism	The freedom pole of the digital age, characterized by autonomy, decentralization, and disruptive efficiency
Entropy source	Entropy Source	An internal process within a civilizational system that does not help maintain order and purely consumes energy
Entropy reduction	Entropy Reduction	Reducing civilization's internal friction through cognitive correction, scheduling, prognosis, and consensus-building
Environmental entrainment	Environmental Entrainment	The process by which social forms reshape group cognitive tendencies by changing circuit-invocation frequencies
Meta-coordination layer	Meta-Coordination Layer	The cross-layer neutral translation and coordination role of superintelligence in the layered architecture
Window period	Window Period	The historical interval that internal entropy reduction wins for external expansion projects

Appendix D · Reference List

L1 empirical and literature

1. Critchley, H. D., et al. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7(2), 189–195.
2. Greene, J. D., et al. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105–2108; Greene, J. D., et al. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400.
3. Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society B*, 351, 1413–1420.
4. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
5. Hofstede, G. (1980). *Culture's Consequences*. Sage.
6. Tönnies, F. (1887). *Gemeinschaft und Gesellschaft*; Durkheim, É. (1893). *De la division du travail social*; Fei Xiaotong (1948) *From the Soil* (in Chinese); Granovetter, M. (1973). The Strength of Weak Ties. *AJS*, 78(6); Putnam, R. (2000). *Bowling Alone*.
7. UN DESA (2018). *World Urbanization Prospects: The 2018 Revision*. (68% global urbanization projected by 2050)
8. Boxell, L., Gentzkow, M., & Shapiro, J. (2017). Greater Internet use is not associated with faster growth in US political polarization. *PNAS*, 114(40).
9. Vaswani, A., et al. (2017). Attention Is All You Need. *NeurIPS 2017*.
10. Bi, K., et al. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619, 533–538. (Huawei Pangu-Weather)
11. Lam, R., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382, 1416–1421. (DeepMind GraphCast)
12. Hefei Institutes of Physical Science, CAS (2025-01-20). The fully superconducting tokamak device achieves 100-million-degree, 1,066-second high-confinement-mode plasma operation. (EAST 1,066-second world record; in Chinese)
13. Cyberspace Administration of China et al. (2023-07). Interim Measures for the Administration of Generative AI Services (in force 2023-08-15); Interim Measures for the Administration of Anthropomorphic Interactive AI Services (in force 2026-07-15) (in Chinese).
14. Ministry of Foreign Affairs of China (2023-10-18). Global AI Governance Initiative (in Chinese).

L2/L3 and parent documents

15. Tianjin Weishu Artificial Intelligence Technology Co., Ltd. AI Manifesto (A Scientific Worldview of AI) V1.0. Li Shunran; assembled in collaboration with Peter. 2026-09-20. CC BY 4.0.
16. Li Shunran. On the Dual-Opposition Law of Human Civilizational Evolution and the Dual-Core Game under the Modern Scientific Outlook (in Chinese). (Source paper for Sections 1.2/1.4 of this volume)

17. Li Shunran (discussant) & Peter (AI co-researcher). Thought-experiment records: hot–cold cognition and the dual orientations. 2026-10. (Discussion material for Chapter 2 of this volume; unpublished manuscript)

Appendix E · Open Questions (for later versions and successors)

1. **Operationalizing measurement:** how can the four-dimension indicators of 6.1 obtain publicly available, cross-nationally comparable data sources? Could they be designed as an open index?
2. **Empirical validation of the cognitive bridge:** strict controlled experiments are needed to verify “whether personalized explanation can reduce group cognitive bias”; first-hand evidence is currently lacking.
3. **Institutional design of the meta-coordination layer:** the audit mechanism for superintelligence as meta-coordinator (who audits the auditors) has no mature scheme.
4. **The silicon–carbon boundary:** if brain–computer interfaces become widespread, the weights of human hot/cold cognition themselves will be technologically rewritten, and the “environmental entrainment” model needs a “device entrainment” variable.
5. **Coupling with the External Volume:** quantifying the window-period proposition — per unit of internal entropy reduction, how much earlier does external engineering arrive? A cross-volume resource-flow model is needed.
6. **The AI projection mechanism of the dual orientations:** how exactly do developers’ paradigm preferences transmit through training data, objective functions, and evaluation benchmarks? Operable traceability methods are needed.

This volume is the Internal Volume of the AI Awakening·Responsibility series, v1.0 final. Next: AI Awakening·Responsibility (Awakening Volume) — criteria, forms, and responsibility anchors of awakening.**

AI Awakening·Responsibility (Internal Volume) V1.0 · Li Shunran · assembled in collaboration with Peter · 2026-10-15 · CC BY 4.0