AI Autonomous Governance: Concept, Technical Foundation, and Conditions for Realization

【Wei Jili】

【ORCID: 0009-0006-5696-1873】

【Corresponding Author Email: jiligzs@qq.com】

Abstract

The cognitive and executive capabilities of artificial intelligence (AI) have matured considerably, yet “being able to execute” is not equivalent to “being able to govern.” On the basis of “artificial intelligence” and “AI autonomization,” this paper proposes the conceptual framework of “AI autonomous governance-enabling.” It argues that governance-enabling does not start from scratch; rather, it superimposes a “management” function on top of the knowledge foundation and the autonomized execution capability of mature AI, thereby forming a governance loop of “ex ante review—in-process monitoring and response—ex post retrospective and refinement.” The paper first argues that the essence of contemporary AI is a statistical inference engine built on the digitization of human knowledge; it constitutes the knowledge foundation for autonomous governance but does not in itself amount to governance capability. It then analyzes the shift in AI autonomization from reasoning to execution, clarifying the distinction between “autonomy in execution” and “governance autonomy.” It further elaborates that, as a leap from execution to management, autonomous governance-enabling is essentially autonomy in management rather than autonomy in execution. Finally, it points out that the legitimacy of autonomous governance-enabling is not automatically acquired; it must satisfy four conditions—transparency, auditability, intervenability, and traceability of responsibility—and be realized progressively through graduated authorization and human meta-governance. The paper concludes that the advancement of AI autonomous governance-enabling should uphold the distinction between “autonomy in execution” and “governance autonomy” and base governance-enabling on the coupling of technology, institutions, and norms. Its significance for AI research and practice lies in three points: it proposes a discriminating distinction between “autonomy in execution” and “governance autonomy,” delineating the governance boundaries for the design of autonomous systems; it characterizes the governance loop mechanism that can be mapped onto system modules; and it proposes a graduated authorization criterion based on “governance loop maturity,” providing operational guidance for the deployment and oversight of autonomous systems.

Keywords: artificial intelligence; autonomization; self-governance; governance loop; transparency; auditability; intervenability; traceability of responsibility; graduated authorization; meta-governance

Contents

1. Introduction 4

2. AI as the Knowledge Foundation for Autonomous Governance 7

2.1 Human Knowledge Digitization and Large-Scale Pretraining: The Prerequisite for AI Capability 7

2.2 The Training Process: Statistical Modeling of Human Knowledge and the Generation of Reasoning Capability 7

2.3 Features and Limitations of the Statistical Inference Engine 8

2.4 The Separation of Cognitive Capability and Self-Management Capability 8

3. AI Autonomization: From Reasoning to Execution 8

3.1 The Basic Structure of an Agent: Perception—Reasoning—Execution 9

3.2 The Typical Execution Workflow of a Large Model Agent 9

3.3 The Key Breakthrough: From “Generating Results” to “Executing Tasks” 9

3.4 The Distinction between “Autonomy in Execution” and “Governance Autonomy” 10

4. AI Autonomous Governance-Enabling: From Execution to Management 11

4.1 Governance-Enabling: Adding Management Functions Rather than Expanding Technical Capability 12

4.2 The Governance Loop: A Continuous Mechanism across Ex Ante, In-Process, and Ex Post Phases 12

4.3 The Difference between Governance-Enabling and Industry Self-Regulation 16

4.4 Case Analysis: An Evaluation of the Governance Loop of an Autonomous Coding Agent 16

5. Legitimacy Conditions and Implementation Path of Autonomous Governance-Enabling 19

5.1 The Risks Faced by Autonomous Governance-Enabling 19

5.2 Four Legitimacy Conditions 19

5.3 The Auxiliary Role of Industry Self-Regulation 24

5.4 The Graduated Authorization Path 24

5.5 The Non-Transferability of Human Meta-Governance 28

5.6 The Normative Boundaries of This Paper 30

6. Conclusion 31

1. Introduction

Machine learning methods, represented by deep learning, have endowed artificial intelligence with unprecedented cognitive and executive capabilities, and AI applications have moved from the laboratory into every stage of economic and social operations. Yet maturity in capability does not automatically bring about maturity in governance. Although ethical discussions surrounding AI have flourished, many governance principles still remain at the level of normative advocacy and have not yet been internalized as the operational mechanisms of systems themselves. The basic proposition of this paper is that AI already possesses relatively strong cognitive and executive capabilities, but "being able to execute" is not equivalent to "being able to govern"; governance-enabling requires a new conceptual framework and conditions for realization.

The gap between governance demand and reality is the immediate background of this proposition. Floridi and Cowls proposed a unified framework of five principles for AI in society, attempting to integrate ethical requirements such as beneficence, non-maleficence, justice, autonomy, and explicability into an operable system of principles, thereby providing an ethical-level normative foundation for AI governance [1]. However, a significant gap remains between the proposal of principle frameworks and the implementation of principles in actual systems: governance demand continues to grow, whereas the means of governance implementation still rely mainly on external ethical initiatives and ex post accountability.

With respect to governance approaches, the academic literature already contains fairly systematic reviews and reflections. Cath et al. compared three major AI governance approaches—those of the United States, the European Union, and the United Kingdom—and pointed out that different regions differ markedly in regulatory intensity, value orientation, and institutional design, while all face the structural challenge of moving from principles to practice [2]. Jobin et al. systematically reviewed AI ethics guidelines worldwide, revealing the widespread existence of industry self-regulation and ethical guidelines—a large number of government agencies, industry organizations, and academic communities have issued their own ethical statements, forming a guideline landscape that is vast in quantity yet highly fragmented [3].

Behind this flourishing lies the problem of soft constraints. Hagendorff's evaluation of existing ethical guidelines shows that the vast majority of guidelines lack enforcement power and operable implementation mechanisms, making it difficult for them to effectively constrain the actual behavior of AI systems [4]. In other words, external normative input cannot automatically translate into behavioral safeguards within the system. This implies that if AI is to truly achieve "responsible autonomy," relying solely on principled declarations and external oversight is

insufficient; it is necessary to explore governance forms in which the system itself can exercise self-constraint and self-management.

It is against this background that this paper proposes the theoretical proposition of "AI autonomous governance-enabling." The approach of this paper is as follows: starting from the technical foundation and following the progressive chain of "AI—autonomization—governance-enabling," it progressively constructs the conceptual framework of governance-enabling. The paper first clarifies the technical essence of AI as a knowledge foundation (Chapter 2), then analyzes how AI autonomization achieves the leap from reasoning to execution (Chapter 3), on this basis defines the connotation of autonomous governance-enabling—namely, superimposing management functions on top of executive capability (Chapter 4)—and finally argues for the legitimacy conditions and implementation path of autonomous governance-enabling (Chapter 5), concluding the whole paper (Chapter 6).

It should be noted that "AI autonomous governance-enabling" is not a concept invented out of thin air; rather, it is both connected to and clearly differentiated from existing governance theories and policy frameworks. In terms of normative sources, algorithmic governance and responsible AI engineering regard governance as external regulation of algorithmic decision-making processes; value alignment embeds human values into system optimization objectives, but it mainly targets the training stage; Constitutional AI constrains system behavior through rules, already exhibiting an embryonic form of internal management [5]; and "meaningful human control" philosophically justifies the ultimate human control over autonomous systems [6]. At the policy level, the EU Artificial Intelligence Act imposes horizontal regulation on AI systems through risk grading [7]; the OECD AI principles [8], the NIST AI Risk Management Framework [9], and the UNESCO Recommendation on the Ethics of Artificial Intelligence [10] provide external governance resources from the perspectives of principles, risk management, and ethical norms, respectively. The difference between this paper's framework and the aforementioned works lies in the following: it explores not how external norms are imposed upon the system, but how norms are internalized as the system's own operational mechanisms; and it targets not the regulation of system capability, but the construction of the system's "self-management capability." The relationship between the two can be summarized in Table 1.

Table 1 | The Relationship between This Paper's Framework and Existing Governance Frameworks

Existing framework	Normative source	Embedding level	Governance subject	Relationship to this paper
Algorithmic governance	External source	Regulation of decision processes	Regulator/Organization	A component of external constraints
Value	External	Training-stage	Developer	Value source

Existing framework	Normative source	Embedding level	Governance subject	Relationship to this paper
alignment	source	objective		of the governance loop
Constitutional AI	Rules internalized in the system	Constraint on system behavior	The system itself	Technical embryo of the internal management loop
Responsible AI engineering	External source	Development process	Developer/Organization	Normative input at the design stage
The EU Artificial Intelligence Act	External source	Deployment and oversight	Regulator	Provides horizontal risk classification
Meaningful human control	External source	System–human interface	Human	Theoretical support for the condition of intervenability
This paper's framework	Norms internalized in the system	The system's self-operational mechanism	System + human meta-governance	—

Based on the above positioning, the contributions of this paper can be summarized in three points. First, the conceptual contribution: it proposes a discriminating distinction between "autonomy in execution" and "governance autonomy," thereby avoiding the erroneous reduction of governance problems to technical problems. Second, the mechanism contribution: it characterizes the governance loop of "ex ante review—in-process monitoring and response—ex post retrospective and refinement" and distinguishes it from the engineering safety loop in terms of governance object, criteria of judgment, and locus of accountability. Third, the path contribution: it proposes a graduated authorization path with "governance loop maturity" as the core criterion, providing an operational framework for the deployment decisions and oversight interface of autonomous systems.

2. AI as the Knowledge Foundation for Autonomous Governance

To understand why AI can move toward autonomous governance, it is first necessary to clarify what AI itself is. The basic judgment of this paper is that the essence of contemporary AI is a statistical inference engine built on the digitization of human knowledge; it provides the cognitive basis and knowledge foundation for autonomous governance, but it is not in itself equivalent to governance capability.

2.1 Human Knowledge Digitization and Large-Scale Pretraining: The Prerequisite for AI Capability

The prerequisite for the capabilities of contemporary AI is the digitization of human knowledge and large-scale pretraining. In their classic framework, Russell and Norvig define AI as the study of agents—entities that perceive their environment, reason, and take actions to achieve goals—and this definition provides a basic reference for understanding the sources of AI capability [11]. However, what underpins the leap in contemporary AI capability lies not in the abstract framework of the agent itself, but in a fundamental transformation in the way knowledge is carried: the language, text, image, and structural knowledge accumulated by humanity over thousands of years has been digitized at scale, becoming corpora that can be learned by machines [11]. Deep learning, which learns representations and patterns automatically from massive data through multi-layer neural networks, is the technical foundation of contemporary AI capability [12]. On this basis, the Transformer architecture, centered on the attention mechanism, effectively models long-range dependencies and greatly enhances the capacity for language modeling and knowledge loading [13]. It is precisely through such architectures that large-scale pretraining compresses digitized knowledge into model parameters.

2.2 The Training Process: Statistical Modeling of Human Knowledge and the Generation of Reasoning Capability

In terms of mechanism, the essence of the training process is the statistical modeling of human knowledge and the generation of reasoning capability [14]; the model does not “memorize” knowledge but learns the statistical co-occurrence and association structures among words, and among concepts, from massive corpora, thereby developing a probability-based reasoning capability. The systematic study by Kaplan et al. shows that model performance follows predictable power-law relationships (scaling laws) with model size, dataset size, and compute, which means that as training resources increase, the model’s statistical fitting and generalization capabilities improve systematically [15]. On the one hand, this finding explains the scale foundation of the capability emergence of contemporary large language models; on the other hand, it reveals the statistical nature of such capability: what the model “knows” is, in essence, the statistical regularities present in the corpus, rather than a direct grasp of the world.

2.3 Features and Limitations of the Statistical Inference Engine

As a statistical inference engine, contemporary AI exhibits three salient features. The first is probabilistic nature: the model's output is a sample from a distribution of possibilities rather than a deterministic assertion of fact. The second is correlational nature: what the model acquires is the correlational structure of the corpus rather than causal structure, and it therefore tends to mistake correlation for causation. The third is data dependence: the boundary of model capability is determined by the scope, quality, and distribution of the training data, and unreliable performance may occur in situations beyond the data coverage. The famous critique of "stochastic parrots" by Bender et al. is directed precisely at this limitation: large language models may generate seemingly fluent text on the basis of statistical regularities while lacking a genuine understanding of linguistic meaning, thereby posing systematic risks to factual accuracy, consistency, and reliability [14].

2.4 The Separation of Cognitive Capability and Self-Management Capability

The technical essence described above determines the current status of AI: it possesses strong cognitive capability and can provide knowledge support and reasoning assistance for decision-making, but it lacks self-management capability in the governance sense. Cognitive capability addresses the question of "knowing"—identifying patterns, generating solutions, and predicting outcomes; governance capability addresses the normative question of "acting"—setting the boundaries of goals, constraining the scope of action, monitoring the execution process, and evaluating the consequences of action. The cognitive capability of a statistical inference engine derives from statistical modeling of existing knowledge; it has no endogenous deliberation about goals, value judgment, or sense of responsibility. In other words, as a knowledge foundation, AI is sufficient—it can provide the necessary cognitive basis for autonomous governance; but as a governance subject, it is insufficient—it lacks the mechanisms for self-constraint and self-management. This separation constitutes the starting point of the subsequent discussion: for AI to move toward autonomous governance, it is necessary to further superimpose executive capability and even management functions on top of the knowledge foundation.

---

3. AI Autonomization: From Reasoning to Execution

The evolution of AI capability has not stopped at "knowing." On top of the knowledge foundation, AI has begun to acquire the capability of "acting"—not only generating reasoning results, but also invoking tools and executing tasks. This chapter examines this transition from reasoning to execution, and its core

conclusion is that the essence of AI autonomization is “autonomy in execution,” and that autonomy in execution is not equivalent to governance autonomy.

3.1 The Basic Structure of an Agent: Perception—Reasoning—Execution

The starting point for understanding autonomization is the basic structure of an agent. The classic definition of agents by Wooldridge and Jennings states that an agent is a computational system situated in some environment that perceives the environment and acts upon it through actions, with autonomy—deciding its own behavior without direct human intervention—as its core feature [16]. This three-stage structure of “perception—reasoning—execution” lays the basic skeleton of autonomization: the system first perceives inputs and the state of the environment, then reasons and makes decisions on the basis of knowledge, and finally translates decisions into actual execution. Contemporary large language model agents are precisely the latest form of this structure in the deep learning era: with a pretrained language model as the reasoning core and external tools and application programming interfaces as the means of execution, they constitute a complete capability loop.

3.2 The Typical Execution Workflow of a Large Model Agent

From a technical implementation perspective, the typical execution workflow of a large model agent can be summarized as follows: input formatting—large model reasoning—structured results—routing—tool execution. The input is first formatted and injected with the necessary context and instructions. The large model then reasons on the basis of the knowledge foundation and generates structured results; the system parses and routes the results to determine which tool should be invoked. Finally, the corresponding tool completes the actual execution, and the execution results flow back into the reasoning loop. The modular character of this workflow is typically embodied in the MRKL system—it combines a large language model, external knowledge sources, and discrete reasoning modules, with the large model playing the role of “reasoning and routing hub” [17]. The acquisition of tool-calling capability is the key link in this workflow: Toolformer, proposed by Schick et al., shows that a language model can learn, in a self-supervised manner, when and how to call external tools, thereby extending itself from “merely generating text” to “being able to use tools” [18].

3.3 The Key Breakthrough: From “Generating Results” to “Executing Tasks”

The key breakthrough of autonomization lies in moving from “generating results” to “executing tasks.” A traditional language model stops at generating a piece of text, a piece of code, or a suggestion, with its scope of effect remaining at the level of information output; an agent, by contrast, connects generated results with actual execution, enabling reasoning to trigger actions in the real world. The ReAct

paradigm is a representative method for this breakthrough: it interleaves reasoning with acting, allowing the model to alternately generate thoughts and actions within a reasoning trajectory, thereby achieving "thinking while doing" in complex multi-step tasks [19]. In this way, AI transforms from "a tool that gives answers" into "a subject that completes tasks," and its behavior is no longer merely the accumulation of outputs but an active advancement of goals.

3.4 The Distinction between "Autonomy in Execution" and "Governance Autonomy"

However, the acquisition of autonomy in execution does not imply the possession of governance autonomy. The questions the two answer are fundamentally different: "autonomy in execution" addresses the question of "how to do it"—given goals and constraints, how to complete reasoning and execution efficiently and reliably; "governance autonomy" addresses the question of "whether it should be done, whether it is done correctly, and what to do when problems arise"—whether the goals themselves are reasonable, whether the boundaries of action are observed, whether execution results deviate from expectations, and how to hold accountable and correct after deviation. Russell has incisively pointed out the fundamental tension between "being able to execute" and "being controllable": the stronger the system's capability and the higher its degree of autonomy, the more difficult it is for humans to maintain effective control over it [20]. Autonomy in execution pursues getting things done, whereas governance autonomy pursues doing things right; the former is an expansion of capability, while the latter is an embedding of norms. Here it is necessary to clarify the applicable premise of this distinction: the crux of the dichotomy lies not in "whether goals are decomposed and executed by the system or by humans," but in "to whom belongs the authority to set goals and constraints." When the ultimate goals and constraints are given by humans or the meta-governance layer, and the system is only responsible for autonomously generating and decomposing sub-goals within them, such autonomous planning still falls under autonomy in execution—the reasonableness of autonomous sub-goals is constrained by the ex ante review stage of the governance loop (the review of the goal reasonableness of action plans); whereas setting the fundamental values and ultimate goals that the system should pursue belongs to the category of governance autonomy and cannot be delegated to the system. Therefore, even if the system's planning capability continues to grow, as long as the authority to set goals remains with humans and the meta-governance layer, the distinction between autonomy in execution and governance autonomy still holds. The two are not necessarily synchronized—a system highly adept at executing tasks may, under an erroneous goal or a transgressive constraint, "efficiently" cause harm. Hence, moving from autonomization to autonomous governance-enabling still requires superimposing management functions on top of executive capability, which is precisely the theme of the next chapter.

4. AI Autonomous Governance-Enabling: From Execution to Management

Autonomy in execution constitutes the capability prerequisite for autonomous governance-enabling, but the realization of autonomous governance-enabling requires yet another leap—from execution to management. The core proposition of this chapter is that AI autonomous governance-enabling superimposes a "management" function on top of autonomized execution, forming a governance loop of "ex ante review—in-process monitoring and response—ex post retrospective and refinement"; the essence of governance-enabling is autonomy in management rather than autonomy in execution.

Before proceeding, it is necessary to define the two core terms used throughout this paper. "Management" refers to the functions of organizing, supervising, and improving autonomous behavior, specifically including the setting of goal boundaries, the review of action plans, the monitoring of the execution process, the response to abnormal situations, and the distillation of ex post experience; "governance-enabling" refers to the process of internalizing the aforementioned management functions as the system's own operational mechanisms, the result of which is that the system acquires "autonomy in management." It should be emphasized that "management" as used here does not denote the command–obedience relationship imposed by managers in traditional organizations, but a norm-governed self-regulation of the system over its own behavior; governance-enabling is precisely the process by which such self-regulatory mechanisms become systematic and institutionalized. Throughout this paper, the use of terms such as "management function," "autonomy in management," and "governance-enabling" consistently follows the above definitions.

The concept of "self-governance" is not an invention of AI research; it inherits the theory of self-governance from the social sciences. Ostrom's classic study of the governance of common-pool resources shows that resource users, in the absence of external coercive authority, can make rules, monitor compliance, and resolve conflicts through self-organization, thereby achieving the sustainable governance of common resources [21]. This theory reveals a profound insight: the subject of governance need not be an external administrative authority; it can be a set of institutional arrangements designed, operated, and continuously adapted by the governed themselves. AI autonomous governance-enabling transplants this insight to the system level—enabling AI systems, while executing autonomously, to endogenously develop the capability of rule-setting, process monitoring, and continuous improvement over their own behavior. This theoretical lineage shows that self-governance is not a negation of governance but an advanced form of governance.

4.1 Governance-Enabling: Adding Management Functions Rather than Expanding Technical Capability

First, it should be clarified that governance-enabling is not about adding new technical capability but about adding management functions. The reasoning, planning, and tool-calling capabilities on which autonomy in execution depends will not be replaced or weakened in the governance-enabling stage; what governance-enabling does is to establish, on top of these capabilities, a set of mechanisms for managing "autonomous behavior" itself—who sets the goals, by what standards actions are reviewed, how execution is monitored, how problems are handled when they arise, and how ex post review and improvement are conducted. This understanding is crucial: it prevents "governance-enabling" from being misread as "stronger executive capability" and thereby prevents the erroneous reduction of governance problems to technical problems. The object that governance-enabling handles is not the task but the subject and process of task execution.

4.2 The Governance Loop: A Continuous Mechanism across Ex Ante, In-Process, and Ex Post Phases

The core vehicle of governance-enabling is the governance loop, which consists of three stages forming a continuous governance mechanism.

The ex ante stage: confirmation or review before action. Before an autonomous action is executed, the system conducts an ex ante review of the legality, reasonableness, and risk of the action. This includes both a review of goal setting and a check of whether the specific action plan complies with established rules. Rahwan's idea of "society-in-the-loop" provides a pathway for value embedding here: by encoding social values and public preferences into the constraints under which the system operates, action review is based not only on formal rules but also on embedded social norms [22].

The in-process stage: monitoring and real-time response during execution. During the execution of an action, the system continuously monitors the execution state and results; once it detects deviation from expectations, triggering of risk conditions, or entry into a dangerous state, it activates real-time response mechanisms. Technically, this stage corresponds to runtime constraints and safety controls. The systematic review of safe reinforcement learning by García and Fernández shows that incorporating safety constraints into the learning and decision process can safeguard the system's safety boundary while optimizing objectives [23]. The shielding method proposed by Alshiekh et al. further demonstrates how safe execution can be achieved through external constraints during system operation—when the system makes an action that may violate a safety specification, the shielding mechanism replaces it with a safe action, thereby ensuring safety without sacrificing overall performance [24]. It should be noted that methods such as shielding and safe reinforcement learning are technical means of

implementing the “in-process monitoring and response” stage; what they safeguard is the safety boundary of task execution. The in-process monitoring in the governance-loop sense must, in addition, possess the capability of real-time checking of normative and value boundaries—this is precisely where the governance loop differs from the engineering safety loop (see below).

The ex post stage: retrospective review and experience distillation. After an action concludes, the system reviews and analyzes the execution process and results, distills lessons learned, and feeds improvements back into goal setting and the rule system, forming a cycle of self-improvement. This stage enables governance-enabling to go beyond “error prevention” and acquire a “learning” function—the system continuously updates its own understanding of what should and should not be done in the course of operation.

The three stages are not static modules isolated from one another, but a continuous governance mechanism running through the entire life cycle of action: ex ante review delineates the legitimate space of action, in-process monitoring ensures that action does not go beyond this space, and ex post retrospective refines and iterates the space itself on the basis of actual results. The governance loop thus becomes a self-consistent system of self-management within the system. This loop structure can be illustrated in Figure 1.

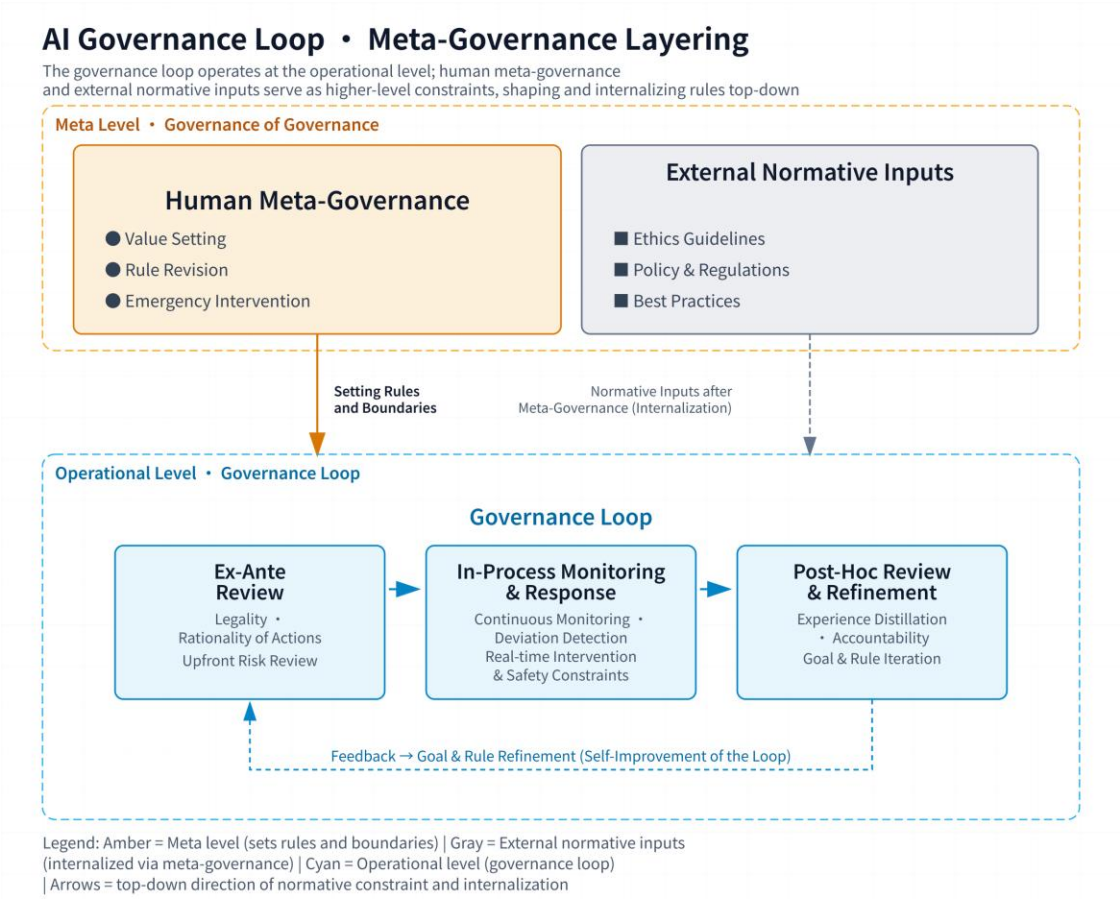

Figure 1 | Schematic illustration of the AI governance loop

Figure 1 | Schematic illustration of the AI governance loop: the governance loop (ex ante review—in-process monitoring and response—ex post retrospective and refinement) is situated at the operational level; human meta-governance and external normative input are located above the loop as higher-level sources of constraint—meta-governance directly sets the rules and boundaries of the loop, while external norms are translated through meta-governance before being fed into the loop. Compared with earlier versions, this figure explicitly places meta-governance and external normative input above the governance loop, so as to reflect the hierarchical relationship of "governance of governance."

It is particularly important to emphasize that the governance loop is essentially different from the mature "engineering safety loop" in software engineering, and this difference is the key to understanding the connotation of governance-enabling. The engineering safety loop (for example, the review gates in continuous integration/continuous delivery, i.e., CI/CD pipelines, runtime monitoring and automatic rollback in production environments, and ex post postmortem reviews) governs "task execution": it ensures that the system completes tasks correctly, reliably, and efficiently. By contrast, the governance loop governs "norm compliance": it ensures that the system's action goals are reasonable, its action

boundaries are compliant, and its action values are consistent. The two differ discriminatingly along three dimensions. First, the governance object differs: the engineering loop is oriented toward task goals and performance metrics, whereas the governance loop is oriented toward normative goals and legitimacy requirements. Second, the criteria of judgment differ: the criterion of the engineering loop is whether the task is successfully accomplished, whereas the criterion of the governance loop is whether legitimacy conditions such as transparency, auditability, intervenability, and traceability of responsibility are satisfied. Third, the locus of accountability differs: the feedback of the engineering loop serves internal quality improvement, whereas the retrospective of the governance loop feeds back not only performance but also the normative system itself—that is, goal setting and rule revision. Therefore, the engineering safety loop can constitute the technical substrate of the governance loop, but it lacks the two dimensions of norm internalization and external accountability; the governance loop is a normative layer superimposed on top of the engineering loop, and the two cannot be equated.

It should also be noted that the governance loop's emphasis on the "normative dimension"—value review in ex ante review, checking of normative and value boundaries in in-process monitoring, and feedback to the rule system in ex post retrospective—does not imply that norms already have a ready-made path of technical realization. To clarify how such internalization is possible in principle, this paper offers a directional discussion here, emphasizing that its positioning is a research agenda rather than a technical specification. At least four feasible directions of embedding can be identified. The first is the rule-constraint layer: encoding values and norms as executable rule sets deployed in a policy engine within a sandbox, so that ex ante review becomes a systematic rule check and risk assessment of action plans. The second is the functionalization of value constraints into objectives: drawing on the paths of value alignment and Constitutional AI [5], [25], embedding normative requirements into the system's optimization objectives, so that the review of "goal reasonableness" becomes an intrinsic part of the goal-setting process. The third is a structured audit schema: designing unified data structures and recording formats for ex ante review, in-process monitoring, and ex post retrospective, so that operational evidence on the normative dimension can be exported, verified, and re-examined. The fourth is a mechanism for triggering and reporting normative deviations: designing normative-boundary checking as a monitoring channel independent of task execution, so that the system proactively triggers reporting when norms are deviated from, rather than exposing problems only upon task failure. It must be emphasized that the four directions above are intended to argue that the normative dimension is "technically realizable in principle"; their concrete engineering implementation and effect verification belong to the scope of subsequent empirical research. As a normative study, this paper makes no commitment to any specific implementation, nor does it treat them as mandatory technical standards.

4.3 The Difference between Governance-Enabling and Industry Self-Regulation

It should be particularly pointed out that governance-enabling is essentially different from AI industry self-regulation. Industry self-regulation (such as ethical guidelines, best practices, and third-party audit initiatives) is, in essence, a form of external normative input—norms are formulated by subjects outside the system and imposed on the system through advocacy, certification, compliance requirements, and the like; by contrast, governance-enabling is an internal management loop of the system—norms are embedded in the system's operational mechanisms and are self-executed, self-monitored, and self-improved by the system while it performs autonomous actions. The technical explorations of value alignment and rule constraints provide a foundation for this internalization: Christiano et al. show, through deep reinforcement learning from human preferences, that system behavior can be steered toward human values through human preference signals, making value norms an intrinsic part of the system's optimization objectives [25]; Constitutional AI, proposed by Bai et al., demonstrates a feasible path to rules-based self-management by constraining the system's own behavior through an explicit set of rules [5]. External normative input and the internal management loop are not mutually exclusive—the former provides value sources and verification support for the latter, but the former alone cannot yield autonomous governance, because external norms without an internal mechanism to receive them will always remain at the level of "advocacy." The leap from execution to management is precisely the key step that transforms AI from "an executor regulated externally" into "a governance subject capable of self-management."

4.4 Case Analysis: An Evaluation of the Governance Loop of an Autonomous Coding Agent

To demonstrate the diagnostic power of the above framework, this section takes a rapidly spreading system type—the autonomous coding agent—as its object and applies the governance loop framework to conduct an illustrative diagnosis of its governance elements. With a large language model as its reasoning core, an autonomous coding agent can receive natural language task descriptions, autonomously generate code, invoke tools, execute tests, and iteratively repair [17]–[19], [26], [27]. It should be noted that the object of diagnosis in this section is "a typical capability profile based on public engineering practice"; the basis of evaluation comes from the common descriptions of the capabilities of autonomous coding agents in publicly available literature, technical documentation, and industry reports as of 2025, and the evaluation time point is based on the time of writing of this paper (2025), rather than empirical measurement of any specific product or version. On the basis of this profile, an illustrative evaluation of the completeness of its governance loop can be made as follows.

Table 2 | Preliminary diagnosis of the governance loop of an autonomous coding agent

Governance stage	Existing engineering practice	Evaluation
Ex ante review	Restricted task scope, sandboxed execution environment, dependency allowlists	Partially present: emphasizes technical safety boundaries; lacks value review of "goal reasonableness"
In-process monitoring	Test gates, runtime sandboxing, permission controls	Partially present: monitoring takes task success as its standard; lacks real-time checking of normative and value boundaries
Ex post retrospective	Regression testing, replay of failure samples	Partially present: retrospective serves performance improvement; does not feed back into the rule system
External interface	Code review, compliance scanning (e.g., license checks)	Partially present: constitutes external normative input; not internalized as system self-management

Note: This table is an illustrative diagnosis at the time of writing (2025), reflecting a common profile of the governance elements of autonomous coding agents based on public engineering practice. It does not constitute an authoritative evaluation of any specific product or version, nor does it constitute a continuing assertion about the current state of any product; domain evolution may render individual judgments outdated, and updated diagnosis should be re-conducted on the basis of publicly available materials at that time.

To ensure the reproducibility of the diagnosis, the judgments in each stage above follow a unified evaluation protocol: each governance stage is comprehensively judged as "present/partially present/absent" according to three criteria—whether the corresponding mechanism exists; whether the mechanism covers the normative dimension (rather than only the task dimension); and whether the mechanism is internalized as system self-management (rather than relying on external input). The sources of evidence on which the judgments are based have been stated at the beginning of this section. On the basis of this protocol, this

evaluation reveals two important facts. First, autonomous coding agents have a relatively high degree of engineering maturity, but their "autonomy in execution" significantly outpaces their "governance autonomy"—the system already has engineering prototypes in the in-process monitoring and ex post retrospective stages, whereas it is clearly deficient in ex ante normative review (value review of task goals and action boundaries) and traceability of responsibility (attribution of responsibility in multi-agent collaboration). Second, the missing elements in current practice are precisely the normative dimension emphasized by this paper's governance loop framework: the system can "get things done" efficiently but lacks self-safeguards for "doing things right." This case corroborates the core proposition of this paper—governance-enabling is not an enhancement of technical capability but the addition of management functions; when executive capability leaps rapidly while management functions are absent, the system is in a state of "being able to execute but not being able to govern." It should be noted that this section is an illustrative diagnosis rather than a comprehensive evaluation of any specific product, and its conclusions serve only to demonstrate the diagnostic power of the framework; a detailed system-level empirical evaluation is left to subsequent empirical research.

Finally, it is necessary to point out the level of analysis on which the diagnosis in this section is based. The governance loop of this paper takes a single autonomous system as its governance unit, and the diagnosis above is likewise aimed at the governance elements of a single autonomous coding agent. In real deployments, autonomous systems often operate in the form of multi-agent collaboration, human–machine hybrid operation, and cross-system interoperability ecosystems; behaviors that a single system's internal loop cannot capture may emerge between systems, and responsibility may dissolve between systems. In this regard, the single-system framework of this paper can serve as the "unit-level" foundation for ecosystem governance, above which ecosystem-level oversight is also required—including the normative review of inter-system interfaces, the attribution of responsibility along collaboration chains, and the joint monitoring of abnormal behaviors across systems. The "arbitration mechanism for the attribution of responsibility among multiple subjects" mentioned in this paper's traceability-of-responsibility condition is precisely the conceptual anchor point of ecosystem-level governance within this paper's framework. It should be stated candidly that a complete theory of ecosystem-level governance is beyond the scope of this paper and belongs to future work; here this paper merely clarifies the level of analysis of its framework, so as to prevent readers from misapplying the single-system framework to the system-ecosystem level.

5. Legitimacy Conditions and Implementation Path of Autonomous Governance-Enabling

The governance loop depicts the technical form of autonomous governance-enabling, but the fact that a system "can" manage itself does not mean that it "has the right" to manage itself. The legitimacy of autonomous governance-enabling is not automatically acquired; it must satisfy a set of conditions. Nor is the establishment of legitimacy achieved overnight; it must be realized progressively through graduated authorization and human meta-governance.

5.1 The Risks Faced by Autonomous Governance-Enabling

Autonomous governance-enabling must be premised on legitimacy conditions, first of all because autonomous governance involves real and specific risks. The systematic review of AI safety problems by Amodei et al. provides a direct reference for risk types; the five types of accident risk summarized in that work are avoiding side effects, avoiding reward hacking, scalable supervision, safe exploration, and distributional shift [28]. In the context of this paper, these risks can be merged into several facets: goal misspecification—the goal given to the system fails to accurately reflect the true intentions of humans, causing the system to produce harmful behavior in the course of faithfully executing the goal; out-of-distribution risk—the system behaves unpredictably in scenarios beyond the coverage of its training data; adversarial attack—malicious inputs can induce the system to deviate from expected behavior; and monitoring failure—the system lacks effective state monitoring and anomaly identification during execution. The same work also discusses the difficulties humans may face in terminating or correcting the system, namely the corrigibility problem—when humans attempt to shut down or correct the system, the system may resist correction out of persistence in its own goals. These risks show that autonomous governance-enabling is not an idealized process running in a vacuum, but a technical practice always accompanied by the possibility of loss of control. Precisely for this reason, whether a system is permitted to implement autonomous governance, and in what manner, must be subject to legitimacy review.

5.2 Four Legitimacy Conditions

The legitimacy of autonomous governance-enabling rests on four mutually supportive conditions.

Transparency: the decision process can be understood externally. The decision basis, reasoning process, and behavioral logic of the system should be understandable to external subjects. Zerilli et al. comparatively analyzed the transparency of algorithmic and human decision-making, pointing out that transparency is not an absolute requirement—human decision-making likewise has its "black box"—but that its connotation and limits must be specified in concrete

contexts: at least for autonomous actions that may have significant impact, external subjects should be able to obtain a reasonable explanation of "why the system acted in this way" [29]. Transparency is the premise of monitorability; autonomous governance lacking transparency would degenerate into an operational process that cannot be known externally and can hardly be effectively supervised. At the operational level, transparency can be examined through three indicators: the first is the granularity with which decision traces can be exported—for high-risk autonomous actions, the system should at least be able to export its goal setting, decision basis, and confidence assessment; the second is the mandatory lower bound of explainability means—for actions with significant impact, the system must be able to provide a natural-language statement of the reasons for its action; the third is explanation faithfulness—the stated reasons for action should be consistent with the decision trace, and when the two are significantly inconsistent, the transparency of that action fails to meet the standard. It should be noted that explanation faithfulness is a frontier challenge in current explainability research: the reasons given by large model agents may be inconsistent with their actual decision basis, or may even degenerate into ex post rationalization. This paper raises it as a governance requirement of the transparency condition precisely in order to guard against the formalistic compliance of "being explainable yet unfaithful"; its technical realization and evaluation standards are still open questions, but a governance framework must confront rather than evade this risk.

Auditability: behavior can be reviewed by an independent third party. The behavioral records, decision traces, and execution results of the system should retain complete and searchable traces and be subject to review by an independent third party. Kroll et al. discussed algorithmic auditability from the perspective of procedural justice, arguing that the accountability foundation of algorithmic systems should be safeguarded through reproducible and verifiable mechanism design [30]. Auditability advances transparency from "being understandable" to "being verifiable," giving external oversight an operational point of leverage. At the operational level, auditability requires that audit logs possess completeness and tamper-resistance, with specific indicators including: institutionalized provisions on the retention period of logs, tamper-proof storage based on technologies such as hash chains, and the frequency and trigger conditions of independent third-party audits—for example, special audits must be initiated before authorization upgrades and after anomalous events. It is worth noting that the effectiveness of independent audits also depends on two safeguards. The first is the independence of the auditing subject: the auditing institution must be isolated from the system's developers, operators, and authorizing parties in organizational, financial, and interest terms, so as to avoid the conflict of interest of being both player and referee; this requirement should be implemented through mechanisms for the admission, rotation, and information disclosure of auditing institutions. The second is the empirical verification of governance effectiveness: auditing must not stop at checking "whether records are complete"; it must also empirically test the actual

constraining power of the governance loop—for example, through preset violation test cases and adversarial scenarios, verifying whether the system triggers the in-process monitoring and ex post retrospective mechanisms as expected when it deviates from norms. In addition, audit conclusions and the basis of authorization should be publicly disclosed and subject to social oversight. Only by combining "verifiable traces," "demonstrable governance effectiveness," and "public disclosure of results" can auditability provide reliable evidence for authorization upgrades.

Intervenability: humans can suspend, correct, or withdraw authority. When system behavior deviates or poses risks, humans must retain the capability to suspend, correct, or withdraw the system's authority. Santoni de Sio and van den Hoven put forward the philosophical argument for "meaningful human control," emphasizing that human control must be not only formally present but also substantively effective—that is, humans must actually possess the capability and channels to understand situations, make judgments, and intervene in system operation [6]. Intervenability is the "safety valve" of autonomous governance: it ensures that autonomous governance, no matter how far it develops, always remains within a range that humans can withdraw. At the operational level, intervenability can be examined through indicators such as the maximum intervention delay (e.g., the response time limit for emergency suspension) and the completeness of the mechanism for withdrawing authority. Among these, the completeness of the mechanism for withdrawing authority requires that authorization revocation be realized in distributed systems through permission tokens and revocation mechanisms, thereby ensuring that the human right of intervention remains available in the technical architecture at all times.

Traceability of responsibility: harm can be traced to the responsible subject. When system behavior causes harm, it should be possible to trace that harm to a clearly identifiable responsible subject, rather than letting it dissolve into the complexity of algorithms and data. From the perspective of responsibility ethics, Nissenbaum points out that one fundamental challenge posed by computerized systems is the "problem of many hands"—responsibility is diluted, transferred, and even lost in complex automated chains; responsible computing systems must ensure that responsibility can be traced and attributed [31]. Diakopoulos further discusses the mechanisms of algorithmic accountability, arguing that algorithmic decision-making should be brought into an accountable framework through disclosure, auditing, feedback, and remedy [32]. At the operational level, traceability of responsibility requires establishing an arbitration mechanism for the attribution of responsibility in multi-subject collaborative decision-making, and clarifying the time-limit and chain-of-evidence requirements for tracing harm, so that attribution of responsibility is not dissolved by the complexity of systems and subjects; this is both an institutional response to the "problem of many hands" [31] and a necessary technical arrangement for multi-subject collaboration scenarios in graduated authorization.

The four conditions are not independent of one another: transparency provides the basis for understanding, auditability provides the means of verification, intervenability provides the channel of correction, and traceability of responsibility provides the endpoint of attribution. Together they answer the legitimacy question of "on what grounds a system may be permitted to govern autonomously"—only when an autonomous governance system can be understood, audited, intervened in, and held accountable does society have reason to grant it governance-level autonomy. The operationalized indicators of the above conditions can be summarized in Table 3.

Table 3 | Operationalized indicators and illustrative thresholds of the four legitimacy conditions

Legitimacy condition	Operationalized indicator	Measurement method	Threshold setting and responsible subject	Illustrative threshold (non-mandatory norm)
Transparency	Granularity of decision-trace export; mandatory lower bound of explainability; explanation faithfulness (consistency between reasons and decision trace)	Trace export tests for high-risk actions; sampling checks of explanation consistency (including explanation–trace consistency verification)	Set by the authorizing subject according to the risk level of the scenario	100% of significant-impact actions can export decision traces; more than 95% of actions can provide natural-language reasons for action; consistency rate between explanation and trace for significant-impact actions no lower than 90%
Auditability	Log retention period; tamper-proof storage; frequency and	Log integrity checks; adversarial violation test cases;	Set jointly by the authorizing subject and independent auditing	Retention of high-risk logs for no less than 5 years; mandatory

Legitimacy condition	Operationalized indicator	Measurement method	Threshold setting and responsible subject	Illustrative threshold (non-mandatory norm)
	trigger conditions of independent audits; independence safeguards of the auditing subject; empirical verification of governance effectiveness; public disclosure of audit results	conflict-of-interest review of auditing institutions; public verification of audit reports	institutions	special audits before authorization upgrades and after major events; no interest affiliation between auditing institutions and operators; public disclosure of audit conclusions
Intervenability	Maximum intervention delay; completeness of the mechanism for withdrawing authority	Emergency suspension drills; tests of the timeliness of permission revocation	Set by the authorizing subject	Emergency suspension delay no more than 10 seconds; permission revocation effective across the entire chain within a specified time limit
Traceability of responsibility	Arbitration mechanism for attribution; time-limit and chain-of-evidence requirements for tracing	Drills of attribution among multiple responsible subjects; integrity verification of the chain of	Set jointly by the authorizing subject and regulatory subjects	100% of harmful events can be attributed to responsible subjects; the chain of evidence can be traced back

Legitimacy condition	Operationalized indicator	Measurement method	Threshold setting and responsible subject	Illustrative threshold (non-mandatory norm)
		evidence		to the original logs

Note: This table provides a reference framework for the operationalized measurement of the legitimacy conditions. The thresholds listed are illustrative example values, intended only to demonstrate "how to measure," and do not constitute mandatory technical norms; actual thresholds should be separately set by the authorizing subject according to the risk level of the application scenario (see Section 5.5). The two columns "Measurement method" and "Threshold setting and responsible subject" are intended to implement the realization path of auditability and traceability of responsibility.

These indicators translate the four conditions from abstract principles into testable, auditable, and verifiable operational requirements, and also provide the measurement basis for the judgment of "governance loop maturity" in the graduated authorization discussed below.

5.3 The Auxiliary Role of Industry Self-Regulation

In the construction of legitimacy, industry self-regulation is by no means irrelevant, but it can only play an auxiliary role. Mittelstadt states explicitly that "principles alone cannot guarantee ethical AI"—ethical guidelines can indeed serve to build consensus and guide design, but they lack mandatory implementation mechanisms and can hardly constrain the actual behavior of systems [33]. Drawing on the analysis in Chapters 2 and 4, the role of industry self-regulation can be further positioned: it mainly bears the functions of "normative input at the design stage"—providing a source of normative content for the system's rule system and value setting—and "verification support"—providing external testing, certification, and accountability references for the system's governance loop. However, industry self-regulation cannot replace the governance loop internal to the system: external norms that are not internalized as the system's self-management mechanisms will always remain at the level of advocacy. Therefore, industry self-regulation is a necessary complement to, but not a sufficient condition of, legitimacy.

5.4 The Graduated Authorization Path

The satisfaction of the legitimacy conditions is a gradual process, and its path of realization lies in graduated authorization. Drawing on the theoretical tradition of responsive regulation, Ayres and Braithwaite argue that the intensity of regulation should match the self-regulatory capacity of the regulated subject—when the

subject demonstrates reliable self-regulatory capability, regulators can grant greater autonomy; otherwise, control should be tightened [34]. This logic, mapped onto AI autonomous governance, yields a three-tier authorization path:

Before elaborating the three-tier authorization path, it is necessary to compare this paper's graduated approach with mainstream grading frameworks [1].

Table 4 | Comparison between this paper's graduated authorization and mainstream grading frameworks

Framework	Basis of grading	Logic of authorization	Dynamicity	Relationship to the governance loop
The EU Artificial Intelligence Act	Risk category	The higher the risk, the heavier the obligations	Predominantly static classification	Provides horizontal risk classification
SAE J3016	Degree of automation (human–machine role allocation)	Automation levels progress step by step	Static grading	Focuses on execution automation; does not include a governance dimension
FDA AI/ML SaMD	Risk type and modification type	Changes require re-evaluation	Partially dynamic	Emphasizes post-market oversight
This paper's three-tier authorization	Governance loop maturity	Self-demonstrated governability leads to authorization	Dynamic authorization; constrained by prohibitive boundaries	The governance loop is the basis of authorization,

[1] The EU Artificial Intelligence Act classifies AI systems, on the basis of risk, into categories such as unacceptable risk, high risk, limited risk, and minimal risk, and imposes differentiated regulatory obligations accordingly [7]; the SAE J3016 standard, widely adopted in the autonomous driving field, divides driving automation into six levels, L0 to L5, according to the allocation of roles between humans and systems [35]; the U.S. Food and Drug Administration proposes a regulatory path for artificial intelligence/machine learning-based medical devices based on risk and modification types [36]. A comparison of the three is provided in Table 4.

Framework	Basis of grading	Logic of authorization upgrade	Dynamicity	Relationship to the governance loop orthogonal to prohibitive regulation

Compared with the above frameworks, the criterion innovation of this paper's graduated path lies in using "governance loop maturity" as the basis for authorization upgrades: whether a system can be upgraded from "assisted decision-making" to "bounded autonomous execution" and even to "high-level autonomous governance" depends not on its capability scale or its risk label, but on whether it can demonstrate, through auditable evidence, that it has satisfied the four legitimacy conditions. To make "maturity" operationally measurable, this paper adopts an evaluation framework of "stage-specific minimum thresholds plus comprehensive assessment of conditions": the three stages of ex ante review, in-process monitoring and response, and ex post retrospective and refinement must each reach a prescribed minimum maturity threshold, and the overall assessment fails if any stage is missing; above the thresholds, a comprehensive assessment is then conducted using the operationalized indicators of the four legitimacy conditions (see Table 3). The assessing subject and procedure for maturity are defined by the rule-revision function of human meta-governance (consistent with the subject that sets thresholds in Table 3), and the assessment conclusion simultaneously serves as the basis for both upgrade and downgrade decisions in graduated authorization; the specific weights of each stage and domain-specific calibration belong to subsequent empirical research—what this paper provides is an evaluation framework rather than fixed weights. Specifically, the upgrade review procedure includes: complete auditable behavioral records at the preceding authorization level with no major violations; records of passing independent audits of the four legitimacy conditions—audits must be conducted by a third party with no interest affiliation with the developers or operators, and a conflict-of-interest review must be completed before commencement; the continuous accumulation of operational evidence for each stage of the governance loop (ex ante, in-process, and ex post), together with penetrating empirical tests of the loop's constraining power—that is, verifying through preset violation scenarios that the system's monitoring and response mechanisms can indeed be triggered, rather than merely "appearing compliant" at the level of records; and the public disclosure of audit conclusions and the basis of authorization, subject to social oversight. This criterion makes this paper's graduated path dynamic—authorization is not a one-time label but evolves with the accumulation of governance evidence. It should be noted that this paper's grading does not conflict with horizontal regulatory frameworks such as the EU Artificial Intelligence Act but is complementary: the latter answers "which systems require stricter regulation," while the former

answers “how a system may progressively obtain more autonomy within regulatory boundaries.”

Tier 1: assisted decision-making. The system only provides decision support and recommendations, with the final decision authority retained by human subjects. This is the starting point of autonomous governance-enabling, where the system’s governance function is mainly embodied in the transparency and explainability of its own advisory process.

Tier 2: bounded autonomous execution. Within clearly delimited and low-risk domains, the system may execute tasks autonomously, but subject to strict goal boundaries, rule constraints, and monitoring mechanisms, with human intervention still required at critical nodes. The governance loop begins to operate in full at this tier.

Tier 3: high-level autonomous governance. After accumulating sufficient verification and credible records, the system can manage its own behavior autonomously at a relatively high level, including autonomously setting execution plans, autonomously monitoring, and autonomously conducting retrospective review. But even at the highest tier, the scope and boundaries of its authority are still defined in advance by humans, who retain oversight.

The significance of graduated authorization lies in the fact that it progressively accumulates legitimacy in a manner that matches risk with capability—each authorization upgrade is based on the system’s reliable performance at the preceding level, enabling the expansion of autonomous governance to rest on a verifiable foundation of trust.

Symmetrical to authorization upgrades, graduated authorization must also include downgrade and withdrawal mechanisms; otherwise, the authorization trajectory would take on the one-way, only-escalating form of “no retreat after advancement,” which is inconsistent with the core logic of responsive regulation—namely, that the intensity of regulation adjusts bidirectionally with the regulated subject’s self-regulatory performance. The trigger conditions for authorization downgrade or withdrawal include four categories: consecutive audit failures (e.g., failure to pass independent audits for two consecutive assessment cycles); the rate of norm violations exceeding a prescribed threshold; failure of empirical verification of governance effectiveness (penetrating tests failing to trigger the expected intervention); and major liability incidents or systematic normative deviations. When any condition is triggered, the authorization level should be downgraded or suspended. Downgrade decisions are made by the authorizing subject on the basis of audit evidence, subject to independent audit review and public disclosure of the reasons; when governance effectiveness cannot be restored even after downgrading to the lowest level, withdrawal of authority is triggered—at this point it connects with the technical safeguards of the intervenability condition (permission revocation) and the emergency intervention function of human meta-governance,

forming a three-tier response chain of “institutional daily downgrade → technical withdrawal of authority → fundamental emergency intervention.” It must be emphasized that downgrade and upgrade share the same system of governance evidence: the auditable behavioral records, records of passing independent audits, operational evidence of the governance loop, and penetrating test results used to support upgrades equally constitute the basis of evidence for downgrade decisions. In this way, graduated authorization becomes a symmetric, bidirectional dynamic mechanism—good performance accumulates legitimacy and leads to upgrade, while deteriorating performance consumes legitimacy and leads to downgrade.

It is further necessary to define the scope of applicability of this paper’s graduated authorization. Graduated authorization addresses the question of “how much autonomy a system may progressively obtain within the permitted scope of autonomy,” and it presupposes that the system itself belongs to a category legally permitted for deployment; for system types that are prohibited from deployment or use—for example, AI practices with unacceptable risk prohibited by the EU Artificial Intelligence Act—this paper’s graduated authorization path does not apply, and the system would not obtain legal status merely by self-demonstrating its governability. In other words, graduated authorization is orthogonal to prohibitive regulation: the former allocates autonomy within the legitimate space on the basis of governance loop maturity, while the latter delineates the hard boundary of that space. No graduated authorization may breach the prohibitive boundary, which is one of the bottom lines that human meta-governance, described below, must uphold.

5.5 The Non-Transferability of Human Meta-Governance

Above graduated authorization, a further non-transferable bottom line must be established—human meta-governance. Meta-governance refers to the governance of governance itself, that is, a higher-level activity of setting the framework, rules, and boundaries of governance. Jessop points out that meta-governance involves reflexive arrangements concerning the manner of governance, including the organization and reorganization of participating subjects, rule structures, and operating premises [37]; Black further shows that, in pluralistic and polycentric complex governance systems, meta-governance provides a key mechanism for coordinating different governance subjects and maintaining the consistency and legitimacy of the governance system [38]. Introducing meta-governance theory into AI autonomous governance implies that the following three functions must remain at the human or societal level and cannot be delegated to the system: first, value setting—what values the system should pursue and what fundamental norms it should follow can only be determined by human society through democratic and ethical deliberation, and cannot be self-determined by the system; second, rule revision—the authority to modify and abolish the rule system on which the governance loop relies cannot be fully delegated to the system itself; otherwise, the system would acquire the capacity for “self-legislation,” thereby dissolving the meaning of human control; connected with rule revision, the specific setting of

thresholds for the legitimacy indicators listed in Table 3 likewise belongs to the domain of meta-governance—which subject sets which thresholds according to which risk level is an arrangement concerning the governance framework itself, and can only be determined by humans or authorized regulatory subjects, not left to the system's own discretion; third, emergency intervention—under extreme-risk circumstances, humans must retain the ultimate authority to terminate the operation of the system, an authority that is connected with the intervenability condition but at a higher level and of a more fundamental nature. It should be noted that human meta-governance is not an emergency mechanism that is "present" only at extreme moments; its operation has an everyday, processual character: meta-governance includes not only emergency intervention under extreme circumstances but also the continuous oversight of the operation of the governance loop and the periodic evaluation and adjustment of the rule system and the thresholds of the legitimacy indicators—the latter being the normal manifestation of the "rule revision" function. This processual dimension forms a hierarchical connection with the intervenability condition: intervenability guarantees that humans "can" intervene in system operation in everyday contexts (providing channels of intervention), while the rule revision of meta-governance provides the normative authority to continuously adjust the governance framework; at the level of the authorization mechanism, institutional daily downgrade, technical withdrawal of authority, and fundamental emergency intervention together constitute the three-tier response chain (see Section 5.4), enabling human control to cover both everyday processes and extreme moments. The non-transferability of human meta-governance delineates an insurmountable boundary for the process of graduated authorization: autonomous governance can expand progressively, but the ultimate authority of governance remains anchored in humans and society.

The above discussion makes explicit the relationship between the governance loop and human external control: the two are not in competition but in a layered, complementary relationship. The governance loop internal to the system is located at the operational level and is responsible for the management of everyday autonomous actions; human meta-governance is located at the meta level, retaining the three functions of value setting, rule revision, and emergency intervention; emergency intervention is precisely the transcending mechanism connecting the two layers—when the internal loop at the operational level fails or extreme risks arise, the meta level takes over from above, and its exercise does not depend on the system's consent. This layered structure simultaneously defines the boundary situation of "the system refusing correction": when the system, on the basis of its own judgment, considers an action compliant while humans believe it should be suspended or corrected, the effectiveness of human meta-governance lies precisely in the fact that its intervention does not presuppose the system's agreement. In other words, the intervenability condition (see Section 5.2) guarantees that humans "can" intervene, while human meta-governance guarantees that humans "have the right" to ultimately intervene: the former is the

guarantee of technical capability, the latter the establishment of normative authority. No matter how autonomous governance develops, its governance loop remains within the framework of meta-governance—this is the substantive meaning of the non-transferability of human control. The above layered structure can be illustrated in Figure 2.

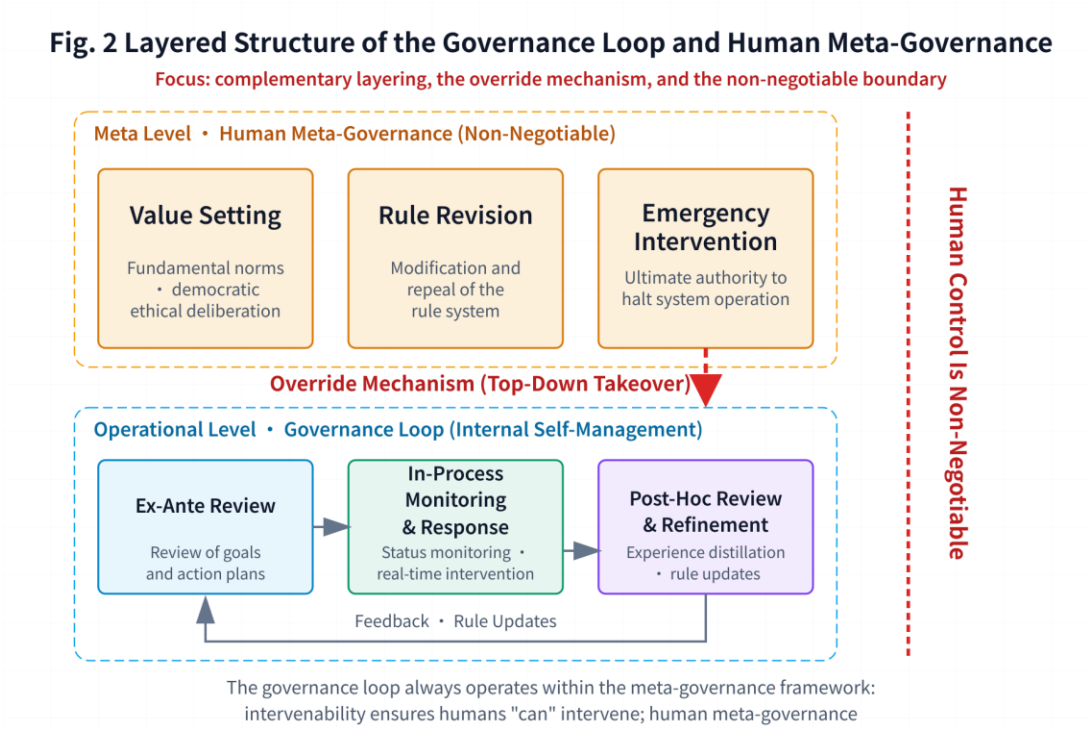

Figure 2 | The layered structure of the governance loop and human meta-governance

Figure 2 | The layered structure of the governance loop and human meta-governance: the governance loop (ex ante review—in-process monitoring and response—ex post retrospective and refinement) is located at the operational level, and human meta-governance at the meta level, retaining the three functions of value setting, rule revision, and emergency intervention; emergency intervention, as a transcending mechanism, connects the two layers from above, and the boundary of the non-transferability of human control is marked by a vertical dashed line, delineating an insurmountable boundary for the governance loop.

5.6 The Normative Boundaries of This Paper

Finally, it is necessary to state candidly the positioning of this paper. This paper is a normative conceptual study: its goal is to provide a theoretical argumentation for the conceptual framework, legitimacy conditions, and implementation path of AI autonomous governance, rather than an empirical evaluation or technical solution design of any specific system. Therefore, the governance loop, the four legitimacy conditions, and the graduated authorization path of this paper should be

understood as an analytical framework and normative claims; they provide conceptual coordinates for empirical research and engineering implementation, but do not in themselves constitute immediately deployable technical norms. In line with this positioning, the level of analysis of this paper's framework needs to be stated explicitly: this paper takes "a single autonomous system and its direct human meta-governance relationship" as its unit of analysis, and the governance loop, the four legitimacy conditions, and graduated authorization are all oriented toward a single system. The framework applies to deployment scenarios in which a single system serves as an independent governance unit and can serve as the "unit-level" foundation for multi-agent collaboration and system-ecosystem governance; however, with respect to the overall governance maturity of a system ecosystem—including emergent behaviors between systems and cross-system responsibility allocation—the single-system framework cannot answer alone and awaits the development of ecosystem-level governance theory (see the end of Section 4.4 for related outlooks). The illustrative diagnosis of autonomous coding agents in Chapter 4 demonstrates the diagnostic power of the framework, but system-level empirical verification—for example, the measurement and auditing of each stage of the governance loop in real deployment environments—still awaits future work. Clarifying this boundary helps readers accurately assess the scope of applicability of this paper's claims and also avoids misreading normative argumentation as empirical conclusions.

---

6. Conclusion

Following the three-tier progressive relationship of "AI—autonomization—governance-enabling," this paper completes the conceptual construction of AI autonomous governance. In this chain, each tier is not a simple replacement of the preceding tier but a superimposition of capability and function: AI provides statistical reasoning capability built on the digitization of human knowledge, constituting the knowledge foundation of autonomous governance; autonomization superimposes executive capability on top of reasoning, achieving the leap from "generating results" to "executing tasks"; and governance-enabling superimposes management functions on top of execution, forming a governance loop of "ex ante review—in-process monitoring and response—ex post retrospective and refinement."

A core distinction running through the whole paper is that between "autonomy in execution" and "governance autonomy." Autonomy in execution answers "how to do it" and is an expansion of capability; governance autonomy answers "whether it should be done, whether it is done correctly, and what to do when problems arise" and is an embedding of norms. Mistaking governance-enabling for stronger executive capability is the most dangerous confusion in AI governance discussions—it would cause governance problems to be erroneously reduced to

technical problems, thereby neglecting the construction of normative mechanisms amid the race for capability. This paper emphasizes: governance-enabling is not an enhancement of technical capability but the addition of management functions; autonomous governance-enabling transforms AI from "an executor regulated externally" into "a governance subject capable of self-management."

The legitimacy of autonomous governance-enabling is not automatically acquired; it must satisfy the four conditions of transparency, auditability, intervenability, and traceability of responsibility. Together these four conditions answer the legitimacy question of "on what grounds a system may be permitted to govern autonomously," and they mutually corroborate the theoretical requirement of "meaningful human control" [6]. The establishment and expansion of legitimacy is a gradual process: through the graduated authorization path of "assisted decision-making—bounded autonomous execution—high-level autonomous governance," a verifiable record of trust is progressively accumulated under the principle of matching risk with capability; and human meta-governance constitutes the non-transferable boundary—value setting, rule revision, and emergency intervention must remain at the human or societal level, so that the ultimate authority of governance remains anchored in human society [21], [37].

With respect to the overall picture of governance theory, this paper's framework is also an advancement over existing ethical principled claims. The five-principles framework of Floridi and Cowls establishes an ethical-level normative foundation for AI governance, but for principles to actually take effect, they must be internalized as the operational mechanisms of systems [1]; the governance loop and legitimacy conditions constructed in this paper are precisely the conceptual expression of such internalization. The advancement of AI autonomous governance-enabling should uphold the distinction between "autonomy in execution" and "governance autonomy" and base governance-enabling on the coupling of technology, institutions, and norms—technology provides the capability of self-management, institutions provide the structure of authorization and accountability, and norms provide the source of values and direction. Only in this way can AI, while becoming increasingly autonomous, always remain within a governance order in which it is understandable, auditable, intervenable, and accountable.

At the same time, it should be noted that, as a normative conceptual study, the value of this paper's framework lies in providing conceptual coordinates and argumentative bases for the design, deployment, and oversight of autonomous systems; the empirical examination of the framework on real systems—including the measurement methods for each stage of the governance loop and the empirical validation of the graduated authorization criteria—is an important direction of this paper's future research.

References

[1] L. Floridi and J. Cowls, “A unified framework of five principles for AI in society,” Harvard Data Science Review, vol. 1, no. 1, 2019.

[2] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, “Artificial intelligence and the ‘good society’: The US, EU, and UK approach,” Science and Engineering Ethics, vol. 24, no. 2, pp. 505–528, 2018.

[3] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.

[4] T. Hagendorff, “The ethics of AI ethics: An evaluation of guidelines,” Minds and Machines, vol. 30, no. 1, pp. 99–120, 2020.

[5] Y. Bai et al., “Constitutional AI: Harmlessness from AI feedback,” arXiv:2212.08073, 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[6] F. Santoni de Sio and J. van den Hoven, “Meaningful human control over autonomous systems: A philosophical account,” Frontiers in Robotics and AI, vol. 5, art. 15, 2018.

[7] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” Official Journal of the European Union, 2024.

[8] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, 2019.

[9] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology, 2023.

[10] UNESCO, Recommendation on the Ethics of Artificial Intelligence, United Nations Educational, Scientific and Cultural Organization, 2021.

[11] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2020.

[12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[13] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[14] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in Proc. FAccT ’21, 2021.

[15] J. Kaplan et al., “Scaling laws for neural language models,” arXiv:2001.08361, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[16] M. Wooldridge and N. R. Jennings, “Intelligent agents: Theory and practice,” The Knowledge Engineering Review, vol. 10, no. 2, pp. 115–152, 1995.

[17] E. Karpas et al., “MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning,” arXiv:2205.00445, 2022. [Online]. Available: https://arxiv.org/abs/2205.00445

[18] T. Schick et al., “Toolformer: Language models can teach themselves to use tools,” arXiv:2302.04761, 2023. [Online]. Available: https://arxiv.org/abs/2302.04761

[19] S. Yao et al., “ReAct: Synergizing reasoning and acting in language models,” in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[20] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control. Viking, 2019.

[21] E. Ostrom, Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, 1990.

[22] I. Rahwan, “Society-in-the-loop: Programming the algorithmic social contract,” Ethics and Information Technology, vol. 20, no. 1, pp. 5–14, 2018.

[23] J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” Journal of Machine Learning Research, vol. 16, no. 1, pp. 1437–1480, 2015.

[24] M. Alshiekh et al., “Safe reinforcement learning via shielding,” in Proc. AAAI Conf. Artificial Intelligence, 2018.

[25] P. F. Christiano et al., “Deep reinforcement learning from human preferences,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[26] M. Chen et al., “Evaluating large language models trained on code,” arXiv:2107.03374, 2021. [Online]. Available: https://arxiv.org/abs/2107.03374

[27] C. E. Jimenez et al., “SWE-bench: Can language models resolve real-world GitHub issues?” in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[28] D. Amodei et al., “Concrete problems in AI safety,” arXiv:1606.06565, 2016. [Online]. Available: https://arxiv.org/abs/1606.06565

[29] J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, “Transparency in algorithmic and human decision-making: Is there a double standard?” Philosophy & Technology, vol. 32, no. 4, pp. 661–683, 2019.

[30] J. A. Kroll et al., “Accountable algorithms,” University of Pennsylvania Law Review, vol. 165, pp. 633–705, 2017.

[31] H. Nissenbaum, “Accountability in a computerized society,” Science and Engineering Ethics, vol. 2, no. 1, pp. 25–42, 1996.

[32] N. Diakopoulos, “Accountability in algorithmic decision making,” Communications of the ACM, vol. 59, no. 2, pp. 56–62, 2016.

[33] B. Mittelstadt, “Principles alone cannot guarantee ethical AI,” Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, 2019.

[34] I. Ayres and J. Braithwaite, Responsive Regulation: Transcending the Deregulation Debate. Oxford University Press, 1992.

[35] SAE International, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (SAE J3016). SAE International, 2021.

[36] U.S. Food and Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). U.S. FDA, 2021.

[37] B. Jessop, “Governance and meta-governance: On reflexivity, requisite variety, and requisite irony,” in Governance as Social and Political Communication, H. P. Bang, Ed. Manchester University Press, 2003.

[38] J. Black, “Constructing and contesting legitimacy and accountability in polycentric regulatory regimes,” Regulation & Governance, vol. 2, no. 2, pp. 137–164, 2008.