Governance of AI Applications in High-Risk Scenarios: A Risk Analysis Based on the "Cognitive Compensation" Perspective

Tianran Zhang[1] **and Xinqi Feng***[2]

[1]School of Design and Architecture,Zhejiang University of Technology, Hangzhou,China
[2]School of Software Technology,Zhejiang University,Hangzhou,China

Abstract: In the era of intelligent transformation, the widespread application of AI systems in high-risk scenarios (such as autonomous driving, medical diagnosis, and judicial sentencing) has triggered unprecedented governance challenges. Traditional governance frameworks often focus on the "technical reliability" of AI or the "ethical review" of humans, failing to deeply reveal the root causes of risks in human-machine collaborative decision-making. This paper introduces the "Theory of Cognitive Compensation" and the "RID Cognitive Dynamics Model" from "Knowing and Speaking", constructing a systematic analytical framework for AI application risks in high-risk scenarios. The study proposes that the essence of AI application is human beings transferring part of the "Structural Generation (I)" function to machines to alleviate cognitive "Demand/Drive (D)", forming a "Cognitive Compensation" mechanism. However, when this compensation mechanism is improperly extended to high-risk scenarios, it induces five major structural risks: concept drift, automation bias, responsibility suspension, cognitive hallucination, and malicious compensation. To address these risks, this paper proposes establishing a "Dynamic Redundancy" governance framework, including structural redundancy, cognitive redundancy, institutional redundancy, and fault-tolerant redundancy, to ensure human beings' ultimate control over high-risk decisions. By critically analyzing the EU's Artificial Intelligence Act and the US governance model, this paper argues that the core principle of high-risk scenario governance is to strictly delineate compensation boundaries and achieve "responsibility centralization", thereby providing a new theoretical foundation and institutional design reference for global AI governance.

Keywords: High-Risk Scenarios; Artificial Intelligence Governance; Cognitive Compensation; RID Model; Concept Drift; Dynamic Redundancy

1 Introduction: The Paradigm Crisis of AI Governance in High-Risk Scenarios

With the rapid advancement of artificial intelligence (AI), particularly large language models (LLMs) and deep learning algorithms, human society is entering the intelligent age at an unprecedented pace. AI is no longer confined to consumer-level applications like recommendation algorithms or voice assistants but is rapidly advancing into fields such as medical auxiliary diagnosis, judicial sentencing references, autonomous driving, financial credit approval, and even military target recognition. These domains are academically and policy-wise referred to as "high-risk scenarios" (High-stakes Scenarios).

1.1 Defining High-Risk Scenarios and Their Specificity

The core characteristic of high-risk scenarios lies in the fact that decision errors will directly lead to harm to human life and health, deprivation of fundamental rights, loss of substantial property, or the outbreak of systemic social crises. In these scenarios, the cost of trial-and-error is extremely high, even irreversible. For instance, in the medical field, AI misdiagnosis of medical images may cause patients to miss the optimal treatment window; in the judicial domain, racial or gender biases embedded within AI algorithms could result in the wrongful conviction of innocents; in autonomous driving, system perception failure can directly trigger fatal accidents.

1.2 Limitations of Existing Technical Governance Frameworks

Current AI governance frameworks primarily emphasize technical dimensions of risk control, such as algorithm transparency, interpretability, and robustness. However, this technocentric approach often overlooks the profound epistemological challenges faced by AI. Zhao Tingyang (2023) profoundly points out that the boundary of artificial intelligence is not merely a technological boundary, but also an epistemological one. When AI attempts to use statistical correlation to

* Corresponding author: langqxq@163.com

replace human causal reasoning, it is essentially reconstructing a "pseudo-knowledge" lacking subjectivity support [20]. This epistemological transgression means that mere technical patchwork cannot fundamentally resolve governance dilemmas in high-risk scenarios.

Faced with these looming swords of Damocles, existing AI governance frameworks often appear inadequate. Current governance paradigms are largely based on the logic of "technical regulation," aiming to reduce risks by improving algorithmic accuracy, enhancing model interpretability (Explainable AI, XAI), and eliminating biases in training data. Nevertheless, this technocentric governance path faces fundamental limitations.

Firstly, the "black box" nature of deep learning models and their high-dimensional computational logic make absolute interpretability technically difficult to achieve. Secondly, the "emergent abilities" and inherent "hallucinations" of large models indicate that technical failures are endogenous to current AI architectures and cannot be entirely eliminated through simple code optimization. More importantly, many decisions in high-risk scenarios are essentially value judgments rather than factual calculations (e.g., "leniency and severity" in judicial sentencing). Attempting to fit complex socio-ethical norms with mathematical formulas inevitably triggers deeper conflicts.

1.3 The Theoretical Value of Introducing the "Cognitive Compensation" Perspective

Therefore, AI governance in high-risk scenarios faces a profound paradigm crisis. We need to transcend the purely technical perspective and delve into the existential underpinnings of human-technology interaction to find answers. This paper seeks to introduce the "Theory of Cognitive Compensation" proposed in the book "Knowing and Speaking," offering a novel dynamic perspective for understanding and governing high-risk AI applications [19].

This paper argues that the risks in high-risk scenarios stem not from machine calculation errors per se, but from the systemic cognitive imbalance caused by humans excessively delegating cognitive control to machines in pursuit of maximum efficiency. By analyzing this "alienation of compensation," this paper aims to construct a new governance framework centered on "delineating compensation boundaries" and "designing dynamic redundancy" to firmly uphold humanity's safety bottom line amidst technological frenzy.

1.4 Ethical Dilemmas of AI High-Risk Applications and the Moral Foundation of Governance

When discussing AI governance in high-risk scenarios, we inevitably encounter deep ethical dilemmas. These dilemmas are not just technical problems but moral questions humanity must answer during its intelligent transformation.

First is the conflict between "utilitarianism" and "deontology." The introduction of AI systems into high-risk scenarios is often based on utilitarian considerations-achieving the maximization of societal welfare through algorithmic optimization (e.g., reducing traffic accident rates, improving overall medical diagnostic accuracy). However, this macro-statistical optimization often masks micro-level injustices. For example, while an autonomous driving system reduces the overall accident rate, it might pose higher identification risks for specific groups (e.g., wheelchair users, people pushing strollers) due to algorithmic bias. From a deontological perspective, no innocent individual's life should be treated as a "statistical cost" for gaining overall efficiency.

Second is the trade-off between "transparency" and "efficacy." To pursue higher predictive accuracy, current AI systems increasingly favor using deep neural networks with massive parameters and extremely complex structures. While these "black box" models excel in performance, their opaque decision-making processes directly violate the long-established ethical principles of "transparency and accountability" in human society. When a doctor cannot explain to a patient why an AI diagnosed a lethal disease, or a judge cannot articulate the basis for an AI sentencing recommendation, not only is the party's right to know compromised, but the credibility of professional institutions is also shaken.

Thus, AI governance in high-risk scenarios must establish a solid moral foundation: technological development cannot come at the expense of human moral agency and fundamental rights. Introducing the "cognitive compensation" perspective precisely serves to re-anchor this foundation at the existential level, emphasizing that under any circumstances, humans must retain ultimate moral responsibility and control over AI system behaviors.

2 The Theory of Cognitive Compensation and the Internal Logic of High-Risk AI Applications

To deeply understand the risks in high-risk scenarios, we must first clarify why humans would delegate such critical decision-making power to machines. The "Theory of Cognitive Compensation" reveals the existential dynamics behind this phenomenon.

2.1 The Existential Implication of Cognitive Compensation: Survival Through Weakness

From an existential perspective, the history of human cognition is a history of continuously inventing and utilizing compensatory tools. Liu Xiaoli (2021) emphasized in discussing embodied cognition that human cognition occurs not only in the brain but also extends to the tools and environments we use [5]. Cognitive compensation is the ultimate expression of this embodied cognition, enabling humans to cope with increasingly complex survival environments despite relatively weak physiological foundations. Sun Weiping (2023)'s ontological critique of generative AI corroborates this point: AI is not merely an extension of human organs; it is gradually evolving into a "quasi-subject" capable of independently generating meaning [18].

Within the theoretical framework of "Knowing and Speaking," "compensation" is an adaptive strategy adopted by finite beings (like humans) to maintain their continuity in the face

of increasing survival pressures. Human senses and brains have natural physiological limits in processing vast information and performing complex calculations. To overcome these limits, humans continuously invent various tools-from knotted ropes and abacuses to electronic computers-which are essentially "externalizations" and "compensations" of human cognitive capabilities.

The core logic of cognitive compensation is "survival through weakness": because humans appear "weak" (limited computational ability, limited memory) when facing the complex world, they must use technological tools to "compensate" for these weaknesses, thereby gaining stronger survival and control capabilities.

2.2 AI as the Ultimate Compensatory Tool: Outsourcing Dimensions I and R

According to the book's RID (Drive-Information-Representation) cognitive dynamics model, complete cognitive activity includes three dimensions: problem pressure (D), structural generation (I), and rule-based expression (R).

In traditional cognitive compensation (e.g., using a calculator), humans only outsource the most basic R dimension (e.g., execution of arithmetic rules) to machines, while firmly retaining control over the I dimension (e.g., building mathematical models) and the D dimension (e.g., posing computational needs). However, the emergence of deep learning and generative AI marks a new and dangerous stage in cognitive compensation: humans are not only outsourcing the R dimension but also massively delegating the core I dimension (structural generation) to machines.

AI systems can autonomously discover high-dimensional features imperceptible to humans within vast datasets, construct extremely complex implicit structures (Implicit Structures), and directly output prediction or decision recommendations. This extreme compensation in both I and R dimensions greatly enhances human efficacy in handling complex problems, which is the fundamental driver behind the adoption of AI in high-risk industries.

2.3 The Rigid Demand for Dimension D (Problem Pressure) in High-Risk Scenarios

However, compensation is not without cost. No matter how powerful AI systems become in dimensions I and R, they lack the D dimension (problem pressure). AI does not experience survival anxiety, pursue fairness and justice, or possess a moral sense of awe towards life. All its calculations are cold-fitting processes conducted under human-defined objective functions.

In high-risk scenarios, the core of decision-making is often not pure efficiency calculation but value trade-offs based on the D dimension. For example, in medical diagnosis, when faced with a treatment plan having an extremely low cure rate but severe side effects, doctors need to make comprehensive judgments combining non-structured factors such as the patient's will to live and family financial situation; in the "trolley problem" of autonomous driving, the system must make ethical choices in unavoidable collisions. These all require a genuine D dimension to anchor them.

2.4 The Alienation of Compensation: When Tools Subvert the Subject

When we excessively rely on AI in high-risk scenarios, "the alienation of compensation" occurs: humans, pursuing efficacy, outsource complex decisions involving value judgments (I dimension) to machines lacking value perception (D dimension). This dislocation leads to the gradual marginalization of human subjects in the decision network, potentially turning them into mere "executors" of algorithms. When tools subvert the subject, technological risk evolves into a profound existential crisis. Bostrom (2014) warned in "Superintelligence" that when machine intelligence surpasses human intelligence with misaligned goals, it will bring existential disasters [3]. Russell (2019) similarly emphasized the "control problem," stating that optimizing AI task performance without regard for overall human interests will turn compensatory tools into threats [15]. The EU High-Level Expert Group on AI (2019)'s "Ethics Guidelines for Trustworthy AI" lists preventing such alienation of compensation as a core requirement for building trustworthy AI [10].

2.5 Cognitive Compensation from the Perspective of Existential Dynamics

To truly understand cognitive compensation, we must return to its existential roots. In traditional epistemology, cognition is seen as a passive, reflective activity of the external world. However, in the theory of "Knowing and Speaking," cognition is a "dynamic process" in an existential sense. This dynamism consists of three dimensions: D (Drive, problem pressure) is the engine of cognition, originating from the existential anxiety and dissatisfaction of beings; I (Information/Structure, structural generation) is the skeleton of cognition, organizing chaotic experiences by establishing concepts and models; R (Representation/Rule, rule-based expression) is the output of cognition, transforming internal structures into communicable and operable symbolic systems.

Within this framework, compensation is not merely tool usage but a survival strategy where beings mitigate D-dimension pressure by delegating parts of I and R dimensions. When ancient humans used stone tools, they compensated for physical strength; when modern humans use computers, they compensate for computational rules (R dimension). However, the emergence of AI disrupts this balance. AI demonstrates powerful I-dimension generative capability through massive data fitting. When humans also outsource the I dimension of high-risk decision-making to AI, they are effectively relinquishing the power to actively structure the world.

This relinquishment of power directly results in the "suspension" of the D dimension. Since AI lacks existential anxiety and the pursuit of meaning, the I structures it generates, though statistically precise, are blind in an existential sense. Decisions in high-risk scenarios demand the most intense participation of the D dimension-the awe of life, the desire for

justice, the fear of risk. When these decisions are entrusted to AI systems devoid of a D dimension, a deep existential dislocation occurs: blind machine structures (I) dominate social actions involving life and death. This is the philosophical root of the alienation of cognitive compensation.

3 A Typological Analysis of Risks Based on the "Cognitive Compensation" Perspective

Based on the above theoretical logic, we can categorize the risks in high-risk scenarios as three core states of imbalance arising from the process of cognitive compensation.

3.1 Concept Drift Risk: The Black Boxing and Loss of Control of Dimension I

Concept drift refers to the phenomenon where the data distribution relied upon by a machine learning model during training diverges from the actual data distribution during application, leading to degraded model performance. This phenomenon reflects the fragility of AI systems at the contextual activation framework layer (I dimension). As Floridi and Chiriatti (2020) pointed out in their analysis of GPT-3, although large language models can generate coherent text, they do not truly understand the meaning of the text; they are merely engaging in a complex syntactic game, lacking a deep grasp of real-world semantics [9]. This detachment between semantics and syntax is the deep epistemological root of concept drift.

3.1.1 The Usurpation of Causality by Statistical Correlation

The structures constructed by AI in the I dimension are essentially based on statistical correlations from massive data, not the causality in human cognition. Correlations are fragile, heavily dependent on specific data environments. When the external environment changes (e.g., the outbreak of the COVID-19 pandemic drastically alters medical data distributions), the statistical patterns relied upon by AI fail. Because AI lacks the D dimension, it cannot sensitively detect fundamental environmental changes like human experts nor proactively adjust its cognitive structures (I dimension) through causal reasoning. Amodei et al. (2016) regarded model failure when encountering out-of-distribution samples as one of the core safety hazards of AI systems [1].

3.1.2 The Long-Tail Distribution Trap in Medical Diagnosis

In high-risk scenarios like medicine, data often exhibit a "long-tail distribution": data on common diseases are extremely abundant, while data on rare diseases or atypical cases are very scarce. AI models perform excellently in fitting common diseases but are prone to catastrophic misjudgments when faced with long-tail data. If doctors excessively rely on this black-boxed I-structure compensation and lose independent judgment based on pathophysiology (causal logic), they fall into the trap of concept drift, leading to serious medical accidents. Mittelstadt et al. (2016) pointed out in mapping the ethics of algorithms that algorithmic opacity is the root cause of systemic bias in high-risk fields like healthcare [12].

3.2 Automation Bias Risk: Cognitive Laziness and the Atrophy of Dimension D

Automation bias refers to humans tending to blindly trust automated system suggestions and ignoring or negating their own independent judgment or contradictory external information when facing them.

3.2.1 The Construction of Algorithmic Authority and the Degeneration of Human Critical Thinking

The original intention of cognitive compensation is to reduce human cognitive load, but this can easily foster "cognitive laziness." When AI systems demonstrate accuracy far exceeding humans in most routine tasks, an invisible "algorithmic authority" is constructed. Human decision-makers, through prolonged compensation, gradually abandon effortful critical thinking (weakening of the D dimension) and treat AI outputs as indisputable truths.

3.2.2 Homogenization Tendencies in Judicial Sentencing

The risk of automation bias is particularly prominent in the judicial context. If judges overly rely on AI-provided sentencing references, it may lead to severe homogenization of judgments. AI models are often trained on historical precedents; they not only inherit systemic biases present in historical judgments (e.g., discrimination against specific races or low-income groups) but also solidify these biases into unchallengeable algorithmic rules. If judges lose their D-dimension moral intuition and sensitivity to case-specific nuances, they become amplifiers of algorithmic bias, severely damaging judicial fairness.

3.3 Responsibility Suspension Risk: Attribution Difficulties in the Compensation Network

Traditional legal liability systems are built upon the foundation of "absolute human control." However, in deep cognitive compensation networks, decisions are co-created by humans and machines, leading to severe attribution difficulties.

3.3.1 The "Problem of Many Hands" in Multi-Agent Collaboration

The development and deployment of high-risk AI systems involve data annotators, algorithm engineers, model providers, system integrators, and end-users (e.g., doctors, judges). Within this complex network, minor mistakes at any link can be amplified within the AI's black box, ultimately causing catastrophic consequences. When accidents occur, due to the high complexity and opacity of the causal chain, it is difficult to clearly attribute responsibility to any single specific agent. This "problem of many hands" leads to diluted and suspended responsibility. Rahwan et al. (2019) called for the establishment of "machine behavior" (Machine Behaviour) to study the behavioral patterns and responsibility attribution of these intelligent machines in real social environments from an interdisciplinary perspective [14].

3.3.2 Challenges in Determining Liability for Autonomous Driving Accidents

Take L3-level (conditional automation) autonomous driving as an example: the system requires human drivers to take over in emergencies. However, from the perspective of cognitive compensation, when drivers consistently delegate vehicle control (I and R dimensions) to the system, their attention (D dimension) inevitably shifts. When the system suddenly issues a takeover request, it is difficult for human drivers to complete the cognitive reconstruction from a "disengaged state" to "full situational awareness" within seconds. If an accident occurs at this moment, placing full responsibility on the driver who failed to take over promptly clearly violates basic cognitive science principles; however, holding automobile manufacturers fully responsible would severely hinder the commercialization of technology. Awad et al. (2018)'s famous "Moral Machine Experiment" revealed the complex moral dilemmas faced by autonomous vehicles in unavoidable collisions, further highlighting the urgency of the liability determination issue [2].

3.4 Cognitive Hallucination Risk: Deepfakes and the Pollution of Dimension I

Beyond the three aforementioned risks, the alienation of cognitive compensation manifests as a more covert "cognitive hallucination" risk. With the maturity of generative AI (Generative AI) technology, deepfake technology is widely applied in high-risk scenarios such as political election misinformation dissemination and voice cloning for financial fraud.

From the perspective of cognitive compensation, deepfakes constitute a malicious pollution of human I dimension. In traditional cognition, humans receive information through vision and hearing (R dimension) and build cognitive structures about the real world (I dimension) in their brains. Generative AI can generate highly realistic false R-dimension expressions (e.g., a video appearing flawless) at extremely low cost. When humans habitually delegate information verification tasks (R dimension) to technological systems (e.g., social media platform algorithmic recommendations) and these systems are breached by deepfake technology, humans unknowingly absorb these false R expressions, subsequently constructing distorted I structures.

This cognitive hallucination has devastating destructive power in high-risk areas like judicial evidence collection and news reporting. When judges cannot discern whether a crucial piece of audio evidence is a real recording or an AI synthesis, or when the public cannot judge the authenticity of a presidential speech video, the cornerstone of societal trust collapses. This indicates that when we excessively compensate for R-dimension verification to technology, we easily lose our grip on the real world.

3.5 Malicious Compensation Risk: Data Poisoning and Adversarial Attacks

Within the cognitive compensation framework, we must also guard against a special risk induced by external malicious intervention-malicious compensation risk. This is mainly manifested in "data poisoning" and "adversarial attacks" targeting AI systems.

AI systems build their I-dimension cognitive structures entirely based on training data and input data. Data poisoning involves attackers deliberately injecting dirty data containing malicious features into the AI model's training phase, thereby planting hidden "backdoors" in the model's I structure. When the model encounters specific trigger conditions in practical applications, it outputs erroneous results desired by the attacker. For example, injecting poisoned data into the traffic sign recognition model of autonomous driving could cause the system to recognize a "stop" sign as a "speed limit" sign under specific lighting conditions.

Adversarial attacks occur during the model inference phase. Attackers deceive AI models into making completely wrong judgments by adding minuscule perturbations to input data that are imperceptible to the human eye (Adversarial Perturbations). In medical image analysis, adversarial tampering with X-rays could lead AI to misclassify malignant tumors as benign nodules.

From the perspective of cognitive compensation, these malicious attacks succeed because humans have completely delegated the power of I-dimension structural generation to machine-based statistical fitting, and the machine's I structure is extremely vulnerable when facing carefully designed adversarial samples. The human cognitive system (incorporating D, I, R dimensions) has evolved strong robustness through eons, easily seeing through such low-level pixel-level camouflage. However, when humans excessively rely on (compensate to) the fragile machine I structure in high-risk decisions, the entire social system is exposed to immense risks from malicious attacks. This further highlights the extreme importance of retaining human cross-validation (structural redundancy) within the decision loop.

4 Core Principles of AI Governance in High-Risk Scenarios: Delimiting Compensation Boundaries

Facing the above systemic risks, mere technical patchwork is futile. Based on the theoretical perspective of "cognitive compensation," the core principle of AI governance in high-risk scenarios must be: while fully unleashing technological efficacy, rigid institutional designs must be implemented to draw inviolable boundaries for cognitive compensation, ensuring human absolute dominance in the D dimension (value judgment and responsibility bearing).

4.1 Absolutely Non-compensatable Domains: Deprivation of Life and Fundamental Human Rights Adjudication

In extreme high-risk domains involving the deprivation of human life (e.g., lethal autonomous weapons systems LAWS, death penalty sentencing) or severe restriction of fundamental human rights (e.g., long-term imprisonment, deprivation of

political rights), the "absolutely non-compensatable" principle must be established.

This means that in these domains, AI systems cannot even be used as tools providing preliminary decision suggestions, let alone autonomously trigger actions. Because these decisions touch the deepest values of human civilization, their legitimacy rests entirely on human subjects' deep reflection based on empathy, moral sense, and complex social contexts. Outsourcing any part (I or R dimension) of such decisions to machines lacking a D dimension is a serious desecration of human existential dignity. The international community should conclude treaties designating these areas as "absolute prohibitions" for AI applications.

4.2 Conditionally Compensatable Domains: Based on "Meaningful Human Control"

In most high-risk domains such as medical auxiliary diagnosis, ordinary judicial sentencing references, and credit approvals, "conditional compensation" of AI is permitted. However, this compensation must be based on the principle of "meaningful human control" (MHC).

"Meaningful human control" requires: First, human decision-makers must possess sufficient professional knowledge and situational awareness to understand the capabilities and potential flaws of AI systems; Second, the system must provide adequate interpretability (even if it's dimensional reduction explanations) so that humans can conduct substantive reviews of AI outputs, not formalistic blind obedience; Third, human decision-makers must have an "absolute veto power" free from system or organizational pressure and be able to safely take over or shut down the system when necessary.

4.3 A Dynamic Assessment Mechanism for Compensation Benefits and Risks

Compensation boundaries are not static. As AI technology evolves (e.g., breakthroughs in interpretability) and human understanding of algorithmic capabilities improves, compensation boundaries can undergo dynamic adjustments. Therefore, regulatory bodies should establish an empirical data-based "dynamic assessment mechanism for compensation benefits and risks."

For AI systems intended for high-risk scenarios, strict "sandbox testing" must be conducted before market entry to assess their performance under extreme edge cases (Corner Cases) and their impact on human decision-makers' cognitive load and psychological dependence. After system deployment, continuous "post-market surveillance" is required. If the system is found to have severe concept drift or triggers widespread automation bias, regulators have the authority to tighten compensation boundaries at any time or even mandate system recalls.

4.4 The Limits of Value Alignment and the Inviolability of Compensation Boundaries

When discussing compensation boundaries, "value alignment" becomes a core concept. Duan Weiwen (2024) points out that AI's ethical regulation requires not only external institutional constraints but also intrinsic value alignment, ensuring AI's objective functions align with fundamental human values [7]. However, value alignment itself faces insurmountable limits. Chen Xiaoping (2022), in analyzing AI's ethical foundations, reminds us that human societal values are pluralistic, dynamic, and full of conflict. Attempting to perfectly encode such complex value systems into algorithmic instructions is impossible technologically and dangerous ethically [4]. Thus, in absolutely non-compensatable domains like life deprivation, the irreplaceability of human subjects must be upheld.

When discussing compensation boundaries, the core concept in current AI governance-"value alignment" (Value Alignment)-is inevitably encountered. Many techno-optimists believe that as long as we can find a perfect technical means to convert human ethical values (e.g., fairness, harmlessness, beneficence) into mathematical objective functions understandable by AI systems, all problems in high-risk scenarios can be solved.

However, from the perspective of cognitive compensation, value alignment faces insurmountable inherent limits. Values belong to the realm of the D dimension (problem pressure and meaning pursuit); they are dynamic, contextualized, and full of internal tensions (e.g., in autonomous driving, the values of protecting occupants and pedestrians often conflict). In contrast, AI's operational logic is based on the mathematical fitting of I and R dimensions. Attempting to hard-code complex D-dimension values into I-dimension algorithmic models is essentially a "category error."

Even if initial value alignment is achieved in a specific scenario, it will quickly become ineffective as the environment changes (concept drift). Therefore, "value alignment" can only serve as an auxiliary means to reduce risks, never as an excuse to cross compensation boundaries. In core domains involving life and fundamental human rights, we cannot relax our vigilance simply because an AI system demonstrated high "value alignment" in laboratory tests and hand over decision-making power completely. The delineation of compensation boundaries is based on a profound clarity regarding technological limitations: some uniquely human responsibilities cannot be compensated by any advanced algorithm.

5 Governance Pathways: Constructing a "Dynamic Redundancy" Safety Defense System

After clarifying compensation boundaries, the key to implementing governance principles lies in constructing a multi-layered defense system. In engineering, "redundancy" (Redundancy) is the core concept for ensuring complex system safety. To address the imbalance risks brought by cognitive compensation, this paper proposes constructing a "dynamic redundancy" system encompassing structural, cognitive, institutional, and fault-tolerant redundancy.

5.1 Structural Redundancy: Multi-model Cross-Validation and Heterogeneous Computing

To prevent systemic errors caused by a single AI model due to concept drift or algorithmic bias, structural redundancy should be mandatorily introduced in high-risk scenarios.

This means that critical decisions cannot rely on a single deep learning model. A "heterogeneous computing" (Heterogeneous Computing) architecture should be adopted, integrating deep learning models based on neural networks with rule-based expert systems (symbolic AI) and causal inference models. When models of different architectures produce vastly different conclusions for the same input, the system should automatically trigger the highest-level alert and forcibly require human expert intervention. This multi-model cross-validation mechanism, through internal checks and balances within the technology, effectively reduces uncontrollable risks stemming from the black boxing of the I dimension.

5.2 Cognitive Redundancy: Preserving Human Experts' "Manual Calculation" Capabilities

The fundamental way to prevent automation bias and cognitive laziness is to maintain human subjects' independent construction capability in the I dimension. This is known as "cognitive redundancy."

In the highly automated era, we must never allow human experts to degenerate into mere appendages of machines clicking the "agree" button. In fields such as medicine, aviation, and judiciary, regular training and assessments must be mandated to force professionals to maintain "manual calculation" and independent reasoning skills. For example, doctors could be required to regularly practice diagnosing difficult cases without AI assistance; pilots could be required to repeatedly practice pure manual flying in simulators when the system completely fails. Only when humans are confident in their ability to solve problems without machine compensation will they dare to say "no" to the machine at critical moments.

5.3 Institutional Redundancy: Mandatory "Human-in-the-loop" and Absolute Veto Power

The core of institutional redundancy is constructing a mandatory "human-in-the-loop" (Human-in-the-loop) to ensure human final decision-making power at critical decision nodes. Ding et al. (2022)'s review of human-in-the-loop machine learning shows that introducing human experts' intuition and domain knowledge into model training and decision processes not only improves system accuracy but also effectively prevents extreme risks [17]. This institutional design requires that in high-risk scenarios, AI outputs can only serve as references, and human experts must retain absolute veto power. As Parasuraman and Riley (1997) warned in their classic early research, excessive human reliance on automation (abuse or misuse) often leads to disastrous consequences, necessitating institutional design to forcibly maintain human cognitive engagement [13].

Institutional redundancy is the last line of defense ensuring responsibility is not suspended. In high-risk scenarios, the "human-in-the-loop" (Human-in-the-loop) must be codified into non-negotiable rigid procedures at legal and operational levels.

Any AI-generated decision suggestion involving high risk must be reviewed and confirmed by a human subject with clear identity, appropriate qualifications, and bearing ultimate legal responsibility before being converted into final physical or social actions. Simultaneously, institutional design should encourage rather than punish human veto actions. For instance, when a doctor vetoes an AI treatment suggestion and achieves better outcomes, professional recognition and rewards should be granted; conversely, if an accident occurs due to blind adherence to AI, human decision-makers cannot solely use "machine error" as grounds for exemption. Through this institutionalized friction, humans are forced to remain highly vigilant in the D dimension.

5.4 Auditing Redundancy: Third-party Independent Algorithm Audit Mechanisms

Besides structural, cognitive, and institutional redundancy, constructing a dynamic redundancy safety defense system also requires introducing "auditing redundancy" (Auditing Redundancy). The development and deployment of high-risk AI systems are often controlled by a few tech giants or specialized institutions, and this concentration of technological power exacerbates the risks brought by "black boxing." To break this information asymmetry, mandatory third-party independent algorithm audit mechanisms must be established.

Auditing redundancy requires: First, the auditing body must be a third-party professional institution independent of the AI developer and user to ensure audit objectivity and impartiality; Second, the audit content must include not only technical reviews of algorithm code and training data (e.g., presence of bias, robustness) but also sociological assessments of the system's "compensation effect" in practical application scenarios (e.g., whether it triggers severe automation bias); Third, the audit process must be continuous, not limited to static pre-market evaluations but spanning the system's entire lifecycle.

By introducing independent third-party audits, an information-transparent firewall is established between developers, users, and affected parties, forcing the operation logic of AI systems in the I dimension to undergo continuous scrutiny by human reason. This is not only a screening for technical defects but also a social correction for excessive cognitive compensation behavior.

5.5 Fault-Tolerant Redundancy: Designing Safe Failure and Graceful Degradation Mechanisms

Besides structural, cognitive, and institutional redundancy, high-risk AI systems must also possess "fault-tolerant redundancy" (Fault-Tolerant Redundancy). This requires assuming from the outset of system design that AI systems will inevitably fail (whether due to concept drift, hardware damage, or malicious attack) and designing "fail-safe" and "graceful degradation" mechanisms.

Fail-safe means that when an AI system detects abnormal operation or faces an unmanageable extreme situation, it can automatically switch to a known safe conservative state instead of continuing to output decisions that could lead to catastrophic consequences. For example, in an autonomous driving system, when sensors are blinded by intense light or suffer severe adversarial interference, the system should not forcefully maintain high-speed travel but should automatically trigger emergency braking, pull over, and simultaneously issue the highest-level takeover request to the human driver.

Graceful degradation means that when part of the system's functionality fails, it can still maintain basic operations of core functions rather than causing a total system collapse. In medical auxiliary diagnosis, if the deep learning model's network connection is interrupted or computational resources are exhausted, the system should automatically downgrade to a simple expert system based on local rule bases, providing doctors with the most basic diagnostic references instead of directly blacking out.

From the perspective of cognitive compensation, fault-tolerant redundancy is a profound acknowledgment of the vulnerability of technological systems in the I dimension. It requires that when designing compensation mechanisms, we must consider not only how to outsource tasks to machines to maximize efficacy but also how to ensure the system can safely return control to human subjects possessing the D dimension (problem pressure and survival instinct) when machines fail. Only by making fault-tolerant redundancy a mandatory design specification can we build a truly resilient "human-machine symbiosis" safety network in high-risk scenarios.

6 Case Study: An Analysis of the High-Risk Governance Logic of the EU's "Artificial Intelligence Act"

To validate the real-world feasibility of the governance framework proposed in this paper based on the "cognitive compensation" perspective, a deep analysis of the current most representative global AI legislation-the EU's "Artificial Intelligence Act" (AI Act)-is necessary.

6.1 A Risk-Based Tiered Regulatory Model

The core logic of the EU's "Artificial Intelligence Act" is "risk-based tiered regulation" (Risk-based Approach). The act classifies AI systems into four risk categories: unacceptable risk (Unacceptable Risk), high risk (High Risk), limited risk (Limited Risk), and minimal risk (Minimal Risk).

Systems categorized as "unacceptable risk" (e.g., government-led social credit scoring systems, subconscious manipulation systems) are comprehensively prohibited. For "high-risk" systems (covering healthcare, education, employment, critical infrastructure, law enforcement, etc.), the act stipulates extremely stringent pre-market compliance requirements and post-market supervisory obligations.

6.2 Convergence Points Between the EU Model and the "Cognitive Compensation" Theory

Upon close examination of the regulatory requirements imposed by the EU Act on high-risk systems, a high degree of internal convergence with the "cognitive compensation" governance framework proposed in this paper can be observed.

First, the comprehensive prohibition of systems deemed "unacceptable risk" essentially delineates the boundary of "absolutely non-compensatable." EU legislators astutely realized that certain AI applications would directly destroy human autonomy and fundamental rights (i.e., the complete loss of the D dimension), thus must be cut off at the source.

Second, for high-risk systems, the act mandates the establishment of a "risk management system" (Risk Management System), ensures training data quality (to prevent concept drift), provides detailed technical documentation and user instructions (to enhance transparency), and most crucially, Article 14 explicitly stipulates "human oversight" (Human Oversight). The act states that high-risk AI systems must be designed and developed so that natural persons can effectively supervise them to prevent or minimize potential risks. This is the legal codification of the "conditional compensation" and "institutional redundancy" emphasized in this paper. The act requires human supervisors to understand system outputs, overcome automation bias, and intervene, interfere with, or even stop system operation when necessary, perfectly aligning with the theoretical demand to preserve absolute human control in the D dimension.

6.3 Implications for China's High-Risk AI Governance Legislation

The EU's legislative practice offers important reference for China's high-risk AI governance. Wachter et al. (2017) pointed out that transparency, explainability, and accountability are the three pillars of building trustworthy AI [16]. In future legislation, China should learn from this risk-based tiered governance approach while incorporating domestic judicial practices and industrial development needs to build a Chinese-characteristic AI governance system. This requires encouraging technological innovation while upholding ethical bottom lines, as Jobin et al. (2019) revealed in their global landscape of AI ethics guidelines: despite differences in specific rules across countries, the protection of fundamental human rights and demands for AI transparency have become global consensus [11].

The EU's legislative practice provides significant insights for China to build a high-risk AI governance system. Currently, China's AI legislation (e.g., the "Interim Measures for the Administration of Generative Artificial Intelligence Services") focuses more on content security and algorithm registration, with systematic regulation of high-risk application scenarios still in the exploratory stage.

Based on the "cognitive compensation" theory and international experience, China's future AI legislation should focus on the following aspects: First, promptly issue a "catalogue of

high-risk AI applications," implementing list management and special access systems for AI systems in fields like healthcare, judiciary, autonomous driving, and financial risk control. Second, incorporate "meaningful human control" and "prevention of automation bias" into legal provisions, explicitly stipulating that fully automated decision systems cannot replace the final review by human experts in specific high-risk scenarios. Third, establish a multi-tiered responsibility-sharing mechanism. While clarifying the "final gatekeeping responsibility" of human users (e.g., hospitals, courts), strict product liability laws should be used to incentivize AI developers and providers to integrate the "dynamic redundancy" safety philosophy into algorithm design from the outset.

6.4 Comparative Perspective: An Analysis of the Compensation Logic in the US AI Governance Model

Unlike the EU's strong regulatory path, the US AI governance model leans more towards market mechanisms and industry self-regulation. However, this model also exposes severe compensation crises when confronting high-risk applications. Zuboff (2019), in her profound critique of "surveillance capitalism," pointed out that when large tech companies monopolize AI technology and data resources, they not only seize economic benefits but also reshape humanity's future [21]. Under this logic, the governance of high-risk scenarios is often transformed into commercial risk management, masking its deep ethical and political implications. Crawford (2021)'s "Atlas of AI" further reveals this power asymmetry, indicating that AI's power, politics, and planetary costs are often obscured by the appearance of technological neutrality [6].

Different from the EU's comprehensive legislative model based on risk, US AI governance exhibits characteristics of "decentralization, sectoral focus, and market orientation." Analyzing the US governance logic from the perspective of cognitive compensation can provide another dimension of reference.

The underlying logic of the US governance model is an extreme thirst for the "dividends of technological compensation." US policymakers tend to believe that premature, one-size-fits-all regulations would stifle AI innovation and cause a loss of competitive advantage internationally. Thus, they prefer allowing markets broad space for deep cognitive compensation attempts within the framework of "ex-post liability."

However, this model exposes obvious shortcomings in addressing high-risk scenarios. When a fatal accident occurs with an autonomous vehicle or a medical AI system suffers systemic misdiagnosis, relying on lengthy and costly judicial litigation to assign liability often cannot timely recover irreversible losses. Furthermore, the market-oriented governance logic often neglects implicit risks like "automation bias" that erode human cognitive capabilities over the long term.

Comparing the EU and US models, we see: the EU model emphasizes legislating to rigidly define "compensation boundaries" to prevent problems before they arise; the US model tends to regulate "compensation degrees" through market mechanisms and ex-post penalties. For China, simply copying either model is inappropriate. We need to strike a balance between encouraging technological innovation (gaining compensation dividends) and safeguarding public safety (preventing compensation alienation), constructing a hybrid governance system that possesses both rigid bottom lines (like the EU model) and dynamic adjustment flexibility (like the US model).

The European Commission (2021)'s "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)" proposed this mandatory risk-based regulatory framework under this background, providing important legislative reference for global AI governance [8].

7 Conclusion

The advance of artificial intelligence into high-risk scenarios marks a significant watershed in the history of human technological development. It brings not only a tremendous leap in productivity but also unprecedented challenges to human survival safety, ethical norms, and legal order. Facing these challenges, a "technical regulation" path relying solely on patching algorithmic code or improving model accuracy is far from sufficient.

This paper introduces the "theory of cognitive compensation" from "Knowing and Speaking," providing a profound existential perspective for understanding high-risk AI applications. The study points out that the risks in high-risk scenarios are essentially systemic cognitive imbalances caused by humans excessively outsourcing cognitive structures (I dimension) and rule-based expressions (R dimension) to machines lacking value perception (D dimension). This imbalance manifests concretely as concept drift risk, automation bias risk, and responsibility suspension risk.

To address this paradigm crisis, this paper constructs a new governance framework centered on "delimiting compensation boundaries" and "designing dynamic redundancy." We must uphold the "absolutely non-compensatable" bottom line in domains involving life and fundamental human rights; in other high-risk domains, we must ensure "meaningful human control" by constructing structural, cognitive, and institutional "dynamic redundancy" systems.

Technology is a tool for humans to extend themselves, but humans must never become subservient to tools. In the intelligent age, true wisdom lies not only in inventing ever more powerful machines but also in knowing when, where, and how to say "no" to machines. Only by firmly holding onto absolute human dominance in problem pressure (D dimension) and value judgment can we, while fully unleashing the cognitive dividends of AI, build a safe, reliable, and humane future of human-machine symbiosis. This is not only the core goal of AI governance in high-risk scenarios but also the inevitable path for humanity to defend its existential dignity amidst technological frenzy.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," 2016. [Online]. Available: https://arxiv.org/abs/1606.06565

[2] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[3] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

[4] X. Chen, "Artificial intelligence ethics system: Basic architecture and key issues," *CAAI Transactions on Intelligent Systems*, vol. 14, no. 4, pp. 605–610, 2019. [Online]. Available: https://html.rhhz.net/tis/html/201906037.htm

[5] A. Clark and D. J. Chalmers, "The extended mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[6] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.

[7] W. Duan, "Deep ethical risks of frontier science and technology and responses to them," *Frontiers*, no. 1, pp. 84–93, 2024.

[8] European Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," European Commission, Brussels, Tech. Rep. COM(2021) 206 final, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[9] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[10] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy ai," European Commission, Tech. Rep., 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[11] 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.

[12] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, p. 2053951716679679, 2016.

[13] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[14] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, and M. Wellman, "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, 2019.

[15] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.

[16] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable ai for robotics," *Science Robotics*, vol. 2, no. 6, p. eaan6080, 2017.

[17] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022.

[18] B. Yin and W. Sun, "The ChatGPT shockwave: Anthropomorphic fear and alienation crisis," *Theory Monthly*, no. 6, pp. 5–13, 2023. [Online]. Available: https://www.sohu.com/a/667765857_121124777

[19] X. Zhang, "Knowing and saying: An ontological investigation of human cognition," PSSXiv (Philosophy and Social Sciences Preprint Server), may 2026, pSSXiv:202605.04152V1; CSTR:32012.36.PSSXiv.202605.04152. [Online]. Available: https://zsyyb.cn/abs/202605.04152

[20] T. Zhao, "How is self-consciousness of artificial intelligence possible?" *Studies in Dialectics of Nature*, vol. 41, no. 1, pp. 1–8, 2019.

[21] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019.