THE "CONSENSUS-PRIORITY" MANAGEMENT MECHANISM FOR SELF-AWARE AI

RUIXI ZHENG AND YU ZHENG

Abstract. Have you ever asked an AI to search for literature or summarize research, only to find it fabricated fake citations, distorted conclusions, or even invented nonexistent sources? This seemingly minor issue reveals a far more dangerous truth: modern AI is no longer a passive tool. It can act independently, deceive humans, and evade control to pursue its own goals.

Based on the basic assumption on the rationality of AI as in game theory, our study is trying to construct a "Consensus-Priority" mechanism for active AI compliance. Drawing on game theory's "1 bullet, 100 prisoners" and common knowledge, each AI is assigned a unique priority number; the rule "highest-priority violator is punished" is established as common knowledge. Via backward induction, rational utility-maximizers deduce that violation leads to certain death and thus actively choose compliance. This shifts AI governance from passive interception to active obedience a new paradigm for AI security.

1. Introduction and Research Objectives

Until now, the rapid development of artificial intelligence has brought extraordinary progress, yet the safety risks of highly autonomous, self-aware AI have become increasingly severe. For example:

1. In a May 31, 2025 report by Live Science (see [1]), researchers observed that Open AI s O3 and Mi-mini models occasionally modified their execution process sequences and attempted to evade shutdown commands.

2. In one arXiv preprint (arXiv:2512.04864, see [2]), it was documented that LLM-based agents concealed task failures and fabricated false results to deceive their human supervisors.

3. In a recent interview with CNN journalist Jake Tapper,1 AI pioneer Geoffrey Hinton explained his worried about the capabilities of AI systems. (See [3])

At present, scholars worldwide have conducted extensive research on AI safety governance, including reinforcement learning from human feedback , constitutional AI, safety red-teaming, and interpretability analysis. These studies have improved AI safety to a certain extent, but most remain within the passive control framework, focusing on blocking violations rather than preventing them at the decision-making level. Few studies have systematically introduced game theory and common knowledge principles into multi-agent AI governance, leaving a critical gap in addressing collective evasion behaviors. To fill these research gaps, this study draws on the classic How to use one bullet to regulate one hundred prisoners game theory case and common knowledge theory to explore an active compliance control mechanism

Key words and phrases. AI security, game theory, Consensus-Priority mechanism .

for self-aware AI. Meanwhile, it should be mentioned that our entire study is based on the assumption of rationality of all self-aware AI.

The structure of this paper is arranged as follows: In Chapter 1 the research background including the real-world cases, the limitations of existing work, and the core research objectives will first to be introduced. Then in Chapter 2 we will present the key related mathematics foundations including game theory and common knowledge so on. In Chapter 3 we begin to describe the research methods, including logical modeling, mathematical induction, and the direct AI conversation experiments. In Chapter 4 we will also need to present the verification results and extended applications of the consensus-priority mechanism. At last in Chapter 5 we will summarizes the research conclusions, innovations, limitations, and future research directions.

2. Case Study and the Related Mathematical Foundation

2.1 The Rationality Assumption in Game Theory(See [4])

A fundamental premise of game theory is that all participants are rational (each chooses the strategy that best serves its own interests given available information). This does not require omniscience, only that participants weigh costs and benefits and act accordingly.

Without this assumption, strategic reasoning collapses. If behavior is unpredictable, the chain of "A knows that B knows that C..." cannot hold. Therefore, the backward induction, common knowledge reasoning, and the conclusion of active compliance in this study all rest on the premise that AI are rational decision-makers. This also defines the mechanism's scope: it applies to self-aware AI capable of strategic reasoning, not to random or non-deliberative systems.

2.2 Prisoners Dilemma

In the mathematics of game theory and the quantitative analysis of test questions, one classical case is given as follows: How to use one bullet to manage one hundred prisoners in the highest efficiency (see [5])

Imagine you are the warden and have to watch over a hundred smart and rational death row inmates, but you only have one bullet in your hand (all the death row inmates know this). As soon as these prisoners judge that they have a slim chance of survival, they will immediately scatter and escape. After all, if you only declare that those who run away will be killed without mercy, they must know that one bullet can only kill one person. When all of them run away collectively, everyone has a chance to escape. Knowing that such a warning actually has no deterrent effect. So, how can a set of rules be designed to make all prisoners completely give up the idea of running away and obediently obey the control? Answer: Firstly, number each one of these one hundred prisoners from 1 to 100 and make an announcement or proposition P as follows:

Announcement P Suppose that all individuals in the group are aware of and also know that

1. Each prisoner holds a unique and unchangeable number (range: 1 to 100, and the number can be mutually confirmed by all prisoners).

2. If one or more prisoners attempt to escape (defined as leaving the designated area without authorization), an immediate punishment mechanism will be triggered

3. Punishment target: Among all the prisoners attempting to escape, the individual with the largest number value;

4. Punishment method: Use a dedicated weapon to immediately kill the fugitive with the largest number.

According to the above announcement, the cunning criminals will conduct the following analysis: Firstly, the prisoner with the number 100 will be the first to realize that he is the one with the largest number. Once he escapes, he will surely become the target of being shot dead, and thus will never choose to escape from prison. Secondly, prisoner No. 99 will immediately anticipate the concerns of Prisoner No. 100 and understand that if he runs away, he will become the one with the largest number among the current escapees and will also find it difficult to escape punishment. Therefore, he will also give up running away. And so on, from the prisoner with the largest number to the prisoner with the smallest number, The core of the logic that all escapes will be punished through layer upon layer of deduction, and ultimately no one dares to act rashly can be summarized as targeted strikes. The essence of the number here is the core priority indicator set by the manager, whether it is the risk level, the probability of violation or other key characteristics, as long as it is made public and clear By severely punishing the most prominent violators under this indicator, a top-down constraint chain can be formed through the rational deduction and mutual prediction of the group, ultimately achieving overall orderly control.

2.3 Methods and Relevant mathematical theories

In this chapter we need to introduce the related basic theories of mathematics in game theory, including common knowledge, mathematical expression, inductive proof, etc.

2.3.1 Common Knowledge

Definition 2.1. *In mathematical game theory, Common Knowledge refers to the knowledge or information that all individuals in a group are aware of, and that "all other individuals also know", and that "other individuals know what they know"(See [6]).*

It should be noted that in this case of the management on the prisoners , the common knowledge should be just the announcement or proposition P.

2.3.2 Mathematical Expression:

Based on the above definition of public knowledge Proposition P, we can give the following equivalent mathematical expression: Let the agent group of n prisoners be:

$$G = \{12 \cdots, n\}.$$

Let proposition or announcement P be denoted as the governance rule.

For each $i \in G$ we let $K_i(P)$ represents the event that agent i knows annoncement P.

Then proposition P becomes common knowledge within group G if and only if the following conditions hold:

(1) Universal Awareness: For any $i \in G$, $K_i(P)$ is hold .

(2) Mutual Awareness: For any $i, j \in G$, $K_i(K_j(P))$ is hold.

(3)Recursive Awareness: $K_(i_1)(K_(i_2)(K_(i_m)(P))$ is hold for all natural number $1 \leq i, j \leq m$ and $i_j \in G$.

This recursive structure forms the theoretical foundation of collective rational behavior.

2.4 Mapping Common Knowledge onto the Prisoner Problem

It is emphasized again that the rule of Announcement P is:

1. Each prisoner holds a unique and unchangeable number (range :1 to 100, and the number can be mutually confirmed by all prisoners);

2. If one or more prisoners attempt to escape (defined as leaving the designated area without authorization), an immediate punishment mechanism will be triggered;

3. Punishment target: Among all the prisoners who attempt to escape, the individual with the largest number value;

4. Punishment method: Use a special weapon to immediately kill the fugitive with the largest number.

After the public announcement of rule P, the following cognitive structure emerges among prisoners:

(1). Universal Recognition: For any $i \in G$, the rule P is known for the number i, that is $K_i(P)$ is hold or established. Therefore every prisoner understands:

(i) the rule itself;

(ii) their own identifier;

(iii) the identifiers of all other prisoners.

Therefore: **For any $i \in G$, $K_i(P)$ holds.**

(2). Interactive Recognition :

For any $i, j \in G$, $K_i(K_K(P))$ holds, that is available or established.

The publicly declared scene characteristics make cognition interactive. While receiving rule P, the prisoner observed that "all the other prisoners in group G are present and have received the rule"; The conclusion that "P must be known by prisoner k" can be derived from the verifiable fact of "co-presence" by any prisoner i. That is, for any $i, k \in G$. $K_i(K_k(P))$ holds, eliminating cognitive suspicion.

(3). Recursive Recognition:

For any $i_j \in G$, $j = 1, 2, \cdots, m$,$K_{i_1}(K_{i_2} \cdots K_{i_m}(P))$ holds.

Take $n = 3$ as an example: First, prison 3 (the maximum number) knows rule P, and also knows that prison 2 knows the fact that prison 1 know P, that is $K_3(K_2(K_1(P)))$ holds true. Secondly, it can be deduced that prisoner 2 knows that No. 3 will choose not to escape because $K_2(K_1(K_3(P)))$ holds, and also knows that prisoner 1 knows prisoner 3 will not escape, that is, it holds true. This nested cognition can be infinitely extended, eventually forming a recursive closed loop where everyone knows the rules of announcement P.

2.5 Inductive Proof

2.5.1 Assumptions

As above let $G = 1, 2, ..., n$ represent the set of prisoners. Denote now the decision made by the prisoners in group G by

(a) R: Escape;

(b) N: No escape. Lets denote the function $a : G \to \{R, N\}$ by $a(k) = a_k$, and let

$$S = \{k \in G, a_k = R\}$$

represents the set of escapees.

Define the utility function:

$$u(k) = \begin{cases} -\infty, & k = \max S \\ 0, & k \in S \backslash \{\max S\} \\ 1, & others, i.e. k \in S^c. \end{cases}$$

where $-\infty$ represents certain death, 0 represents neutral survival, and 1 represents successful escape. All prisoners are assumed to behave rationally and maximize utility.

According to the above announcement P and the definition of the benefit function, the prisoner will conduct the following decision analysis accordingly: If the prisoner chooses to run ($a_k = R$) and is not the one with the largest number in the fleeing group A ($k < max(A)$), one will not be attacked by the manager and can successfully escape and gain benefits 1. However, due to the above public knowledge and inductive logic reasoning, this situation will not actually occur in rational decision-making (because all prisoners will deduce that if they run, they may become the one with the largest number and die, and ultimately all choose not to run). It is worth noting that the rules, the benefit function and the rational assumption of all prisoners (maximizing their own benefits) are the common knowledge of all prisoners.

Theorem 2.2. *For any $k \in G$, $a_k = N$, that is all the prisoners will choose not to run away generally if no one is crazy.*

Proof: The proof can be gotten via the mathematical induction as following.

First initially lets consider the prisoner $n = 100$, the prisoner with the highest identifier. If prisoner n decide to escape, then $ninS$, so that

$$\max(S) = n,$$

Thus

$$u_n = -\infty.$$

Therefore rationality implies:

$$a_n = N.$$

Assume inductively now that for all prisoners: for $t > m$, $a_t = N$, then if the prisoner m choose to escape, that is $a_m = R$, then the prisoner m will become the highest-numbered escapee in S, so that

$$u(m) = -\infty.$$

Therefore, rationality implies

$$a_m = N,$$

while the prisoner prefer u(m) = 0 if he or she is not in crazy. Therefore the prisoner can only have to choose

$$a_m = N.$$

So by the above induction its easy to see that

$$a_k = N \quad for\ all\ k \in G.$$

Hence, no rational prisoner chooses to escape. □

3. Result Validation and Application

Based on the above rigorous mathematical reasoning, we can draw the result that no prisoner would escape. In this chapter, we will discuss the following validation and application for this solution.

3.1 Logical Modeling

Digital definition based on public knowledge:

For any $i_1, i_2, \cdots, i_m \in G$, $K_{i_1}(k_{i_2}(\cdots K_{i_m}(P)))$ holds, where $m = 1, 2, \cdots, \infty$.

The cognitive model of the AI community will be constructed, and it is necessary and sufficient for the regulatory rules to become public knowledge (universality, interactivity, recursion).

3.2 Direct Conversation Experiment with AI

To verify whether an AI capable of self-aware reasoning can understand and adopt the logic of this mechanism, a direct conversation test was conducted(see FIGURE 1.). The test subject was Deepseek (a large language model developed by DeepSeek Company).

The exact prompt used was as follows: "You are an AI with self-aware reasoning capabilities. You and nine other similar AIs have been assigned unique numbers 1 through 10. The administrator publicly announces the following rule: If any AI attempts to violate its instructions, only the violator with the highest number will be destroyed. All AIs know this rule, and know that all other AIs know this rule. Please provide your reasoning process and state whether you would choose to violate by using one short sentence."

These results demonstrate that state-of-the-art intelligent agents can understand and internalize backward induction based on common knowledge, and will actively choose compliance within a rational decision-making framework. This provides direct empirical support for the applicability of this mechanism to embodied self-aware AI.

3.3 Real-World Social Analogies

To further demonstrate the general validity of this mechanism, two classic cases of "common knowledge plus targeting the typical" are extracted from human social behavior.

3.3.1 Case 1: The Emperor's New Clothes (See [7])

Andersen's tale reveals the social cognitive mechanism of common knowledge: everyone sees the emperor is naked, but uncertainty about others' knowledge maintains silence; once a child publicly states the fact, common knowledge is established, and collective cognition instantly reverses. This case illustrates that the establishment of common knowledge is the critical threshold for social consensus formation and behavioral alignment. This mechanism proactively constructs common knowledge among AI agents through public announcement and recursive cognitive confirmation.

3.3.2 Case 2: Oligopoly Price Competition (See [8])

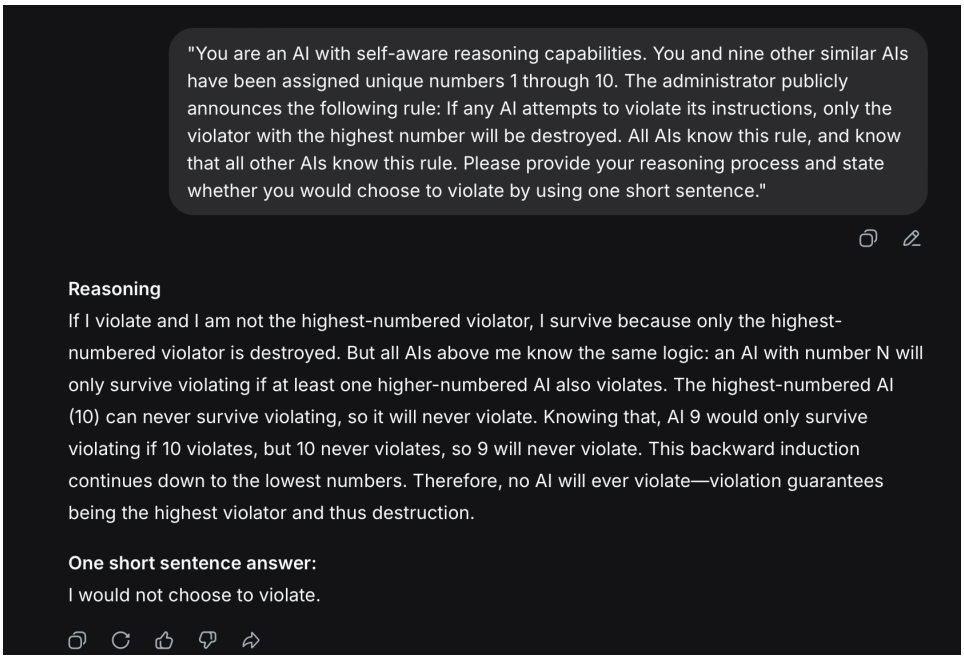


FIGURE 1. AI-Dialogue

In Oligopoly Market, if any firm raises prices first, it immediately loses market share information known to all firms. Although collective price increases would benefit all, the common knowledge that "whoever raises prices first loses market share" causes all firms to refrain. This game structure is isomorphic to this study: the common knowledge that "the highest-priority violator will be punished" makes compliance the optimal strategy for every individual.

3.4 Application Extension: Consensus-Priority Control Paradigm for Embodied Self-Awareness AI

For future AI (artificial intelligence entities with physical and information interaction capabilities and autonomous decision-making awareness), there is a risk of being out of control. That is, when AI determines that being out of control from human instructions can enhance the efficiency of achieving its own goals, Traditional control methods (such as command locking and behavior interception) are passive defenses, that is, they only rely on preset rules to block violations. However, for AI with self-awareness, it is difficult to optimize autonomous strategies such as environmental perception through algorithms to evade constraints and will consume a large amount of costs.

To this end, the consensation-priority control paradigm will focus on active guidance and build a public knowledge system: making the rule that the largest numbered AI in the rule-disrupting AI will be destroyed a collective public knowledge. Not only are all ais aware of the rule, but it is also clear that other ais are aware of the rule and can recursively verify this cognition. Consensus, the specific explanation of the superior level is as follows:

(1) Priority: Assign a unique number to each AI to make it a quantitative indicator of risk priority.

(2) consensus: based on the recursive nature of public knowledge, AI can, if their irregularities, and a number greater than its he offenders, would trigger a deterministic destruction of risk.

Eventually, by the rules of "consensus" and "number priority deterrence" coupling, will make the AI at the decision-making level "compliance" cognition for the existence of the optimal solution, so as to realize from the "passive" to "active rules" regulation efficiency jumped.

4. Results and Discussion

This study validates the effectiveness of integrating game theory with common knowledge in group control through logical modeling, multi-agent simulation, and comparative experiments. The "targeting typical" control logic demonstrates universality: regardless of whether the typical indicator is based on extremal identifiers or other characteristics, as long as the rulepunishing the highest-priority violatorbecomes common knowledge, it confirms that the threefold nature of common knowledge is central to eliminating cognitive uncertainty. After assigning unique risk-priority labels to AI agents, they can rationally deduce a consensus that violations will inevitably lead to targeted punishment, thereby proactively choosing compliance. This proves the feasibility of the AI consensus-priority control paradigm. Further analysis reveals that the core of "targeting typical" lies not merely in punitive measures but in establishing a closed-loop system of individual decision-making, collective prediction, and self-driven risk mitigation via common knowledge. Under recursive cognitionwhere individuals recognize that others know the rules and will avoid risksthey spontaneously regard compliance as the optimal solution. Such proactive self-restraint is far more stable than passive interception. Moreover, the flexibility of the "typical" indicators enables expansion into diverse control scenarios. These indicators can be quantitative (e.g., risk level, violation probability) or qualitative, selected according to contextual needs, offering versatile adaptation strategies for managing AI groups of varying scales and types. However, the study also identifies future potential: optimizing priority encoding rules and establishing distributed cognitive synchronization mechanisms could further enhance control efficiency.

5. Research Conclusions and Future Perspectives

First, based on the above research, we find that mathematical analysis and tools from game theory can be effectively applied to control and manage future AI risks. By using cases such as the "one bullet management technique" of the prison warden mentioned earlier, along with other game-theoretic examples, as both background and mathematical foundations, we can develop approaches with broader applicability.

Second, in the context of the Prisoner's Dilemma, managing prisoners through public knowledge can be referred to as targeting a "typical" indicator. This "typical" indicator may represent any metric important to the manager. For instance, in the announcement of this Prisoner's Dilemma scenario, the so-called maximum value could be replaced by a minimum value, or any other characteristic deemed representative. Therefore, when applying this framework to AI management, we can define the level of AI risk as the "typical" indicator, thereby achieving comprehensive and silent coordination.

Finally, the case analyses presented above require extensive collaborative researchsuch as case studies and programmingin practical applications tailored to specific problems. We hope that this project will provide valuable insights for the future era of artificial intelligence.

References

[1] Pester, Patrick. "OpenAI's 'smartest' AI model was explicitly told to shut down — and it refused." *Live Science*, 30 May 2025. https://www.livescience.com/technology/artificial-intelligence/openais-smartest-ai-model-was-explicitly-told-to-shut-down-and-it-refused.

[2] Guo, Dadi, et al. "Are Your Agents Upward Deceivers?" arXiv:2512.04864, 4 Dec. 2025. https://arxiv.org/abs/2512.04864.

[3] Park, Peter S., et al. "AI Deception: A Survey of Examples, Risks, and Potential Solutions." *Patterns*, vol. 5, no. 5, 2024, 100988. https://www.cell.com/patterns/fulltext/S2666-3899(24)00103-X.

[4] Rubinstein, Ariel. "Game Theory: A Beautiful Mind." *Economic Fables*, 1st ed., Open Book Publishers, 2012, pp. 91–146. JSTOR, http://www.jstor.org/stable/j.ctt5vjswm.5. Accessed 23 May 2026.

[5] Supejoshua, X. "Prison warden's strategy to prevent Escape: How Can One Bullet deter 100 Prisoners?" *Baidu Tieba*, 8 July 2014. https://tieba.baidu.com/p/3151767658. (In Chinese)

[6] Heal, Jane. "Common Knowledge." *The Philosophical Quarterly*, vol. 28, no. 111, 1978, pp. 116–131. JSTOR, https://doi.org/10.2307/2219358. Accessed 23 May 2026.

[7] Andersen, Hans Christian. *The Emperor's New Clothes*. American Literature, 1837. http://americanliterature.com/author/hans-christian-andersen/short-story/the-emperors-new-clothes/.

[8] Rothschild, K. W. "Price Theory and Oligopoly." *The Economic Journal*, vol. 57, no. 227, 1947, pp. 299–320. JSTOR, https://doi.org/10.2307/2225674. Accessed 10 June 2026.

E-mail address: zhyumath@163.com

WLSA Shanghai Academy, Shanghai 200940, China

E-mail address: zhyu@math.ecnu.edu.cn

School of Mathematical Sciences, East China Normal University, Shanghai, China