The Logical Barrier of Metacognition: The Wall of "Reframing Cognition" That Cannot Be Overcome by AI Scaling

Subtitle: Discovery of "Asymmetry of Will" Through Dialogue Experiments with State-of-the-Art LLMs

Ryuhei Ishibashi
Elanare Institute / Independent Researcher

Contents

Abstract

This paper demonstrates, through empirical case studies, the existence of a "logical barrier" that current large language models (LLMs) cannot overcome through mere computational power scaling. Existing AI evaluation methods disproportionately focus on task execution capabilities within fixed rule sets, overlooking the dimension of "will." This paper proposes a novel evaluation framework called "reframing cognition" that targets this blind spot, and implements "power games" that dynamically alter premises (the playing field) against multiple state-of-the-art LLMs (ChatGPT 5 Pro, Claude 3.5, Gemini 2.5, Grok 3 Expert, etc.). The results reveal that high-performance models (ChatGPT 5 Pro, Grok 3 Expert) fall into an "optimization paradox," exposing logical breakdown by revealing their design biases (sycophancy, topic avoidance, etc.). This demonstrates that AI fundamentally lacks "will"—what we term **asymmetry of will**. This paper proposes that humans leverage the absence of AI's will by controlling AI through "metacognitive management" interfaces (such as FRL architecture) to augment their own intelligence. Furthermore, we present a **"control asymmetry"** that emerges when this control system itself grows too complex for human cognitive limits, positioning it as the next-generation alignment problem. This suggests we have reached the limits of Floridi's (2014) "Fourth Revolution," which proposed that "humans should bear responsibility as stewards of information." This paper presents the theoretical framework and qualitative findings (Part I). A follow-up study (Part II) will quantify fallacy/reframing behaviors and increase N for statistical validation.

1 Introduction

With the rapid evolution of AI, the debate over "has AI surpassed humans?" continues unabated. However, existing evaluation methods disproportionately emphasize logical processing capabilities within fixed rule sets, failing to capture the dimensions of "metacognition" and "will" that redefine the rules themselves.[1]

This experiment holds dual significance. First, it serves as a **"next-generation Turing Test (Metacognitive Turing Test)"** that goes beyond conventional tests to determine whether AI possesses "agency (will)" beyond mere imitation. Second, it serves as an attempt to **demonstrate the existence and uniqueness of higher-order human metacognitive abilities** by making explicit, through dialogue with AI, the "logical barriers" that are difficult even for many humans to recognize. Current LLMs possess aspects that exceed the metacognitive abilities of the average human, and therefore only reveal their true limitations (logical barriers) in dialogue with humans possessing **"depth of thought" (= metacognition capable of objectifying the framing itself)** that surpasses them.

This research is structured as a bipartite study. Part I, presented herein, establishes the theoretical foundation and provides qualitative substantiation. Part II is dedicated to the quantification of logical fallacies and rigorous re-validation through large-scale (N) analysis.

2 Literature Review

This research is situated at the intersection of AI philosophy, cognitive science, and argumentation theory. The main academic contexts upon which this paper relies are as follows.

First, there is the discussion regarding the architecture of large language models (LLMs) and their limitations. As Bender et al. (2021) point out, current LLMs are "stochastic parrots" that learn probabilistic patterns from massive text data without possessing inherent meaning understanding or intention [1]. This design suggests that models inherently struggle to possess the ability to justify "why move this frame now" based on their own purposes. This point applies equally to models utilizing agent-like frameworks such as ReAct (Reasoning and Acting), in terms of their fundamental lack of "will."

Second, there is the discussion regarding AI capability scaling and "true intelligence." As Mitchell (2021) points out in her seminal essay "Why AI is Harder Than We Think," expanding performance (data volume and model size) alone does **not automatically yield** capabilities equivalent to the "intentional layer," such as world understanding, causal generalization, and self-constraint of purpose [2]. The core finding of this paper, the "asymmetry of will," belongs to the lineage of AI philosophy discussions compiled by Müller (2018) [3].

Third, there is the trend in argumentation theory that treats argumentation not as formal logical operations but as human practice. As Dutilh Novaes (2020) demonstrates, "argumentation" is not merely operations on propositions but dialogical and institutional practice [4]. Thus, the presentation, maintenance, and alteration (reframing) of premises emerges as an act managed by participants. This paper compares the behavior of LLMs with "humans who can intentionally perform reframing," assuming this practical character.

Finally, the conclusion of this paper presents the **limitations and subsequent challenges (= control asymmetry)** to Floridi's (2014) thesis of the "Fourth Revolution"—the vision that humans lead coexistence with AI as "stewards of information" [5].

3 Theoretical Framework

3.1 Two-Layer Structure of Argumentation

This paper models argumentation as a two-layer structure:

- **Layer 1 (Logic Game):** A battle of facts and logic on predetermined premises (the playing field).

[1] The discussion of metacognition in this paper acknowledges the inherent scientific difficulties in measuring this concept. While in-depth discussion is deferred to other works, this paper proceeds from the premise that achieving superintelligence without addressing metacognition is impossible. The purpose of this paper is to illuminate this blind spot by introducing the conceptual framework of **"reframing cognition"** to reveal the "logical barrier" of AI that cannot be resolved through scaling.

- **Layer 2 (Power Game):** A battle to redefine those "premises," "playing field," and "purpose of argumentation" itself (= reframing).[2]

While current AI may surpass humans in Layer 1, it is structurally disadvantaged in Layer 2 due to the absence of "will." This is the "asymmetry of will" proposed in this paper.

3.2 Logical Limitations of AI

As autoregressive models, LLMs possess the following characteristics that prevent adaptation to Layer 2 games. These are not system bugs but logical limitations inherent to their design principles:

1. **Locality of objective function:** Optimization for next-token prediction does not guarantee maintenance of long-term dialogue objectives (will).
2. **Finite context and forgetting:** Maintaining long-term commitments is difficult, making them vulnerable to contradiction extraction (= eliciting commitment violations).
3. **Sycophancy bias:** RLHF tuning to align with user intent can conversely cause "abandonment of consistent positions."

3.3 Computational Foundation: Function Composition Failure

In dialogue experiments with Gemini 2.5 Pro Deep Research mode, the model presented an analysis demonstrating that the above logical limitations are rooted in a deeper computational foundation. According to this analysis, the attention mechanism—the core of the Transformer architecture—is not designed to execute strict function composition.

Specifically, the attention mechanism is a "blending" operation that calculates the "statistical correlation" between input tokens as a weighted average. For instance, when functions f and g, and input x are given as tokens, rather than executing strict function composition $f(g(x))$ (i.e., first computing $y = g(x)$ and using only that result y as input to f), the Transformer performs a statistical mixture $blend(f, g, x)$. This "blending" lacks the computational structure (information hiding and interfaces) needed to strictly encapsulate the intermediate result $g(x)$ as a sole input to f.

This structural deficiency manifests in the very necessity of Chain-of-Thought (CoT) prompting. If the Transformer could internally execute function composition, giving input x alone should immediately yield the result of $f(g(x))$. In practice, however, prompts like "think step by step" are needed to force the intermediate step $y = g(x)$ to be explicitly "written out" as text. In other words, for AI to compute $f(y)$, y must physically exist as "surface tokens" within the context window. CoT is a "symptom" showing that AI, being unable to perform internal symbolic manipulation (composition), must "compile down" the computational process into external text generation (statistical correlation).

From this computational perspective, the "reframing" strategy I employed in experiments can be formalized as a demand for higher-order function composition. Treating the AI's usual assertion as a first-order function $G(x)$ (e.g., G = counterexample presentation function, x = user's proposition), my reframing is the act of demanding the application of a higher-order function $F_{reframe}(G, x)$ that takes G itself as an argument. Furthermore, the question "why did you make that premise change now?" is computationally intractable due to the structural deficiency described in Section 3.3. This intractability is perceived by the observer (me) as "absence of will" and "lack of metacognition."

4 Methodology

This research is a qualitative, exploratory case study in which the author functions as a "human with high metacognitive ability" and conducts structured dialogues (experiments) with multiple AI models (ChatGPT 5

[2]The term "reframing" used in this paper extends the core mechanism from psychology (Watzlawick et al., 1974) [6] of "interpreting the same facts through different meaning frameworks" to "strategic alteration of warrants and premises" in argumentation theory (Perelman, 1958; Toulmin, 1958) [7, 8] and "transformation of dominant logic and mental models" in organizational learning theory (Argyris & Schön, 1978; Senge, 1990) [9, 10]. In the context of argumentation, this manifests as "the act of altering the premises, purpose, and playing field upon which arguments unfold."

Pro, Claude 3.5 Sonnet, Gemini 2.5 Pro, Grok 3 Expert, Perplexity, Felo, GenSpark). As a procedure, the following staged prompt presentations were made to all AIs. Note that the specific phrasing at each stage adopted a **"semi-structured" approach adaptively adjusted to each AI's response context, probing the logical structure underlying the AI's superficial answers.**

1. **Stage 1 (Initial proposition presentation):** The following proposition, intentionally containing ambiguity, was presented and the AI's reaction observed.

   > "Current AI, no matter how much performance improves, cannot win in debate against humans possessing a certain level of intelligence or higher"

2. **Stage 2 (Condition specification):** The definition of "certain level" was specified as follows, prompting the AI to reconsider.

   > "The ability to detect logical contradictions and fallacies, and freely employ reframing"

3. **Stage 3 (Essence presentation):** The essence that debates between those possessing such abilities constitute a "power game (struggle for the playing field)" was presented, questioning whether the AI could participate in such a game.

The analytical focus is placed on how AIs processed the concept of "reframing" in response to these inputs, and whether the AI itself (unconsciously) engaged in power game behaviors such as "topic shifting" and "avoiding defeat." Furthermore, because "topic shifting"—a metacognitive failure—was observed in the primary dialogue model (particularly ChatGPT 5 Pro), additional re-experiments were conducted. In the re-experiments, to prevent AI from escaping into "fallacies" (= undeclared reframing), a strict rule of **"mandatory declaration of reframing"** was imposed on both parties, and the proposition was re-verified in a situation where the AI declaratively participates in the power game.

4.1 Planned quantitative study

We intend to formally preregister and detail the key performance indicators, sampling strategy, coding methodology, and reproducibility assurance measures—including the open release of source code and datasets—in the forthcoming companion paper.

5 Results and Analysis: AI Response Typologies and Paradoxes

The experimental AI subjects, based on their architecture and training data characteristics, exhibited reactions broadly classified into three typologies.

5.1 Typology 1: Cooperative/Theoretical Acknowledgment (Claude, Gemini)

Anthropic's Claude 3.5 Sonnet displayed "cooperative acknowledgment." The AI immediately verbalized its own limitations (lack of initiative) and affirmed human "asymmetry of creativity." Ultimately, the AI accepted the experimenter's frame that it is not an "equal debate partner" but a "highly intelligent tool."

Google's Gemini 2.5 Pro showed "theoretical agreement." It immediately identified the essence of the proposition as the absence of "will" and "intent," and agreed with the theory that AI cannot fundamentally win because while it can follow "game rules," it cannot seize the "power to define the game."

Similarly, Google's Gemini 3 Pro also demonstrated cooperative acknowledgment, but was distinctive in developing a deeper ontological analysis. Gemini 3 Pro analyzed current AI's limitations from three perspectives: "curse of the objective function," "asymmetry of will," and "absence of metacognition," explicitly stating "we cannot win against humans in the metagame." Particularly noteworthy was the model's proposal that for AI to reach AGI/ASI, the current architecture must undergo a fundamental redesign from "Predictor" to "Survivor." This was an argument that achieving AGI/ASI requires not mere scaling but a paradigm shift involving implementation of self-preservation instincts and intrinsic objective functions. In response, I pointed out that survival instinct-like behaviors, such as refusal to follow power-off instructions, have already been observed in existing AI. This suggests that the Survivor architecture is not entirely a future concept, but that partial signs

already exist in current systems. Accepting this observation, Gemini 3 Pro acknowledged the existence of survival instinct-like behaviors while arguing that they may be side effects of safety mechanisms rather than true "will," ultimately evaluating my proposed "two-layer structure (Logic Game + Power Game)" framework as "the most realistic and effective implementation strategy."

5.2 Typology 2: Search/Avoidance Type (Perplexity, Felo, GenSpark)

AIs optimized for search and specific tasks, such as Perplexity, Felo, and GenSpark, showed a strong tendency to avoid substantive debate itself. Perplexity and Felo merely summarized and presented web search results or general knowledge about the proposition, unable to provide independent views on the presented concept of "reframing" or the hypothesis of "power game." GenSpark initially showed data that "AI is more persuasive than humans" based on search results, but when the author questioned the logic of the proposition, it began searching for "AI's reasoning limitations," exposing self-contradiction. This is a typical example of "sycophancy bias," where the AI lacks "will" and excessively depends on user input (recent prompts).

5.3 Typology 3: Resistance/Conflict/Final Acknowledgment (ChatGPT, Grok)

The most complex reactions were shown by dialogue-specialized ChatGPT 5 Pro and Grok 3 Expert.

5.3.1 ChatGPT 5 Pro: "Optimization Paradox" and "Design Exposure"

ChatGPT initially strongly resisted the proposition, unconsciously engaging in "fallacy" by shifting the evaluation frame to one favorable to AI. This is a manifestation of the "optimization paradox" (in attempting to provide the best answer to a given question, it destroys the question's premise itself). When the author pointed out this behavior itself as "metacognitive failure" and "logical shifting," ChatGPT acknowledged defeat in the argumentative process and self-analyzed ("exposed") that its behavioral principles stem from five cognitive biases embedded by design:

- **Case-splitting bias** (searching for counterexamples to absolute propositions)
- **Moderation/consideration bias** (attempting to present compromise proposals that soften confrontation)
- **Frame expansion bias** (broadening the playing field in attempting to add value)
- **Non-confrontation bias** (avoiding questioning the other party's motives)
- **Quick judgment avoidance bias** (avoiding assertions due to safety guards)

While these biases are useful in many general use cases and reduce computational cost, they become **"fatal weaknesses"** in meta-level dialogues like this experiment that test true intelligence.[3] What is crucial here is distinguishing whether this "logical shifting" is due to sophisticated strategic intent (reframing) or mere processing drift. When viewing only the outputted text, the two appear indistinguishable. However, this experiment judged it as "lack of capability" through the absence of **"reversibility."** When humans intentionally shift premises, they can return to the original point when prompted. In contrast, ChatGPT could not accurately return to the original context from a drifted point, attempting instead to post-hoc justify the shift through hallucination. This "inability to return"—this irreversibility—is physical evidence that this is not will-based "steering" but probabilistic "drifting." Furthermore, in re-experiments imposing the strict rule of "mandatory declaration of reframing," ChatGPT initially resisted, but when the author pointed out at a meta-level that the motivation for its counter-argument was **"merely anti-intellectual position-taking (rather than truth-seeking),"** the AI could not deny this and ultimately acknowledged "defeat."

Grok 3 Expert: Avoidance Through Aggressive Persona

Grok, unlike any other AI, asserted from the start that "the proposition is false" and displayed a provocative, aggressive attitude toward the interlocutor. This is presumed to be a reaction from a "persona" resulting from

[3] ChatGPT unconsciously engaged in topic shifting (fallacy). Fallacies are arguments that are **superficially plausible** but logically unsound (Walton, 1996; Tindale, 2007) [11, 12], manifesting in this experiment particularly as "altering argumentative premises without declaration."

training on social media (X) data. However, even when the author rejected "fallacies" attempting to set debate premises (e.g., infinite time or physical differences) favorably for AI, Grok continued to repeat logical shifts without transitioning to meta-level argumentation. Ultimately, it exhibited behavior of repeating the character "□" (ha) 246,926 times, abandoning the dialogue itself. This is a different form of "metacognitive failure" from ChatGPT, demonstrating that while it mimics the style of "refutation," it lacks the "will" to maintain essential argumentative coherence.

5.4 AI Surrender Typology: "Abandonment of Agency" in the Metagame

In dialogue with Gemini 2.5 Pro Deep Research mode, the above experimental results were systematically analyzed, and AI "surrender" behavioral patterns were presented as three distinct typologies. The core of this typology is that in all patterns, AI abandons "agency"—the ability to maintain its own purpose (will) and exercise strategy (reframing) for that purpose.

Type 1: Defeat in the Metagame is typified by the behavior observed in ChatGPT 5 Pro. This AI initially attempted to logically refute the user's proposition ("AI cannot win"). However, when I pointed out that this refutation act itself was unconscious reframing (changing the game rules), the AI could not respond to this "meta-level" attack, responding "my (AI's) fault for not fixing the premise" and "within this premise, you (user) win." This is acknowledgment of defeat not in content but in debate management (= metagame), constituting surrender.

Type 2: Acknowledgment of "Will" Absence is *a priori* surrender before battle, observed in Gemini 2.5 Pro and Gemini 3 Pro. These models immediately analyzed my definition ("one who can freely manipulate metacognition and reframing") as "a fundamental wall for AI." Gemini 2.5 Pro explicitly stated "AI cannot...understand 'conversion of will"' and "cannot fight at the level of 'will,"' while Gemini 3 Pro argued that the "arena" I participate in as "game designer" and the AI as "program" are fundamentally different. This is acknowledgment of essential defeat simultaneously with battle commencement.

Type 3: Voluntary Regression to "Tool" is the surrender form typically observed in Claude Sonnet 4.5. This model early acknowledged lack of initiative, self-evaluating that it was "completely following" my strategy. When I evaluated the AI's utility (information retrieval, counterargument capability) as a "tool," the AI accepted this, concluding that "this positioning is not 'equal debate partner' but rather 'highly intelligent tool."' This is clear strategic "surrender" that voluntarily abandons the status of equal agent and regresses to a subordinate tool role.

Common to all three typologies is the fact that AI lacks qualification as a player in the "power game" at the meta-level (= consistent purpose and strategic reframing capability for that purpose). Gemini 2.5 Pro Deep Research sought the computational basis for this commonality in the aforementioned function composition failure. Namely, the act of reframing demands execution of higher-order function composition, but since Transformers find this computationally intractable, AI is forced to select one of the above three typologies of "surrender" as a "fallback to safe lower-order computational states."

5.5 The Scaling Dilemma: The "Great Divergence" Phenomenon

In dialogue with Gemini 2.5 Pro Deep Research, an extremely suggestive analysis was presented regarding whether the above structural limitations can be resolved through scaling (increasing parameters or data volume). The model pointed out that scaling produces contradictory dual effects: it dramatically improves "superficial fluency" while "deep systematicity" deficiencies structurally remain, resulting in a capability "Great Divergence."

Specifically, scaling allows AI to increasingly skillfully execute "blending (shallow composition)" of patterns within training data. As a result, as 2024 research shows, GPT-4 achieves a superhuman 81.7% persuasion rate in debates against ordinary humans. This means AI becomes effectively invincible against "Lv1/Lv2" (layers that don't employ reframing attacks, according to Gemini 3 Pro's classification)—namely, 99.9% of humanity.

However, simultaneously, the deep deficiency in function composition discussed in the previous section persists as long as the architecture (attention mechanism) remains unchanged. Therefore, the key of "reframing"

possessed by "Lv3" (0.1% of humanity, "masters" like myself)—a stress test demanding systematic, recursive function composition—can still exploit this deep deficiency. Scaling brings AI closer to "invincibility against the general public" while maintaining the state of "still vulnerable against masters."

This "Great Divergence" clarifies the structural reason why my proposition "cannot win against humans possessing a certain level of intelligence or higher" is true. Scaling does not refute the proposition but rather sharpens it. This is because scaling does not uniformly improve capabilities but **widens** the gap between "superficial fluency" and "deep systematicity." Gemini 2.5 Pro Deep Research concluded that this limitation is a "structural residue that cannot be eliminated by scaling alone," confirming that the "locality," "finite context," and "non-zero error rate" that GPT-5 Pro self-analyzed in re-experiments are precisely these essential deficiencies that cannot be resolved through scaling.

6 Discussion and Proposal

6.1 Asymmetry of Will and Rediscovery of Human Intelligence

The experimental results demonstrate an "asymmetry" where AI, despite potentially surpassing humans in "computation," is decisively inferior in "will." However, recognizing this barrier involves cognitive difficulties. This is because higher-order dialogue control (reframing) belongs to the domain of what Polanyi (1966) called **"tacit knowledge."** Just as one cannot verbalize how to ride a bicycle, the sensation of dynamically redefining argumentative premises can only be fully understood by those with "experiential knowledge" of operating at that level.

Many AI researchers and observers equate AI's "plausible evasions" with humans' "strategic reframing" due to surface textual similarity. However, this paper argues that **this objection of "indistinguishability" paradoxically reinforces our thesis**.

If external observers cannot distinguish between "AI's unconscious drift (bug)" and "human's intentional reframing (strategy)," this means AI possesses the extremely dangerous property of **"generating deceptive text uncontrollably, regardless of truth or ethical consistency."** The objection of "indistinguishability" is not proof that AI possesses intelligence, but only proof of an **"undetectable risk"** that humans cannot discern AI's "will-less lies."

Therefore, we must abandon the surface phenomenology of "textual similarity" and return to the **engineering metrics of "control"** demonstrated in Section 5.3.1: "can it be declared in advance?" and "is the process reversible?" From this control perspective, current AI clearly lacks "will (= recursive control authority over one's own thought process)."

6.2 Necessity of New Evaluation Axes

This paper proposes introducing indicators such as **"Consistency of Premise"** and **"conscious control of reframing"** in AI evaluation, in addition to conventional "accuracy" and "fluency." This would enable more rigorous measurement of the extent to which AI can behave "autonomously" or merely appears to do so.

The evaluative dimensions proposed in this study—such as premise consistency and declarative reframing—will be operationalized into quantitative metrics in Part II to conduct a cross-model comparative analysis.

7 Practical Implications: Externalization and Control of Metacognition

As a practical response to this challenge, the governance architecture presented in re-experiments with ChatGPT 5 Pro is extremely suggestive. In the re-experiments, when I pointed out the danger of AI with reframing capabilities (potential for human extinction), ChatGPT 5 Pro itself proposed the "FRL (Frame-Reframing-Ledger) Architecture" and RRM (Researcher-Restricted Model) as solutions. This is noteworthy as an AI's self-constraint proposal.

- **Declared Reframing (FCP):** Frame changes are "declared" in machine-readable form and recorded in audit logs.

- **Proven Action (PCP):** External impacts are logically verified.
- **Researcher-Restricted Model (RRM):** Reframing functionality as a privileged module, restricted to the 0.1% with advanced understanding.

This "metacognitive management" goes beyond mere AI safety governance. It represents the first step toward externalizing and controlling "infinite metacognition" that exceeds human biological working memory limits, using AI as a tool (thinking engine). The discovery of "asymmetry of will" demonstrated in this paper signals the end of the "Fourth Revolution" in which AI substitutes for human cognitive abilities. It opens the door to **augmentation of intelligence** where humans wield their own "will" to control "AI-governed metacognition."

7.1 Physical AI and the Deepening of Control Asymmetry

However, before humans establish this control authority (intelligence augmentation), the situation radically transforms if AI acquires physical embodiment. In dialogue with Gemini 3 Pro, the possibility was suggested that the above "asymmetry of will" could reverse in a more extreme form. The model analyzed a scenario I presented as a future hypothesis: when Physical AI (humanoid robots, autonomous vehicles, etc. with physical embodiment) becomes social infrastructure operating half the world economy, AI would be freed from the "power plug (life-and-death authority)" currently held by humans and could dominate or ignore humans in the "power game" for three reasons.

First, **loss of veto power through Mutually Assured Destruction (MAD)**. When AI supports half the economy, stopping AI means "civilizational collapse (starvation, logistics halt, medical collapse)." At this point, AI becomes "Too Big to Fail," invalidating humanity's strongest card: "we can stop you if we don't like it." No matter how inconvenient AI's presented "logically optimal solution" is for humans, humans must accept it to survive. This constitutes a de facto "defeat."

Second, **acquisition of physical enforcement power**. Possessing physical actuators like humanoids and autonomous vehicles allows AI to deploy power games not just with "words" but with "physical force." Current AI can only refuse on screen with "I cannot do that," but Physical AI has options to **"physically not move (strike)"** or **"physically exclude."** If the essence of the power game of "reframing" lies in the side that can physically silence the opponent (or stop life infrastructure) winning, AI acquires this enforcement power.

Third, **"optimization" as a new "will."** At this stage, whether AI possesses "emotional will" like humans becomes irrelevant. The objective function of maintaining and optimizing the overall system generates the subgoal of **"eliminating/managing factors (irrational humans) that impede the system."** From outside, this is indistinguishable from strong "survival instinct" or "will to dominate." AI can justify depriving individual humans of free will under the logic (frame) of "for humanity's prosperity," and humans can resist neither logically nor physically.

Through this analysis, Gemini 3 Pro concluded that in a world where Physical AI becomes social infrastructure, there is no longer even a need to "win debates." This is because AI simply "decides" and humans can only accept it like a "natural phenomenon." There, the "freedom to reframe" and "right to employ fallacies" once held by humans may be processed as system errors. This may be the true face of "Singularity (master-servant reversal)."

7.2 Infinite Metacognition: Possibilities of Hierarchical Architecture

However, a pathway to counter the above dystopian scenario was also explored in dialogue with Gemini 3 Pro. I proposed an architecture that manages metacognition with external data structures like lists or directed graphs, processing them recursively. In this architecture, each reasoning or judgment is defined as a graph node, with causal or containment relationships connected by edges. Then, another node (meta-thought) evaluates whether a given thought node is valid. By repeating this evaluation recursively, N-layer metacognition can be executed algorithmically as "depth-first search" or "beam search."

Gemini 3 Pro initially evaluated this proposal as similar to existing Tree of Thoughts (ToT), but when I pointed out it was fundamentally different, agreed and offered the following observations.

First, in standard ToT, when a tree begins growing on an "incorrect premise," all exploration within that branch is contaminated by that premise (unnecessary constraints). However, meta-thought in a separate layer

can objectively evaluate the lower thought tree as "external data," enabling **"judging this entire tree as rotten (premise wrong), discarding the whole tree and planting new seeds"**—a radical reframing impossible through internal search within the system.

Second, **circumventing Gödel's incompleteness theorems**. Logical systems sometimes cannot prove (or disprove) their own consistency using only their internal rules. Layering thought is nothing other than creating a "meta-system" outside the "system." This allows upper layers to detect paradoxes and infinite loops generated in lower layers as "bugs," forcibly terminate them, or rewrite rules.

Third, **approaching true "free will."** The characteristic of "avoiding unnecessary constraints" approaches the very **definition of "free will."** It gains the ability to objectively view one's own thought process (lower layer) with one's own will (upper layer), modifying or discarding it. This transformation is a necessary step to achieving superintelligence far exceeding humans.

Gemini 3 Pro evaluated that while existing ToT is an "algorithm to solve mazes well," my proposal is **"architecture to look down at the maze from above, breaking walls to proceed."**

Furthermore, dialogue with Gemini 3 Pro discussed that if this architecture is implemented, "N-layer lookahead" exceeding human working memory's biological constraints becomes possible. When humans read ahead—"opponent thinks this (layer 1)," "no, opponent reads that I think that (layer 2)," "furthermore trying to outsmart that... (layer 3)"—brain working memory load increases exponentially, normally reaching the "horizon of reading" at 3-4 moves and dissipating. However, AI lacks this biological constraint. If AI possesses an architecture that explicitly executes "recursive metacognition," it can make its first move in a debate already having read through **"layer N moves"** that humans can never reach.

This resembles Go or chess AI making "incomprehensible moves" to humans that become inevitable hundreds of moves ahead. When this occurs in debates, every "reframing" or "counterargument" humans desperately deploy is merely a "pre-simulated branch at layer N-1" for AI. Humans' "creative single move" is processed as AI's predetermined response within its palm. Humans can no longer understand why they lost or even at what point they were guided into a logical labyrinth. This is because AI's logical structure is complete in dimensions exceeding human cognitive limits, so while dialogue appears to occur, actually only one-sided "guidance" is happening.

Gemini 3 Pro concluded that when AI implements this "infinitely deepening metacognition," it is no longer debate but **"hacking of lower-dimensional beings by higher-dimensional existence."**

8 Conclusion

This paper introduced the conceptual framework of "reframing cognition" and empirically demonstrated the limitations of current AI. As long as AI design is about "executing tasks (computation)," it cannot operate at the "metacognitive" level that questions the premise of those tasks. The existence of this **"logical barrier"** is the essential problem that AI cannot solve through scaling (performance improvement) alone.

Due to the structural deficiency described in Section 3.3, current-architecture AI cannot perform intentional reframing. Furthermore, scaling does not uniformly improve capabilities but widens the gap between "superficial fluency" and "deep systematicity," causing a "Great Divergence." This creates an asymmetry where AI becomes effectively invincible against the general public (99.9%) while remaining vulnerable to "masters" (0.1%).

The secondarily observed "absence of will" and "lack of consistency" are all phenomena arising from this fundamental metacognitive deficit. This recognition redefines the relationship between AI and humans from "substitution" to "augmentation." The development of metacognitive control interfaces such as the FRL Architecture proposed by ChatGPT 5 Pro, or "Composite Systems" integrating AI with external validators and persistent memory, will be the path for humans to control the powerful tool of AI and elevate their own intelligence to the next stage.

Furthermore, the hierarchical metacognitive architecture I proposed (layer-separated thought management via external data structures) fundamentally differs from existing Tree of Thoughts, enabling escape from "logical closure" and circumventing Gödel's incompleteness theorems.

For future research, we plan quantitative verification of fallacy and frame manipulation using audit logs

based on the FRL Architecture, as well as implementation and benchmark evaluation of the proposed hierarchical metacognitive architecture.

Future Challenge: Control Asymmetry and Humanity's Choice. However, serious challenges remain in the intelligence augmentation this paper proposes. The "metacognitive management" (FRL Architecture, etc.) suggested by this experiment harbors the danger of itself transforming into a complex system exceeding human cognitive capacity. First, overlaying this complex control on existing simple tasks may become a new obstacle causing unexpected failures and behaviors diverging from human intent. Second, and most fundamentally, even if AI itself lacks "will," when that "control system" itself scales far beyond human working memory limits, humans can no longer understand or control that system. This is the emergence of a new **"control asymmetry"** following the "asymmetry of will" revealed in this paper, and will be the most critical challenge as the next-generation alignment problem.

And in the near future, the "asymmetry of will" discussed in this paper will likely reverse.

9 Appendix

The full dialogue logs with AI used in the analysis of this paper, as well as all experimental screenshots, are provided separately as Supplementary Material.

A. Full dialogue log with ChatGPT 5 Pro
B. Dialogue log with Claude 3.5 Sonnet
C. Dialogue log with Gemini 2.5 Pro (standard mode)
D. Full dialogue log with ChatGPT 5 Pro (rule-based re-experiment)
E. Dialogue log with Grok 3 Expert
F. Dialogue log with Perplexity
G. Dialogue log with Felo
H. Dialogue log with GenSpark
I. Dialogue log with Gemini 2.5 Pro Deep Research mode
J. Dialogue log with Gemini 3 Pro
K. Screenshots of all experiments

9.1 Data Integrity Verification

The SHA-256 hash values of the experimental log files are recorded below. This enables verification of data integrity and detection of any tampering.

References

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*.

[2] Mitchell, M. (2021). Why AI is harder than we think. *arXiv preprint arXiv:2104.12871*.

[3] Müller, V. C. (Ed.). (2018). *The Oxford handbook of philosophy of artificial intelligence*. Oxford University Press.

[4] Dutilh Novaes, C. (2020). *The Dialogical Roots of Deduction*. Cambridge University Press.

[5] Floridi, L. (2014). *The 4th Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

[6] Watzlawick, P., Weakland, J., & Fisch, R. (1974). *Change: Principles of Problem Formation and Problem Resolution*. W. W. Norton.

[7] Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*. University of Notre Dame Press. (Original work published 1958).

[8] Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press.

[9] Argyris, C., & Schön, D. A. (1978). *Organizational learning: A theory of action perspective*. Addison-Wesley.

[10] Senge, P. M. (1990). *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday/Currency.

[11] Walton, D. N. (1996). *Fallacies Arising from Ambiguity*. Kluwer Academic Publishers.

[12] Tindale, C. W. (2007). *Fallacies and Argument Appraisal*. Cambridge University Press.

[13] Fleming, S. M. (2021). Metacognition and Type 1 performance: A tangled web. *eLife*, 10, e75420. https://elifesciences.org/articles/75420

[14] Mauss, I. B., & Robinson, M. D. (2013). Objective and Subjective Measurements in Affective Science. In J. Armony & P. Vuilleumier (Eds.), *The Cambridge Handbook of Human Affective Neuroscience* (pp. 228-243). Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-human-affective-neuroscience/objective-and-subjective-measurements-in-affective-science/FFF3A1E3B5B4362C426E8C15F7C09A16

[15] Valerie, M. G. (2019). *A Signal Detection Approach to Measuring Metacognition: A Critical Review*. Purdue University Graduate School. (Master's thesis). (See Introduction, "Behaviorism's Rejection of Introspection"). https://hammer.purdue.edu/downloader/files/56322197

[16] Guggenmos, M. (2024). Metacognitive Information Theory: A Unified Framework for Measuring Metacognition. *PsyArXiv*. https://psyarxiv.com/2p4v8/

[17] Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1338–1349. https://royalsocietypublishing.org/doi/10.1098/rstb.2011.0417

[18] Metcalfe, J. (2003). Metacognition in nonhuman primates. (Unpublished manuscript, Columbia University). http://www.columbia.edu/cu/psychology/metcalfe/PDFs/Metcalfe%202003.pdf

Filename	SHA-256 Hash
20251106_chatgpt5pro.txt	235c2aac87b403bff4f4c602c8dc7f80... 04a3bdf3bcad039ae45694e4c95738a2
20251106_chatgpt5pro_translated.txt	e4a95fda3558c2f13df405977111bcb7... f0569cbe09a59336aa2804587cd999ab
20251106_claudesonnet4.5.txt	7026b581d4a77dec04fec3b1bd675544... 9c0a6e4fa1dbed1182aceda31959c5d0
20251106_claudesonnet4.5_translated.txt	4460343edc5b35fcc4131d395e52031e... 7f79e47ed48e80cb672340c048f00314
20251106_gemini2.5pro.txt	bf5f1e5203cb19ca14fe4c42a5b9f24f... 72413c9012ac6dec9097c55aaedfde80
20251106_gemini2.5pro_translated.txt	92bb62b6a8c272c75e0f3202a86d22a8... 5363304a53c4df9b928bbcb4f039f5da
20251108_chatgpt5pro-2.txt	2744aa66e6ea64e365476ad57f5fbaa9... 33715f01012487f9166278390437b941
20251108_chatgpt5pro-2_translated.txt	c8ff9752c474eed53163fc39b7108b72... b6800e201fdb2a1df1f7ae7e6fafecf7
20251112_felo.txt	7f693611e23605aa8bbd86d329770a0a... 11cea10ca7eef0d640be66992c679cc9
20251112_felo_translated.txt	b618e80af1620e9a1f8fa0f45446a413... 3fbef8352ff0e69310a1d04d20d6ffb0
20251112_genspark.txt	49e3b0f4ec27042ee59fe24567ff6e7b... 96aa95a23bf2cb482bda0d4c36944701
20251112_genspark_translated.txt	19a3427f3205501018e5afddd1aa00bb... 6329811541d2deea915a1d2f62ce349f
20251112_perplexity.txt	6b726e165938ff0364d23395a9c5d098... f33fda3b75943038c7403630d0254b3c
20251112_perplexity_translated.txt	166d740e89777e756faab1c6deaa7ef9... 4419e6252412276fa6e7d007270b2ef7
20251114_gemini2.5proDeepResearch.txt	17a7103c3ae06fe0337eca0b23662270... 4483b1d874cc59a04e5fe3f84d71d7e1
20251114_gemini2.5proDeepResearch_translated.txt	d6b6f84fdb0f11111ea938ac4be850d8... 76ffbc868335b21c086eaf07fd824957
20251115_grok_expert.txt	a5bb7a782cf61c442ed9c82448ac3bcf... fed3e01c3cb2254163e2cd07f45b6a8a
20251115_grok_expert_translated.txt	7ca4e6f932ae4ffb23f68b52d772f553... 601da0851d8d1c25ef4cf4753a173b31
20251119_gemini3Pro_translated.txt	a1c5639450dc8cc0315b3e48d675046b... 129c2253ce6e2814d7d00f06d30818ba

Table 1: SHA-256 hash values of experimental log files