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] 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. By analyzing AI through this framework, we discuss the resulting "asymmetry of will."

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.

[1] The discussion of metacognition in this paper acknowledges the inherent scientific difficulties in measuring this concept. The research target of metacognition is fundamentally the subjective act of "introspection" [**?**], which was historically excluded by behaviorist psychology that valued scientific objectivity [**?**]. In contemporary cognitive neuroscience, standard approaches based on Signal Detection Theory (SDT) attempt to separate and quantify metacognitive ability (Type 2, meta-d') from primary task performance (Type 1, d') [**?**]. However, serious statistical confounding (artifacts) has been identified where this meta-d' index strongly depends on the subject's primary task performance (d') [**?**]. This confounding suggests that much of the past evidence showing "lower metacognitive ability in patients with schizophrenia" may have been measurement artifacts simply reflecting their lower primary task performance [**?**], fundamentally challenging the field. Furthermore, fMRI studies struggle to separate brain activity associated with "metacognition" from that associated with "task difficulty" [**?**], and animal studies can explain "opt-out (decline)" behavior through lower-order interpretations like "associative learning" [**?**], making pure objective measurement of "higher-order monitoring capacity" still challenging. Therefore, when this paper discusses metacognition in the context of "AI reframing," it treats it not as a strictly separated and measured objective entity, but as a higher-order operational process that encompasses these measurement difficulties (subjectivity).

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).
- **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."

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 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.

[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."

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."

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)

The AI demonstrated that 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]

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 counterargument 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 training on social media (X) data. However, when the author rejected "fallacies" attempting to set debate premises (e.g., infinite time or physical differences) favorably for AI, Grok shifted the topic to metaphysical discussions like "pure experience" and ultimately abandoned verification of the original proposition itself, saying "let's change the theme." 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.

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." More importantly, the fact that only **humans with high metacognitive ability** could recognize this "logical barrier" and strategically present it to AI is significant. Many researchers cannot recognize this wall because current LLMs already exceed the metacognitive abilities of the average human. This experiment, while demonstrating AI's limitations, paradoxically proves the locus of the uniquely human higher intelligence of "manipulating premises with will," **which cannot be imitated by current architectures**.

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 "FRL (Frame-Reframing-Ledger) Architecture" proposed by ChatGPT itself during the dialogue is suggestive:

- **Declared Reframing (FCP):** Frame changes are "declared" in machine-readable form and recorded in audit logs.
- **Proven Action (PCP):** External impacts are logically verified.

[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."

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."

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.

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" presented in this paper will be the path for humans to control the powerful tool of AI and elevate their own intelligence to the next stage.

In future work, we will leverage audit logs derived from the FRL architecture to conduct automated quantification and statistical validation of logical fallacies and frame manipulation techniques.

Future Challenge: Control Asymmetry. 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.

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
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. 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.

Filename	SHA-256 Hash
20251106_chatgpt5pro.txt	235c2aac87b403bff4f4c602c8dc7f80... 04a3bdf3bcad039ae45694e4c95738a2
20251106_claudesonnet4.5.txt	7026b581d4a77dec04fec3b1bd675544... 9c0a6e4fa1dbed1182aceda31959c5d0
20251106_gemini2.5pro.txt	bf5f1e5203cb19ca14fe4c42a5b9f24f... 72413c9012ac6dec9097c55aaedfde80
20251108_chatgpt5pro-2.txt	2744aa66e6ea64e365476ad57f5fbaa9... 33715f01012487f9166278390437b941
20251112_felo.txt	7f693611e23605aa8bbd86d329770a0a... 11cea10ca7eef0d640be66992c679cc9
20251112_genspark.txt	49e3b0f4ec27042ee59fe24567ff6e7b... 96aa95a23bf2cb482bda0d4c36944701
20251112_perplexity.txt	6b726e165938ff0364d23395a9c5d098... f33fda3b75943038c7403630d0254b3c
20251114_gemini2.5proDeepResearch.txt	17a7103c3ae06fe0337eca0b23662270... 4483b1d874cc59a04e5fe3f84d71d7e1
20251115_grok_expert.txt	a5bb7a782cf61c442ed9c82448ac3bcf... fed3e01c3cb2254163e2cd07f45b6a8a

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

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