Qeios

Peer Review

Review of: "Is DeepSeek a Metacognitive AI?"

Regio Marcos Pinto Abreu Filho[1]

1. PMERJ, Brazil

Overall editorial verdict

This paper is **interesting, readable, and timely**, but it is **conceptually overextended**. Its core topic—the possible relation between DeepSeek-R1/R1-Zero reasoning behavior and metacognition—is worth discussing. The problem is that the manuscript often moves too quickly from **behavioral analogy** to **cognitive attribution**. It repeatedly treats "self-verification," "reflection," "thinking longer," and the so-called "aha moment" as near-evidence of metacognition, when the stronger and safer interpretation is:

> DeepSeek-R1-Zero exhibits **metacognition-like behavioral patterns** under reinforcement learning, but the available evidence does not establish metacognition in the psychological sense.

That distinction is the entire paper's make-or-break issue.

My rating:

doi.org/10.32388/2O4EXM

Dimension	Score
Clarity/readability	**4.0 / 5**
Timeliness	**4.2 / 5**
Conceptual ambition	**4.0 / 5**
Technical precision	**2.6 / 5**
Citation quality	**2.5 / 5**
Mathematical/computational rigor	**2.0 / 5**
Philosophical rigor	**2.8 / 5**
Educational relevance	**3.5 / 5**
Overall scientific strength	**3.0 / 5**

Qeios-style rank: 3.0 / 5.

Recommendation: Major revision, but not rejection.

It is publishable as a **perspective/commentary essay** if its claims are narrowed. It is not yet strong enough as a rigorous research article.

Reader-friendly summary

The paper argues that DeepSeek-R1 and DeepSeek-R1-Zero may show early forms of machine metacognition because reinforcement learning led them to display behaviors such as self-verification, reflection, longer reasoning chains, and the famous “aha moment.” The author then connects this to education, arguing that if AI systems are beginning to simulate metacognitive behavior, human education should prioritize metacognitive skills even more strongly.

The basic idea is good. The weakness is that the paper often uses words like **self-awareness**, **reflection**, **metacognitive control**, and **awareness of cognitive processes** too strongly. DeepSeek-R1-Zero did show interesting reasoning behaviors after reinforcement learning, and the DeepSeek technical report itself says that R1-Zero was trained through large-scale RL without SFT as a preliminary step and developed behaviors such as self-verification, reflection, and long CoTs. But this is still evidence of **observable**

reasoning strategies, not direct evidence that the model monitors its own mental states in the human metacognitive sense. (arXiv)

So the paper has a strong essayistic thesis, but it needs a stricter conceptual boundary:

> Do not ask "Does DeepSeek have metacognition?"
> Ask "Which operational components of metacognition are functionally approximated by DeepSeek-style reasoning models?"

That would make the paper much stronger.

Major strengths

1. The topic is genuinely timely

DeepSeek-R1 and R1-Zero are important because they made reinforcement-learning-based reasoning highly visible. The DeepSeek-R1 report states that R1-Zero was trained with large-scale RL without SFT as the preliminary step, while DeepSeek-R1 added cold-start data and a multi-stage training pipeline to improve readability and performance. (arXiv)

That makes the paper's topic current and relevant. A discussion of whether RL-trained reasoning models simulate metacognitive functions is not trivial. It sits at the intersection of AI reasoning, cognitive science, education, and philosophy of mind.

2. The educational angle is valuable

The final move—connecting AI reasoning advances to metacognitive education—is one of the paper's better contributions. Even if DeepSeek does not possess metacognition, its ability to simulate planning, self-checking, and strategy-switching makes a strong educational point: humans should be trained not only in content knowledge but in monitoring, self-evaluation, error correction, and adaptive strategy selection.

That is a defensible conclusion, even if the AI-metacognition claims are weakened.

3. The paper correctly distinguishes DeepSeek-R1-Zero and DeepSeek-R1

The manuscript notes that R1-Zero used RL without preliminary SFT, while R1 used cold-start SFT and subsequent RL stages. That distinction is important. The DeepSeek technical report supports this: R1-Zero is described as applying RL directly to the base model, while R1 uses cold-start data, reasoning-oriented RL, rejection sampling, and further training. (arXiv)

This is one of the technically stronger parts of the manuscript.

4. The paper correctly mentions CoT faithfulness concerns

The manuscript cites Anthropic's "Reasoning Models Don't Always Say What They Think." That is highly relevant because any claim about metacognition based on chain-of-thought must confront the possibility that CoT is not a faithful report of the model's actual computation. Anthropic's paper reports that CoTs often reveal the use of reasoning hints at low rates, often below 20%, and argues that CoT monitoring is promising but insufficient to rule out undesired behavior. (arXiv)

This is exactly the kind of caveat the paper needs. The problem is that the manuscript mentions it but does not fully integrate its implications.

Major weaknesses

1. The central concept of "metacognition" is under-operationalized

The paper quotes general definitions of metacognition, but it does not build a rigorous operational framework. It uses terms such as:

metacognition;

self-awareness;

self-reflection;

self-verification;

self-improvement;

control;

monitoring;

strategy adjustment;

awareness of cognitive processes.

These are not equivalent.

In human cognitive science, metacognition usually involves at least two separable components:

Monitoring: estimating one's own uncertainty, error likelihood, knowledge state, or performance.

Control: changing behavior based on that monitoring, such as allocating more time, seeking help, switching strategies, or revising an answer.

The paper should create a table like this:

Metacognitive component	Human meaning	Possible LLM analogue	Evidence in DeepSeek?	Strength
Error monitoring	Knowing when one might be wrong	Uncertainty-sensitive verification or longer CoT	Indirect	Weak/moderate
Strategy control	Changing approach based on monitoring	Revising solution path during generation	Behavioral examples	Moderate
Confidence calibration	Confidence tracks accuracy	Calibration curve / Brier score	Not shown	Missing
Meta-memory	Knowing what one knows	Self-estimated knowledge limits	Not shown	Missing
Introspective access	Access to own mental process	Faithful CoT	Challenged by CoT-faithfulness literature	Weak

Without this, the paper's use of "metacognition" remains rhetorically attractive but scientifically loose.

2. The manuscript anthropomorphizes the model

Several formulations are too strong:

> “self-awareness reminiscent of human thought processes”
>
> “actively engaged in the ability to reflect on its own problem-solving strategy”
>
> “awareness of its cognitive processes”
>
> “metacognitive completeness”
>
> “machines may have surpassed humans in these characteristics”

These claims are not adequately supported. They should be replaced with:

“metacognition-like behavior”

“functional analogue of monitoring/control”

“self-verification-like behavior”

“strategy-revision behavior”

“observable policy-level adjustment”

“not evidence of phenomenal awareness or human-like introspection”

The paper’s strongest safe claim is:

> DeepSeek-R1-Zero displays **functional patterns analogous to some components of metacognitive control**, especially self-verification and adaptive allocation of reasoning length.

The paper’s current stronger claims about awareness and machines surpassing humans are not justified.

3. The “aha moment” is overinterpreted

The “aha moment” is rhetorically compelling, but it is not rigorous evidence of metacognition. In the DeepSeek report, the “aha moment” is presented as an observed behavior during RL training, where the model appears to learn to allocate more reasoning effort and reconsider its approach. That is interesting. But it could be explained without invoking metacognition:

reward shaping favors longer or more structured reasoning;

models learn text patterns associated with checking and revision;

RL selects outputs with better solution trajectories;

the model imitates or amplifies reasoning-like discourse patterns already present in pretraining;

longer CoT gives more opportunity for correction without implying self-awareness.

So the paper should say:

> The “aha moment” is best interpreted as evidence of emergent strategy-revision behavior under RL, not direct evidence of metacognition in the psychological sense.

That would make the argument much stronger.

4. The paper relies too much on weak secondary sources

The references include Medium posts, LinkedIn articles, BGR, Chess.com, and the Financial Times. Some of these are acceptable for public reception/context, but they should not carry the technical argument. The technical backbone should be:

DeepSeek-R1 technical report;

DeepSeek-V3 technical report;

GRPO source paper or DeepSeekMath/GRPO paper;

OpenAI o1/o3 technical or system-card sources;

Anthropic CoT faithfulness paper;

cognitive-science metacognition literature;

RL/test-time scaling literature;

calibration and uncertainty-monitoring literature.

Right now, the paper’s technical argument leans too heavily on popular accounts. This lowers the scientific rank.

5. The paper has almost no math or formal model

This is acceptable if the paper is a commentary, but not if it is presented as a research article about metacognition in AI. The paper talks about RL, PPO, GRPO, reward-driven mechanisms, self-improvement, and strategy optimization, but it does not formalize any of these.

At minimum, it should add a simple formal distinction:

Let a model produce a reasoning trajectory:

[

z_{1:T}

]

and answer:

[

y

]

given prompt:

[

x.

]

A metacognition-like process would require a monitoring variable:

[

m_t = M(z_{1:t}, x)

]

and a control policy:

[

c_t = C(m_t, z_{1:t}, x)

]

that changes future reasoning:

[

z_{t+1:T} \sim \pi_\theta(\cdot \mid x, z_{1:t}, c_t).

]

Then the paper can ask whether DeepSeek behavior gives evidence for (m_t) and (c_t), or only for surface-level text patterns that look like monitoring/control.

This would immediately improve mathematical rigor.

Mathematical and technical rigor review

1. PPO versus GRPO explanation is broadly right but incomplete

The manuscript says PPO requires a critic/value model and that GRPO avoids the separate critic by using group-based reward baselines. That is broadly aligned with the DeepSeek report, which says GRPO avoids a critic model and estimates the baseline from group scores. (arXiv)

But the paper should avoid saying GRPO is simply "more effective" as a universal claim. It is more accurate to say:

> GRPO is computationally attractive for large-scale reasoning RL because it removes the need for a critic model of comparable size and uses relative group scores to estimate the baseline.

The current language risks implying that GRPO is generally superior to PPO across tasks, which is not established.

2. "RL over SFT" is oversimplified

The paper repeatedly suggests DeepSeek prioritizes RL over SFT. This is true for R1-Zero, but not for R1 as a whole. DeepSeek-R1 uses cold-start data, two SFT stages, RL stages, rejection sampling, and preference alignment. The DeepSeek report explicitly says R1 incorporates cold-start data and a multi-stage pipeline. (arXiv)

Correction:

> DeepSeek-R1-Zero is the cleanest case for studying RL-first reasoning. DeepSeek-R1 is better described as an RL-centered but hybrid pipeline, not a pure replacement of SFT.

3. "AI can learn reasoning without human assistance" is too strong

DeepSeek-R1-Zero did not learn in a vacuum. It started from DeepSeek-V3-Base, which was pretrained on massive human-generated and synthetic corpora. It also used hand-designed reward structures and benchmark-style tasks. So the phrase "without human assistance" is misleading.

Better:

> DeepSeek-R1-Zero shows that a pretrained base model can acquire stronger reasoning behavior through RL without a preliminary SFT reasoning dataset.

This is precise.

4. “Self-improvement” is ambiguous

The paper uses self-improvement several times. In AI safety and ML, “self-improvement” can imply a model autonomously improving its own weights or training process. DeepSeek-R1-Zero did not autonomously rewrite itself; it was optimized by an external RL training loop.

Better term:

> training-induced improvement under reinforcement learning

or:

> policy improvement induced by RL optimization

Avoid “self-improvement” unless explicitly defined as “improvement of model behavior through an externally run RL loop.”

5. “Metacognitive completeness” is unsupported

The phrase “metacognitive completeness” is not defined and is too speculative. Remove it or define it formally. The paper does not have enough evidence to discuss whether machines may achieve complete metacognition.

Suggested replacement:

> “Whether such systems could ever satisfy stronger criteria for machine metacognition remains an open theoretical and empirical question.”

6. “Necessary component of coherent information processing” is overclaim

The conclusion says metacognitive control through RL indicates that it is a necessary component of coherent information processing. That is not shown. DeepSeek’s behavior may suggest

monitoring/control-like mechanisms are useful, not necessary.

Better:

> "These results suggest that monitoring- and control-like behaviors may be useful for coherent long-form reasoning, but they do not establish necessity."

7. "Machines may have surpassed humans in these characteristics" is unjustified

The paper says the extent to which machines may have surpassed humans in these characteristics remains open. Even as speculation, this is weak because the paper has no metric comparing machine and human metacognition.

Remove or reframe:

> "It remains unknown whether machine systems can match humans on calibrated metacognitive monitoring and adaptive control across open-ended contexts."

Logical structure audit

Current logic

The manuscript's implicit reasoning is:

Metacognition involves monitoring and controlling one's cognitive processes.

DeepSeek-R1-Zero shows self-verification, reflection, and longer reasoning.

Therefore DeepSeek may show metacognition-like behavior.

Therefore DeepSeek may represent a new stage in machine reasoning.

Therefore human education should prioritize metacognition.

This is partly valid, but step 2 → step 3 needs tightening.

Stronger version

The rigorous argument should be:

Human metacognition can be decomposed into operational components: monitoring, control, error detection, strategy revision, and confidence calibration.

Some DeepSeek-R1-Zero behaviors—self-verification, revision, extended reasoning, and adaptive reasoning length—are **functional analogues** of some monitoring/control components.

However, available public evidence does not show calibrated introspective access, faithful self-report, conscious awareness, or human-like metacognition.

Therefore DeepSeek should be treated as evidence for **metacognition-like control behavior**, not metacognition itself.

This still matters educationally because the human advantage may shift from content recall to reflective strategy control, critical thinking, verification, and adaptive learning.

That version is much stronger and less vulnerable.

Citation and source-quality audit

Strong references

The strongest technical references are:

DeepSeek-R1 technical report.

DeepSeek-V3 technical report.

Anthropic CoT faithfulness paper.

OpenAI o3/o4-mini announcement, if used only for descriptive context.

Flavell on metacognition.

DeepSeek-V3 is relevant because it provides the base model architecture/context: 671B total parameters, 37B activated per token, MLA, DeepSeekMoE, auxiliary-loss-free load balancing, and 14.8T pretraining tokens. (arXiv)

OpenAI o3/o4-mini is relevant only as contextual evidence that reasoning models and RL-trained tool use became a broader industry direction; OpenAI states that o3 and o4-mini were trained to reason about when/how to use tools and that o3 development continued scaling RL. (OpenAI)

Weak references

These should be removed or demoted:

LinkedIn article on DeepSeek.

Medium article on DeepSeek metacognition.

BGR article on “aha moment.”

Chess.com article, unless only used as a public-historical illustration.

Popular press sources, unless clearly marked as journalistic context.

Missing references

The paper needs stronger cognitive-science and ML references on:

Nelson & Narens metacognition framework;

confidence calibration;

uncertainty monitoring;

self-regulated learning;

metacognitive control;

chain-of-thought faithfulness;

test-time scaling;

process supervision;

RLHF/PPO technical source;

GRPO/DeepSeekMath source.

The absence of Nelson & Narens is especially noticeable. A paper about metacognition should not rely only on Flavell and general definitions.

Reader and magazine friendliness

The paper is easy to read, but it currently reads more like a **magazine essay** than a scientific article. That is not necessarily bad for Qeios, but the genre should be clear.

To make it more reader-friendly and scientifically safer, add a box near the start:

What this paper argues: DeepSeek-R1-Zero shows behaviors that functionally resemble some components of metacognitive control, especially self-verification and strategy adjustment.

What this paper does not argue: That DeepSeek is conscious, self-aware, introspective, or metacognitive in the full human psychological sense.

Why it matters: If AI systems increasingly simulate monitoring/control behaviors, education should prioritize human metacognitive skills: error monitoring, self-regulation, critical evaluation, and adaptive strategy use.

This would prevent misreadings.

Section-by-section audit

Abstract

The abstract is clear but too strong.

Problem phrase:

> "it is evident that the interactions between the system's monitoring and control processes are both present..."

This is not evident. It is inferred.

Better:

> "The published behavior of DeepSeek-R1-Zero suggests functional analogues of monitoring and control, such as self-verification, revision, and adaptive allocation of reasoning length."

Problem phrase:

> "simulate behaviours based on self-reflection"

Better:

> "simulate behaviors that resemble self-reflection at the output or reasoning-trace level."

Section 1

Good background, but too many broad statements about “new generation of AIs.” Also, the cost comparison with ChatGPT/OpenAI should be handled cautiously because direct cost comparisons are often incomplete. DeepSeek-V3 reports 2.788M H800 GPU hours, but this does not automatically mean that a total all-in cost comparison with OpenAI systems is straightforward. (arXiv)

Section 2

The metacognition section needs the most strengthening. It should include a metacognitive framework with operational components. It should distinguish:

cognition;

metacognition;

self-regulation;

consciousness;

introspection;

confidence calibration.

Right now, these are blurred.

Section 3

The DeepSeek-R1/R1-Zero distinction is useful. But the paper should be more precise:

R1-Zero: RL directly on base model, no preliminary SFT.

R1: cold-start SFT + RL + rejection sampling + further SFT/RL.

R1’s performance comparability to OpenAI-o1-1217 should be described as benchmark-specific, not global equality.

Section 4

The RL/GRPO discussion is probably the paper’s best technical section, but it should include either a formula or a more precise explanation.

A simple description:

[

A_i = r_i - \frac{1}{G}\sum_{j=1}^{G} r_j

]

where each response (i) in a group is reinforced according to its reward relative to the group baseline. That would help readers understand the core idea of GRPO without overcomplicating the paper.

Section 5

The conclusion is too speculative. It should remove:

“there is no doubt”

“awareness of cognitive processes”

“necessary component”

“metacognitive completeness”

“machines may have surpassed humans”

Replace with:

> “DeepSeek-R1-Zero provides a compelling case of metacognition-like behavior at the level of generated reasoning traces and training-induced strategy revision. However, current evidence does not establish human-like metacognition, introspective awareness, or faithful self-monitoring.”

Suggested revised thesis

The whole paper should be reframed around this:

> **DeepSeek-R1-Zero does not prove machine metacognition, but it provides a useful test case for distinguishing three levels: apparent metacognitive language, functional metacognition-like control, and full psychological metacognition. The available evidence supports the first two only weakly-to-moderately, and does not support the third.**

This would make the paper conceptually rigorous.

Suggested ranking by article type

Article type	Fit
Short perspective essay	**Good: 4.0 / 5**
Qeios conceptual commentary	**Acceptable: 3.5 / 5**
Rigorous AI/cognitive science review	**Weak: 2.5 / 5**
Technical ML paper	**Not suitable: 1.8 / 5**
Education/opinion piece	**Good: 3.8 / 5**

The paper should be submitted/framed as a **perspective/commentary**, not a technical research article.

Most important corrections

Correction 1

Replace:

> DeepSeek demonstrates metacognition.

With:

> DeepSeek demonstrates metacognition-like functional behaviors, especially self-verification, strategy revision, and adaptive reasoning-length allocation.

Correction 2

Replace:

> awareness of its cognitive processes.

With:

> observable output-level patterns resembling monitoring and control.

Correction 3

Replace:

> AI can learn reasoning without human assistance.

With:

> a pretrained base model can improve reasoning performance through RL without a preliminary supervised reasoning dataset.

Correction 4

Replace:

> RL is necessary for coherent information processing.

With:

> RL appears to incentivize behaviors that resemble monitoring and control in some reasoning tasks.

Correction 5

Replace:

> machines may have surpassed humans in metacognition.

With:

> no current evidence establishes that machines match humans in calibrated metacognitive monitoring across open-ended contexts.

Final assessment

This is a **promising but overclaiming conceptual manuscript**. Its topic is good, its writing is accessible, and its educational motivation is valuable. However, its central concept—metacognition—is not operationalized with enough rigor, and several claims anthropomorphize DeepSeek beyond what the technical evidence supports.

With moderate-to-major revision, it could become a good Qeios perspective article. The revision should not try to prove that DeepSeek has metacognition. It should instead make a more precise and valuable claim:

> DeepSeek-style RL reasoning models force us to distinguish between human metacognition, functional metacognition-like control, and mere metacognitive language.

That would be a real contribution.

Final rank: 3.0 / 5

Recommendation: Major revision, with strong potential as a perspective/commentary.

Declarations

Potential competing interests: No potential competing interests to declare.