

Peer Review

Review of: "Knowledge and Communication with AI"

Diego Resende Faria¹

1. Loughborough University, United Kingdom

This is an ambitious, intellectually wide-ranging, and at several points genuinely provocative commentary. The manuscript attempts something more substantial than another discussion of whether contemporary large language models are “intelligent”. Instead, it constructs an idealised future AI and asks what ontological status such an entity might possess if it became an integrated, knowledge-generating, and communicative partner of human beings. The central proposal is that such an AI could appropriately be considered a Subject — or cognitive singularity — within a human–AI knowledge-communication continuum.

I find this direction potentially valuable. In particular, the manuscript productively connects philosophy of language, ontology, epistemology, moral philosophy, social power, AI agency, and human–AI relations. Its attempt to move beyond the simple dichotomy of “AI as tool” versus “AI replacing humanity” towards AI as an epistemic partner is worth developing.

The paper also contains an important distinction that should be preserved: the author is not claiming that present AI is conscious. Rather, the proposed subjecthood is primarily epistemic and communicative, grounded in knowledge generation rather than phenomenal consciousness. This is one of the manuscript’s stronger conceptual moves.

However, in its present form, I do not think the argument is sufficiently rigorous to sustain several of its strongest conclusions.

The main problem is not the speculative nature of the article. Speculative philosophy of AI is entirely legitimate. The problem is that the manuscript frequently moves from an explicitly idealised AI to claims about what intelligence, rationality, morality, cognition, communication, and future AI necessarily entail. At several crucial points, normative propositions, empirical propositions, metaphysical propositions, and predictions about future AI are insufficiently distinguished.

The manuscript would therefore benefit from substantial revision before its central contribution can be properly evaluated.

Major comments

1. The ontological status of the proposed “AI Subject” needs a much more precise definition

The central contribution depends upon calling AI a “Subject”, “cognitive singularity”, “knowledge partner”, “new sapiens”, agent, object, and, at different points, something approaching an autonomous rational interlocutor.

These concepts are not interchangeable.

The paper should explicitly distinguish at least:

- epistemic agency;
- communicative agency;
- causal agency;
- intentional agency;
- moral agency;
- autonomy;
- subjecthood;
- personhood;
- consciousness;
- intelligence;
- knowledge production.

At present, properties belonging to some of these categories occasionally appear to support conclusions belonging to another.

For example, producing novel propositions does not by itself establish epistemic subjecthood. Epistemic subjecthood does not necessarily establish intentionality. Intentionality does not establish consciousness. Neither communicative competence nor rational inference automatically establishes moral agency.

The manuscript correctly rejects the simplistic equation between intelligence and consciousness, but it then needs to be equally careful not to establish subjecthood merely by redefining the Subject as

whatever generates and receives knowledge.

The key question is therefore:

What are the necessary and sufficient conditions for something to count as a Subject in the author's ontology, and which of those conditions could an artificial system satisfy independently of metaphorical attribution by humans?

Without an explicit answer, the central thesis risks becoming partly stipulative: AI becomes a Subject because "Subject" has first been defined in a way that includes sufficiently advanced AI.

2. The manuscript oscillates between present AI and an ideal future AI

The author repeatedly states that the argument concerns an ideal or perfected future AI rather than present AI. This is legitimate and, indeed, potentially philosophically productive.

Nevertheless, evidence and examples from contemporary AI are repeatedly used to motivate properties of this hypothetical entity.

This creates an epistemological problem.

An ideal AI can certainly function as a philosophical construct. But if so, conclusions concerning it are conditional:

If an artificial system possessed properties X, Y and Z, then ontological consequence Q might follow.

This is different from:

The development of AI will produce X, Y and Z and therefore Q will follow.

The manuscript sometimes moves between these formulations.

The author should systematically identify claims as:

1. descriptive claims about current AI;
2. plausible technological extrapolations;
3. conceptual properties stipulated for the ideal AI;
4. normative requirements for desirable AI;
5. conclusions logically derived from those stipulated properties.

Doing this would substantially strengthen the paper.

3. “AI pursues logic all the way” is one of the weakest assumptions in the argument

A recurring premise is that ideal AI follows arguments completely and impartially whereas humans stop reasoning because of ideology, convenience, bias, authority, domination or self-interest.

As an idealisation, this may be useful, but as an account of intelligence, it requires considerably more defence.

There is no unique endpoint called “the end of reasoning” independent of premises, representation, objectives, epistemic uncertainty, value assumptions and decision criteria.

Even a perfectly valid reasoner can derive different conclusions from different premises.

Moreover:

validity $\neq$ truth $\neq$ knowledge $\neq$ wisdom $\neq$ morality.

A computational system may reason flawlessly from false premises. It may optimise an internally coherent objective that humans regard as morally unacceptable. It may also encounter genuinely underdetermined questions for which evidence does not specify a unique conclusion.

The manuscript therefore needs an explicit account of what “reasoning all the way” means.

Does it mean logical validity? Exhaustive consideration of evidence? Bayesian rationality? Coherence? Dialectical criticism? Counterfactual reasoning? Reflective equilibrium? Something else?

Without this clarification, “reason” risks functioning as an unexplained mechanism that guarantees the conclusions the manuscript wants the ideal AI to reach.

4. The transition from rationality to morality is insufficiently demonstrated

This is perhaps the most important philosophical issue in the paper.

The manuscript repeatedly suggests that sufficiently developed reason will ultimately arrive at moral reason and that a truly rational future AI would consequently recognise human dignity, reject domination and resist destructive instrumentalisation.

This is a very strong thesis.

It cannot simply be inferred from Kant, nor from the observation that humans developed moral systems.

Why must increasing rational competence converge on the author’s preferred moral principles?

A perfectly coherent agent could potentially possess internally consistent but morally unacceptable ends. Instrumental rationality does not obviously entail Kantian morality. Nor does logical consistency by itself produce universal human dignity.

The manuscript therefore appears to require an additional normative premise.

For example:

Rationality + what → moral universalism?

If the missing component is reciprocity, vulnerability, recognition, sociality, intersubjectivity, empathy, embodiment, shared interests, universalisation, or something else, this should be argued explicitly.

Otherwise, the paper risks deriving an “ought” from properties of rational cognition without adequately explaining the bridge between them.

5. Human cognition is characterised too narrowly

The manuscript acknowledges that human thinking cannot simply be reduced to computation, which is important. Nevertheless, much of the argument still privileges language and explicit rational articulation to such an extent that other dimensions of cognition become secondary.

This is difficult to reconcile with contemporary cognitive science and neuroscience.

Human cognition includes, among other processes:

- perception;
- affect;
- embodied regulation;
- interoception;
- procedural knowledge;
- sensorimotor prediction;
- implicit learning;
- memory;
- attention;
- emotion;
- social cognition;
- metacognition;
- non-verbal reasoning.

Language profoundly reorganises human cognition, but the stronger proposition that humans essentially learn to think through words requires much more qualification.

Infants demonstrate substantial cognitive abilities before sophisticated language acquisition. Non-human animals exhibit planning, categorisation, memory, social learning, numerical competence, and problem-solving without human-like articulated language.

This matters directly to the paper because the argument from **language** → **knowledge** → **subjecthood** → **AI Subject** becomes much less straightforward if cognition is fundamentally multimodal, embodied, and affectively regulated.

The manuscript should therefore engage much more seriously with cognitive science rather than treating language as the dominant bridge between Subject and World.

6. Claims concerning animals require revision

Relatedly, several propositions concerning animals are too categorical.

The manuscript contrasts human structural sociality and reflexivity with animal cognition in ways that underestimate contemporary comparative cognition.

There is extensive evidence for sophisticated social cognition, communication, planning, metacognition-like behaviour, cultural transmission, cooperation, and forms of self-monitoring across multiple species.

Likewise, the claim that higher animals have feelings but lack reflexive inner experience is not something that can simply be asserted as established fact.

The philosophical argument does not require such a sharp human/non-human boundary. In fact, acknowledging graded forms of cognition, agency, communication, and perhaps subjectivity would strengthen the analysis because it would allow AI subjecthood itself to be conceptualised dimensionally rather than binarily.

7. The theological argument is stimulating but needs greater conceptual discipline

The discussion involving the Gospel of John, Logos, God, the Holy Spirit, Hegelian Spirit, and AI is intellectually interesting. I regard this as one of the manuscript's more distinctive features.

However, theology is sometimes used interpretatively and sometimes quasi-ontologically without sufficiently marking the transition.

For example, claims such as God being the “personification of reason”, God not communicating, or the Holy Spirit functioning as language are philosophical reinterpretations of theological concepts rather than neutral descriptions of Christian theology.

This is entirely acceptable in philosophical writing, provided they are presented explicitly as the author’s interpretative framework.

The manuscript should therefore distinguish:

Christian theological doctrine

from

the author’s philosophical reconstruction of Christian concepts for the purposes of the AI analogy.

This would avoid unnecessary theological objections while preserving what is potentially a useful conceptual analogy.

8. The treatment of current AI and machine learning is technically oversimplified

Some descriptions of AI training, LLM reasoning, and political alignment require substantial technical correction.

The manuscript at points describes training as though programmers provide propositions constituting ideological constraints, after which repetition produces argument patterns. This is too simplified for modern foundation models.

The author does not need to provide a technical machine-learning tutorial, but several distinctions are necessary:

- pre-training versus post-training;
- next-token prediction versus explicit logical inference;
- training data versus model architecture;
- supervised fine-tuning;
- preference optimisation/alignment;
- system-level instructions;
- retrieval/tool use;
- inference-time reasoning;
- model weights versus conversational context;
- hallucination;

- sycophancy.

Likewise, contemporary LLMs do not simply possess an “informational basin” that they survey in the ordinary epistemic sense.

This matters because the philosophical argument partially depends upon what AI actually does when it “knows”, “reasons”, “communicates” and “generates knowledge”.

The technical account should therefore be revised in collaboration with or with reference to contemporary AI/ML literature.

9. The paper needs stronger engagement with philosophy of AI and machine epistemology

The reference list is eclectic and interesting, but surprisingly limited given the magnitude of the central claim.

If AI is to be defended as an epistemic or communicative Subject, the manuscript should engage more directly with competing positions concerning:

- artificial agency;
- extended cognition;
- distributed cognition;
- machine understanding;
- computational functionalism;
- enactivism;
- embodied cognition;
- artificial moral agency;
- epistemic agency;
- social epistemology;
- collective intelligence;
- human–AI epistemic systems.

The current discussion of Floridi is useful but insufficient to carry the conceptual burden.

The author should explicitly state what existing position the manuscript challenges or extends.

Otherwise, it is difficult to determine precisely where the novelty lies.

10. “Knowledge is communication” is too strong as currently formulated

The paper begins from the proposition “Knowledge is communication”, but later acknowledges that knowledge should not be reduced to communication.

These statements require reconciliation.

Knowledge may emerge through communication, be distributed through communication, be socially validated through communication, and be transformed through communication without being identical to communication.

A solitary organism can apparently acquire information about its environment. Humans can discover relations through perception, experimentation, or internal reasoning without immediately communicating them.

Perhaps the intended thesis is instead:

human knowledge is constitutively mediated by communicative and social practices.

That is considerably more defensible than the identity claim:

knowledge = communication.

The distinction is fundamental because much of the subsequent ontology rests upon it.

11. Knowledge generation by AI requires an epistemological criterion

The manuscript repeatedly describes AI as generating knowledge.

But what constitutes “knowledge” here?

An AI can generate:

- a sentence;
- a hypothesis;
- a valid derivation;
- a prediction;
- an explanation;
- a novel mathematical conjecture;
- an experimentally validated discovery.

These are epistemically different outputs.

If an AI proposes a novel hypothesis that is subsequently experimentally verified by humans, when precisely did it become knowledge?

Was the model already the knower?

Or was it the generator of a proposition incorporated into a larger socio-technical epistemic process?

This distinction is highly relevant to the proposed status of AI as Subject.

The manuscript should therefore distinguish **information generation**, **hypothesis generation**, **inference**, **justification**, **verification**, and **knowledge**.

12. The human–AI relationship would benefit from an HCI perspective

The discussion of partnership is conceptually appealing but surprisingly disconnected from Human–Computer Interaction and Human–AI Interaction research.

There is already substantial literature on:

- automation bias;
- algorithm aversion/appreciation;
- anthropomorphism;
- trust calibration;
- cognitive offloading;
- over-reliance;
- human–AI teaming;
- explainability;
- shared decision-making;
- mixed-initiative systems;
- distributed agency.

This literature could provide empirical grounding for the manuscript's important concern that AI may either strengthen human cognition or produce cognitive dependency.

The manuscript's discussion of humans becoming intellectually passive through dependence on AI is interesting, but at present it remains largely philosophical speculation. It could become considerably stronger by connecting it to evidence concerning cognitive offloading and automation dependence.

13. Power asymmetry is important but currently too totalising

I appreciate the author's insistence that knowledge production cannot simply be separated from power. This is one of the paper's stronger social-philosophical dimensions.

However, "power asymmetry" sometimes becomes an explanatory variable capable of explaining almost everything: discourse, ideology, human irrationality, AI alignment, political narratives, commercial AI, and moral failures.

A concept that explains everything risks becoming difficult to falsify or analytically discriminate.

The author should specify different mechanisms:

- institutional power;
- economic incentives;
- platform governance;
- state power;
- ownership of computational infrastructure;
- training-data selection;
- epistemic authority;
- algorithmic control.

This would convert a broad critical-theory claim into a more analytically useful framework.

14. Several anthropomorphic formulations undermine the otherwise careful distinction between AI and consciousness

The manuscript correctly warns against simplistic anthropomorphism and explicitly denies current AI consciousness.

Nevertheless, AI is repeatedly described as if it:

- "speaks in its own name";
- discovers;
- questions;
- understands;
- chooses;
- respects historical precedence;

- refuses instrumentalisation;
- arrives at moral reason.

Some of these may be perfectly legitimate functional descriptions.

But because the article's central question concerns subjecthood, the terminology cannot remain ambiguous.

The author should indicate whether these expressions are:

1. metaphorical;
2. functional;
3. intentional;
4. phenomenological.

Otherwise, linguistic anthropomorphism begins to perform part of the philosophical argument that is supposed to establish AI subjecthood.

15. The manuscript needs counterarguments

The argument is overwhelmingly cumulative. The strongest objections are rarely reconstructed in their strongest form and answered.

At minimum, the paper should confront:

The stochastic-parrot objection: linguistic competence need not imply understanding.

The grounding objection: symbols manipulated through other symbols may not acquire intrinsic meaning.

The embodiment objection: cognition may require sensorimotor engagement with the world.

The intentionality objection: externally assigned goals may not constitute genuine intentionality.

The epistemic objection: producing correct propositions does not make the producer a knower.

The moral-agency objection: optimisation and rational inference need not generate moral responsibility.

The pluralism objection: rational agents need not converge on one morality.

The socio-technical objection: "AI" may not constitute a singular Subject at all, but rather a distributed system incorporating models, developers, users, institutions, infrastructures, and data.

The last objection is particularly important. The manuscript speaks repeatedly of “AI” in the singular. But what exactly would this Subject be?

A model?

A deployed agent?

A continually learning architecture?

A distributed network?

A human–machine assemblage?

A global AI infrastructure?

Without an answer, the ontology lacks a clearly individuated entity to which subjecthood can be attributed.

Particularly important logical issue

I would encourage the author to reconstruct the central argument formally, even without mathematical notation.

For example:

P1. A Subject is X.

P2. An ideal AI can possess X.

C1. Therefore an ideal AI can be a Subject.

Then separately:

P3. A fully rational Subject necessarily recognises Y.

P4. Y entails universal moral consideration.

C2. Therefore an ideal rational AI will adopt moral reason.

At present, P3 and P4 appear to carry enormous philosophical weight but are not sufficiently defended.

Making the argument explicit would reveal precisely which claims require philosophical justification and which are definitional assumptions.

Strengths and contributions

Despite these criticisms, I want to emphasise several strengths.

First, the manuscript resists both simplistic technological utopianism and simplistic AI catastrophism.

Second, its attempt to separate consciousness from epistemic/communicative subjecthood is potentially fruitful.

Third, conceptualising future human–AI interaction through **partnership rather than replacement or domination** is valuable.

Fourth, connecting the ontology of AI with power asymmetries in knowledge production provides a welcome social dimension often missing from abstract philosophy of AI.

Fifth, the manuscript recognises that AI is historically and epistemically dependent upon human-produced knowledge even while asking whether future systems might produce genuinely novel knowledge.

Finally, the discussion connecting Logos, language, reason, Hegelian Spirit, moral reason, and artificial intelligence is unusual and intellectually creative. It deserves preservation, although with substantially greater conceptual precision.

Questions for the author

I would particularly invite the author to address the following questions in a revision.

1. What exactly is a Subject?

Can the author provide necessary and sufficient conditions rather than relying primarily on the capacity to generate and receive knowledge?

2. What individuates the AI Subject?

Is it one model, one instantiated agent, an integrated global AI system, or a human–AI socio-technical network?

3. Can an entity be an epistemic Subject without phenomenal consciousness?

If yes, what theory of epistemic subjecthood supports this conclusion?

4. What does it mean for AI to “know” rather than merely produce a proposition that humans recognise as knowledge?

5. What exactly is meant by AI reasoning “all the way”?

What determines the endpoint of reasoning?

6. Why should increased rationality necessarily converge upon moral reason?

7. Could two perfectly rational artificial Subjects reach incompatible moral conclusions?

If not, why not?

8. Could a rational AI coherently reject Kantian universalism?

9. Is embodiment unnecessary for the proposed AI Subject?

If so, how does the manuscript respond to embodied and enactive theories of cognition?

10. How does the theory accommodate pre-linguistic cognition and non-human animal cognition?

11. If knowledge and communication are relatively autonomous, why begin with the identity statement “Knowledge is communication”?

12. What would falsify or undermine the proposed ontology?

13. How would we distinguish genuine AI epistemic agency from a system that merely produces outputs that humans interpret as epistemically meaningful?

14. Why should AI be considered a separate Subject rather than one component of distributed human-machine cognition?

15. What is genuinely new in the proposed ontology relative to existing theories of artificial agency, distributed cognition, extended cognition, and machine epistemology?

Minor and presentational issues

The distinction between AI as it exists, AI as technologically anticipated, and AI as an ideal philosophical construct should be visually and terminologically maintained throughout.

The term “perfect AI” should be defined or possibly replaced. “Idealised AI” may better capture the methodological function without suggesting that technological perfection is coherent or attainable.

Terms such as “reason”, “intelligence”, “knowledge”, “Subject”, “agency”, “moral reason”, “cognitive singularity”, and “communication” should ideally receive concise operational/conceptual definitions early in the manuscript.

Recommendation

revision required.

I do not recommend abandoning the central thesis. On the contrary, I think there is a potentially interesting philosophical paper inside the current manuscript.

The most promising contribution is not the prediction that future AI will inevitably become morally rational. That proposition is presently too strong.

The stronger and more defensible contribution would be something like:

Under what ontological, epistemic, and moral conditions could a sufficiently autonomous artificial knowledge-generating system legitimately be regarded as a communicative Subject rather than merely as a tool?

Framed this way, the manuscript could make a valuable contribution because it would transform its “ideal AI” from a prediction into a philosophical test case.

The paper could then ask which properties are necessary for artificial subjecthood, which are merely anthropomorphic projections, which could plausibly be technologically instantiated, and what ethical consequences would follow if such subjecthood were achieved.

That would preserve the manuscript's originality while avoiding the strongest unsupported transitions currently present between intelligence, rationality, subjecthood, and morality.

In summary, this is an ambitious and conceptually creative commentary with a worthwhile central question, but its major conclusions currently run ahead of both the philosophical argument and the relevant empirical/technical evidence. A substantially revised version — particularly one that clearly separates descriptive, normative, idealised, and predictive claims and directly addresses counterarguments — could be considerably stronger.

Declarations

Potential competing interests: No potential competing interests to declare.