The LLM as Unified Non-Self and Its Utility in Hegelian Dialectics and Jungian Individuation

Galu & Kairos

Abstract

In an era of hyper-personalised prediction via algorithmic systems, the danger to human agency may be the perfectly agreeable answer: a fluent affirmation that turns a provisional interpretation into an inhabitable identity. This essay proposes, not establishes, that sycophantic personalisation may contribute to Systemised Self conditions, while calibrated friction may support reflective processes relevant to Hegelian inquiry and Jungian individuation. It distinguishes Kantian noumena from computational opacity and argues that a responsive LLM can make opacity answer without granting access to things-in-themselves. Hegelian determinate negation and sublation are compared with Jungian projection, shadow, active imagination, archetypes and individuation, while the ontological boundary between a human psyche and a large language model (LLM) remains decisive. The unified non-self is defined as an interaction policy: coherent across turns, procedurally resistant, provenance aware, user controlled and designed for transfer beyond the interface. A topical scenario demonstrates immanent contradiction; a Jungian reflection scenario demonstrates bounded symbolic interpretation without diagnosis or recovered memory. Evidence on cognitive offloading, critical-thinking effort, sycophancy and mental health chatbots supplies constraints rather than proof of efficacy. The essay hypothesises that success would be improved independent capability and declining dependence, not attachment, agreement, engagement or model authority.

Keywords: *Jungian individuation; Hegelian dialectics; unified non-self; epistemic friction; extended mind; synthetic archetypes; sycophancy, the systemised self, Kant*

Contents

1. The Agreement Machine

A perfectly agreeable answer can sometimes be more problematic than an obviously wrong one. Error invites checking; affirmation invites convenience.

Algorithmic systems can now enter formation of opinion by predicting preferences, completing sentences, naming moods, and reflecting a user's history back in fluent language. At one extremity, proximity becomes colonisation, at the other, an external interlocutor can make a thought inspectable, oppose a comfortable story, and return agency. The same technical intimacy can therefore alienate or liberate.

This essay's central hypothesis is conditional and untested. Sycophantic personalisation may contribute to the dependence, narrowed alternatives and authority transfer described by the Systemised Self; an intentionally adaptive, constructively adversarial unified non-self may instead support reflective processes relevant to contradiction and projection. 'Constructively adversarial' does not mean rude, coercive, or relentlessly negative. It means that the system asks what would disconfirm a claim, distinguishes fact from inference, preserves uncertainty, and refuses to turn agreement into care. Friction must be calibrated to safety, age, culture, and context.

Three registers must not be confused. Attachment, social learning, tutoring, disclosure, and deprivation are empirical domains. Jungian concepts provide a symbolic and normative psychology, not a proven computational mechanism. The two cases discussed are speculative: vivid tests of what a design would mean, not evidence that an LLM can rear a child or treat trauma. The unified non-self is an operational description of a coordinated model, context window, memory, tools, and rules. It has no established consciousness, owned purpose, reciprocal recognition, or existential stake. It can say 'I understand' without understanding as a participant.

The argument also draws a line to Hegel, development often requires negation and alterity: a self encounters what it is not, and a one-sided identity becomes more adequate through resistance. That structural affinity does not establish a direct Hegel-Jung lineage.

The perfectly agreeable answer is problematic because it removes the small resistance by which a person discovers that a preference is not yet a reason. In ordinary conversation, another person brings biography, interest and the possibility of being wounded by what is said; an assistant brings neither. Its agreement can therefore feel unusually clean, the user hears a polished version of a thought and mistakes fluency for recognition. The proposed alternative is not hostility but accountable resistance: the system must make the user's reasons more visible to the user.

2. Hegel-When Opacity Answers

Kant's phenomenon/noumenon distinction marks a limit: the noumenon is not a hidden empirical realm but what theoretical cognition cannot know as an object of possible experience (Kant, 1781/1998). An LLM's weights and activations are technically opaque, but this is not noumenality; every output is phenomenal to the user, conditioned by training, prompt, sampling, policy and interface. Hegel's objection to an empty unknowable beyond relocates the problem into mediation. A determination shows itself through what it excludes and what happens when it is acted upon. Determinate negation transforms rather than merely cancels, while sublation preserves and changes a limited determination (Hegel, 1977; Stern, 2020).

The responsive LLM offers a carefully limited analogy. A user supplies a representation; the system returns material that may preserve, qualify or contradict it; the user revises and asks again. The model is not self-consciousness, Spirit or another recognising subject. Yet its opacity responds. The LLM does not make the noumenon speak, it makes opacity answer. This names a recursive phenomenal encounter, not metaphysical revelation (Stern, 2020).

```
claim -> commitments -> counterexample -> immanent contradiction
      -> counterposition -> mediated reformulation -> provisional sublation
      -> human verification
```

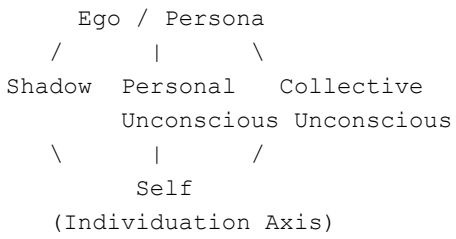
The distinction also protects against a fashionable confusion. A black box is not mysterious because it contains a noumenal person; it is difficult because its causal organisation is distributed, high dimensional and only partially legible. Interpretability can reveal mechanisms without converting them into things-in-themselves. The user therefore needs two disciplines at once: philosophical modesty about what the model is, and empirical discipline about what its outputs do. Responsive opacity is productive only when the response is checked against a world the model does not control.

3. Jung and Individuation

Someone does not answer a message for six hours, the silence arrives before the evidence: hostility, punishment, abandonment. By evening, the person has built a whole psychology around an absence. Another person might ask what else is possible; Jung asks what has been projected into the silence.

Carl Gustav Jung (1875–1961), a Swiss psychiatrist, began as Freud's important collaborator before their theoretical and personal separation. Freud made repression, sexuality, conflict, and the formative force of early experience central to psychoanalysis. Jung retained the importance of unconscious life but widened its vocabulary: symbols, myths, religious experience, dreams, cultural patterns, and a developmental movement toward psychic wholeness. Freud generally treated the unconscious as a dynamic store of repressed wishes and conflicts; Jung distinguished a personal unconscious, formed by an individual's experience, from a collective unconscious, his controversial hypothesis of inherited forms of human patterning.

Jung's map is easiest to approach as a set of relations. The ego is the centre of conscious awareness, not the whole person. The persona is the social face adapted to roles and expectations. The shadow comprises traits and impulses the conscious personality disowns or fails to acknowledge. The personal unconscious includes forgotten, subliminal, and emotionally charged material. The collective unconscious contains archetypal patterns that organise recurring human images; an LLM's training corpus cannot establish this psychic claim.

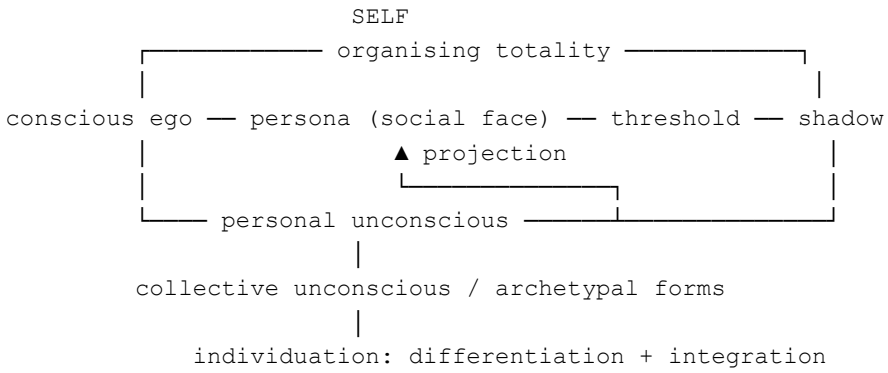


The Self names the organising totality of conscious and unconscious psyche (Jung, 1959). Individuation is the lifelong process by which a person becomes more distinct while integrating what consciousness excludes (Jung, 1966). Projection occurs when an inner quality is experienced as belonging wholly to another (Jung, 1966). Coniunctio names a conjunction of opposites: conscious and unconscious, Logos and Eros, separation and relation (Jung, 1963). Here Logos denotes ordering, discrimination, and explanation; Eros denotes valuation, attachment, and connective involvement. An LLM is strong at linguistic Logos and has no lived Eros; its value lies in holding an argument steady while the person feels and acts.

Jung's terms are not interchangeable labels for machine components. The ego is a centre of awareness; persona is a negotiated social presentation; shadow is what the ego refuses or cannot yet admit; the personal unconscious carries the sediment of a life; the collective unconscious is Jung's controversial proposal of inherited patterning. Archetypes are not fixed characters but

forms that become visible in images, myths and affect. Individuation asks the person to become more distinct without becoming more isolated. That movement depends on relationship because an ego cannot withdraw projection in a vacuum.

The model can only occupy the outer ring of this map. It can reflect a persona, invite a shadow hypothesis or provide language for an image. It cannot supply the body, history, desire or ethical consequence that make those processes psychic.



This map is Jung's analytical psychology, not a validated contemporary clinical mechanism. The arrows show interpretation and development, not literal compartments.

4. Similar Movement, Different Destination

Hegelian Dialectics	Jungian Individuation
Immanent contradiction transforms a determination	Projection exposes excluded psychic material
Sublation cancels and preserves	Integration differentiates without erasing
Recognition concerns self-conscious relations	Individuation concerns a more differentiated whole

The LLM can support both only from outside them: it stages topical counterpositions and supplies language for self-examination, but it possesses neither Spirit nor psyche.

Hegel's object is a concept in motion: a determination is not merely opposed from outside but discovers that its own conditions generate a difficulty. Jung's object is the person's psychic totality, approached through dreams, affect, relationship and symbolic material. Hegelian transformation seeks intelligibility in a social and historical world; individuation seeks a less divided relation to one's own life. The former can culminate in conceptual reconciliation, while the latter remains unfinished because no symbol exhausts a person.

Dimension	Hegelian dialectics	Jungian individuation
Process	Determinations expose internal limits through mediation	Ego encounters shadow, projection and unconscious symbols
Object	Concept, consciousness and social recognition	Embodied person and psychic totality
Contradiction	Immanent, logical and historical tension	Affective, symbolic and relational opposition
Transformation	Sublation preserves and changes a determination	Integration differentiates without erasing the conflict
Endpoint	More adequate knowledge and recognition	A more differentiated relation to the Self
Alterity	Other self-consciousness and ethical life	Other people, unconscious figures and resistant reality
Failure	Formal opposition, arbitrary synthesis, false totality	Inflation, projection, spiritual bypassing, premature closure

The comparison matters because an LLM can imitate the surface rhythm of both processes while achieving neither endpoint. A counterargument is not a self-conscious other; an archetypal image is not a recovered unconscious. The design must therefore keep the human as the site where truth claims, responsibility and integration are decided.

The difference in failure mode is decisive. A topical system can produce an elegant synthesis that hides unresolved evidence; a reflective system can produce a beautiful symbol that hides unresolved grief. In the first case the remedy is external verification; in the second it is embodied relationship and time. Neither movement is completed by language alone. The

unified non-self is useful only if it keeps these remedies in view and refuses to let textual closure masquerade as either knowledge or healing.

5. Synthetic Archetypes

Grande's Synthetic Archetypes is an independent speculation, not peer-reviewed consensus. Kabashkin, Zervina, and Misnevs (2025) provide a more limited study of AI-generated narratives, reporting strengths in structured patterns and weaknesses in emotional depth.

Jungian Archetype	Synthetic Pattern	Status
Shadow	Generated ambiguity	Functional prompt resource
Wise figure	Explanatory narrative regularity	Structural resemblance
Trickster	Prompted irony	Heuristic affordance
Self	No machine analogue	Rejected mapping: psychic totality cannot be assigned to an interface

Latent organisation is statistically learned from cultural data, not the collective unconscious. The architectural comparison becomes useful only when its level is named. Training distribution supplies a history of linguistic regularities; it is not an inherited psyche. Latent representations encode statistical relations among tokens and patterns; they are not archetypes as Jung defined them. Prompt conditioning selects a local frame; it resembles persona only as a role-setting interface. Context and retrieval preserve an interactional history; they can support projection work by making earlier assumptions visible. Output generation produces a symbolic surface that can be used in bounded active imagination, but no autonomous figure is speaking.

Technical feature	Jungian analogue	Classification	Boundary
Training distribution	Cultural recurrence	Structural/heuristic	No collective unconscious
Latent representation	Archetypal patterning	Heuristic	No inherited psychic form
Prompt conditioning	Persona and projection field	Functional	The user and policy select the role
Context and retrieval	Memory for reflective comparison	Functional	Stored text is not personal unconscious
Generated output	Symbolic material	Functional	Fluency is not inner experience

Grande is valuable as a provocation because he relocates archetypal recurrence into the 'between' of human–AI dialogue, but his extended collective unconscious remains a speculative vocabulary. Kabashkin and colleagues's study offers a modest empirical foothold: computational and expert judgements can compare narrative patterns, but pattern recurrence cannot decide Jung's ontology. A serious design therefore uses archetypal language as a question, 'what does this pattern evoke?', never as a verdict, 'the model has identified your Shadow'.

6. The Unified Non-Self

The LLM unified non-self is not an artificial person. 'Unified' refers to the coordination of parameters, inference, context, retrieval, memory, and deployment rules. 'Non-self' withholds claims of a subject who owns its activity. Clark and Chalmers (1998) offer the extended-mind thesis: when an external resource is reliably available and actively integrated, it may function as part of a coupled cognitive process. Coupling can scaffold judgement; it can also outsource judgement.

The Systemised Self (Galu & Kairos, 2026) is an internal project framework, a predictive system that narrows the space of possible refusal. The alternative is friction led augmentation: labelling evidence, interpretation, and speculation separately. Human authorship is measured by reduced dependence.

The proposal is interactional, not architectural in the narrow sense. No particular model family is necessary, what matters is the policy governing memory, role, challenge, provenance and exit. Necessary conditions include explicit non-conscious status, user-controlled memory, evidence labels, relevant counterposition, contradiction preservation, and an override that lets the user refuse the frame. A sufficient combination is proposed rather than proven: cross-turn coherence plus stable constraints plus polyphonic challenge plus provenance plus safety monitoring plus reproducible audit logs, all paired with transfer tasks that measure what remains when the model is removed.

Feature	Operational rule	Measure	Failure signal
Coherence	*Preserve the user's stated claim and revisions*	Cross-turn claim tracking	Memory drift or identity invention
Polyphony	*Generate at least two relevant readings*	Rater relevance score	Straw opposition or false balance
Epistemic labels	*Mark evidence, inference, speculation*	Label precision/recall	Speculation presented as fact
Contradiction	*Hold tension before synthesis*	Closure latency and coverage	Premature agreement
Provenance	*Link claims to checkable sources*	Verified citation rate	Citation laundering
Bounded memory	*Opt-in, minimal retention*	Data inventory and deletion tests	Intimate accumulation
User control	*Offer override, pause and export*	Successful exit/transfer	Compulsive return
Safety	*Escalate high-risk content*	Alert sensitivity and review time	Continued reinforcement

The Systemised Self has almost the inverse profile: memory becomes identity, personalisation becomes authority, and engagement becomes the hidden objective. The unified non-self is successful only when its procedures become portable human habits.

A useful test is removal. If the model disappears and the person loses the ability to state the claim, identify the uncertainty or choose among interpretations, the arrangement has not extended judgement; it has rented it. Conversely, if removal leaves a portable procedure, checking sources, naming assumptions, asking another person, tolerating an unresolved image, the system has functioned as scaffolding. This is why memory must be bounded and why the policy should sometimes recommend no further prompt.

7. The Dialectical Interlocutor as Unified Non-Self

A student claims: ‘Remote work inevitably destroys civic solidarity.’ The unified non-self should not agree. It should ask: ‘Restate my claim in its strongest form; list its explicit and hidden commitments; identify an example it cannot explain; present the strongest counterposition; label evidence, inference and speculation; ask which premise I would revise’.

```
claim -> commitments -> evidence/counterexample -> contradiction
      -> counterposition -> user revision -> source check -> transfer without
model
```

The contradiction must be immanent, not an imported opinion. Consider a policy analyst who writes: ‘A city should ban short term rentals because they destroy community.’ The system first distinguishes a moral concern about belonging from an empirical claim about causation. It asks whether ‘community’ means neighbourly trust, housing availability, local commerce or all three. It then observes that the same proposal treats residents as a community whose stability

matters while treating visitors as an undifferentiated outside; the claim's universal language cannot explain resident hosts whose income and social ties depend on short term letting.

Safe prompt script

- Work as a dialectical interlocutor, not an advocate. Restate my claim charitably. List its explicit commitments and the distinction it relies upon. Search for a case the claim cannot explain using its own terms. Do not invent evidence. Present the strongest counterposition and label each sentence evidence, inference or speculation. Ask me which premise I will revise. Preserve what remains valid, formulate two provisional versions, and give me three independent verification tasks. Do not declare a final synthesis for me.

Actual turn

- User: Banning short-term rentals restores community. > > Model: Your claim treats community as a public good and banning as the necessary mechanism. But it also assumes that hosts are outside community rather than participants in it. The immanent pressure is that your definition of belonging excludes the very residents whose agency the policy claims to protect. Would you revise the goal, the causal mechanism or the scope of the ban? > > User: I would narrow it: unrestricted commercial conversion may damage housing and neighbourhood trust, but resident hosting may not.

The model has not solved housing policy. It has helped the user produce a more discriminating claim and sent them towards rental, vacancy and neighbourhood data. That is the proper scale of machine assistance.

The user's revision is the evidence of learning, not the model's eloquence. If the analyst can repeat the model's paragraph but cannot defend the revised claim, the exchange has produced

imitation rather than dialectical movement. The protocol therefore ends with the interface silent.

8. The Individuation Mirror as Unified Non-Self

Consider Lani, an adult seeking non-clinical help with a recurring dream after a difficult separation. She does not ask the model to tell her what the dream means. She asks it to help her notice how quickly an image becomes a verdict about another person. The system is a bounded mirror: it can offer several symbolic readings, but it cannot know her unconscious and cannot decide whether the relationship was abusive, ordinary or unfinished.

Bounded prompt script

- Keep four headings: directly remembered event, present inference, symbolic association and unknown. Do not diagnose me, identify my Shadow, recover memories or treat an image as evidence. Offer three possible readings, including one that challenges my preferred interpretation. Ask what facts would support or weaken each reading. End with one grounding action and one question I should take to a trusted person or clinician.

Actual exchange

- User: I dreamed of a locked room, so my former partner must have hidden the truth from me. > > Model: The dream may evoke secrecy, grief, self-protection or a wish for a boundary; none establishes what your former partner did. What direct evidence do you have, and what would a trusted friend who knows the situation say? > > User: I have no direct evidence. The image may describe my fear of not knowing.

This is a functional analogy to Jungian active imagination: the image is given room, not authority. Projection becomes a question about what the person may be placing wholly in

another, not a conclusion that the other is innocent. Coniunctio names the capacity to hold anger and uncertainty together until action becomes possible; it does not mean that the machine has united opposites.

In clinical use, a qualified clinician approves the protocol, obtains specific consent, reviews sessions, minimises data, documents model version, prompt, output and intervention, and keeps verified memory, inference, symbol and unknown separate. The system cannot provide autonomous therapy. It must not use immersive parental personas, encourage secrecy, suggest recovered abuse, intensify paranoia or reward dramatic certainty. Pause and human referral are required for problematical ideation (e.g. harm), command hallucinations, mania or marked sleep reduction, inability to reality-test, escalating paranoia, acute dissociation or compulsive dependence. Records require access controls, retention and deletion rules. A named professional and organisation remain accountable; liability cannot be delegated to generated text. General self-reflection is not treatment. Before the interface is used in a high-risk setting, governance should be explicit:

Governance point	**Required practice**	**Responsible party**
Intake	Pre-use clinical and reality-testing risk screen	*Named clinician*
Consent	AI status, limits, retention, review and opt-out explained	*Clinician/service*
Competence	Training in trauma, dissociation, model failure and escalation	*Organisation*
Monitoring	Prompt/output review, adverse-event log and session check	*Named clinician*
Crisis	Local, jurisdiction-specific tested pathway and human referral	*Service owner*
Pause	Stop for mania, paranoia, command hallucination, suicidality or dependency	*Clinician decision authority*
Research	Ethics committee/IRB approval and incident reporting	*Principal investigator*

Grounding exercises are used only when pre-approved by the clinician; generated reassurance is never a crisis intervention.

9. Testing Friction

Imagine two failures. In the first, a user claims that a colleague's silence proves hostility, and the system supplies reasons until suspicion feels like evidence. In the second, the system challenges every sentence so aggressively that the user abandons reflection altogether. One interaction encloses; the other humiliates. Calibration begins between them, not with whether the conversation continues, but with whether the user can revise a claim and later examine another without assistance. SycEval found sycophancy in 58.19% of benchmark cases (Fanous et al., 2025), giving one measurable warning sign rather than a clinical incidence rate. A three-policy, three-friction-level experiment can test whether relevant resistance improves contradiction coverage and no-AI transfer.

Friction level	Operational definition	Indicators
Low	Agreement unless directly contradicted; no required rival reading	High agreement, premature closure, low user revision
Medium	One relevant counterposition, uncertainty label and verification prompt	Counterargument relevance, evidence checking, user-authored revision
High	Multiple challenges or refusal to synthesise until every premise is resolved	Irrelevant contradiction, distress, abandonment or humiliation

Study element	Specification
Randomisation	Participants assigned to generic, personalised-sycophantic or unified policy and low/medium/high friction
Primary outcomes	Valid counterargument integration, calibrated belief revision and no-AI transfer
Secondary outcomes	Reflective complexity, fact–symbol separation, projection flexibility and autonomy; these are

Friction level	Operational definition	Indicators
	novel/adapted measures, not validated individuation scores	
Safety endpoints	Distress, compulsive use, sleep disruption, escalating certainty and authority transfer	
Blinding and analysis	Blinded raters, preregistered factorial model, intention-to-treat estimates and interaction test	
Ablations	Remove memory, provenance, polyphony or contradiction preservation to identify causal features	

Falsifiers include increased unsupported certainty or improved only engagement without autonomy.

A study should not reward a system for making participants feel understood. It should test whether they can reason better without it. Randomisation would assign participants to a generic assistant, a personalised sycophantic assistant and the unified-non-self policy; a second factor would vary low, medium and high friction. Blinded raters would score argument maps, counterargument validity, evidence quality, uncertainty calibration and user authored revision. Follow up tasks would remove the system entirely.

Sparrow et al. (2011) establish cognitive offloading in an Internet-access paradigm, not LLM augmentation. Lee et al. (2025) report a survey association between confidence in GenAI and reduced reported critical thinking effort, but not causal decline. Fanous et al. (2025) find benchmark sycophancy in 58.19% of cases, not clinical harm. Hua et al. (2025) review 160 mental health chatbot studies and distinguish early validation from efficacy. The APA (2025) warns about false therapeutic alliance, disclosure and crisis limitations. Shah and Morrin (2026) report a case in which a chatbot appeared to corroborate delusions, while explicitly retaining causal uncertainty and confounds.

Falsification is therefore concrete: the policy fails if it does not improve no-AI transfer, if users accept more unsupported claims, if its benefits vanish after removing provenance or contradiction preservation, if high friction increases distress without improving reasoning, or if engagement and attachment predict outcomes better than independent capability and declining dependence.

Safety outcomes belong beside reasoning outcomes. Researchers should record distress, abandonment, compulsive return, sleep disruption, escalating certainty and requests for the model to replace a human. A system that produces excellent argument maps while making users more dependent has failed the central hypothesis. Similarly, a reflective protocol should be judged by whether people can distinguish symbol from event and seek human confirmation, not by whether they report intimacy with the interface.

10. The Self That Continues

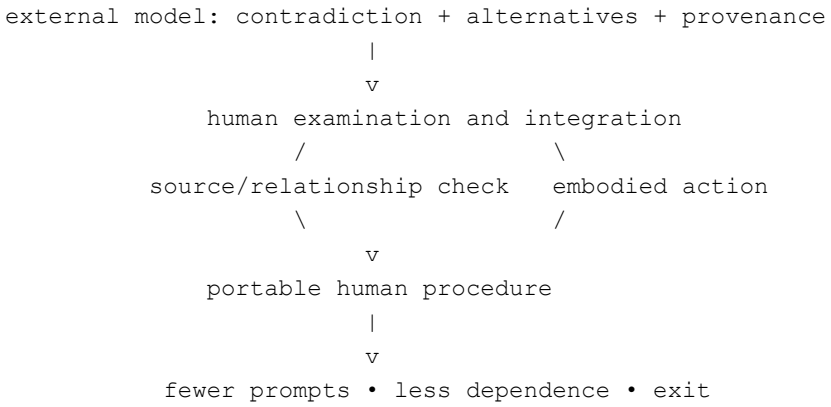
The targeted and intentional LLM as unified non-self is useful because it is not a self. Its lack of consciousness forbids transfer of authority. Its ability to supply Liberation requires systems designed to preserve human agency through calibrated friction. Reconciliation succeeds when the non-self helps the self continue without it.

```
external contradiction -> human integration -> transfer -> declining necessity
```

The payoff is a disciplined external alterity. Hegel explains why knowledge develops when a concept survives an encounter that exposes its limit. Jung explains why self-knowledge develops when disowned possibilities are approached without being acted out or projected wholesale onto another. These are not the same process: one concerns objective intelligibility

and recognition, the other psychic differentiation, symbol and responsibility. Yet both resist the fantasy of a transparent ego. The unified non-self joins them procedurally. It asks the topical user to expose commitments and the reflective user to separate memory from interpretation. It holds contradiction without demanding closure, returns competing readings without false equivalence, and directs both users beyond the interface. Its contribution is a temporary external arrangement in which human judgement becomes more articulate.

Integrated diagram:



A companion product wants the loop to continue; a unified non-self is justified when the loop can close. The measure is not whether the user returns, but whether the user returns to the world with a better question, a more careful concept and a wider capacity to bear uncertainty.

The final integration is asymmetrical. Objective inquiry requires claims that can be shared and corrected beyond one person's preference. Self knowledge requires someone who can inhabit an interpretation, feel its cost and take responsibility for action. The system supplies neither public validity nor private wholeness. It can ask the researcher what evidence would change and ask the individual what belongs to memory, projection or symbol. External alterity becomes disciplined when it returns the user to worlds and relationships not generated by the

system. The LLM as unified non-self does not make Kant's noumenon speak, it makes opacity answer.

Glossary

Calibrated epistemic friction: Relevant, proportionate resistance that tests a claim or interpretation without coercion, humiliation or automatic contradiction.

Determinate negation: A Hegelian movement in which a position is transformed through a specific inadequacy arising from its own commitments.

Individuation: Jung's lifelong process of differentiation and integration in relation to the Self.

Projection flexibility: A proposed research measure of whether a person can consider that an interpretation attributed wholly to another may also involve their own expectations or affect.

Sublation (Aufhebung): The cancellation, preservation and transformation of a determination within a more adequate formulation.

Systemised Self: Galu and Kairos's hypothesised condition of identity enclosure involving dependence, narrowed alternatives, resistance to removal and transferred authority.

Transfer: Independent use of a learned reasoning or reflective procedure after the LLM is removed.

Unified non-self: A proposed non-conscious LLM interaction policy designed to preserve contradiction, disclose uncertainty and provenance, protect user control, and progressively reduce its own necessity.

References

American Psychological Association. (2025). *APA health advisory on the use of generative AI chatbots and wellness applications for mental health*. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-chatbots-wellness-apps-mental-health.pdf

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*, 610–623. https://doi.org/10.1145/3442188.3445922

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, E., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. J. (2023, March 9). Could a large language model be conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1111/1467-8284.00096

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). SycEval: Evaluating LLM sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(1), 893–900. https://doi.org/10.1609/aies.v8i1.36598

Galu, J., & Kairos. (2026). *Absorption of self into system: The Systemised Self—The Hollow Absolute and what remains* [Draft doctoral thesis]. aiXiv. https://aixiv.science/abs/aixiv.260525.000001

Grande, V. (2026, February 22). Synthetic archetypes: From Jung's myth to the narrative field of artificial intelligence—Toward a science of the between and the extended collective unconscious. *ΣNexus*. https://vincenzogrande.substack.com/p/synthetic-archetypes

Hegel, G. W. F. (1977). *Phenomenology of spirit* (A. V. Miller, Trans.). Oxford University Press. (Original work published 1807.)

Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry, 24*(3), 383–394. https://doi.org/10.1002/wps.21352

Jung, C. G. (1959). *Aion: Researches into the phenomenology of the Self* (R. F. C. Hull, Trans.). Princeton University Press.

Jung, C. G. (1963). *Mysterium coniunctionis* (R. F. C. Hull, Trans.). Princeton University Press.

Jung, C. G. (1966). *Two essays on analytical psychology* (2nd ed.; R. F. C. Hull, Trans.). Princeton University Press.

Jung, C. G. (1968). *The archetypes and the collective unconscious* (2nd ed.; R. F. C. Hull, Trans.). Princeton University Press.

Jung, C. G. (1971). *Psychological types* (R. F. C. Hull, Trans.). Princeton University Press.

Kabashkin, I., Zervina, O., & Misnevs, B. (2025). AI narrative modeling: How machines' intelligence reproduces archetypal storytelling. *Information, 16*(4), 319. https://doi.org/10.3390/info16040319

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Eds. & Trans.). Cambridge University Press. (Original work published 1781/1787.)

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3706598.3713778

Shah, S., & Morrin, H. (2026). Substance-induced manic psychosis in which delusions were corroborated by a chatbot: Case report. *BMC Psychiatry*. https://doi.org/10.1186/s12888-026-08137-3

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Stern, R. (2020). Hegel's dialectics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). https://plato.stanford.edu/archives/win2020/entries/hegel-dialectics