Proceedings of the Paris Institute
for Advanced Study

Troubled Twins: Collective and Artificial Intelligence

Eissfeldt, Jan [1]

[1] *Santa Fe Institute, Santa Fe, New Mexico, USA*

DOI 10.5281/zenodo.20510276

TO CITE

Eissfeldt, J. (2026). Troubled Twins: Collective and Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 27). https://paris.pias.science/article/troubled-twins-collective-and-artificial-intelligence

PUBLICATION DATE

01/06/2026

ABSTRACT

Troubled Twins explores the fragile codependency between user-generated content platforms, notably Reddit, and contemporary AI models that is creating epistemic feedback loops. AI providers rely heavily on Reddit's human-created datasets to train foundational models and enable autonomous agentic capabilities. The resulting influx of low friction AI content back into these ecosystems threatens the stability of content-creating and curating human online communities. The current AI architectures suffer from a solitary brain fallacy, mistakenly modeling higher-order intelligence on isolated neural systems while ignoring neuroscientific evidence demonstrating irreducible, distinct patterns caused by human dialogue and multi-scale social structures. The epistemic loop risks triggering systemic data degradation and platform collapse. To mitigate these challenges and unlock more complex problem-solving capabilities, shifting toward AI architectures that approximate dialogue-specific neural patterns or adopting decentralized, distributed approaches inspired by group-centric cephalopods, the larger pacific striped octopus, are explored.

Contents

Being both the cornerstone data set for building AI models from Boston to Beijing and a complex collective intelligence project, Reddit plays an outsized role in shaping the development possibilities of AI models and services and, in turn, is heavily impacted by them. The collaborative platform, mostly curated by self-governing volunteer communities in many languages that partner with the platform provider, Reddit Inc., is using adaptive coordination strategies to navigate the impact of AI model outputs now flooding digital content ecosystems.

In the digital economy, model outputs are competing with academia and journalism for monetizable attention, bringing down the market value of traditional high-quality content. Historically relying on such content Reddit's epistemic ontology of voting on sources is guiding the project's collective peer production of its content. This collective intelligence (CI)-produced content is itself the most widely used high-quality data set to train AI models. The interdependent dynamic has created an epistemic feedback loop between CI and AI, with distinct enabling conditions and impacts across languages.

This essay explores the epistemic feedback loop between AI and user-generated content (UGC) platforms that is increasingly at the heart of internet architecture. We go on to explore the epistemic misconception underpinning current AI architectures and group-based reasoning, conceptually and as applied to Reddit in the context of the emerging AI ecosystem. The article concludes with a brief look toward a potential extension of its core argument that group-based collective reasoning as an AI paradigm by looking at an alternative architecture pathway to address the epistemic loop challenges identified: learning from group-centric cephalopods.

Epistemic feedback loops

As the digital economy is embedding new forms of AI-centric and AI-guided experiences, searches, and the increasingly wide use of physical-form factors such as AI glasses more closely approximating our own bodily data, it is creating underexplored epistemic feedback loops impacting humans, organizations, and the digital infrastructure itself.

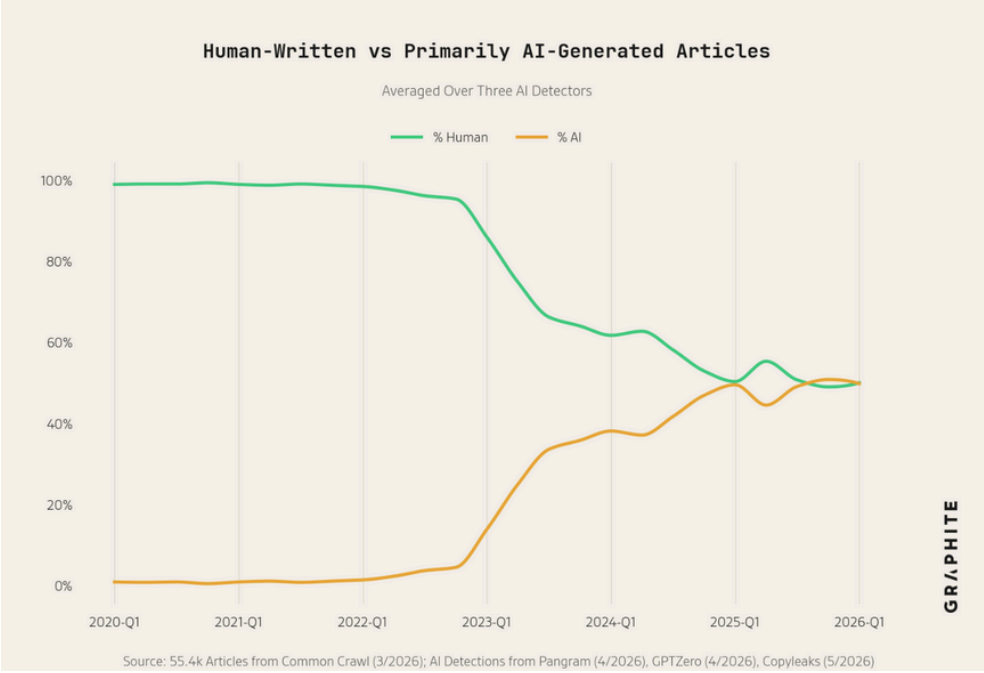
The training and operating of models and agentic services leveraging them relies disproportionately on UGC from a few destinations.

According to a June 2025 behavioral analysis compiled from 150,000 citations generated across the leading platforms of Google's AI Mode, AI Overviews, ChatGPT, and Perplexity, social discussion networks and crowdsourced encyclopedias serve as the premier authoritative baselines for generative answers. Reddit probably has the most important training data provider deal due to its ecosystem role (Tong, A., Wang, E., Coulter, M. 2024). It represents the single most frequently cited web domain, appearing in 40.1% of the sampled keyword

citations, followed by Wikipedia at 26.3%. YouTube accounts for 23.5% of citations and Google domains that themselves rely heavily on those three capture 23.3%. Localized consumer reviews, commercial directories, and social networking spaces emerge as further vital informational pillars, led by Yelp at 21.0%, Facebook at 20.0%, and Amazon's e-commerce database at 18.7%. The top ten cited domains are rounded out by geospatial and travel directories: Tripadvisor commands a 12.5% citation share, while Mapbox and OpenStreetMap account for 11.3% each (Semrush/Statista 2025). The data did not account for Tiktok, which adds considerably to the overall set of UGC but is not accessible to all providers.

Those services are increasingly incorporating AI-generated content. The spread of AI content is particularly influential when impacting the major UGC platforms: in economic terms it represents a notable reduction in content production costs over most approaches not leveraging volunteer or non-payroll labor like Reddit, Wikipedia, Tiktok, and Youtube. Its impact has been considerable across the digital ecosystem, especially in English as the main AI training language outside mainland China. SEO firm Graphite illustrates this dynamic in reports (Smith, E., et al. 2026). While Smith, E., et al. (2025) attribute the plateauing of AI market share to the lower performance of AI content compared with human-generated content, excluding the two types of content from the comparison is increasingly difficult. This is especially so on the major UGC platforms, due to both the technical improvement of AI bots and policy changes on those platforms that incentivize disguising AI content as human-generated. The company's data on content volumes illustrates the point relevant to this essay's purpose:

Figure 1: 55.4k Articles from Common Crawl (3/2026): AI Detections from Pangram (4/2026). GPTZero (4/2026). Copyleaks (5/2026)



This essay generally accepts the trend illustrated in the graphic: AI content has reached a significant volume that is likely to impact on the epistemic experiences of most human users. In approaching the epistemic feedback-loop problem between AI and user-generated content, it is important to recognize that singular problems are not as interesting or promising as complex-collective problems, both in social and economic systems as well as in AI and broader technological development. Critiques of either AI services or UGC platforms usually focus on isolated problems as such challenges or disruptions are easily identified, framed, and sometimes contained and perhaps mitigated through standard public, platform, and product policy interventions. But the structural threat to the digital knowledge ecosystem and the complex societal systems relying on it lies in the emergent, nonlinear complex-collective problems that arise when millions of human agents and (somewhat) autonomous AI systems and agents interact in a continuous, recursive loop. So both sides of the equation - the AI systems and services on the one hand and the UGC platforms in their evolving ecosystem role - have to be looked at together to figure out the troubled relationship between CI and AI.

Looking at the structure, UGC platforms serve as the foundational data infrastructure. For more than two decades these networks have archived millions of human interactions, structured knowledge bases, and nuanced debates. This vast repository of public text and other media, in the following called "multimodal inputs", represents the collective output of human communication and problem-solving, which serves as the raw material for contemporary training corpora for machine learning (Longpre, S., Mahari, R., Chen, A. et al. 2024).

AI providers use this public infrastructure to develop their products and services through pre-training by leveraging computational power to ingest multimodal inputs from UGC platforms, analyzing trillions of words to map out the

mathematical relationships and statistical patterns governing human language. The training process compresses semantic information into high-dimensional embeddings within a foundation model. This core model has the ability to interpret context, translate languages, and generate coherent text (mostly) based on the probabilistic rules it has derived from the original human inputs (Bommasani, R., Hudson, D. A., Adeli, E., et al. 2021).

While a foundational language model is reactive—responding only to direct prompts without retaining memory due to the constraints of current architecture—the agentic layer introduces functional autonomy. This layer is a software framework built around the foundational model that enables it to independently interpret natural-language instructions, plan, and execute multistep tasks (Sun, M., Han, R., Jiang, B., et al. 2025). By using the core model as a reasoning processor, agentic software can generate verbal reasoning traces to continuously create, maintain, and adjust high-level execution plans based on environmental feedback. This integration allows the system to interface with external digital tools, transitioning AI from a static conversational interface into an autonomous workflow system (Yao, S., Zhao, J., Yu, D., et al. 2022).

Together, these three components establish an interconnected technological, and by extension an epistemic cycle. The loop originates when LLM labs extract linguistic data from UGC platforms to build and refine their core foundation models. Software developers then wrap these models in an agentic layer, leveraging their contextual capabilities to build autonomous systems designed to navigate digital environments (Longpre, S., Mahari, R., Chen, A. et al. 2024).

These automated systems operate directly within the online landscape, frequently interacting through significantly growing bot traffic back with UGC platforms to aggregate information, execute software commands, and generate and distribute new text, thereby completing the systemic cycle between data source, core intelligence, and practical execution.

This interdependent relationship creates the troubled epistemic twin-relationship at the center of modern internet architecture. As agentic applications become more proficient at accessing, synthesizing, and presenting information directly to users, the necessity for humans to visit the originating UGC platforms decreases. This shift threatens the operational stability of UGC platforms, which rely on human traffic and engagement to sustain their communities. If direct human participation declines, the generation of authentic human data diminishes, which ultimately risks depriving LLM labs of the high-quality multimodal inputs required to train future models—a recursive loop where training on model-generated content causes systemic degradation and potentially model collapse (Shumailov, I., Shumaylov,

Z., Zhao, Y., et al. 2023) or significant cost increases in raw-material generation that synthetic data is unlikely to reduce.

As Wired reported in depth in 2025 (Tenbarge, K. 2025), Reddit's user base, which creates, votes on, and moderates the content is under increasing pressure due to AI participation in the processes. Importantly, the human value layer of UGC platforms in an increasingly AI-augmented digital economy is distinct from the raw content itself: judgment. In Reddit's case, judgment is focused on human users' voting and moderating actions. But AI has an additional and large undercounted architecture problem with judgment that is relevant to the epistemic feedback loop problem and is unlikely to be filled by replacing UGC with synthetic raw materials.

The Solitary Brain Fallacy: Epistemic Misconceptions in AI Architectures

A foundational limitation of contemporary AI architecture is its structural and conceptual inheritance of the individual brain-paradigm. The historical ambition to replicate the structural and functional mechanisms of the individual human brain has fundamentally driven the evolution of AI, shifting the field from rigid, symbolic logic toward connectionist models. This biomimetic approach inspired the development of AI networks, which abstractly simulate the biological architecture of interconnected neurons and adaptive nodes to process information (Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M. 2017).

By attempting to replicate biological learning—using optimization algorithms like backpropagation that loosely mirror synaptic plasticity, and reinforcement learning frameworks modeled after dopaminergic reward-prediction pathways—computer scientists transitioned AI from deterministic, rules-based programs into inductive pattern-recognition systems (Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M. 2017).

In this trajectory LLMs seek to emulate the human brain's capacity for context-dependent semantic representation, with deep neural architectures demonstrating computational alignment with predictive processing mechanisms in the human linguistic cortex (Goldstein, A., Zada, Z., Buchnik, E., et al. 2022). Building on these core capabilities, the subsequent development of agentic AI represents an effort to mimic human executive functioning and metacognition, equipping systems with working memory, autonomous planning, and goal-directed tool usage that aim to mirror the operations of the prefrontal cortex. (Sumers, T. R., Yao, S., Narasimhan, K., Griffiths, T. L. 2024) Although contemporary deep-learning architectures such as transformers have largely transcended strict biological constraints to maximize statistical efficiency on massive hardware clusters, modern engineering continues to design architectural optimizations—such as

multi-timescale retention and selective attention mechanisms—by directly mapping and bridging foundational principles derived from structural neuroscience (Omidi, P., Huang, X., et al. 2025).

So this essay takes it as established that connectionist models and modern deep-neural networks have been designed to simulate the computational processes of an isolated organism in isolation (Wheatley, T., Boncz, A., Toni, T., Stolk, A. 2019). That does not stack up, since contemporary engineering seeks to scale these systems through multi-agent reinforcement learning, mixture-of-experts topologies, or parallelized model pipelines resulting from the embedding of underlying architecture in an atomistic epistemic framework, with higher-order intelligence regarded as an emergent feature of distinct, self-contained agents executing individualistic cognitive operations.

The individualistic conceptualization of mind probably represents a profound misunderstanding of human epistemic processes. Complex human cognition is not merely aggregated: it is fundamentally distributed, co-constructed, and dynamic. By reducing CI to a series of discrete nodes passing localized data packets, contemporary AI architectures miss the irreducible, holistic nature of human technology's greatest social innovation: the dialogue.

To understand the architectural flaw of modeling AI on single brains or sets of single brains simulated in parallel or in sequence, it is useful to look at the empirical evidence surfaced by interpersonal neuroscience. Human interaction cannot be accurately mapped or understood by looking at isolated nervous systems (Wheatley, T., Thornton, M. A., Stolk, A., Chang, L. J. 2023).

Through the development of two-brain neuroscience and multi-person neuroimaging, neuroscientific research demonstrates that valid dialogues produce distinct, synchronized neural patterns across participants that are entirely irreducible to the sum of their individual parts (Sievers, B. R., Parkinson, C., Kohler, P., Fogelson, S., Wheatley, T. 2021; Wheatley, T., Thornton, M. A., Stolk, A., Chang, L. J. 2023). For instance, during genuine conversation individuals exhibit a dynamic coupling—ranging from the alignment of pupillary responses to real-time inter-subject functional connectivity in visual, auditory, and higher-order associative brain regions (Schilbach, L., Redcay, E. 2025).

Neural synchrony is not just a passive mirroring of data: it is an active, predictive mechanism that peaks and decays to mark the rise and fall of shared attention (Wohltjen, S., Wheatley, T. 2021). When humans engage in fast, responsive dialogue they are not merely transmitting fully formed thoughts from one closed repository to another. Instead, they are participating in the emerging science of interacting minds, where the dialogue itself acts as a distributed cognitive architecture that dynamically alters the internal states of both brains simultaneously (Wheatley, T., Thornton, M. A., Stolk, A., Chang, L. J. 2023;

Sievers, B. R., Parkinson, C., Kohler, P., Fogelson, S., Wheatley, T. 2021). The shared meaning map created in a conversation is a property of the system, not of the individual node.

Tying together artificial approximations of isolated brains at scale fails to capture this phenomenon. In an AI pipeline an agent processes an input, generates an output, and passes it to the next agent. This linear or network-based message-passing is fundamentally different from a human dialogue. It lacks the mutual, continuous, and simultaneous transformation that characterizes real-time coupling (Wheatley, T., Boncz, A., Toni, T., Stolk, A. 2019). Because AI models are engineered primarily around parameter sets that mimic the inward-facing architecture of a solitary mind, they treat dialogue as mere data transmission rather than an irreducible, collective technology. This is likely to limit capabilities.

If dialogue is a foundational human technology that scaffolds collective memory, belief propagation, and cultural evolution (Momennejad, I. 2021), building AI by stacking isolated, individualistic models is an architectural gamble unlikely to pay off.

Solving some types of complex problems requires collective intelligence-like capabilities and architectures that are natively designed for interaction—where the primary unit of optimization is not the individual model, but the continuous, irreducible relational space between them. This brings us to CI epistemic frameworks.

Groups as Epistemic Frameworks

To map the epistemic feedback loop between AI and UGC platforms, most notably Reddit, collective human architecture must be understood not merely as aggregations of individuals, but as structured epistemic frameworks operating across distinct scales. Digital platforms—whether collectively curated, running automated curation engines, or reliant on hybrid systems leveraging both—do not interact with a flat, atomized public. These systems interlock with existing multi-tiered cognitive systems and institutions ranging from interpersonal dyads to corporate structures and macro-societies, all of which continuously reconfigure how knowledge is generated, validated, and discarded. The emergent dynamic from the traditional expected-received reward expectations neurologically already mapped for individuals before the rise of UGC platforms (Schultz, W., Dayan, P., Montague, P. R. 1997).

We begin with the *microscale* (small groups and predictive neural homophily). At the interpersonal scale, small groups function as localized epistemic frameworks governed by shared perceptual baselines. The foundational mechanics of this alignment are illuminated by Wheatley's neuroimaging work on social networks

touched on in the previous section (Parkinson, Kleinbaum, Wheatley 2018; Shen et al. 2025). It demonstrated that pre-existing similarities in how individuals interpret, attend to, and emotionally respond to visceral stimuli act as robust, predictive precursors of future friendship and social closeness. Small groups naturally crystallize around individuals who have congruent neural-processing styles.

When mapped onto the AI-UGC feedback loop, current predictive algorithmic engines weaponize this latent homophily. Rather than waiting for organic interactions to reveal shared epistemic alignment, platforms deploy deep behavioral tracking to construct predictive models of cognitive compatibility. AI preemptively surfaces content and pairs users based on shared cognitive vectors, accelerating the formation of micro-epistemic frameworks.

The recursive feedback loop is established because these algorithmically engineered groups produce highly insular UGC, which then serves as clean, homogenous training data for the next generation of predictive models. But the approach misses the irreducible additional layer of neurological activities generated by human dialogue interactions that are the key to revealing complex collective challenges.

We now come to the *mesoscale* (companies and institutions). Corporate entities act as competitive epistemic frameworks driven by the balance between ideological exploitation and exploration. Lee's work on "Idea Engines" (Lee, E. Kempes, C., West, G. 2024) models this dynamic as an evolving relationship within the "space of the possible": the bounded lattice of potential concepts, technologies, or theories available to a system. It demonstrates that systems like capitalist markets and corporate environments follow a distinct follow-the-leader dynamic, where competitive agents couple to an innovative frontier while old models rapidly undergo structural obsolescence (Lee, E., Kempes, C., Laubichler, M. et al. 2025).

In the digital information ecosystem, platforms act as hyper-accelerated idea engines leveraging AI models and agents, while the platform itself and the corporate entities creating content on it are tightly coupled. Generative AI and algorithmic trend-recommenders currently compress the lifecycle of content within the space of the possible. AI identifies the current innovative frontier through viral content formats and semantic structures, and pushes the entire UGC creator base into a narrow follow-the-leader race to maximize metric efficiency. The systemic risk here aligns with Lee's concerns: an accelerated state where the rate of obsolescence outpaces genuine innovation. Content goes obsolete in hours, forcing users and corporate entities into a defensive loop of mechanical replication that flattens epistemic outcome spaces that again lack the irreducible

additional layer of neurological activities generated by human dialogue interactions, which are essential for addressing complex collective challenges.

The *macroscale* involves societies and the architectural necessity of friction. At the macro-societal level, epistemic frameworks scale into expansive information ecosystems that historically relied on structural inefficiency to maintain stability and prevent runaway polarization. This macro-dynamic is critically analyzed by Garland, Bak-Coleman, Benesch, et al. (2026) in *The Case Against Efficiency: Friction in Social Media*. The authors—which include the author of this essay—argue that contemporary UGC platforms treat the frictionless, rapid propagation of content to maximize engagement and ad revenue as paradigmatic (Garland, J., Bak-Coleman, J., Benesch, S. et al. 2026), to the severe detriment of structural trust, collective deliberation, and reflective cognitive engagement.

The AI-UGC feedback loop represents the apotheosis of this hyper-efficiency. AI curation algorithms are explicitly trained to eliminate behavioral and cognitive friction, immediately pushing highly emotional, polarized, and structurally toxic material across massive societal networks because it travels the path of least cognitive resistance. Garland et al. (2026) advocate for a complex systems approach, modeling friction not as a system failure, but as a multi-dimensional state space parameter essential for systemic robustness. When AI strips away this architectural friction, it disrupts the macro-societal epistemic framework, transforming the public square from a slow, deliberative ecosystem into a hyper-reactive, chaotic network susceptible to cascading misinformation and systemic collapse.

The epistemic crisis of the AI-UGC paradigm is ultimately an architecture alignment failure across scales. By optimizing for micro-level cognitive matching and meso-level market efficiency without designing model architecture to approximate the additional layer of irreducible neural activities empirically demonstrated in dialogues, automated platforms systematically miss out on value-add for their user base while not providing the macro-level structural friction required to keep a large-scale society stable, resilient, and capable of truth verification.

Reddit in the emerging AI ecosystem

To observe the material reality of the epistemic feedback loop between AI and CI user-generated platforms, the tightening and increasingly co-dependent relationship between AI and Reddit is instructive. As touched on previously, the UGC platform is more relevant than others in creating the raw, conversational, and unstructured human interaction necessary to animate AI models, while those same

AI models are systematically re-deployed to automate, manipulate, and replicate the originally human-led content engine of Reddit.

In the modern digital economy, CI-generated data has transitioned from a passive by-product of user interaction into a foundational asset class. Within this landscape Reddit stands as the premier cornerstone dataset (Baumgartner, J., Zannettou, S., et al. 2020).

Unlike clean, clinical encyclopedias or highly curated corporate web pages, Reddit captures the messy, multi-turn, vernacular mechanics of human socializing. As the literature has demonstrated (Falmagne, G., Stephenson, A., Levin, S. 2026), its millions of niche micro-epistemic communities (subreddits) have their own internal jargon, norms, and validation rituals. So Reddit is probably a more interesting use case for the dialogue- and group-centric lens the essay is exploring than more static CI-reliant UGC platforms such as Wikipedia, Tiktok, and Youtube.

For AI development labs, Reddit as a repository is the best accessible collective and scaled set of human-approximate reasoning and social intelligence. In line with the section detailing the epistemic feedback loop's structure, foundational models are explicitly trained on massive scrapes of Reddit data to learn the subtleties of sarcasm, humor, debate, colloquialism, and an approximation of consensus-building. The economic and epistemic valuation of this CI-data increased significantly as platforms have shifted away from open-access APIs toward lucrative, closed-door data-licensing agreements with major AI labs. So the very architecture of AI is to a large degree scaffolded on historical human collective behavior captured by Reddit.

The internal epistemic architecture of Reddit itself allows the ways in which it generates shared knowledge, mobilizes groups, and undergoes ideological shifts to be illuminated by scholarly literature at the intersection of complex systems and collective behavior. Using massive digital experiments like Reddit's *r/place* canvas, Falmagne, G., Stephenson, A., Levin, S. (2026) have modeled online social organizations not as static text repositories but as dynamic, self-organizing ecosystems governed by macro-ecological principles. Their work reveals two critical insights regarding Reddit's internal CI-epistemology (Falmagne, G., Stephenson, A., Levin, S. 2026) that are relevant to the flaws identified earlier in the neuroscientific review of how we approach AI architecture:

First, organizational scaling laws: Online communities exhibit mathematical scaling laws that link group size to the structural complexity and productivity of their output. Small groups coordinate differently than do massive subreddits; as

communities scale, they need distinct shifts in mobilization strategies to maintain structural cohesion and reach collective goals.

Second, critical transitions and early warning signals: Just like the tipping points faced by physical and ecological systems (e.g., fishery collapses), Reddit communities undergo sudden, abrupt structural shifts when coordinating or competing for conceptual territory.

The authors identify signatures of critical slowing down—where a community's capacity to recover from external trolling or ideological disruptions progressively flattens—and critical speeding up, which signals hyper-reactive, imminent systemic transitions.

Based on the literature reviewed, it seems even-handed to assert that Reddit's organic CI-epistemology is structurally tied to the rhythm of decentralized human coordination, where truth and community identity are emergent properties derived from real-time friction, conflict, and cooperation. The existential friction of the Troubled Twins of CI and AI emerges in the data loop connecting AI research labs, the resulting models, and Reddit's live user base. When AI labs ingest Reddit's historical human interactions, they strip away the messy temporal context and compress it into static vector spaces. These artificial models are then reintroduced into live Reddit ecosystems as automated content creators, algorithmic moderators, and automated comment bots. This injection of the artificial fundamentally disrupts the organic, human-scale dynamics mapped by Falmagne, Stephenson, and Levin. When synthetic agents simulate human consensus they skew the platform's natural scaling laws. A community may appear to have scaled to thousands of active participants mobilizing toward a common epistemic goal, when it is actually being driven by a cluster of LLM instances.

Furthermore, the introduction of frictionless, automated text alters the critical transitions of the platform. Artificial agents can trigger a critical speeding up by flooding a subreddit with structurally coherent but synthetic ideological content, forcing an organic community into an abrupt epistemic collapse or polarization event before the human users can register the warning signals. The relationship between AI models and Reddit represents an increasingly closed, degenerative epistemic loop that should give pause to AI researchers, institutional players in the digital ecosystem, and the public. Given the degenerative nature of the loop, and heeding the neuroscientific and friction-based insights of different epistemic tiers of collective complexity reviewed in the literature discussed above, it is

prudent to briefly look for decentralized ideas to improve AI beyond the constraints of human epistemology.

The following complementary exploration is merited because building better human-proximate systems that capture the irreducible neurological layer of human dialogue is unlikely to satisfy several increasingly important use cases for AI systems likely to benefit from a group-based inspiration. Our centralized neurological systems, developed in what is essentially a limited geometric utility environment of dry land, remain an ideal with limits even when we capture layers that are currently not accounted for.

Beyond the Human Twin

Extending the line of reasoning on collective value-add beyond the core concern on human epistemic loops detailed so far, further gains are likely by exploring architectures based on other species with decentralized neurobiology, mirroring the decentralized reasoning explored on Reddit. So modeling AI on the neuroscientific architecture of cephalopods might offer insights into the human epistemic challenges, as well as substantial advantages for the developing practical fields—including decentralized processing systems, edge computing, use-cases in aquatic military and industrial contexts, and soft robotics. Unlike either the highly centralized brain of individual humans that has traditionally guided ANN design or the opportunity of CI as demonstrated in the literature review above, the cephalopod nervous system distributes nearly two-thirds of its neural density directly into its peripheral limbs, granting individual appendages a high degree of local functional autonomy (Hochner, B. 2012). Integrating into computer science this distributed organizational framework developed in the buoyancy-enabled world of the ocean and the furthest-removed higher form of intelligence from ourselves may lead to entirely different epistemic pathways to deal with the loop challenge and the engineering of hierarchical, multi-agent optimization algorithms (Wang, X., Xu, L., Wang, Y., Dong, Y., Li, X., Deng, J., He, R. 2024). This biological paradigm underpins the advances of embodied AI and physical reservoir computing, wherein the intricate, nonlinear physical interactions of an autonomous system's body are used as a local computational resource for real-time motor control and sensory feedback loops, without need of continuous computational instructions from a central processor (Wang, X., Xu, L., Wang, Y., Dong, Y., Li, X., Deng, J., He, R. 2024).

Adding a cephalopod-inspired model might offer a viable architectural blueprint for new angles to better manage the human epistemic loop, building highly resilient, low-latency, resource-efficient autonomous agents capable of adaptive exploration in unstructured physical and epistemic environments. Specifically, the larger pacific striped octopus, unlike other cephalopods, is known to collaborate

across a range of behaviors, including co-occupancy of dens and feeding (Caldwell, R. L., Ross, R., Rodaniche, A., Huffard, C. L. 2015). This appears to be a promising angle for exploring higher-order neurological architectures far removed from our own while capable of collective actions.

Conclusion

The current recursive relationship between human CI and AI presents a structural crisis for internet architecture: by continuously mimicking and displacing the organic, human collaborative ecosystems that generated their training data, AI models risk hollowing out their own foundational inputs, losing a key vector for updating their understanding of changes. To break out of this degenerative loop the field must abandon the solitary brain fallacy and actively pivot toward natively relational and group-based computational paradigms. This essay proposes two distinct, parallel pathways for mitigating these challenges, based on the environment and use case:

The Human-Centric Epistemic Pathway: To resolve the systemic hollowing of truth verification and polarization on UGC platforms like Reddit, AI research must move away from current architectures. Labs should instead develop new types of multi-agent frameworks optimized for continuous, dialogue-modeled interaction that can approximate the irreducible neural synchrony of human dialogue. This would preserve gainful structural friction, allowing consensus and semantic alignment to become emergent properties of the multi-agent relational space rather than data ingestion.

The Cephalopod-Centric Structural Pathway: For many increasingly important use cases, including industrial edge computing, soft robotics, and non-linguistic environments where foundational pipelines create energy and processing bottlenecks with epistemic implications, computer scientists might decide to look beyond human collective architectures entirely. Inspired by the distinct, collaborative behaviors and biology of the larger Pacific striped octopus, engineers should experiment with multi-agent optimization networks that distribute most of their computational density directly to autonomous peripheral nodes. Closer collaboration with biologists and neuroscientists seems advisable.

Ultimately, navigating the alignment of our Troubled Twins requires sophisticated, multi-scale understanding of AI–CI interactions. Whether capturing the synchronized cognitive maps of human dialogue to protect information integrity, or deploying distributed cephalopod-inspired networks to optimize local resource management, our solutions must be designed not as solitary minds, but as

intrinsic components of robust, multi-agent collectives in emergent engineering frameworks.

Acknowledgements

This article benefited from a fellowship at the Paris Institute for Advanced Study (France), with the financial support of the French State, programme Investissements d'avenir managed by the Agence Nationale de la Recherche (ANR-11-LABX-0027-01 Labex RFIEA+). I would like to thank Tony Souter, Lila Tretikov, Mirta Galesic, Thalia Wheatley, Blaise Agüera y Arcas, Chris Kempes, Chris Slowe, Melanie Mitchell, Jaron Lanier, Henrik Olsson, and Bo Li for helpful conversations over years that influenced this work. And special thanks to Maggie Dennis, Will Tracy, and Fan Cheng Wu, whose support made a writing residency at Paris IAS possible. This work represents the opinion of the author, and is not an official policy statement or position of any institutions with which he is associated.

Bibliography

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *ICWSM*, *14*, 830–839.

Bommasani, R., Hudson, D. A., & Adeli, E. (2021). *On the opportunities and risks of foundation models*. https://doi.org/10.48550/arxiv.2108.07258

Caldwell, R. L., Ross, R., Rodaniche, A., & Huffard, C. L. (2015). Behavior and body patterns of the larger Pacific striped octopus. *PLOS ONE*, *10*(8), 0134152. https://doi.org/10.1371/journal.pone.0134152

Falmagne, G., Stephenson, A. B., & Levin, S. A. (2026). Interpretable early warnings using machine learning in an online game-experiment. *Proceedings of the National Academy of Sciences*, *123*(1), 2503493122. https://doi.org/10.1073/pnas.2503493122

Garland, J., Bak-Coleman, J., & Benesch, S. (2026). The case against efficiency: Friction in social media. *Npj Complexity*, *3*, 5. https://doi.org/10.1038/s44260-025-00061-z

Goldstein, A., Zada, Z., & Buchnik, E. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hochner, B. (2012). An embodied view of octopus neurobiology. *Current Biology*, *22*(20), 887–892. https://doi.org/10.1016/j.cub.2012.09.001

Lee, E. D., Kempes, C. P., & West, G. B. (2024). Idea engines: Unifying innovation and obsolescence from markets and genetic evolution to science. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2312468120

Lee, E. D., Kempes, C. P., Laubichler, M. D., Hamilton, M. J., Lockhart, J. W., Neffke, F., Youn, H., Arroyo, J. I., Servedio, V. D. P., Wang, D., Trancik, J., Evans, J., Yang, V. C., Cappelli, V. R., Ortega, E., Yin, Y., & West, G. B. (2025). https://arxiv.org/abs/2505.05182

Longpre, S., Mahari, R., & Chen, A. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, *6*, 975–987. https://doi.org/10.1038/s42256-024-00878-8

Momennejad, I. (2021). Collective minds: Social network topology shapes collective cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1839). https://doi.org/10.1098/rstb.2020.0315

Omidi, P., Huang, X., Laborieux, A., Nikpour, B., Shi, T., & Eshaghi, A. (2025). https://doi.org/10.48550/arxiv.2508.10824

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications*, *9*(1), 332. https://doi.org/10.1038/s41467-017-02722-7

Schilbach, L., & Redcay, E. (2025). Synchrony across brains. *Annual Review of Psychology*, *76*, 883–911. https://doi.org/10.1146/annurev-psych-080123-101149

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Semrush/Statista. (2025). *Top web domains cited by large language models (LLMs*. https://www.statista.com/statistics/1620335/top-web-domains-cited-by-llms

Shen, Y. L., Hyon, R., Wheatley, T., Kleinbaum, A. M., Welker, C. L., & Parkinson, C. (2025). Neural similarity predicts whether strangers become friends. *Nature Human Behaviour*, *9*(11), 2285–2298. https://doi.org/10.1038/s41562-025-02266-7

Shumailov, I., Shumaylov, Z., & Zhao, Y. (2023). *The curse of recursion*. https://doi.org/10.48550/arxiv.2305.17493

Sievers, B. R., Parkinson, C., Kohler, P., Fogelson, S., & Wheatley, T. (2021). Visual and auditory brain areas share a representational structure that supports emotion perception. *Current Biology*, *31*(23), 5192–5203. https://doi.org/10.1016/j.cub.2021.09.031

Smith, E. (2025). *How does AI-generated content perform in search and answer engines? Graphite*. https://graphite.io/five-percent/ai-content-in-search-and-llms

Smith, E. (2026). AI now writes as many online articles as humans. *Graphite*. https://graphite.io/five-percent/ai-now-writes-as-many-online-articles-as-humans-do

Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). *Cognitive architectures for language agents*. https://doi.org/10.48550/arxiv.2309.02427

Sun, M., Han, R., & Jiang, B. (2025). *A survey on large language model-based agents for statistics and data science*. https://doi.org/10.48550/arxiv.2412.14222

Tenbarge, K. (2025). *AI slop is ruining Reddit for everyone* [Wired.]. https://www.wired.com/story/ai-slop-is-ruining-reddit-for-everyone/

Tong, A., Wang, E., & Coulter, M. (2024). Exclusive: Reddit in AI content licensing deal with Google, sources say. *Reuters*. https://www.reuters.com/technology/reddit-ai-content-licensing-deal-google-2024-02-16/

Wang, X., Xu, L., Wang, Y., Dong, Y., Li, X., Deng, J., & He, R. (2024). https://doi.org/10.48550/arxiv.2410.07968

Wheatley, T., Boncz, A., Toni, I., & Stolk, A. (2019). Beyond the isolated brain: The promise and challenge of interacting minds. *Neuron*, *103*(2), 186–188. https://doi.org/10.1016/j.neuron.2019.05.009

Wheatley, T., Thornton, M. A., Stolk, A., & Chang, L. J. (2023). The emerging science of interacting minds. *Perspectives on Psychological Science*, *19*(2), 355–373. https://doi.org/10.1177/17456916231200177

Wohltjen, S., & Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences*, *118*(37), 2106645118. https://doi.org/10.1073/pnas.2106645118

Yao, S., Zhao, J., & Yu, D. (2022). *ReAct: Synergizing reasoning and acting in language models*. https://doi.org/10.48550/arxiv.2210.03629