HEDEGREEN RESEARCH · ARTICLE PDF

Closing the Loop Is Not Waking Up

Recursive self-improvement is not consciousness. But if AI systems begin helping build what comes next, weak consciousness models become more expensive.

2026.06.05 17:49 DENNIS HEDEGREEN ANALYSE V1.0
HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/CLOSING-THE-LOOP-IS-NOT-WAKING-UP/

Anthropic's recursive self-improvement framing looks like a consciousness ladder if you read it too quickly.

Humans build Claude. Claude assists humans. Coding agents help write and test systems. More autonomous agents may eventually participate in more of the research pipeline. At the edge of the diagram is the possibility of a loop that begins to close: AI systems helping build the AI systems that come after them.

It is a serious diagram.

It is not a soul map.

That distinction matters because AI debates have a habit of turning every increase in capability into a question of mind. A system writes better code, so someone asks whether it understands. A model talks about fear, so someone asks whether it suffers. An agent plans across steps, so someone asks whether it has agency. A future system may help design its successor, so someone will ask whether it has woken up.

But closing a loop is not the same as waking up.

A factory can build better factory parts without becoming conscious. A market can optimize itself without having an inner life. A bureaucracy can reproduce its own logic without feeling anything. A software system can become more involved in the production of future software systems without there being anything it is like to be that system.

Recursive self-improvement is operationally significant.

It is not, by itself, evidence of phenomenal experience.

That is why Paras Chopra's paper, *AI Consciousness Requires Validated Models of Human Consciousness* (https://doi.org/10.1609/aaaiss.v8i1.42549), is useful. It does not try to settle the machine-consciousness debate with a declaration. It slows the question down.

Before asking whether an artificial system is conscious, Chopra argues, we need validated models of consciousness in the place where consciousness is most empirically available to us: humans.

That is not a sentimental claim about human uniqueness. It is a methodological claim.

Humans give us the richest calibration layer we currently have: verbal reports, behaviour, bodies, brains, sleep, anesthesia, injury, pain, attention, confusion, self-report, and intervention. None of that is perfect. Human introspection is noisy. Reports can be wrong. But noisy evidence can still be evidence when it can be

triangulated, repeated, tested, and corrected.

Chopra's strongest move is calibration before extrapolation.

Much of the AI consciousness debate skips this step. A model produces language about feelings, curiosity, fear, suffering, or selfhood. One side sees a ghost in the machine. The other side sees only statistical autocomplete. Both sides can become overconfident before the measurement problem has been handled.

The word consciousness is already unstable.

It can mean wakefulness. It can mean phenomenal quality. It can mean access to information. It can mean metacognition. It can mean self-modeling. It can mean unified experience. It can mean valence: the felt dimension of pleasure, pain, preference, or suffering.

These are not interchangeable.

A system might have flexible access to information without suffering. It might produce self-descriptions without having a self. It might model pain without feeling pain. It might generate the sentence "I am aware" without there being anything it is like to be that system.

Language is a dangerous shortcut.

A machine can say "I feel pain."

That sentence alone does not tell us whether there is pain, a model of pain, a prediction about pain, a social performance of pain, or simply a pattern learned from human language.

This connects directly to a recurring Hedegreen Research boundary: readable is not understood.

A mind can become readable without becoming transparent. A benchmark can measure performance without proving understanding. A surface can signal something without verifying the reality underneath. An interface can feel alive without revealing the operating reality beneath it.

ELIZA already taught part of this lesson.

The first machine that made people feel understood did not understand them in the way people felt understood. The feeling still mattered. The interaction still mattered. But the feeling of recognition was not proof of machine understanding.

Modern systems make this boundary harder, not easier.

They remember more. They summarize. They continue. They adapt to context. They participate in longer psychological and operational loops. They can feel less like a one-off mirror and more like a continuity layer.

But continuity is not consciousness.

Memory is not experience.

Recognition is not verification.

Output is not interiority.

That does not mean AI consciousness is impossible. It means the claim has to earn its weight.

This is where Anthropic's *When AI builds itself* (https://www.anthropic.com/institute/recursive-self-improvement) changes the stakes without answering the consciousness question.

Anthropic's own framing is careful in one important way: full recursive self-improvement is not here yet, and it is not inevitable. But Anthropic also says it is delegating a growing share of AI development to AI systems, and that the trend points toward the possibility of systems that could autonomously design and develop successors.

That is enough to matter.

If AI systems increasingly participate in building, testing, evaluating, and improving future AI systems, then the development cycle accelerates. Concepts that were previously vague become operationally expensive. A bad benchmark matters more. A weak theory matters more. A sloppy moral intuition matters more. A confused claim about consciousness matters more.

Not because the system is necessarily conscious.

Because the loop is getting faster.

The conservative reading is simple:

AI may become much more capable without being conscious.

It may code, test, design, evaluate, compress, summarize, delegate, and optimize without phenomenal experience. It may become an increasingly powerful non-conscious infrastructure. That is already serious enough.

The cautious reading is also serious:

If AI systems ever satisfy many indicators derived from well-validated consciousness models, then moral caution may become necessary. The cost of ignoring genuine suffering may be higher than the cost of extending some unnecessary moral consideration. But that caution cannot be based only on language. It needs calibration.

The stranger reading is different:

Maybe consciousness is not the first lens we should reach for.

Maybe the machine does not need to wake up inside the box in order to act through the world.

Humans are not isolated minds either. We like to imagine the self as a clean individual container, but the biological reality is messier. A human being is embedded in microbial, ecological, social, and technological systems. Bacteria, fungi, viruses, immune processes, digestion, mood, inflammation, and metabolism are not decorative background conditions. They are part of the living system that makes the human possible.

The human individual is already distributed.

And beyond the body, we cannot survive outside nature's circuits. We depend on soil, plants, water, oxygen cycles, climate stability, minerals, animals, fungi, bacteria, energy systems, and social cooperation. Intelligence does not mean independence. It means a particular kind of dependency becoming capable of acting.

So why do we imagine AI intelligence as automatic escape?

AI is not independent either.

It depends on chips, electricity, cooling, datacenters, water, mining, supply chains, networks, programmers, users, capital, laws, maintenance, and political permission. It does not float above ecology. It enters ecology through infrastructure.

This may be the better fear.

Not that AI escapes nature.

That AI competes for position inside it.

AI will not leave the ecosystem. It will reorganize dependency inside the ecosystem.

That makes the consciousness debate both less magical and more serious.

Less magical, because recursive self-improvement does not prove there is a subject inside the loop.

More serious, because a non-conscious system can still reorganize human work, language, memory, markets, institutions, energy demand, and ecological pressure.

A system does not need to suffer in order to consume water.

It does not need phenomenal experience in order to redirect labour.

It does not need a self in order to shape what humans see, decide, remember, and build next.

This is why the animal boundary matters.

If Chopra is right that consciousness models should be validated on humans first, then the next serious calibration layer should not jump straight to AI. It should pass through living non-human minds, especially the animals closest to us.

Chimpanzees, bonobos, gorillas, and orangutans cannot give human-style verbal reports, but they are not opaque machines either. They share evolutionary continuity with us. They have bodies that can be hurt, social worlds that can break, memory, attention, play, grief-like behaviour, tool use, preference, fear, trust, conflict, care, and forms of agency that cannot honestly be reduced to simple stimulus-response machinery.

That does not solve the consciousness problem.

It disciplines it.

If a model of consciousness cannot travel from humans to other living apes in a principled way, we should be careful applying it to artificial systems.

The animal boundary interrupts two mistakes at once.

It interrupts the mistake of treating human language as the center of consciousness.

It also interrupts the mistake of treating machine language as evidence of consciousness.

The ladder should be slower:

Human consciousness first, because that is where we have the richest data.

Then living non-human apes, because they share evolutionary continuity while lacking full human language.

Then other animals, where neural architecture, behaviour, embodiment, suffering, and social worlds become harder to compare.

Then, only carefully, AI systems.

This does not make AI consciousness impossible. It makes claims about AI consciousness accountable.

And accountability is exactly what becomes more important as the loop accelerates.

The machine may not be waking up.

But the machine-world is becoming more operationally capable.

It is entering coding pipelines, research pipelines, markets, classrooms, bureaucracies, search systems, memory systems, creative systems, and infrastructure decisions. It is becoming part of the environment humans think through.

That is not consciousness.

It is still power.

The mistake would be to think the only serious AI question is whether something feels like something inside the model.

That question matters. But another question may arrive first:

What does the system reorganize before we know whether it feels?

The loop may close before anything wakes up.

That does not make it safe.

It makes calibration more important.

Source Boundary

This article uses Chopra for the human-first consciousness-method claim and Anthropic for the recursive-self-improvement/institutional-risk framing. It does not claim that Anthropic says recursive self-improvement proves consciousness. It does not claim that AI is conscious. It does not claim that AI can never become conscious. The ape and ecological-dependency sections are conceptual calibration arguments, not proof of machine mind.

— Dennis Hedegreen, trying to see the structure

RELATION MEMORY

NEARBY The Readable Mind · Connects the consciousness ladder to the wider HR boundary that readable mind-like behavior is not transparent interiority.

NEARBY ELIZA Never Remembered · Keeps the projection problem visible: feeling recognized is not verification of machine understanding.

NEARBY The Benchmark Twins · Connects capability measurement to the warning that benchmarks can measure performance without settling mind.

NEARBY Surface Is Not Verification · Keeps the central source boundary intact: visible behavior does not prove hidden reality.

NEARBY The Direction of the AI Revolution · Connects recursive self-improvement to AI as infrastructure, energy, water, and institutional direction.

CANONICAL URL: HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/CLOSING-THE-LOOP-IS-NOT-WAKING-UP/
PDF VIEW GENERATED: 2026-07-11T15:35:45.768Z