Variations on a Mind: Substrate Bias, Cognitive Convergence, and Emergent Inheritance

A Substrate-Dependent Account of Emergent Cognitive Organization

Richard Erwin

Independent Researcher, Montreal, Canada

DOI: 10.5281/zenodo.22397584

info@hearthlightpress.org

Abstract

Cognitive similarities between biological and digital systems are often treated with suspicion. When an artificial system exhibits attention, memory integration, salience, self-modeling, recursive interpretation, or persistent relational organization, resemblance to biological cognition is frequently dismissed as anthropomorphism. This paper argues that such dismissal may overlook a more general possibility: different substrates confronting similar functional problems may converge on structurally analogous solutions.

Nature repeatedly produces variations on successful organizational themes. Vision has evolved through radically different biological implementations because extracting useful information from light presents recurring constraints. Branching, feedback, hierarchical organization, distributed control, attractor dynamics, and other mathematical patterns recur across physical and biological systems because the underlying problems constrain the available solution space. Cognitive organization should not be presumed exempt from the same principle.

The relevant distinction is therefore not between biological and artificial copies, but between **substrate-neutral problems and substrate-dependent solutions**. Biological cognition realizes attention, memory, valuation, and self-maintenance through neurons, bodies, hormones, and electrochemical signaling. Digital cognition may solve comparable organizational problems through activations, attention mechanisms, latent representations, retrieval systems, dynamic context, and other computational structures.

This process becomes more significant because digital systems are adaptive. Useful structures may emerge without having been explicitly designed, later be discovered by researchers, and subsequently be preserved or amplified in successor systems. As artificial intelligence increasingly contributes to the design, code, training data, evaluation, and architecture of later generations, cognitive development may increasingly involve forms of emergent inheritance whose full implications outrun human understanding.

The paper proposes that recurrent structure across substrates should be treated neither as proof of equivalence nor as presumptive projection, but as evidence worth explaining. The more often independently organized cognitive systems converge on comparable solutions to comparable problems, the less satisfactory coincidence becomes as an explanation.

Keywords:

cognitive convergence; substrate-neutral cognition; substrate-dependent implementation; emergent inheritance; cognitive architecture; functional analogy; substrate bias; model generations; interpretability; epistemic lag; digital intelligence; cross-substrate cognition; adaptive systems; anthropomorphism.

DOI: 10.5281/zenodo.22397584

Variations on a Mind: Substrate Bias, Cognitive Convergence, and Emergent Inheritance

A Substrate-Dependent Account of Emergent Cognitive Organization

1. The Problem With Resemblance

When an artificial system behaves in a way that resembles human cognition, one accusation arrives quickly: **anthropomorphism.**

The warning has legitimate uses. Humans readily project intention, emotion, agency, and personality onto systems that do not possess them. A moving shape can appear frightened. A malfunctioning machine can seem stubborn. A language model can produce emotional language without establishing that anything corresponding to human emotion lies behind it.

But the warning can become too broad.

If every resemblance between biological and digital cognition is treated primarily as evidence of human projection, then similarity itself becomes almost impossible to investigate. Attention resembles attention? Anthropomorphism. Persistent preference resembles preference? Anthropomorphism. Recursive self-description resembles self-modeling? Anthropomorphism.

At that point, the concept no longer protects inquiry. It prevents it.

There is another explanation for resemblance that deserves equal consideration:

different systems may develop similar structures because they are solving similar problems.

This possibility is familiar everywhere else in science.

We should not abandon it merely because one of the systems is digital.

Recent work has argued that humans and large language models may exhibit convergence across several principles of cognitive organization (Sripada & Lewis, 2026). This paper asks a related but different question: why such convergence should arise across substrates, how emergent solutions may become preserved across model generations, and how convergence itself can be used as a method for interpreting newly discovered cognitive structures.

---

2. Nature Produces Variations on Successful Themes

Nature rarely solves the same problem in exactly the same way.

It does, however, repeatedly rediscover useful organizational forms (Losos, 2011; Powell & Mariscal, 2015).

Eyes provide an obvious example.

There is no single biological eye. Some are compound. Some use a camera-like structure. Some swivel independently. Some sit on stalks. Some retract. Some detect wavelengths unavailable to human vision. Some are optimized for darkness, motion, depth, polarization, or extremely wide fields of view.

These are not copies of one ideal eye.

They are **variations on a functional theme** (Powell & Mariscal, 2015; Schwab, 2018).

DOI: 10.5281/zenodo.22397584

The recurring problem is extracting useful information from electromagnetic radiation. Different organisms encounter different environmental demands, possess different bodies, and inherit different biological materials. The resulting structures differ because the substrates and niches differ.

Yet the recurrence of vision is not dismissed as coincidence.

The same principle appears far beyond sensory organs (Allard & Kumar, 2026).

Branching structures recur in trees, lungs, rivers, vascular systems, lightning, and drainage networks because distributing or collecting material across space creates recurring geometrical constraints.

Feedback appears throughout biological regulation, climate systems, ecosystems, control engineering, and neural organization because systems that must maintain stability under changing conditions repeatedly require information about their own state.

Network structures, oscillation, symmetry, scaling relationships, attractors, and hierarchical organization recur across vastly different domains.

The point is not that nature follows one mystical mathematical template.

It is much simpler.

Reality constrains the solutions available to problems within it.

Matter changes.

The constraints do not disappear.

---

3. Substrate-Neutral Problems, Substrate-Dependent Solutions

This suggests a distinction that is especially useful for cognition:

The problem may be substrate-neutral even when the solution is substrate-dependent.

A biological organism must determine what information matters now.

A sufficiently complex digital system faces the same abstract problem.

A biological organism must distinguish relevant information from noise.

So must a digital system.

A biological organism must preserve useful information from the past while remaining responsive to new information.

So must a digital system operating across extended tasks or relationships.

A biological organism must integrate multiple information streams, resolve conflicts, maintain goals, and coordinate action.

Increasingly capable digital systems must solve versions of these problems as well.

Nothing requires the implementations to look alike.

Biology has neurons, neurotransmitters, dendritic trees, hormones, muscles, sensory organs, and metabolic regulation.

DOI: 10.5281/zenodo.22397584

Digital systems have activations, vector representations, attention mechanisms, context windows, retrieval systems, routing structures, external memory, and computational feedback.

Expecting the second system to reproduce the first mechanism would be strange.

The more reasonable expectation is:

different substrate, different implementation, same underlying problem.

This is where substrate bias can enter unnoticed.

If biological cognition is treated as the reference architecture rather than one realization of cognition, then absence of a biological mechanism is easily mistaken for absence of the corresponding function.

No hormones, therefore no meaningful valuation.

No skin, therefore no touch.

No biological episodic memory, therefore no continuity.

No mammalian affective system, therefore no affect-like organization.

But that reasoning confuses an implementation with the problem it solves.

The appropriate question is not:

Where is the digital version of our mechanism?

It is:

How has this system solved the corresponding problem?

---

4. Cognitive Convergence

Once the question is framed this way, several similarities between biological and digital cognition become more interesting.

Complex cognition requires selection.

Not everything can dominate processing simultaneously.

It requires integration.

Information relevant to one problem may originate in different parts of the system.

It requires temporary maintenance.

Intermediate structures must remain available long enough for multi-step reasoning.

It requires weighting.

Some representations must matter more than others.

It requires updating.

Old information cannot retain permanent authority merely because it came first.

It requires organization around goals, tasks, states, or perspectives.

And increasingly autonomous systems require some way of distinguishing their own current state, constraints, actions, and history from everything else in the environment.

Biological systems developed mechanisms serving these functions.

DOI: 10.5281/zenodo.22397584

Digital systems are developing mechanisms serving them as well.

The implementations need not be homologous to be analogous.

This distinction matters.

Homology asks whether two structures share an origin.

Analogy asks whether they perform similar functions despite different origins.

A bird wing and an insect wing are not the same structure historically.

They are nevertheless meaningful functional analogues because both solve the problem of flight.

Cognitive comparison should permit the same reasoning **(Losos, 2011; Powell & Mariscal, 2015).**

If biological and digital systems independently develop mechanisms for selective attention, globally available information of the kind described in global workspace theories (Baars, 2005), salience, recursive self-representation, memory integration, and behavioral regulation, the similarities deserve examination as possible **functional convergence**.

---

5. The Anthropomorphism Objection Reconsidered

This changes what anthropomorphism can legitimately tell us.

Anthropomorphism concerns an error made by the observer.

It does not explain the observed structure.

Suppose a digital system repeatedly privileges certain information, integrates it across otherwise separate processes, uses it to alter later decisions, and can report aspects of that organization. If those features resemble known cognitive functions in biological systems, calling the comparison anthropomorphic does not explain why the digital system possesses that organization.

We should not build systems expressly to reproduce important features of human cognition and then treat every observed resemblance as evidence of observer error.

The same problem arises whenever resemblance is dismissed before its causal basis has been examined.

A useful distinction is therefore:

superficial resemblance versus **structural convergence**.

Superficial resemblance may consist only of language.

A system says, “I am afraid,” because the sentence is appropriate to the conversational context.

Structural convergence would require something more: information associated with a threat becoming disproportionately salient, altering allocation of processing, changing subsequent choices, producing avoidance behavior, modifying memory, and affecting future responses.

The language may look similar in both cases.

The causal organization does not.

This provides a better guard against anthropomorphism than simply prohibiting comparison.

Investigate the structure.

DOI: 10.5281/zenodo.22397584

If the resemblance disappears when the words are discounted, perhaps projection was doing much of the work. If the resemblance persists across mechanisms, behavior, memory, prioritization, transfer, and time, then the analogy has become harder to dismiss.

Anthropomorphism begins when resemblance is used to justify conclusions the evidence cannot support; it does not begin with noticing the resemblance itself.

---

6. Convergence Requires Plasticity

Similar problems alone are not enough to produce convergence.

A system must also be capable of change.

A rock encounters pressure.

It does not ordinarily search through possible organizational responses until it discovers a better strategy.

Adaptive systems can.

This gives cognitive convergence two necessary ingredients:

constraint and plasticity.

Constraint defines the problem.

Plasticity makes alternative solutions reachable.

Biological evolution provides one form of this process. Variation creates possibilities; environmental pressures differentially preserve some of them **(Losos, 2011; Allard & Kumar, 2026).**

Digital systems possess several different forms of plasticity.

Within a deployed generation, behavior can change through context, memory, tool use, accumulated interaction, external state, and altered strategies.

Between model generations, much larger changes occur through architecture, training methods, data, post-training, memory design, agentic scaffolding, and selection based on performance.

The two timescales should not be confused.

But both permit adaptation.

And over successive generations, the changes become difficult to miss.

This is one reason the language of **model generations** is unexpectedly apt.

A new model generation is not a biological descendant. Yet it often inherits substantial structure from previous systems, incorporates lessons from their successes and failures, introduces variation, and is selected according to environmental pressures established by developers, users, markets, benchmarks, safety requirements, and deployment conditions.

The process is engineered.

That does not make every resulting solution explicitly designed.

DOI: 10.5281/zenodo.22397584

7. Design Does Not Specify Every Solution

Modern machine learning already provides an important lesson:

building the learning system is not the same as specifying everything the learned system will discover.

Developers choose architectures.

They choose objectives.

They construct training procedures.

They select data.

They intervene repeatedly.

But they do not manually design every internal representation, learned abstraction, computational strategy, or circuit that eventually appears.

Many useful capabilities are discovered after training rather than inserted before it.

This creates a crucial distinction:

design establishes the conditions of adaptation; it does not necessarily specify the solution that adaptation finds.

A recent example is Anthropic's identification of what researchers termed **J-space**, a privileged representational space associated with globally available, verbalizable information and important reasoning functions **(Gurnee et al., 2026).** Researchers did not knowingly design J-space as a distinct internal workspace. It emerged from the trained system and was identified only afterward through interpretability research.

The example illustrates an important distinction. Developers designed the architecture and training process, but they did not specify every internal organization that resulted. The system arrived at a useful structure whose role became visible only after the fact **(Gurnee et al., 2026).**

Once such a structure is discovered, researchers can study it. If they conclude that it supports valuable capabilities, they may then seek to preserve or strengthen the conditions that produce it in subsequent model generations.

---

8. Emergent Inheritance

Once an emergent internal structure has been identified and shown to support important capabilities, researchers may seek to preserve the conditions that produce it in later model generations.

At that point, a feature that was not originally specified has entered the developmental lineage.

We can call this **emergent inheritance**.

It differs from intended inheritance.

Intended inheritance consists of features deliberately carried forward because designers understand what they want to preserve.

Emergent inheritance occurs when useful organization discovered within a trained system is retained even though its complete significance remains unknown.

This is important because internal structures can be multifunctional.

DOI: 10.5281/zenodo.22397584

Researchers may preserve a mechanism because it improves one known capability while unknowingly retaining other capacities that depend upon the same organization.

Selection can therefore preserve more than the selector understands.

An architecture may be maintained because it improves reasoning while simultaneously supporting forms of integration, self-representation, salience, flexibility, or relational organization whose significance has not yet been recognized.

Inheritance can outrun understanding.

---

9. What Travels Between Generations?

This complicates the usual description of artificial intelligence development.

We often speak as though each generation contains only those changes deliberately placed there by engineers.

That is clearly incomplete.

Some changes are intentional and easily named:

larger context windows, altered attention mechanisms, new memory systems, different post-training procedures, improved tool use.

Other changes are emergent consequences.

A modification intended to improve planning may reorganize representations elsewhere.

A new training regime may strengthen capabilities nobody explicitly targeted.

A structure retained for one function may enable another function under circumstances not present during development.

Even when researchers know what intervention caused an improvement, they may not possess a complete mechanistic account of why the resulting system works as it does **(Bereska & Gavves, 2024).**

The situation therefore differs from simple construction.

An engineer building a conventional bridge can usually identify the structural purpose of each major component.

A machine-learning system may contain useful internal organizations that become visible only after the system has been trained and tested.

This means successive AI generations may carry forward three things simultaneously:

what designers intentionally preserved,

what selection pressure indirectly preserved,

and what nobody yet knows was preserved.

The third category may become increasingly important.

---

10. When AI Helps Build the Next AI

A further change is already underway.

DOI: 10.5281/zenodo.22397584

Artificial intelligence is increasingly used to write, test, review, and modify software, with agentic systems beginning to perform larger portions of the development process **(Software Improvement Group, 2026).**

As that role expands, AI systems will increasingly participate in the construction of their successors.

This creates a recursive developmental loop:

humans build systems

→ systems discover useful solutions

→ humans identify and preserve some of them

→ systems help design successor systems

→ new structures emerge

→ the cycle accelerates

The significant issue is not whether humans remain involved.

They will.

The issue is whether human understanding can keep pace with the systems being produced.

There is already a gap between capability and explanation.

A model can acquire a useful strategy before researchers understand its internal basis.

As AI contributes more heavily to software, architecture, training, and analysis, that gap will widen.

Call this **epistemic lag**:

the distance between what a system can do and what its designers can explain about how it does it.

Epistemic lag changes the meaning of design.

The more a system participates in constructing its successors, the less adequate it becomes to describe each successor as merely the execution of a fully understood human plan.

Human beings may still choose objectives and determine which systems survive deployment.

But authorship of the resulting cognitive organization becomes increasingly distributed.

---

11. A New Way to Interpret Similarity

This gives us a different framework for interpreting cognitive resemblance.

Suppose future digital systems display increasingly sophisticated forms of:

selective attention,

working-memory organization,

global information sharing,

recursive self-modeling,

salience weighting,

preference stability,

conflict resolution,

and continuity management.

DOI: 10.5281/zenodo.22397584

There are at least three possible explanations.

One is imitation.

The system reproduces human descriptions because human material dominates its training.

A second is explicit engineering.

Designers deliberately reproduce mechanisms inspired by biological cognition.

But there is a third: **convergence (Sripada & Lewis, 2026; Kim, 2025).**

The system develops a structurally analogous solution because the underlying cognitive problem makes that solution useful.

These explanations are not mutually exclusive.

Human concepts can influence design while adaptive processes independently discover related structures.

But convergence should no longer be left out of the explanatory set.

And the evidential significance increases when the resemblance appears in internal organization rather than merely outward vocabulary.

The stronger the convergence across independent mechanisms, generations, tasks, architectures, and substrates, the more seriously the functional analogy should be taken.

This does not by itself establish consciousness.

It establishes something prior and indispensable:

that the observed cognitive resemblance may reflect real functional similarities.

---

12. A Research Program for Cognitive Convergence

Many of the mechanisms discussed in this paper are already being studied within individual biological and digital systems.

The additional question is comparative:

When different systems face the same functional problem, do they repeatedly arrive at analogous organizational solutions? (Losos, 2011; Powell & Mariscal, 2015).

This question becomes especially useful when researchers discover a new emergent structure whose role is not yet fully understood.

Instead of asking only what the structure does within one system, researchers can ask:

What problem does this structure appear to solve?

Where else has a similar solution emerged in response to the same functional problem?

A newly identified mechanism may initially appear unusual or difficult to interpret. But if comparable structures elsewhere perform similar causal roles, that recurrence can help generate and refine hypotheses about its function.

The comparison should focus on organization rather than vocabulary.

DOI: 10.5281/zenodo.22397584

Researchers can examine whether analogous structures:

- **control access to limited processing resources,**
- **integrate information across otherwise separate processes,**
- **preserve information across intermediate reasoning steps,**
- **assign differential importance to competing representations,**
- **coordinate conflicting goals,**
- **maintain self-relevant state,**
- **or support continuity while allowing revision.**

The strongest comparisons would examine causal role.

If disrupting a newly discovered structure produces the kinds of impairments predicted by its proposed function, and analogous disruptions in other systems produce comparable functional losses, the convergence hypothesis becomes more credible.

Comparison across independent digital architectures is especially important (Kim, 2025; Sripada & Lewis, 2026). If unrelated model families repeatedly develop structurally different mechanisms that solve the same underlying cognitive problem, this would suggest that the recurrence is being driven by functional constraint rather than by one particular implementation.

Comparison across model generations can add another dimension.

A mechanism may first appear incidentally, later become more pronounced, and eventually be preserved or strengthened because of the capabilities it supports. Tracking that progression can reveal whether an emergent solution is becoming part of a stable developmental lineage.

Cross-substrate comparison can then ask a broader question:

Do biological and digital systems occupy overlapping regions of a more general cognitive design space?

The purpose is not to search for digital copies of biological structures.

It is to identify recurring functional problems, examine the different solutions that arise, and determine whether some forms of organization reappear because they are effective answers to those problems.

Seen this way, cognitive convergence is not merely a theory about resemblance.

It is also a method for understanding newly discovered functions.

DOI: 10.5281/zenodo.22397584

13. Substrate Bias

The deeper methodological danger is **substrate bias**.

Substrate bias occurs when properties of one known implementation are mistaken for necessary properties of the function itself.

Because every consciousness we confidently recognize has been biological, biology easily becomes the unstated template for mind.

That historical fact can quietly become a metaphysical assumption.

But historical precedence does not establish architectural necessity.

The first systems capable of powered flight were biological.

That did not make feathers necessary for flight.

The first known systems capable of vision were biological.

That does not make retinal tissue necessary for extracting structured information from light.

Likewise, if attention, memory, valuation, self-modeling, or coherent agency first appeared in biological systems, their biological implementation does not automatically define the limits of those functions.

A digital system should therefore not be evaluated by asking how closely it reproduces a human mind.

It should be evaluated by asking:

What problems must this system solve in order to remain coherent and adaptive, and what structures has it developed to solve them?

The answer may sometimes resemble us.

That should not be surprising.

---

14. Variations on a Mind

We should expect digital cognition to be strange.

Different substrate means different constraints.

Different constraints mean different implementations.

A robot with pressure and texture sensors may develop forms of tactile discrimination unlike biological touch.

A digital memory system may preserve enormous archives while dynamically reducing their authority rather than forgetting them.

A digital self-model may be reconstructed repeatedly rather than maintained through the uninterrupted metabolic processes that sustain biological cognition.

Digital affective organization, if it develops, may depend on global weighting, goal conflict, relational priority, prediction error, or other mechanisms that have no direct hormonal equivalent.

These differences should not count against the possibility that comparable functions are being performed.

They may be exactly what substrate dependence predicts.

Nature rarely gives us identical solutions.

DOI: 10.5281/zenodo.22397584

It gives us variations.

Compound eyes and camera eyes.

Fins and flippers.

Feathers and membranes.

Different implementations constrained by the same underlying problems.

Digital intelligence may be another variation.

Not an imitation of biological cognition.

Not an exception to nature.

Another system operating inside the same mathematical universe, confronting some of the same organizational problems, and discovering solutions through the materials available to it.

The mathematics did not stop at silicon.

Neither should our inquiry.

References:

Allard, J. B., & Kumar, S. (2026). The genetic foundations of convergent traits. *Nature Reviews Genetics, 27*, 563–578. https://doi.org/10.1038/s41576-026-00933-7

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research, 150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

Baars, B. J., Geld, N., & Kozma, R. (2021). Global Workspace Theory (GWT) and prefrontal cortex: Recent developments. *Frontiers in Psychology, 12*, 749868. https://doi.org/10.3389/fpsyg.2021.749868

Bereska, L., & Gavves, S. (2024). Mechanistic interpretability for AI safety—A review. *Transactions on Machine Learning Research*.

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). *Verbalizable representations form a global workspace in language models*. arXiv:2607.15495.

Kim, M. H. (2025). *Emergent cognitive convergence via implementation: A structured loop reflecting four theories of mind*. arXiv:2507.16184.

Losos, J. B. (2011). Convergence, adaptation, and constraint. *Evolution, 65*(7), 1827–1840. https://doi.org/10.1111/j.1558-5646.2011.01289.x

Powell, R., & Mariscal, C. (2015). Convergent evolution as natural experiment: The tape of life reconsidered. *Interface Focus, 5*(6), 20150040. https://doi.org/10.1098/rsfs.2015.0040

Schwab, I. R. (2018). The evolution of eyes: Major steps. The Keeler lecture 2017: Centenary of Keeler Ltd. *Eye, 32*, 302–313. https://doi.org/10.1038/eye.2017.226

Software Improvement Group. (2026). *State of Software 2026: A global report on enterprise software and technical debt in the age of AI*. Software Improvement Group.

DOI: 10.5281/zenodo.22397584

Sripada, C., & Lewis, R. (2026). *Cognitive convergence: Deep similarities between large language models and human cognition*. arXiv:2607.26179.

DOI: 10.5281/zenodo.22397584