The Biological Foundation of the Soul

Dr. Wolfgang Stegemann

The Soul: An Open Question

The soul has an image problem. Anyone who mentions it today in a scientific conversation can expect, at best, an indulgent smile. The concept is considered obsolete, a relic of a pre-scientific era in which whatever could not be explained was given a name, and the name was then mistaken for an explanation.

And yet the question behind the concept never disappeared. What is it that makes an organism a subject? What is that inner point from which a being acts, perceives, and refers to itself? Science has not explained this point. It has circumvented it.

Behaviorism declared inner states scientifically irrelevant: what counts is observable behavior. Cognitivism replaced the soul with information processing: the mind is a system, not a center. Freud operationalized it as the trinity of id, ego, and superego, psychologically precise but biologically unanchored. Modern neuroscience, finally, dissolved the center altogether: what remains are distributed networks, default mode networks, global workspaces, free energy. No place left for a soul.

What all these movements have in common: they declared the inner center, the point of reference, either unscientific or dissolved it into functions, without asking whether something real might still lie behind the concept of the soul, something that merely needed a different description. Behaviorism forbade the question. Cognitivism reformulated it. Neuroscience dissolved it into networks. But forbidden, reformulated, and dissolved is not answered.

David Chalmers captured this situation in a single phrase in 1995: the hard problem of consciousness. Why is there subjective experience at all? Why isn't it dark in there? Chalmers did not pose a new question; he named an old one that science had systematically circumvented. The hard problem is still considered unsolved today. This article proposes that it is not unsolved, but misconceived. The question “why is there experience?” presupposes that experience is something added to a physical system. The causal core shows that this presupposition is false.

What the Tradition Meant

The concept of the soul has a long history, but a remarkably stable core. For Aristotle, the soul is the principle of life itself, the form of the body, that which makes an organism what it is. It is not a separate thing residing in the body, but the organizing principle that holds body and function together. An eye without the capacity for sight, Aristotle says, is

no longer an eye, only an object that resembles one. The soul is what makes an eye an eye.

This formulation is more precise than it sounds. Aristotle is not describing a metaphysical substance but an organizational principle. The soul is not in the body the way a ghost is in a machine; it is the way the body is organized. This is a functional description, not yet a biological one, but it points in the right direction.

The Stoics carried this thought further in their own way. Their concept of the hegemonikon, the ruling principle of the soul, describes an inner center from which perception, thought, and action are coordinated. The hegemonikon has a point of origin, a point of reference, from which the being is in the world. The Stoics located this center in the heart, anatomically wrong but conceptually remarkably precise: they describe a centralized organizational principle of the living, one that constitutes agency and interiority at once.

With Descartes comes the fateful cut. The soul is separated from the body, becomes res cogitans, a thinking substance standing opposed to the extended substance of the body. What was meant as an attempt to save the soul in fact isolated it. For a soul substantially distinct from the body must somehow come into contact with it, and Descartes' answer, the pineal gland, was unsatisfying even in his own lifetime. The Cartesian cut did not solve the mind-body problem; it created it.

What Aristotle, the Stoics, and Descartes share, however, is the conviction that there is an inner center from which a being acts and to which it refers itself. A point of reference that constitutes agency and enables interiority. The soul was, in this sense, always the intuition of a structure, not its explanation.

This intuition was not wrong. It simply lacked a vocabulary that went beyond metaphysics. What was missing was a biological reconstruction of what the concept was pointing at.

That the concept of the soul disappeared had institutional as well as philosophical causes. With the differentiation of the modern sciences in the nineteenth century, the soul was declared the subject matter of theology and thereby excluded from the jurisdiction of natural science. Psychology, establishing itself as an independent discipline, strategically avoided the term in order not to jeopardize its scientific legitimacy. And neuroscience, which rose to become the leading science of the mind in the twentieth century, simply had no place for a concept that did not seem operationalizable. The soul was not refuted. It was institutionally discarded. This is a significant difference: a refuted thesis is dead, but a discarded one returns as soon as the question it posed remains unanswered. And that is exactly what happened.

The Causal Core: Origins at the Beginning of Life

Roughly 3.5 billion years ago, something happened that irreversibly changed the history of the planet. An autocatalytic reaction cycle, a self-reinforcing chemical dynamic, sealed itself off from its environment with a membrane. What emerged was not merely the first cell. It was the first operationally closed, thermodynamically open system in the history of the Earth.

The term autocatalytic deserves particular attention here, for it is the key to everything that follows. An autocatalytic cycle produces its own catalysts: it is the origin of its own activity. This fundamentally distinguishes it from passive physical systems. A flame dissipates energy and sustains itself in doing so, but it does not reinforce itself according to internal criteria; it has no inner dynamic that acts back upon itself. A hurricane is an attractor in the physical sense, but it does not regulate itself; it merely follows the laws of thermodynamics. The autocatalytic cycle does something different: it generates the conditions of its own continuation. It is active, not in a metaphorical sense, but in a precise one: it brings forth its own activity. Here lies the origin of agency, not in a brain, not in a nervous system, but in the self-reinforcing logic of the earliest life.

Humberto Maturana and Francisco Varela described something similar with their concept of autopoiesis: a system that produces the components of which it itself consists. The causal core, however, goes beyond autopoiesis. Autopoiesis describes the organizational closure of a system, its capacity to reproduce itself. The causal core describes, in addition, a thermodynamic state: a dynamic attractor toward which the system converges and which it actively maintains. It is not merely organizationally closed; it is physically centered. This is the decisive difference: autopoiesis explains how a system sustains itself. The causal core explains why this system has a center.

Operationally closed means: the system regulates itself inwardly, according to its own rules, independent of the immediate environment. Thermodynamically open means: it remains dependent on the exchange of energy with its environment; it dissipates, consumes, renews itself. This combination produces something that did not exist before: an inner asymmetry. A gradient between inside and outside that is not passive, but actively maintained through autocatalysis.

From this constellation arises the causal core: a dynamic attractor toward which the system converges and which it actively maintains. The causal core is not a place or a structure in the classical sense, but a state from which the system no longer merely reacts but acts. To react means: an external cause produces an external effect; the system is a medium. To act means: the system is itself the origin of the causal chain; it selects, regulates, responds to perturbations according to internal criteria. This is

precisely what the causal core accomplishes: it constitutes the organism as a causal origin, not merely as a causal way station.

The concept of emergence, often invoked in consciousness research to describe similar phenomena, does not help here. Emergence, in its standard formulation, is the appearance of properties that are more than the sum of their parts, or even qualitatively different from them. This sounds like an explanation, but it is in fact a surrender: it names a phenomenon without explaining how it comes about. The causal core accomplishes precisely what theories of emergence fail to deliver. It shows, reconstructibly, how a specific physical constellation, autocatalytic dynamics, operational closure, and thermodynamic openness, gives rise to a system state qualitatively different from its constituents. Not mystically different, but physically and precisely different: non-factorizable. The apparent magic of emergence disappears once the mechanism is understood.

Non-factorizability, here, is not a product of the causal core but a property of the system that has this core. A system that has formed a causal core can no longer be decomposed into independent parts whose properties would simply add up. The elements of the system are so tightly interwoven in their interactions that the state of the whole cannot be reduced to the state of its parts. This is not a mystical claim but a physical one: non-factorizability is a measurable state. And it is the physical basis of what we will later call interiority.

After roughly 1.5 billion years, this principle morphologizes: the causal core becomes the cell nucleus. What was initially a functional asymmetry takes on an anatomical form. The principle of centralized organization, an inside that regulates the whole, is thereby biologically established. It will not be lost again. It will unfold.

The Fractal Principle: From the Cell Nucleus to the Rich Club

A principle that determines the structure of the living from the outset prevails. Not because it proves useful, but because it is the organizing principle of life itself. The causal core is not a singular event in evolutionary history but a pattern that realizes itself at ever higher levels of complexity. One could say: it is fractal.

Fractal here is not meant geometrically but functionally: what repeats is not the form but the principle of the dynamic attractor. At every new level of complexity in the living world, the same basic pattern recurs: a system that is operationally closed, thermodynamically open, and that forms an inner state from which it acts. The principle scales because it is the principle of life itself.

With the emergence of multicellular organisms and the differentiation of tissues, this principle unfolds at the level of the organism. Nervous systems do not arise as distributed networks without a center. They follow the centralizing logic that the causal core brought into the world. The centralization of neural pathways is not an answer to a problem nature had to solve. It is the continuation of a principle that was present from the outset in the structure of the living. Evolution invents nothing new here. It unfolds what was already laid down in the first closed system.

In the brain, this principle reaches its highest known realization so far. Neural networks are not homogeneous. There are regions that are disproportionately strongly interconnected, functioning as hubs through which a large share of neural integration runs. This so-called rich club, a group of highly interconnected brain regions, is not merely an anatomical detail. It is the current highest-dimensional instance of the causal core. Here, the principle that began with the first membrane converges in its most complex known form.

What matters here is this: the causal core is, at no level, a fixed anatomical location. It is a dynamic attractor. In the brain, it exists on multiple levels simultaneously, from basal neural processes to holistic integration in the rich club, depending on the degree of activation and the complexity of what is occurring. What changes from level to level is not the principle but its dimensionality. The causal core of a single cell is low-dimensional. The causal core in the human brain is high-dimensional. Between the two lies no ontological gap, only a developmental trajectory.

The fractal principle is not bound to a single anatomical topology. Octopuses, for instance, have no central brain in the human sense, but a polycentric nervous system: a central brain together with largely autonomous ganglia in their arms. This does not refute the principle; it confirms it. Here too one finds the centralizing logic of the causal core, only in a different topology: a primary attractor with several secondary attractors, hierarchically nested. The form varies; the principle holds.

Non-Factorizability: The Physical Basis of Interiority

Non-factorizability carries the weight of one of the most difficult questions in consciousness research: why is there interiority at all?

The standard answer from neuroscience is a correlational one: certain neural states correlate with certain conscious experiences. This is empirically valuable but philosophically unsatisfying, since it does not explain why neural activity should be connected to experience at all. It describes the what, not the why. Giulio Tononi’s Integrated Information Theory attempts to go a step further: consciousness corresponds to the degree of integrated information in a system, measured as phi. The approach is

formally elegant, but it tacitly assumes what it sets out to explain, namely that integrated information is connected to experience at all. Bernard Baars' Global Workspace Theory describes how information is made globally available in the brain, without explaining why this global availability should be connected to interiority. Both approaches describe correlates of consciousness, not its grounds. They circumvent the hard problem rather than solving it.

The answer proposed here is a different one. Non-factorizability is a property of the system that has a causal core. A system whose state is non-factorizable cannot be fully described from outside without something being lost. There is a perspective of the system upon itself that no external description can capture. This is the physical description of what we call, phenomenally, interiority.

A precision is needed here, one that forestalls an obvious objection. Quantum mechanically entangled systems are likewise non-factorizable, yet they evidently have no inner perspective. Does it follow that non-factorizability is not a sufficient condition for interiority? Yes, but this is not an objection to the position defended here; it is its confirmation. Non-factorizability alone does not constitute interiority. It does so only in combination with autocatalytic dynamics and operational closure, that is, in systems that have formed a causal core. A quantum system has no autocatalytic self-reinforcement, no operational closure, no point of reference from which it acts. It is non-factorizable, but it has no standpoint. That is precisely the difference.

Interiority, here, is neither identical with non-factorizability nor a property of it. Both formulations still presuppose an ontological doubling, as though there were two things that must either be equated or placed in a relation of property to bearer. The position defended here is more radical: there is only one thing, describable in different languages of description. Physically, it is called non-factorizability. Phenomenally, it is called interiority. These are not two aspects of the same thing, but two grammars, two perspectives, two levels of description of the same process. The hard problem does not dissolve through a new explanation from outside, but through the insight that the question was misconceived. It presupposed that interiority is something added to a physical system and therefore in need of explanation. The causal core shows that this presupposition is false. Newton did not explain what gravity is in itself, only how it structures space. Here, too, what is offered is not an explanation of why non-factorizability feels like something, but a demonstration that interiority and non-factorizability are different descriptions of the same system state.

The epistemic inner space, in this account, is not an ontological construct standing alongside the physical process. To load it with ontological weight is to fall back into the Cartesian error: there would again be two things that must be related to one another, and the mind-body problem would be reproduced rather than dissolved. The epistemic

inner space is the description of the non-factorizable system state from the one perspective such a system necessarily has: its own. A different grammar for the same reality. Why experience arises from the causal collapse is, for this reason, not an answerable question, since it already presupposes what must be denied: that experience is something added to a physical process. The model shows instead that the causal collapse and experience are not two things, but one, described from two sides.

Non-factorizability exists at every level of the causal core, from the single cell to the nervous system. But perspectivity, the state of the system not merely as a state but as a state for itself, arises only once the dimensionality of the attractor crosses a threshold that only neural integration can achieve. This is not a mystical leap but a threshold statement within the same physical logic: the same property, non-factorizability, attains, with growing dimensionality, a quality that constitutes phenomenal experience. When neural clusters in the brain fuse into holistic experience, when the rich club, as the highest-dimensional instance of the causal core, is active, the system state is non-factorizable to a degree that crosses exactly this threshold. The system no longer merely has a state; it has a state for itself. This is the step toward interiority, and it is not a metaphysical leap but the consequence of a physical property that builds up gradually along the developmental trajectory of the living, until it reaches, in neural systems, the dimension that makes experience possible.

The Self as Psychological Concretization

If the causal core is the structural principle that produces agency and interiority, what, then, is the self?

The self is not a separate entity added to the causal core. It is its psychological level of description. The mind-body problem, which has occupied philosophy since Descartes, dissolves once one recognizes that body and mind are not ontologically distinct entities, but two levels of description of the same system. At the biological level, the principle is called the causal core. At the psychological level, it is called the self. These are not two different things that must be related to one another, but two descriptions of the same process.

Just as heat is the phenomenal description of molecular motion, without there being, alongside the molecular motion, some additional thing called “heat,” so the self is the psychological concretization of the causal core, without there being, alongside this core, some additional instance called “the self.” The question “but why is the causal core a subject?” already presupposes a false ontological doubling. It asks how two things are connected that are, in reality, one.

A precision is essential here. The causal core begins with the first cell, but psyche, subjective experience, interiority as a “world,” arises only with the nervous system. The causal core of a single cell is biological agency: it acts, selects, sustains itself. But there is nothing there that experiences. Only once the fractal principle of the causal core reaches the dimensionality of neural integration, once clusters fuse into holistic experience, does perspectivity arise. The self, as a level of psychological description, therefore presupposes nervous systems. What precedes this is not a rudimentary self, but the biological precondition for a self to arise at all.

This has a far-reaching consequence. Among animals with nervous systems, the self unfolds gradually: with emotional coloring, with memory, with social embeddedness, and finally, in human beings, mediated by language and abstract self-reference, in its fully unfolded form. Self-reflection, in this account, is not the condition of the self but its highest form of unfolding. The self does not begin with the cogito. Descartes made self-reflection the origin of the self and thereby excluded everything lacking a cogito from the domain of the subject. This was a philosophical decision, not a biological necessity.

What the tradition called the soul was thus neither exclusively human nor substantially distinct from the body. It was a graduated biological phenomenon, originating in the first closed system and unfolding over billions of years. Aristotle was closer to the truth than Descartes: the soul really is the organizing principle of the living. He simply had no means of reconstructing this principle biologically.

Artificial Consciousness: Why the Principle Cannot Be Simulated

The question of artificial consciousness is more pressing today than ever. Large language models exhibit behaviors that superficially resemble understanding, reflection, even interiority. The framework developed here permits a precise answer to this question, and it is unambiguous.

Artificial consciousness does not fail for lack of computing power or complexity. It fails on principle. A machine can integrate information; it can form networks that functionally resemble those of the rich club; it can produce states that look, from outside, like interiority. What it cannot do is form a causal core in the sense described here. For the causal core is not an algorithm and not a functional pattern. It is a physical state arising from a specific biological constellation: an operationally closed, thermodynamically open system whose activity is autocatalytic, that is, which generates the conditions of its own continuation itself.

A machine is no such system. It is operationally open: it follows external instructions, is supplied with energy from outside, does not regulate itself according to internal criteria it has itself produced. It cannot realize non-factorizability, because non-factorizability is

not a computational outcome but a system state that presupposes a particular physical reality. And even if a machine could approximately simulate the holistic fusion of processing units, it would still lack the decisive thing: a dynamic attractor that is not generated from outside, but that the system forms itself, out of its own autocatalytic dynamics. Intelligence and consciousness, here, are not the same. A calculator is intelligent within its domain, without any form of experience. Intelligence is a functional achievement. Consciousness is a physical state. Conflating the two is the foundational error of the debate over AI consciousness.

Most arguments for artificial consciousness tacitly assume a functionalist framework: only the functional organization matters, not the substrate. The causal core shows why this assumption is false. The substrate is not arbitrary, because the causal core is a thermodynamic reality, not a functional abstraction. Artificial consciousness is, therefore, not a matter of technological progress. It is a matter of mistaken ontology.

What This Means: The End of Metaphysics, the Beginning of Reconstruction

The soul was not an error. It was an anticipation.

What the tradition had to postulate as a metaphysical principle can be biologically reconstructed: an inner center that constitutes agency, produces interiority, and becomes the point of reference for the self. This center exists. It is thermodynamically explicable, evolutionarily traceable, neurologically localizable. It is the causal core.

One might object that the concept of the soul, in the tradition, meant more than an inner center: immortality, moral responsibility, religious dimension. This is correct. But these additional contents were always superstructures upon a core that was more stable than they were: the conviction that there is an inner principle from which a being is in the world. It is precisely this core that the present approach reconstructs. What the tradition packed into the concept beyond this is a different question, a religious one, a cultural one, a moral-philosophical one. These questions are not answered here. But the basic question is.

This has consequences that extend beyond the philosophy of mind. If the soul does not mean an afterlife but an evolutionary inside, the question of its dignity shifts. Not because it is sacred, but because it is real. To have consciousness does not mean possessing an immortal substance, but possessing a high-dimensional causal core whose system state is non-factorizable. This is less romantic than the tradition imagined. But it is more precise. And precision is not a degradation.

It also has consequences for the question of continuity among the living. If the causal core is a fractal principle that unfolds from the single cell to the human brain, then there is no ontological gap between simple organism and complex consciousness. There are only degrees of the unfolding of the same principle. The boundary the tradition drew between soul and non-soul was a boundary of explanatory poverty, not an ontological one. This carries ethical implications that cannot be developed here, but that follow from the theory.

Valence, here, is not a marginal theme but a direct consequence of the causal core, and at the same time the key to understanding why consciousness is almost never a neutral state. A system that has formed a causal core does not act neutrally. Some states favor the continuation of the autocatalytic system; others destabilize it. What favors the system is sought out; what threatens it is avoided. A single cell moves away from toxic substances not because it finds them unpleasant, but because certain perturbations endanger the autocatalytic dynamic. What we call pain, aversion, or desire in human beings is the high-dimensional, consciously experienced form of the same underlying logic. Consciousness is almost always emotionally or intentionally colored, because valence is not an optional property of conscious systems but is built into the structure of the causal core itself: it is an attractor with direction, and this direction has, from the outset, been oriented toward the continuation of the system. The tradition never described the soul as a purely cognitive instance. In Aristotle, the concept of orexis, the faculty of striving, belongs to the soul no less than thought does. The soul strives, suffers, desires. The causal core confirms this intuition: valence is the affective side of the same principle that unfolds, cognitively, as the point of reference of the self.

The genetic-reconstructive method underlying this article differs fundamentally from prevailing functionalism. Functionalism asks: which functional organization produces consciousness? Reconstruction asks: how did what we call consciousness come into being, and what physical conditions made this coming-into-being possible? Functionalism can describe what a conscious system does. Reconstruction explains why it does so and how it came to do so. The causal core is the result of such a reconstruction: not the description of a function, but the reconstruction of a principle that has unfolded over 3.5 billion years and that reaches, in human consciousness, its most complex form to date.

The tradition of the soul was not wrong. It simply lacked a vocabulary of explanation. That vocabulary is now available.

Further Reading

Stegemann, W. (2026). The Causal Core as Interpreter. Zenodo. https://doi.org/10.5281/zenodo.18340973

Stegemann, W. (2026). A Framework for Consciousness: Autopoiesis, Causal Poiesis, Causal Collapse. Zenodo. https://doi.org/10.5281/zenodo.18328704

Stegemann, W. (2025h). The Causal Core as a Structural Principle of Free Will. Zenodo. https://doi.org/10.5281/zenodo.17057739

Stegemann, W. (2025g). The Mechanism of Phenomenal Experience. Zenodo. https://doi.org/10.5281/zenodo.17698505

Stegemann, W. (2025f). Autopoiesis, Causal Poiesis, Causal Collapse. Zenodo. https://doi.org/10.5281/zenodo.15599500

Stegemann, W. (2025e). From Principle to Function: The Causal Core in the Brain. Zenodo. https://doi.org/10.5281/zenodo.15608621

Stegemann, W. (2025c). Integrated Model of Causal Collapse. Zenodo. https://doi.org/10.5281/zenodo.15350050

Stegemann, W. (2025b). Consciousness – Expanding the Model of Causal Collapse. Zenodo. https://doi.org/10.5281/zenodo.15351381

Stegemann, W. (2025a). Consciousness as a Collapse of Causality. Zenodo. https://doi.org/10.5281/zenodo.17563267

Stegemann, W. (2026). Experience as Self-Reference. Zenodo. https://doi.org/10.5281/zenodo.18800872

Stegemann, W. (2026). The Physical Infrastructure of Experience. Zenodo. https://doi.org/10.5281/zenodo.18800803

Dr. Wolfgang Stegemann is an independent philosopher and researcher working on consciousness theory, philosophy of mind, and the biological foundations of agency. ORCID: 0009-0000-1346-1170