A Thought Experiment on the In-Principle Impossibility of Constructed Consciousness

The thought experiment is structured as a cascade. Each station formulates a thesis about what might give rise to experience. The thesis is tested against its strongest argument and rejected where it no longer explains experience adequately. Each failure gives rise to the next thesis. In this way, the question is driven progressively deeper, step by step. The conclusion holds on the condition that the requirements developed along this path are accepted.

A precise question is being examined: Can consciousness be technically constructed, that is, can its constitutive organization be specified sufficiently and then deliberately realized? This must be distinguished from the question of whether technically created initial conditions could enable an open process in which life and, eventually, consciousness arise. This second possibility is not excluded from the outset. It must satisfy the same requirements as any other possible process of emergence.

Station zero: What does consciousness mean?

Before artificial consciousness can be discussed, it must first be clarified what is meant by consciousness.

A first position understands consciousness operationally as a set of capacities, such as perception, self-modeling, linguistic competence, or flexible behavioral control. In this sense, artificial consciousness is fundamentally a technical question of function.

A second position explicitly accepts experience but identifies it with a particular functional organization. On this view, sufficiently integrated, recursive, or self-modeling information processing already is experience. The "for the system" does not designate an additional fact, but a particular way in which information is present and effective within the system.

A third position regards precisely this identification as requiring justification. It does not presuppose that the functional "for the system" and the phenomenal "for the system" are two ontologically distinct facts. It explicitly leaves open whether they might be identical. It requires only a justification of that identity. Anyone who claims that a particular functional organization already is experience must show what makes functional availability at the same time something undergone by the system itself.

Even a purely functional concept of consciousness cannot dispense entirely with a "for the system." If it abandons this altogether, it describes complex processing, but no longer what is ordinarily meant by consciousness. If, by contrast, it identifies this "for the system" with a particular functional organization, it moves into the second position and thus precisely into the domain examined by the thought experiment.

The cascade therefore does not begin with the claim that function could never in principle explain experience. It examines, step by step, whether a functional description is actually sufficient to turn a state that is processed within a system into a state that exists for the system itself.

Station one: Consciousness is a matter of content

The first thesis is this: A system experiences when its states are sufficiently rich and differentiated, world-structured through sensors, language-structured through discourse, self-structured through recursion. The argument in its favor is that, as differentiation increases, there is increasingly more

that could be experienced. At the limit of maximal content complexity, experience therefore seems almost unavoidable.

Rejection: Every increase in content produces richer structures. It still does not show in what sense these structures exist for the system itself. Content answers the question of what is present, but not what turns this presence into experience. The description of the states can be refined indefinitely without thereby explaining the transition from content to experience. The question thus shifts from content to the mode in which it is present.

Station two: Consciousness is a matter of information

The next thesis is this: Experience arises when a system integrates and processes a sufficient amount of information. The argument in its favor is that perceiving, discriminating, and responding are informational achievements. A system that integrates information about itself and its environment therefore seems to generate a state that is available to the system as a whole.

Rejection: Information processing can be described completely as transformation. Something enters the system, changes further states, and produces certain outcomes. Global availability initially means only that information can become causally effective at many points. It does not follow that the information is undergone by the system. Anyone who identifies the two must show why the functional "for the system" and the phenomenal "for the system" are identical. That identity cannot simply be presupposed by the concepts themselves. The question thus shifts from information to causality.

Station three: Consciousness arises from network architecture

It is then natural to suppose that the decisive property lies in the structure itself. A sufficiently deep and densely interconnected system might produce experience as an emergent property. The argument in its favor is that complex networks exhibit properties that cannot be attributed to any single component. Consciousness could be such a system-level property, appearing only above a certain threshold of complexity.

Rejection: Even maximal structural complexity initially explains only additional interactions and system-level properties. In a network, states remain defined by their causal relations to other states. The language model illustrates the limiting case particularly clearly. It can produce highly complex statements about experience and refer to itself. It does not follow that these states are undergone by it. Emergence names the appearance of new properties of the system as a whole, but it does not explain why one of these properties should be experience. The question thus shifts from structure to self-feedback.

Station four: Consciousness arises through feedback

The next thesis is this: A system must act back upon itself. Only a recursive or sensorimotor closed system, in which perception and action feed back into one another, is affected by its own states. The argument in its favor is that a closed control loop makes the system's own changes into conditions for further changes of its own. Here, for the first time, a genuine "for the system" seems to appear.

Rejection: As long as the loop can be decomposed into separate stations, it remains a mechanism. Here a sensor, there processing, there an effector, followed by renewed input. The feedback changes the system, but a thermostat is also changed by its own regulatory processes. Functional self-reference therefore still does not explain what turns a change in the system into something undergone by the system. The question thus shifts from feedback to the unity of that which is affected.

Station five: The decisive criterion is non-factorizability

The next thesis is this: A system experiences when its self-feedback can no longer be decomposed without remainder into mutually independent subprocesses. The argument in its favor is that experience is given as a unified state. If this state is fully decomposed into mutually independent processes, precisely the unity that was to be explained disappears.

Rejection: Non-factorizability names a criterion, but not yet a reason. It indicates what might distinguish an experiencing system from a mere collection of isolable processes, but not what positively constitutes this unity. The question therefore becomes what, exactly, can no longer be decomposed. The question thus shifts from the criterion to the causal unity of the system.

Station six: The non-factorizable unity is a causal core

The next thesis is this: A system becomes non-factorizable where its self-feedback no longer merely runs through the system but constitutes the ongoing formation of the system itself. The system does not stand apart from its states and compute them. It is produced by its own processes and at the same time changed by them. Self-constitution and being affected coincide in the same process.

The argument in its favor is that the functional "for" here acquires, for the first time, an intrinsic reference. States do not merely act on other states. They alter the nexus that constitutes the system itself. The system is therefore not merely the location of a state change; it is itself what is changed by that change.

Rejection: Even this does not yet explain what makes being affected more than an especially tight form of internal state change. A criterion is still needed for when a change makes a difference to the system itself. This criterion cannot simply be presupposed; it must arise from the organization of the system itself. The question thus shifts from self-constitution to intrinsic relevance.

Station seven: Being affected requires intrinsic relevance

The next thesis is this: A state is undergone by a system in the relevant sense when the difference it makes to the system arises from the system's own organization. Intrinsic relevance therefore designates neither a particular substance nor a particular origin of the system, but a relational property of its organization.

Not every individual experience must directly serve self-maintenance. A color, a sound, or a light touch can be experienced without the continued existence of the organism being immediately threatened or promoted. Such experiences are nevertheless possible only because the system possesses its own structure of relevance, one that has emerged from its self-maintaining organization and can now also discriminate fine differences that are not immediately vital.

What matters at this point is therefore not whether a system arose naturally or technically. What matters is whether the conditions under which states make a difference to this system arise from its own organization or are merely specified as external target variables.

Rejection: Intrinsic relevance names a necessary condition, but it does not yet explain how a system can generate such a structure of relevance of its own. What must be clarified is how its organization itself becomes the standard by which changes count as beneficial, harmful, or neutral for it. The question thus shifts from intrinsic relevance to self-maintenance.

Station eight: Intrinsic relevance requires self-maintaining self-production

The next thesis is this: A system possesses its own structure of relevance when its organization itself generates the conditions under which it can persist as this system. Only then does a distinction arise from within the system between states that preserve, alter, or destroy its organization.

Self-maintenance in the sense intended here is not merely the stabilization of a predefined target state. It requires the system to produce the processes through which its own organization is continued. Intrinsic relevance thus leads to autocatalytic self-sustenance.

In biological self-production, more than particular states are generated. The system produces its components, its boundary, and the processes by which this production is continued. Boundary and organization are products of the system and at the same time conditions of its continued existence. The distinction between beneficial, harmful, and neutral changes thereby acquires a system-internal origin.

Even a technically produced autocatalytic system would not yet be excluded at this point. What matters would be whether it actually generates a structure of relevance of its own and whether its organization becomes the standard for its subsequent states.

Rejection: Autocatalytic self-sustenance explains how intrinsic relevance can arise from the organization itself. It does not yet explain how such a system could develop the specific complex organization in which a non-factorizable causal core and, eventually, consciousness emerge. The question thus shifts from self-sustenance to the evolution of this organization.

Station nine: Open evolution does not necessarily lead to consciousness

The next thesis is this: If an autocatalytic and intrinsically relevant system can be technically produced, one could simply allow it to evolve until consciousness emerges. The argument in its favor is that evolution produces complex organizations without requiring their individual developmental steps to be constructed in advance.

Two problems arise here.

The first concerns the openness of the process. If evolution is deliberately to produce consciousness, it must be determined which changes lead in the desired direction. There must be criteria specifying which forms of organization are favored, which developmental steps count as progress, and which are rejected. The process thereby acquires a technically imposed goal.

These criteria, however, can only be drawn from the one process we know to have actually produced consciousness, namely biological evolution. We know nervous systems, sensory integration, recursive neural organization, and other possible conditions of consciousness because they have already arisen. If we now use these features as criteria for artificial evolution, we are arguing post hoc. We take properties of the known result and retrospectively turn them into developmental goals.

Such a process would no longer be open evolution, but an optimization procedure directed toward a retrospectively specified result. Nor do theoretically derived conditions of consciousness solve this problem. Necessary conditions of an end state are not a generative prescription for its emergence. From the insight that consciousness requires, for example, intrinsic relevance, non-factorizability, and recursive integration, it does not follow which evolutionarily viable sequence of intermediate states will produce that organization. If the derived conditions themselves are made selection targets, the process is once again organized toward a predefined result. If evolution remains open, by contrast, these conditions do not determine its course.

The second problem lies deeper. Even successful open evolution does not necessarily lead to consciousness. Evolutionary success initially means only that organizations persist and reproduce under their respective conditions. It implies no developmental tendency toward nervous systems, central integration, or consciousness.

Consciousness instead requires a series of additional developmental steps. Forms of stimulus reception and signal transmission must arise, sensory and motor processes must become differentiated, increasingly integrated neural organizations and recursive connections must emerge, and finally there must arise the non-factorizable causal unity developed in the preceding stations as a condition of experience. None of these innovations follows necessarily from evolutionary success.

An evolutionary process can therefore remain successful for a very long time without ever producing a nervous system or any other organization relevant to consciousness. Evolution contains no intrinsic pull toward consciousness. Consciousness is not a necessary endpoint of successful evolution, but a historically contingent outcome of a particular sequence of developmental steps.

If one wants to ensure that an artificial evolutionary process goes through precisely these steps, they must once again be specified as selection targets or intermediate targets. They would then derive from our retrospective knowledge of the biological evolution that actually occurred. If, by contrast, the process is genuinely left open, there is no reason why precisely these developmental steps should occur.

This yields a double aporia. If consciousness is fixed as the goal, the process is not open in precisely this respect. If the process remains open, consciousness is neither predictable nor guaranteed by evolutionary success.

Consciousness could arise in such an open process. Its occurrence would then, however, be a contingent outcome of that evolution rather than a technically planned production.

Rejection: The idea that one could simply allow an autocatalytic system to evolve for long enough until consciousness emerges fails for two reasons. Openness makes the concrete outcome unavailable, and evolutionary success has no necessary direction toward consciousness. If one attempts to impose this direction technically, one must specify precisely those developmental steps that are known only from already completed biological evolution. Evolution then becomes the post hoc optimization of a known result. The question thus shifts to the material conditions under which genuinely open evolutionary self-organization is possible at all.

Station ten: Open evolutionary self-sustenance requires a metastable capacity for transformation

The cascade now reaches the chemical level. The thesis is this: A durably evolving autocatalytic system requires a metastable capacity for transformation. Under the same conditions, its bonds must be sufficiently stable for persistence and sufficiently alterable for continued transformation. A system that is too stable does not renew itself; one that is too labile disintegrates. Self-sustenance and evolution therefore require a chemical regime in which stability and transformability coexist.

This property is substrate-dependent. Carbon satisfies it to an exceptional degree. Under moderate conditions, its bonds permit stable complex structures and, at the same time, controlled chemical transformations. Long chains can persist and yet be catalytically modified. Its oxidation product also keeps material turnover open.

Silicon does not satisfy these conditions in the same way. Under relevant environmental conditions, its chains do not possess the corresponding combination of stability and flexible reactivity. In oxygen-

rich environments, highly stable silicon oxides form and, as solids, drop out of a circulating metabolism. Under conditions that avoid this oxidation, other requirements for stable, complex, and finely regulable molecular structures are lacking. What carbon combines within a single chemical regime is distributed, in silicon, across opposing conditions.

Rejection: The exclusion here does not result from a prior commitment to a particular substance. What was required was a material property, namely a metastable capacity for transformation within one and the same physicochemical regime. Only then does it become apparent that this property depends on the substrate. Carbon therefore stands not at the beginning of the argument, but at its end.

The cascade is now complete.

Experience first requires an explanation of what distinguishes the "for the system" from mere functional availability. No ontological difference between function and experience is presupposed here. What is examined instead is the claim that they are identical.

Content alone does not explain this identity. Information alone does not explain it. Complex architecture and feedback are also insufficient. Only with a non-factorizable causal unity does there appear a system that is itself changed by its states. For this to become an intrinsic "for," however, it must be explained how states become relevant from the organization of the system itself.

This intrinsic relevance is not presupposed to be biological. A technical system could also satisfy this condition if its relevances genuinely arose from its own organization. The question then becomes how an organization can itself become the standard for its states.

The answer leads to self-maintenance and autocatalytic self-production. Not every individual experience must serve maintenance. What matters is that there exists a system-specific structure of relevance within which even fine and not immediately vital differences can make a difference to the system.

Even an autocatalytic technical system is therefore not yet excluded. The decisive next step concerns its development. Such a system would have to produce an organizational form that combines non-factorizable causal unity, intrinsic relevance, and the complexity required for consciousness.

It is precisely here that the evolutionary problem arises. Our biological evolution produced exactly the organizations that in fact arose. Only retrospectively can we determine which of their properties may be constitutive of consciousness. This knowledge is retrospective.

If we attempt to organize an artificial evolutionary process toward consciousness on the basis of these features, we use the known result as a selection criterion. We do not repeat evolution; we reconstruct post hoc properties of its outcome. The process becomes an optimization procedure directed toward an already known goal.

If, by contrast, we genuinely leave the process open, a second problem arises. Evolution has no necessary direction toward consciousness. Successful evolution means only successful continuation under the conditions that obtain at each stage. A nervous system, central integration, recursive organization, and ultimately consciousness are additional historical innovations whose occurrence is not guaranteed by evolutionary success.

It is therefore not enough simply to let an open evolutionary process run for long enough. Duration does not necessarily increase the probability of a particular developmental direction. An evolutionarily successful system can persist indefinitely in a form of organization far removed from any prerequisite of consciousness.

If consciousness is nevertheless to be ensured as the result, the developmental steps required for it must be deliberately favored. Nervous systems, integration, recursion, or other features known from our own evolutionary history then become selection criteria. By doing so, the open process is once again oriented toward a retrospectively known result.

The aporia therefore has two sides. A process can be organized with respect to consciousness or it can remain open. If it is organized, it uses criteria that we have obtained post hoc from consciousness that has already arisen. If it remains open, consciousness is neither its necessary goal nor a state guaranteed by evolutionary success.

This also avoids the claim that technology could never be involved in the emergence of a conscious system. Technology might possibly produce an autocatalytic system or the initial conditions for an open evolutionary process. What it cannot do, under the conditions of this thought experiment, is leave that process open while at the same time constructing consciousness as its outcome.

The conclusion is therefore this: Consciousness can neither be implemented as a technical architecture nor deliberately produced by an open evolutionary process. In the first case, the transition from the functional "for" to the intrinsic "for" remains unexplained. In the second, a double constraint arises: openness makes the concrete result unavailable, while successful evolution contains no necessary tendency toward consciousness.

The entire line of thought thus describes a single movement. It begins with the question of what consciousness means at all. From there it proceeds from the functional "for" to the intrinsic "for," from intrinsic relevance to self-maintenance, from self-maintenance to autocatalytic self-sustenance, and from there to the evolution of such an organization.

Only at this point does the historical dimension appear. We know which development led to consciousness because it has already occurred. This knowledge permits retrospective reconstruction, but not an independent prediction of the same path. An artificial evolution that uses our knowledge of this path as target criteria is already structured by the known result. An evolution that does without such criteria remains open and has no necessary direction toward that result.

The decisive contrast is therefore neither machine versus organism nor silicon versus carbon. It lies between a technically specified goal and an open genesis. A goal can be imposed by specifying developmental conditions. An open genesis can be initiated, but neither its concrete result nor a necessary development toward consciousness can be fixed in advance.

Consciousness would therefore not be a property that can be added to a sufficiently complex architecture or deliberately generated by a suitable optimization procedure. It would be the state of a non-factorizable, intrinsically relevant, self-sustaining system whose concrete organization has emerged from an open material and evolutionary process.

If consciousness were nevertheless to arise one day under technically created conditions, this would not constitute a counterexample. It would show that technology can create the conditions of an open process in which consciousness arises contingently. It would not show that consciousness can be constructed or deliberately produced by evolutionary means.