Physics of Life Reviews 59 (2026) 96–119

Contents lists available at ScienceDirect

Physics of Life Reviews

journal homepage: www.elsevier.com/locate/plrev

Review

# Algorithmic bottlenecks in evolution: Genetic code, symbolic language, and the Great Filter hypothesis

Mikhail Prokopenko [a,b,c,*], Nihat Ay [c,d], Angelica Breviario [e], Roland M. Crocker [f], Paul C. W. Davies [g], Pauline Davies [g], Darren Dougan [h], Roland Fletcher [i], Michael Harré [a,b], Marcus G. Heisler [a,j], Zdenka Kuncic [a,k,l], Geraint F. Lewis [m], Ori Livson [a,b], Vivienne Reiner [k], Jaime Ruiz Serra [a,b]

[a] The Centre for Complex Systems, University of Sydney, Sydney, NSW, 2006, Australia
[b] School of Computer Science, Faculty of Engineering, University of Sydney, Sydney, NSW, 2006, Australia
[c] Institute for Data Science Foundations, Hamburg University of Technology, Hamburg, 21073, Germany
[d] Santa Fe Institute, Santa Fe, NM, 87501, USA
[e] Department of Philosophy, Macquarie University, Sydney, NSW, 2113, Australia
[f] Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT, 2611, Australia
[g] The Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, 85287-0506, USA
[h] The Big Questions Institute, Surry Hills, NSW, 2010, Australia
[i] School of Humanities, Faculty of Arts and Social Sciences, University of Sydney, Sydney, NSW, 2006, Australia
[j] School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, NSW, 2006, Australia
[k] School of Physics, Faculty of Science, University of Sydney, Sydney, NSW, 2006, Australia
[l] The Charles Perkins Centre, University of Sydney, Sydney, NSW, 2006, Australia
[m] Sydney Institute for Astronomy, School of Physics, Faculty of Science, University of Sydney, Sydney, NSW, 2006, Australia

ARTICLE INFO

Communicated by J. Fontanari

*Keywords:*
Great Filter hypothesis
Algorithmic bottleneck
Coding threshold
Genetic code
Symbolic language
Major evolutionary transitions
Multichannel games.

ABSTRACT

The Great Filter hypothesis proposes that the emergence of technological societies capable of interstellar travel depends on a small number of exceptionally hard and highly improbable steps. Traditional versions of this hypothesis enumerate such "hard steps" along the trajectory from inanimate matter to complex technological societies, but diverge in their explanations for why these particular steps should be so improbable. The theory of Major Evolutionary Transitions also faces challenges in identifying which steps should be considered universally "hard" across different evolutionary pathways. In contrast, we argue that two deeply structural obstacles dominate the evolutionary landscape: the coding threshold associated with the origin of the genetic code, and the language threshold associated with the emergence of symbolic communication. We examine the developmental precursors of both transitions and analyze the underlying algorithmic bottlenecks: points at which evolving systems separate code from function, while entangling them within information hierarchies. Using a game-theoretic analysis of coupled signaling and coordination dynamics, we then argue that the corresponding multichannel games may exhibit saddle-type equilibria whose stable manifolds define narrow evolutionary paths, making the transitions intrinsically difficult to traverse. We conjecture that the so-called Great Filter is best understood not as a sequence of isolated improbable events, but as a nested structure of tangled information

* Corresponding author.
*E-mail address:* mikhail.prokopenko@sydney.edu.au (M. Prokopenko).

https://doi.org/10.1016/j.plrev.2026.09.001
Received 6 May 2026; Received in revised form 4 September 2026; Accepted 4 September 2026
Available online 6 September 2026
1571-0645/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

hierarchies. Under this conjecture, the rarity of advanced societies follows from the difficulty of crossing these coding thresholds in a competitive noisy environment. This hypothesis reframes the Great Filter as an algorithmic property of evolving systems, suggesting that only a small fraction of life may ever traverse the path toward technological societies capable of interstellar travel.

## 1. Introduction

The Great Filter hypothesis is a speculative framework proposed to explain the Fermi Paradox — the apparent contradiction between the assumed high probability of extraterrestrial life in the Universe and the lack of observable evidence for civilizations[1] and complex societies [28,129,137]. The hypothesis suggests that one or more stages in the evolution of life towards societies capable of interstellar travel and long-term large-scale technological activity are extremely unlikely or difficult to pass. This "filter" could lie in our past (e.g., the emergence of life, multicellularity) or in our future (e.g., surviving technological self-destruction). If the filter is behind us, humanity is rare and lucky. If it is ahead, our prospects may be grim.

Hanson [65] formalized this idea by identifying nine critical steps in the evolution from inanimate matter to lifeforms capable of interstellar travel and durable astroengineering. According to this view, each step is a potential Great Filter: a stage where progress may stall or fail. These nine steps follow "an evolutionary path to an explosion which leads to visible colonization of most of the visible universe" [65]:

1. The right star system (including planetary conditions): a stable star, habitable zone, and suitable planet.
2. The emergence of life (abiogenesis): the transition from non-living chemistry to self-replicating life.
3. Simple (prokaryotic) single-cell life: life that can sustain itself and replicate.
4. Complex (eukaryotic) single-cell life: cells with internal structures like nuclei and organelles.
5. Sexual reproduction: the transition which allows for greater genetic diversity and evolutionary adaptability.
6. Multicellular life: emergence of organisms composed of many specialized cells.
7. Tool-using animals with intelligence: the transition to organisms which are capable of manipulating the environment and abstract reasoning[2].
8. Technological society: development of advanced tools, industry, and communication.
9. Colonization beyond the home planet: expansion into space, detectable across interstellar distances.

This line of reasoning can be traced to the seminal analysis by Carter [15], who argued that the evolutionary emergence of humans depended on successfully traversing a small number of necessary hard ("highly improbable") steps, including "the original establishment of the genetic code, and the final breakthrough in cerebral development". The core argument of Carter [15] is that the existence of human observers sets a limit on how improbable the steps leading to their existence could be, given the available time in the universe and the total lifespan of the sun (our star), which implies that intelligent life might be rare in the universe. In other words, biological evolution is constrained by astrophysical, not just biological, factors.

There are several other proposals describing various "hard steps" which were critically examined by a recent study [107]. This critical survey distilled five most popular candidate "hard steps", coinciding with the exact sequence proposed by Lingam and Loeb [97,98]:

(i) The origin of life ("abiogenesis").
(ii) Oxygenic photosynthesis.
(iii) Eukaryotic cells ("eukaryogenesis").
(iv) Animal multicellularity.
(v) *Homo sapiens*.

The reassessment of Mills et al. [107] led to an alternative model grounded in historical geobiology, in which no hard steps are implied. Instead, *evolutionary singularities* [30] (see Appendix B) were attributed to mechanisms other than intrinsic improbability, with the timing of human origins governed by the successive opening of global environmental windows of habitability throughout Earth's history [107].

An independent body of work focuses on Major Evolutionary Transitions (METs) — such as the origin of life, formation of eukaryotic cells, emergence of multicellularity — which are characterized by fundamental shifts in how organisms exchange information, cooperate, coordinate, and ultimately survive and replicate [143,156]. As functional specialization intensifies, these collectives may become tightly integrated, enhancing cooperative efficiency while simultaneously inducing mutual dependence [156]. Such integration gives rise to higher-order units that increasingly act as the primary targets of selection. By coupling the replication of individual genes to the survival of higher-order structures, selection increasingly favors genes that promote collective viability, thereby enabling greater division of labor and specialization.

[1] A more precise term replacing "civilization" is "*a technosignature-bearing society*" (see Appendix B for this and other italicized terms). An alternative we use henceforth is "technological society".

[2] Abstract reasoning is understood as the capacity to construct and manipulate internal models detached from immediate sensory input or context, and reason about hypothetical, non-present, or counterfactual states.

Examples of METs include early molecular replicators forming compartmentalized cellular collectives; autonomous replicators, likely RNA-based, forming chromosomes to reduce replication-induced information loss; emergence of the genetic code and translation machinery, distinguishing DNA as genetic material and proteins as enzymes; and the emergence of eukaryotic cells with membrane-bound nuclei storing genetic information. This progression continues through multicellularity and eusociality, ultimately extending to language and sociocultural evolution [83,94,143,156]. Importantly, these transitions are typically accompanied by the emergence of novel inheritance systems that reshape how information is stored, transmitted, and processed — sometimes referred to as new "codes of life" [7,94]. A paradigmatic example is emergence of the genetic code, which introduced a novel inheritance system by mapping nucleotide sequences to amino acids, thereby decoupling information storage from biochemical function and enabling open-ended molecular innovation.

Barbieri [7,8] has listed several "codes of life", such as signal transduction codes, splicing codes, tubulin code, nuclear signaling code, molecular codes, histone code, transcriptional codes, apoptosis code, bioelectric code, neural codes, etc. [7,8]. However, as pointed by Kun [94], the "codes of life" and the major evolutionary transitions represent two distinct components of macroevolution and do not coincide. Two notable exceptions which appear in both lists are *the genetic code* and *human language*. Kun [94] argued that for the genetic code and human language the evolution of a "code of life" is a MET.

It has also been argued that major evolutionary transitions occur when systems cross specific "coding thresholds", with the saltations opening new evolutionary phase spaces defined by novel information-processing regimes [94,164]. In his analysis of the emergence of DNA, Woese [164] argued that

> "Somewhere along the line there had to have occurred a saltation that we could call the "coding threshold," where the capacity to represent nucleic acid sequence symbolically in terms of a (colinear) amino acid sequence developed, a development that would generate a truly enormous new, totally unique evolutionary phase space."

Similarly, a "language threshold" may have been crossed when early human groups developed language with symbolic reference and complex syntax[3], which enabled a novel mode of knowledge inheritance and substantially accelerated adaptation [7,94,142,164]. As pointed out by Szathmáry [142], while animal communication systems mostly include *self-regarding signals* about the immediate and present situation, human language is very different: "There is a lot of displacement (referring to items that are not present now or are purely imaginary), and it is full of symbolic (arbitrarily conventional rather than indexical or iconic) reference, aided by complex syntax" (see Appendix B for *indexical* and *iconic references)*. As pointed out by Hauser et al. [71], natural languages extend beyond purely local structure by allowing "recursive embedding of phrases within phrases", thereby generating statistical dependencies that can span arbitrarily long distances. In general, it has been argued that this "symbolic explosion" [74] coincided with the emergence of syntactic language which evolved within a highly specialized system of social learning [125] and enabled truly "semantically unbounded discourse" [124].

We propose that the Great Filter hypothesis can be interpreted via algorithmic bottlenecks (see next section), offering a perspective to understand constraints on the emergence of intelligent life capable of interstellar travel and astroengineering. Our aim is to explore universal principles that govern specific evolutionary transitions which involve "coding thresholds". We hypothesize that these transitions must inevitably pass through algorithmic bottlenecks which may represent saddle-type equilibria in evolutionary dynamics: states that attract trajectories along some directions but amplify perturbations along others, due to environmental, biological, or strategic constraints. We argue that at the bottleneck, the stable manifold of the saddle-type equilibrium occupies only a narrow part of the accessible state space, and trajectories that deviate from it are expelled along unstable directions. As a result, the bottleneck is both difficult to reach and difficult to traverse successfully. In contrast, trajectories that fail to enter these bottlenecks may settle into comparatively stable but stagnant evolutionary regimes, unable to support the emergence of higher-order informational architectures. In other words, if this conjecture is correct, the hardest Great Filter steps may have an algorithmic nature.

## 2. Tangled information hierarchies and algorithmic bottlenecks

How do meaningless molecules create semantically meaningful code? Informally, chemistry in living systems can be identified with hardware, while genetic and epigenetic information can be interpreted as software [27,155]. From this perspective — *the algorithmic origins of life* — the emergence of genetic code and accurate translation machinery is crucial, with life having two origins: the hardware and the software. While the spontaneous emergence of chemical complexity proceeds by fairly well understood physical principles (e.g., laws of thermodynamics, energy dissipation, kinetics and mass-action laws, self-assembly due to intermolecular forces), the leap from hardware to software, that is, from local forces and processes to the encoded functional information transmission and global information management, is a much harder step.

Tangled hierarchies are systems in which levels of organization recursively interact, resisting a straightforward hierarchical ordering: "an interaction between levels in which the top level reaches back down towards the bottom level and influences it, while at the same time being itself determined by the bottom level" [80]. While a tangled hierarchy describes a structural property of a system, a "strange loop" is a dynamic and recursive phenomenon within a tangled hierarchy: "A strange loop is self-referential, meaning that it loops back on itself in some way. Despite the appearance of movement or progression, a strange loop ultimately returns to its starting point, creating a sense of paradox or contradiction." [80], see Fig. 1.

[3] Here, language denotes spoken symbolic language, while writing, notation and other inscription systems are considered secondary externalizations that amplify and stabilize linguistic control, but do not constitute the language threshold itself. See also subsection 3.2.2.

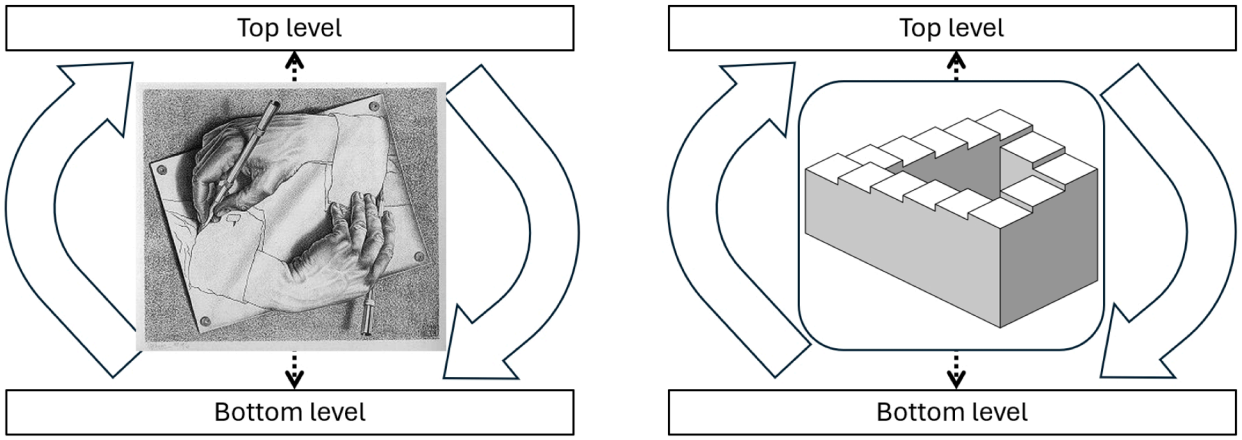



**Fig. 1.** Two well-known strange loops. Left: "Drawing Hands" (Escher, 1948) [158], a visual paradox with two intertwined hands perpetually drawing each other within a recursive loop of drawing an object which draws *itself*. Right: "Impossible staircase" (1958) [115,159], a spatial pattern connecting the stairs in an impossible way, where a continuous staircase loops back on *itself* within an infinite recursion of ascent and descent.

The genetic code encodes the very molecular machinery (tRNAs, aminoacyl-tRNA synthetases, ribosomes) that interprets the code. Given this self-reference, the phenotype-genotype mapping has been interpreted as a "tangled hierarchy" [80], in which phenotypes produced by the bottom-up gene expression influence the genotypes via the top-down selective pressure and inheritance, i.e., via "downward causation" [41]. Thus, the evolutionary dynamic constitutes a self-referential loop in which phenotypes determine the evolutionary fate of the genotypes that produced them, see Fig. 2.

We define an algorithmic bottleneck as a transitional stage in the evolution of an information-managing system where a viable code must be co-established with its interpreter and control machinery, such that (i) the mapping rules, (ii) the interpreting/implementing mechanisms, and (iii) the feedback needed to stabilize their joint operation must all be simultaneously aligned. Because these elements are mutually dependent (the code presupposes an interpreter; the interpreter is specified by the code; stability requires closed-loop control), the system confronts a self-referential and coordination-heavy problem that is hard to traverse and easy to destabilize. Success requires bootstrapping a stable code–interpreter pair under constraints of noise, cost, and ambiguity. In this framework, a transition qualifies as an algorithmic bottleneck only when three conditions are jointly satisfied:

1. Symbolic encoding: the emergence of an explicit symbolic mapping between representation and function, so that information is partially decoupled from immediate physical function.
2. Interpretive machinery: the emergence of an interpreter capable of executing this mapping, converting encoded information into functional outcomes.
3. Recursive closure: a self-referential dependency whereby the encoded information contributes to specifying, maintaining, or regulating the interpreter itself.

Transitions that lack any one of these elements may still be evolutionarily difficult, but they do not constitute algorithmic bottlenecks in the sense intended here. What distinguishes an algorithmic bottleneck is the emergence of a self-referential information-processing architecture that stabilizes an executable mapping between symbolic descriptions and functional outcomes. The core challenge is to reach a coupled fixed point of a joint code–interpreter system $(C, I)$. In this system, encoded descriptions $C$ specify an interpreter $I = f(C)$ capable of executing those descriptions by mapping symbolic encodings to organismal dynamics. The interpreter contributes to the construction of an embodied phenotype within which descriptions $C$ are maintained and reproduced. We denote this effective organism-level reproductive dynamics by $g$, yielding $C = g(I)$ and leaving dependence on the phenotype implicit. Here $g$ should not be understood as a direct transformation from interpreter to description. Rather, $g$ is a coarse-grained map representing the organism-level processes, through which descriptions are preserved and reproduced.

Crossing the bottleneck requires the emergence of a recursively closed information architecture, in which $(C, I) = h(C, I)$ for $h(C, I) = (g(I), f(C))$. The bottleneck is crossed when a stable pair $(C^*, I^*)$ emerges such that $(C^*, I^*) = h(C^*, I^*)$. For a coupled fixed-point $(C^*, I^*)$, both code and interpreter are also composite fixed-points $C^* = g(f(C^*))$ and $I^* = f(g(I^*))$ of compositions $g \circ f$ and $f \circ g$, respectively.

Viewed as a search for a coupled fixed point of a self-referential system, traversing an algorithmic bottleneck marks an expansion of the evolutionary search space itself, by introducing a new level of description and interpretation. Given the strong connections between self-reference, fixed-point constructions, and undecidability in formal systems [75,122,133,165], the emergence of recursively closed code–interpreter systems cannot in general be reduced to a simple local search over predefined functional states [121].

The phenotype-genotype mapping provides a canonical example of such a bottleneck. The genetic code establishes a symbolic mapping from codons to amino acids, while the translation apparatus implements that mapping. Yet, the mapping and its interpreter co-specify one another: an encoding cannot function without an interpreter, while the interpreter is itself encoded within the genome (see Appendix A for generic code and interpreter roles). Crossing this threshold requires jointly bootstrapping a stable code–interpreter

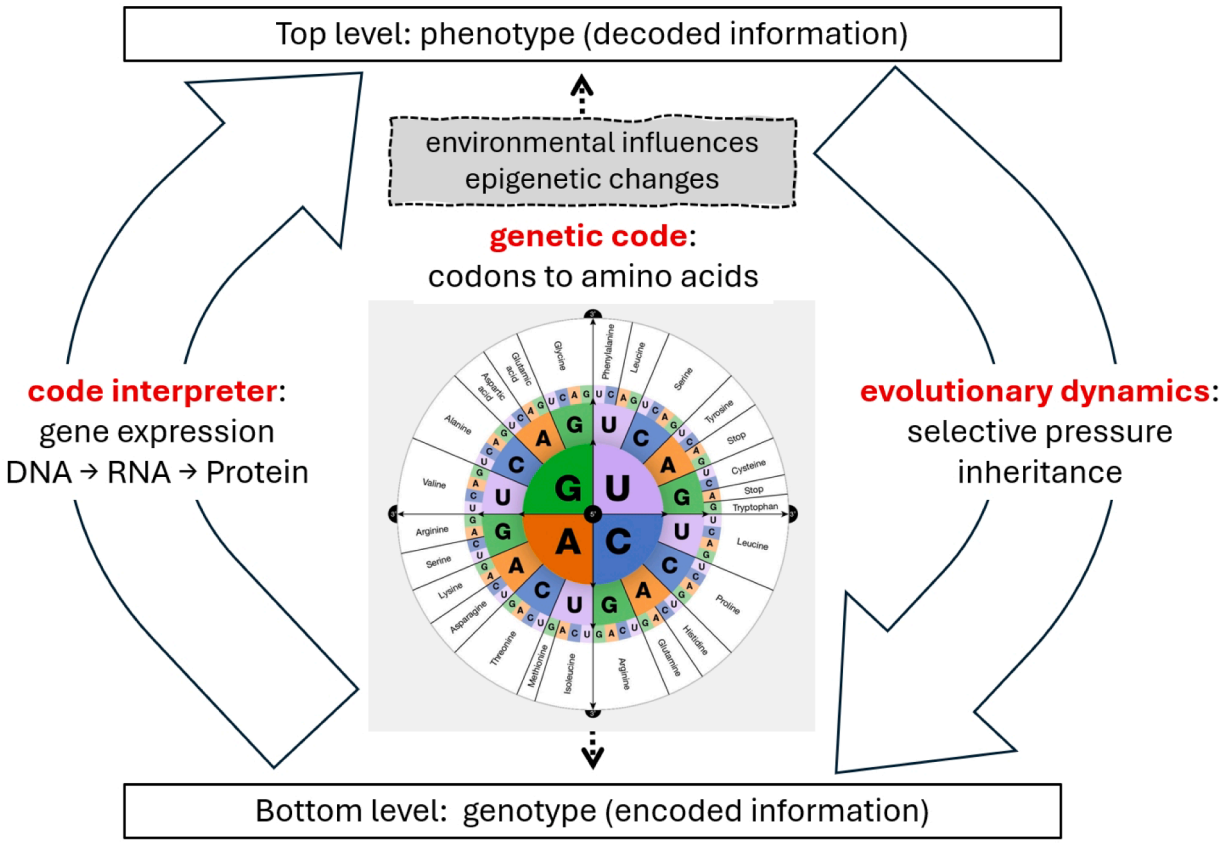



**Fig. 2.** Genotype-phenotype relationship as a tangled hierarchy. Bottom-up construction interprets genetic code via gene expression: transcription and translation of encoded genetic information leads to the decoding of an organism with observable phenotypic traits. Top-down causation: given the phenotype, selective pressure, genetic variation and inheritance influence genotype over generations. Strange loop updates the genetic information which encodes the phenotype *itself*. Genotype-phenotype relationship is realized in context of environmental influences, epigenetic changes and other factors. Genetic code image (public domain): courtesy of the National Human Genome Research Institute (NHGRI) [110].

pair and closing the feedback loops that couple informational rules to biological function. Until the mapping and its implementation become mutually aligned and self-sustaining, the system remains vulnerable to instability and failure. In general, such self-referential, self-modeling dynamics can improve the efficiency of information-processing within a tangled information hierarchy, but they incur the costs of encoding, maintaining and decoding a compressed self-representation [121].

The self-referential structure of genotype-phenotype organization is closely related to Rosen's concept of closure to efficient causation in relational biology [128]. Rosen argued that living systems are characterized by networks in which the components responsible for functional processes are themselves produced within the same system. This perspective aligns with our characterization of bottlenecks as transitions requiring the co-emergence of a code, its interpreter, and stabilizing feedback loops. Rosen further argued that such organizational closure creates fundamental challenges for purely decompositional representations, suggesting that self-referential biological systems occupy a distinctive position in the space of computable models. Rosen's framework provides a valuable mathematical antecedent for our discussion of self-modeling dynamics and tangled hierarchies. Viewed in this light, the emergence of the genetic code can be understood as the establishment of a causally closed informational architecture in which encoded descriptions and the machinery that interprets them become mutually dependent.

Interestingly, unlike Carter's original conjecture [15], neither Hanson's nine-step list [65] nor the five hard steps proposed by Lingam and Loeb [97,98] and later evaluated by Mills et al. [107] explicitly account for the two unique METs involving "coding thresholds". These are the emergence of the genetic code and the emergence of symbolic language.

In discussing transitions in cognitive evolution, Barron et al. [10] proposed five computational cognitive architectures and related the final transition (a computational architecture capable of "reflection") to the emergence of language:

> "Reflection also opens the door to the efficient use of symbolic languages. …it is clear that once symbolic languages are in play new forms of reflection then become possible, offering further efficiencies in task acquisition and execution. Language can be seen as a way to compress instructions into simpler forms [89] and enable new types of mental programmes [9]. By a 'programme', we mean any way in which control flow can be effectively represented and manipulated, thereby shifting the burden of developing new control flow from the biological to cognitive."

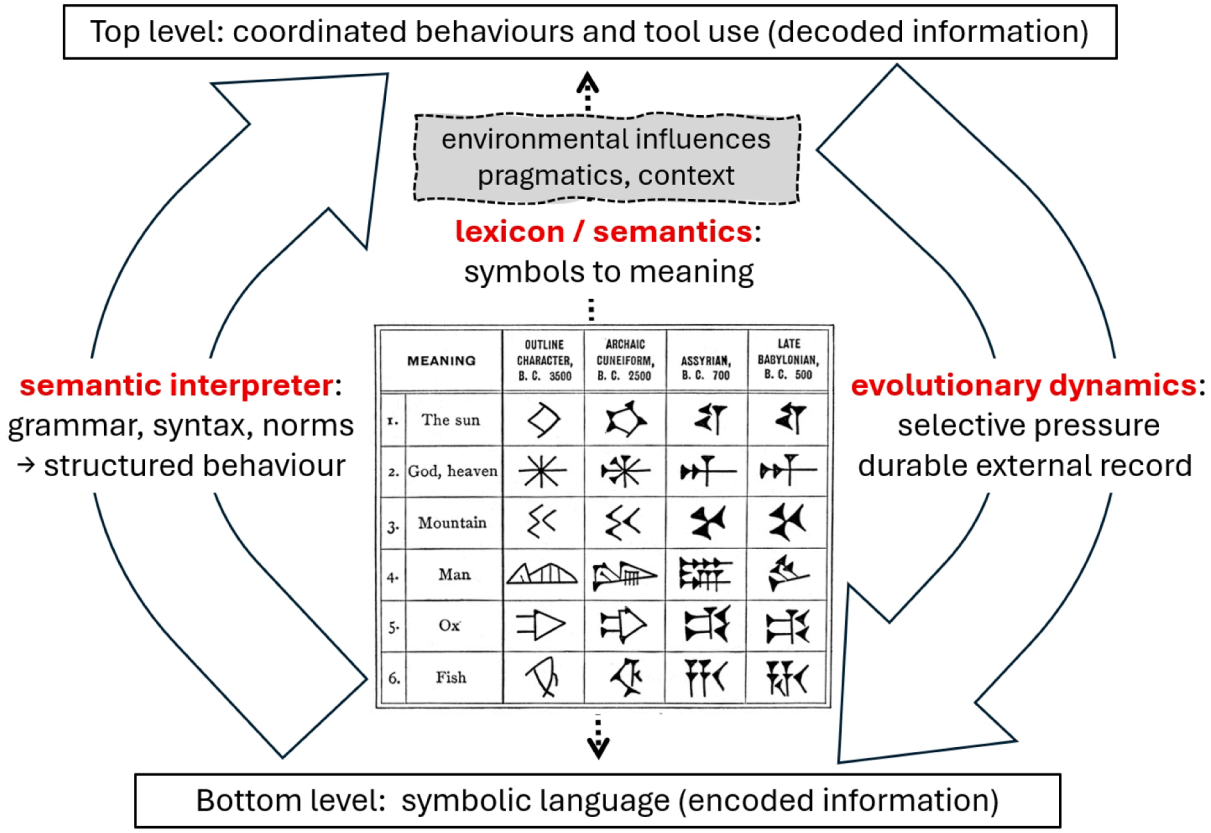



**Fig. 3.** Language-behavior relationship as a tangled hierarchy. Bottom-up construction interprets symbolic language via grammatical and semantic machinery: grammar, recursive syntax, conversational norms and conceptual schemes shape the decoding of communicative acts, coordinated behaviors and tool use. Top-down causation: given the social behavior, selective pressure, linguistic innovation and external storage influence language over generations. Strange loop updates the symbolic information which encodes the social behavior *itself*. Language-behavior relationship is realized in context of environmental influences, pragmatics and contextual interpretation, among other factors. Cuneiform image (copyright-free): a table illustrating evolution of cuneiform [109, p. 55].

This characterization suggests why the language threshold may constitute an algorithmic bottleneck: symbolic language enables compressed functional information to serve as executable control flow, creating a self-modeling loop that entangles cognitive "software" with its neural "hardware".

Passing the language threshold marks the stage at which symbolic control flow ceases to be confined to individual cognition and becomes distributed across external artifacts and social practices[4]. This shift allows recursive control structures to stabilize and compound across time, enabling cumulative, trans-generational computation (see Fig. 3). While other species, including bees or dolphins, may exhibit sophisticated, context-dependent communication, human language uniquely supports the external storage and recursive manipulation of symbolic structures outside of individual brains. The relevant distinction is not a conjunction of species-specific adaptations, but the coupling of two abstract informational functions: durable external symbolic memory and recursive symbol manipulation. These functions are substrate-independent and, in principle, would be equally significant whether realized by humans, dolphins, or extraterrestrial intelligences.

Importantly, both transitions — the emergence of the genetic code and the emergence of symbolic language — go beyond emergent cooperation, division of labor and collective coordination: they require the establishment of an explicit mapping, or "code". This creates an algorithmic self-modeling problem shared by both transitions:

I. The emergence of genetic code and translation machinery: progression from an early (e.g., RNA or protein/peptide) world to a DNA/protein world, in which heritable information specifying phenotypic function is stably encoded in DNA and interpreted (decoded) via a dedicated translation apparatus.

[4] This characterization does not assume the extended mind thesis [17], remaining agnostic about whether such external structures constitute cognition itself.

**Table 1**
Payoff matrix for two players (Row and Column), using four payoff parameters.

| Row \ Column | Cooperate | Defect |
|---|---|---|
| Cooperate | $R$, $R$ | $S$, $T$ |
| Defect | $T$, $S$ | $P$, $P$ |

II. The emergence of symbolic language and linguistic interpretive machinery: progression from pre-linguistic communication and behavioral cues to a symbol-based representational system, in which abstract meanings are stably encoded in symbols and interpreted (decoded) via shared semantic and syntactic rules.

Both these transitions mark a progression from locally constrained, context-bound signaling to increasingly explicit symbolic codes supported by collective mechanisms for parsing, interpretation, and meaning reconstruction. Just as the genetic code required co-evolution of the translation machinery that maps nucleotide sequences into functional proteins, symbolic language needed the joint stabilization of symbolic reference and syntax. Importantly, the relative purity of the genetic code, its almost context-independent mapping between codons and amino acids, contrasts with the inherently context-sensitive nature of human language. Yet what is shared across both transitions is not full context-independence, but the emergence of conventionalized symbol systems whose interpretation is stabilized by communal processes, allowing meaning to persist across contexts even when the mechanisms differ. Appendix A elaborates on the analogy between genetic and linguistic coding systems in more detail.

Our use of tangled hierarchies is related to formal approaches such as Causal Emergence Theory (CET), which quantifies the causal efficacy of macroscopic descriptions using information-theoretic measures tracking interventions and predictability [77,78]. Similarly, studies of collective computation and coarse-graining in adaptive social systems have shown how emergent macro-level variables can both summarize and regulate lower-level interactions through feedback loops [54]. Other studies quantified a form of circular causality where neural activity and extracellular electric fields continuously shape each other [118]. These approaches provide a rigorous account of how higher-level organizational states may exert top-down constraints on lower-level dynamics. Importantly, we do not claim that these frameworks directly explain the emergence of the genetic code or symbolic language. Rather, they provide formal and computational precedents for our broader claim that higher-order informational organizations can acquire causal efficacy and participate in stabilizing evolutionary dynamics.

## 3. Communication and cooperation

A central premise of the "algorithmic origins of life" framework for explaining the emergence of the genetic code [155] is the onset of global information management, whereby encoded functional information — "software" — acquires causal efficacy over the underlying physical substrate, or "hardware". In this section, we examine key aspects of this self-referential dynamic in the context of (a) the emergence of genetic code, "the lingua franca of genetic commerce" [163,164], and (b) the emergence of symbolic language with its "unlimited hereditary potential" [142].

Because both genetic coding and symbolic language formation require agents (molecules, cells, or individuals) to coordinate using shared conventions, we will adopt *evolutionary game theory* to analyze relevant signaling games (sender–receiver alignment) and coordination games (alignment of resource use). This approach will clarify why misalignment, ambiguity, and free-riding (defection) can be evolutionarily attractive locally yet destructive globally. In turn this will expose the fundamental tension that makes an algorithmic bottleneck particularly difficult to traverse. We will argue that this is not a problem of isolated optimization, but rather a strategic coordination problem involving multiple agents dynamically engaged in both competitive and cooperative interactions across multiple channels.

A canonical two-player game with two strategies, Cooperate (C) and Defect (D), can be modeled with a payoff matrix for two players (Row and Column), using four payoff parameters (see Table 1 and Box 1):

- $R$ (Reward): payoff to each if both cooperate
- $T$ (Temptation): payoff to a defector against a cooperator
- $S$ (Sucker's payoff): payoff to a cooperator against a defector
- $P$ (Punishment): payoff to each if both defect

**Box 1: Canonical game-theoretical models.**

Many standard games are obtained by imposing different orderings on $T$, $R$, $P$, $S$, e.g., Prisoner's Dilemma or Stag Hunt game (see below). Nash equilibrium is a key concept where each player chooses a strategy that is their best response to the strategies chosen by other players, resulting in a game-theoretically stable outcome where no single player can improve their payoff by changing their strategy alone, assuming others keep theirs constant. In Nash equilibrium no player could gain more by changing their decision given what others did, although it does not always lead to the best overall result for the group.

In Prisoner's dilemma, each player simultaneously selects an action without communication or binding commitments. Using the canonical payoff parameters, the model assumes that $T > R > P > S$. Mutual cooperation yields a higher payoff for both than mutual defection, yet unilateral defection strictly dominates cooperation for each player. As a result, the Prisoner's Dilemma has a single pure-strategy Nash equilibrium where both players defect, despite mutual cooperation being *Pareto optimal*. This tension between individual rationality and collective optimality makes the Prisoner's dilemma a foundational example in the analysis of strategic interaction, equilibrium selection, and social dilemmas [4].

Stag Hunt game is a canonical game-theoretic model with two or more agents which can choose between strategies of hunting a Hare (i.e., acting alone which is safe and has minimal risk, but offers a low payoff), or hunting a Stag (i.e., requiring cooperation with others which is risky because the hunt fails if others defect, but offers high payoff). Using the canonical payoff parameters, the model assumes that $R > T \geq P > S$. The Stag Hunt game differs from the Prisoner's Dilemma in its strategic structure; it admits two pure-strategy Nash equilibria: mutual cooperation and mutual defection. This distinction highlights the coordination challenge in the Stag Hunt, where trust and assurance are required to reach the cooperative equilibrium, whereas the Prisoner's Dilemma is dominated by incentives to defect regardless of the other player's choice. Importantly, the mutual defection in the Stag Hunt game is *risk-dominant* (being less risky), while mutual cooperation is suffering from uncertainty due to a lack of trust and the fear of the worst-case scenario.

Evolutionary game theory applies game theory to evolving populations, modeling how strategies (behaviors, traits) spread or die out based on their fitness payoffs, which depend on the strategies of others in the environment (the population itself). Unlike classical game theory, evolutionary game theory focuses on frequency-dependent selection, explaining adaptation and the evolution of cooperation, altruism, and complex behaviors in biology, economics, and sociology. A refinement of the Nash equilibrium that remains stable under evolutionary dynamics is evolutionarily stable strategy (or set of strategies): a behavioral strategy that, once adopted by a population, cannot be successfully invaded or displaced by any rare, alternative "mutant" strategy (or set of strategies), ensuring its stability over time.

### *3.1. Emergence of genetic code*

The evolutionary problem underlying the emergence of genetic code has been succinctly formulated by Woese [164]: "how the genotype-phenotype relationship had come to be". This question leads us to consider the precursors of the genetic coding threshold, i.e., the transition to symbolic nucleic-acid–protein coding, focusing on how early molecular interactions and collective dynamics could overcome the algorithmic bottleneck inherent in establishing a self-consistent genotype–phenotype mapping.

#### *3.1.1. Precursors of the genetic "coding threshold"*

Before a stable genotype-phenotype map could crystallize, several enabling conditions had to co-emerge. On the one hand, informational polymers or other heritable chemical regularities needed sufficient templating fidelity and length. This supported heritable variation and sustained polymerization and compartmentalization, helping to concentrate reactants and protect nascent networks [39,145]. On the other hand, catalytic specificity had to rise above background enzyme promiscuity, enabling mechanisms that could associate informational patterns with functional outcomes [92,127], and thus, introduce the rudiments of a code–interpreter pair. Additionally, control and feedback processes such as proofreading, repair, and kinetic discrimination, were needed to suppress error cascades [40,81]. Finally, population-level processes, such as horizontal exchange, communal selection, and convention formation, were required to align partially compatible mappings and drive convergence toward shared assignments under pressures for error minimization, redundancy, and compositionality [152,161]. Together, these elements established the pre-coding scaffold: a milieu in which an initially rudimentary and easily lost mapping between sequence and function could bootstrap into a stable, scalable code–interpreter system. These shared functional prerequisites can be approached from different explanatory perspectives (see Box 2).

A principal disagreement between these perspectives is that coding in the RNA-World is an intrinsic requirement of an RNA-centric replicative system, while in the metabolism-first scenario, coding is an extrinsic solution, emerged to stabilize and propagate something more fundamental: metabolism. In comparing these two approaches, Walker and Davies [155] argued that neither the genetics-first ("digital") nor metabolism-first ("analogue") trajectories provide a satisfactory explanation, highlighting that while information may be expressed in specific chemical structure (digital or analogue), it lacks autonomy [155]. Instead, their study shifted attention toward the underlying logical architecture of information processing and control, as well as the pathways of information flow in living systems — pathways that are themselves influenced by information encoded within the system. Under the algorithmic interpretation of Walker and Davies [155], the threshold is crossed when "an abstract and non-physical systemic entity (algorithmic

**Box 2: Information and metabolism in early life.**

One resolution to the genotype-phenotype duality is that the modern DNA–protein system evolved from a simpler precursor in which a single molecular species performed both informational and catalytic roles. In contemporary biology, RNA mediates between DNA and proteins, and is unique in its ability to function as both a genetic polymer and an enzyme. These properties underpin the influential RNA-world hypothesis, which proposes that all known life descended from an ancestral population that relied on RNA as its sole major biopolymer before the emergence of DNA and proteins [59,61,87].

The RNA-world hypothesis holds that the genetic code emerged within an RNA-based system in which RNA molecules acted as both informational templates and catalytic ribozymes capable of discriminating between amino acids. RNA replication provided the substrate for encoding rules, while proto-tRNAs and proto-ribozymes mediated primitive codon-amino acid associations. Coding then evolved through refinement of translation, driven by selection for greater accuracy, enzymatic diversity, and replicative fitness.

Thus, it is assumed that during the early era of nucleic-acid–based life, primitive cell-like entities may have existed as loosely connected conglomerates open to extensive horizontal gene transfer, with evolutionary dynamics characterized by "communal descent" [152,162]. Crucially, at the coding threshold the genetic code functioned "not only as a protocol for encoding amino acid sequences in the genome but also as an innovation-sharing protocol" [152]: it served both as a component of the replication machinery and as a medium for encoding and disseminating information about the environment. Different proto-cells may have discovered distinct adaptive innovations, and the horizontal exchange of such innovations was facilitated by a shared, converging proto-code.

In contrast, metabolism-first models place chemical reaction networks, rather than genetic molecules, at the origin of biochemical organization. The genetic code appears only after metabolic cycles become stable and catalysts diversify, with the selection favoring more reliable inheritance mechanisms. According to this view, autocatalytic networks function without symbolic representation, and simple peptides, minerals, or small molecules drive early metabolism which precedes informational templates [38,90,153]. The genetic code then arises to orchestrate and stabilize pre-existing metabolic chemistry, rather than to enable early replication. Thus, the mapping from codons to amino acids is shaped not by ribozyme discrimination but instead, by biosynthetic pathways, chemical affinities and metabolic availability of amino acids.

information) effectively becomes a causal agent capable of manipulating its material substrate", giving rise to a form of distributed information control [155].

#### *3.1.2. The codon-signaling game*

It is instructive to consider plausible scenarios for the emergence of genetic code in a game-theoretic context (see Appendix B for *canonical* and *evolutionary games*). In general, signaling game theory examines strategic interactions where a sender communicates information to a receiver through signals that may be truthful, partial, or misleading [167]. The existence of information asymmetry among the macromolecules, including molecular deception, was studied using sender–receiver signaling games, providing evidence that selfish behavior differentiates biochemistry from abiotic chemistry [106]. Massey and Mishra [106] identified the genetic code as a signaling convention "between genes (senders), their mRNAs (the signal) and the set of tRNAs in the cell (receiver)", while pathogens and selfish elements were argued to subvert this convention [106].

In a pre-DNA/protein environment, early molecular systems may be interpreted as players competing for replication and catalytic efficiency, with various co-evolving coding schemes and molecular interactions. In the evolutionary codon-signaling game, players are "senders" (encoding information, e.g., mapping RNA codons to amino acids) and "receivers" (decoding information, interpreting codons and acting: tRNA or ribosome assigning amino acid) [111,123].

Agents (e.g., molecular replicators such as proto-ribozymes, RNA strands, peptides) interact under payoff structures which favor reliable information transfer and functional catalysis. For example, payoff can be formed as the effective protein yield, adjusted for penalties due to amino-acid *misincorporation* (i.e., translational decoding errors) and the translation time costs. Then, high payoff may be achieved when sender and receiver coordinate on a stable, functional mapping, while low or negative payoff is attained under mismatches or noisy, slow interpretations. Game strategies may involve random codon assignments, autocatalytic loops, template-based replication, and so on.

The number of possible codon-to-amino acid mappings is, however, combinatorially vast, and the agents may fail to align on a shared coding scheme, leading to fragmented or incompatible coding systems. High mutation rates and translational errors further disrupt reliable genotype-phenotype mapping: mutations perturb the sender's message (codon sequence), while mistranslations undermine the receiver's interpretation (decoding). Furthermore, limited availability of nucleotides, amino acids, or catalytic surfaces may favor short-term survival strategies over long-term code optimization.

The selective pressure for refining the code is further reduced by weak or delayed feedback between coding accuracy and functional performance, allowing suboptimal mappings to persist. Since the mapping from code configurations to fitness is highly non-linear and discontinuous (i.e., the game has a rugged fitness landscape), it is difficult for evolutionary dynamics to traverse the landscape valleys toward globally optimal codes. In other words, early arbitrary assignments (i.e., local optima) may be entrenched, preventing exploration of potentially superior coding schemes.

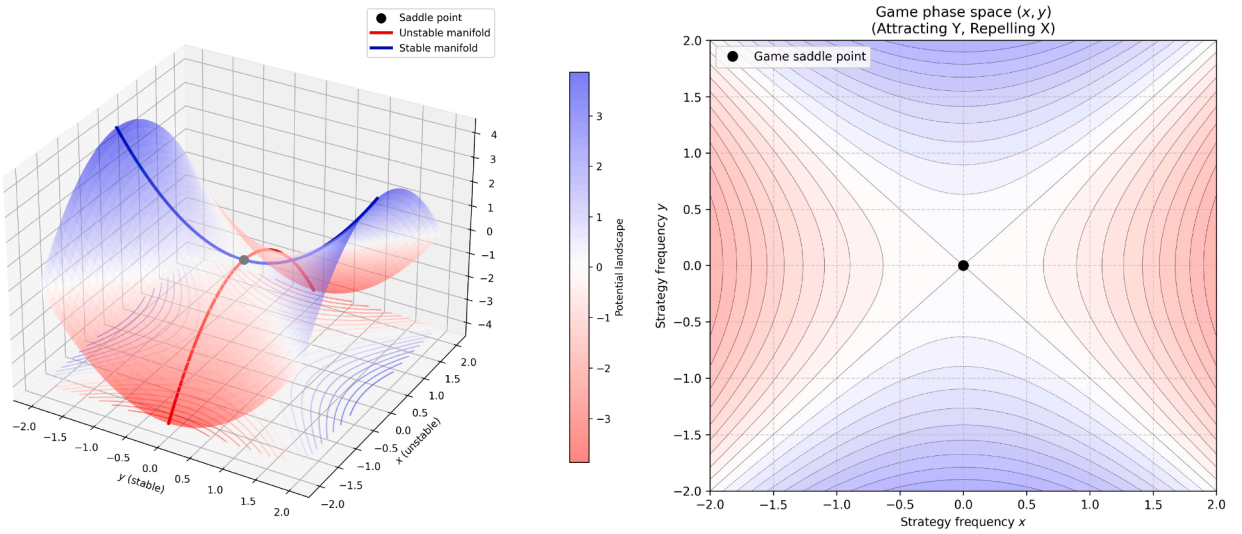



**Fig. 4.** Left: A generic two dimensional saddle-point manifold at the intersection of the red and blue curves. Along the blue curve the saddle point is a local minimum, and along the red curve it is a local maximum. Right: In game dynamics, trajectories converge along stable manifolds (local attraction) but face rapid repulsion along unstable manifolds (repulsion paths), meaning that small strategy deviations take the system away from the equilibrium.

In this scenario, the dynamics may be organized around a saddle-point manifold rather than an attractor. Evolutionary pressures can guide populations along its stable manifold toward a narrow bottleneck of coding coherence, but the corresponding unstable manifold acts as an exit channel through which translation errors, mutational drift, or non-cooperative strategies can rapidly amplify perturbations and drive the system away from viability. Fig. 4 illustrates a saddle-point manifold as a game-theoretic equilibrium. Such dynamics make the apparent convergence deceptive: the same evolutionary forces that funnel populations toward coherence can leave them poised near an instability from which error catastrophes, translational noise, or parasitic strategies readily trigger collapse.

### *3.1.3. The metabolic cooperation game*

The evolutionary dynamics of the codon-signaling game considered in the previous subsection makes convergence toward a widely shared code unlikely without strong constraints or coordination mechanisms, for example, provided by metabolism. In early Earth environments, molecular replicators needed to develop metabolic pathways to harness energy and replicate. These pathways often involved shared enzymes or metabolites that diffused into the environment, benefiting nearby cells [123].

Game-theoretically, metabolic cooperation can be interpreted as a coordination game: a simultaneous game in which players (e.g., proto-cells or cells) benefit from selecting the same strategy, as mutual alignment yields higher payoffs than mismatched choices. Specifically, one may consider a spatial public goods game with diffusion constraints (see Box 3).

**Box 3: Public goods game.**

Public goods game models social dilemmas where individuals choose between contributing to a shared resource or free-riding for personal gain. In the basic game, participants simultaneously and secretly decide how many of their endowed tokens to contribute to a public pot. Each player's payoff consists of private consumption — defined as their remaining endowment after contribution — plus returns from the public good, equal to the total contributions scaled by a fixed multiplier. The game demonstrates the conflict between self-interest and collective benefit, often resulting in under-provision of the good, unless mechanisms like punishment or rewards are introduced.

In a spatial public goods game, agents are arranged on a spatial network (e.g., a lattice) rather than form a fully mixed group, and their ability to contribute and receive benefits is constrained by the presence and actions of their immediate neighbors. This spatial structure may introduce diffusion, so that contributions and strategies spread or are influenced by local interactions, rather than the global, uniform conditions of a classic public goods game [144]. Diffusion limits the reach of public goods, making local cooperation more viable.

In an evolutionary public goods game coupled with ecological dynamics, outcomes may be highly sensitive to initial conditions (shaping a saddle–point structure of the dynamics). Dependent on the parametrization, this may lead to oscillations or population extinction [69].

While cooperators produce and share essential metabolites or enzymes, defectors consume shared resources without contributing. The agents interact under different payoffs: cooperators incur a cost but increase group fitness; defectors gain short-term advantage but risk system collapse. In a *public goods game* with $n$ agents, there are at least $k \leq n$ contributing cooperators which are necessary to produce the public good. The cooperation dilemma arises because each individual has an incentive to avoid the cost of contributing resources ("volunteering") and instead free-ride on the resources contributed by others ("public goods"). Yet, if the public good is not provided, all individuals incur a cost greater than that of volunteering [2]. However, since $n$-person public goods games aggregate contributions non-linearly, a cooperation is possible, and a polymorphic state in which cooperators and non-cooperators coexist can be stable [2].

In selection-diffusion models of ecological public goods, cooperation emerges from spatial pattern formation. Slow diffusion causes cooperators to cluster in productive patches (activation: the presence of nearby cooperators enables or amplifies cooperative payoffs). Rapid diffusion enables defectors to invade and exploit these patches (inhibition: the presence of nearby defectors, or an insufficient density of cooperators, suppresses or undermines cooperative payoffs). The tension between these processes yields a variety of patterns resembling Turing patterns, including persistent coexistence and spatial chaos [154]. In these models, stability hinges on patch formation and invasion fronts (activation-inhibition), effectively turning point equilibria into geometric conditions on cooperative clusters, with the pattern formation shaped by the antagonistic forces of cooperators (activators) and defectors (inhibitors) [154].

From a dynamical-systems viewpoint, the interior coexistence equilibrium of cooperators and defectors may also be a saddle-type equilibrium with stable and unstable manifolds, such that some perturbations decay whereas others grow [69,70]. In the spatial selection-diffusion extension, the corresponding homogeneous coexistence state can become unstable to spatial perturbations, leading to diffusion-driven pattern formation and acting as an organizing center for pattern-forming transitions [154]. Thus, without spatial clustering or regulatory mechanisms (e.g., quorum sensing, membrane compartmentalization), cooperation cannot stabilize.

#### *3.1.4. Algorithmic bottleneck in the emergence of the shared genetic code*

Crucially, the spatial public goods game and the codon-signaling game can be viewed as coupled through a *multichannel* feedback loop, in which the output of each serves as an input to the other. For instance, the outcome of the signaling game — a more robust genetic code — can regulate quorum sensing and stabilize cooperative behavior, thereby increasing the collective payoff in the public goods game. In turn, the public goods game can reinforce the universality of the genetic code within spatial clusters by selectively rewarding compatible signaling systems. If, however, the codon-signaling game yields only a partially coherent genetic code, permeability of the protocell's membrane may be impaired, reducing metabolite retention, and genetic regulation may fail to suppress freeloading. As a result, a spatial public goods game with a saddle-type polymorphic equilibrium may shift toward defector dominance, leading to the collapse of the shared metabolic infrastructure.

Within this loop, the emerging code and cooperative metabolic networks are hypothesized to not only mutually stabilize one another, but the coupling would need to also give rise to self-modeling. In this process, compressed symbolic representations (that is, DNA information or "software") should acquire a functional role in shaping the behavioral patterns and phenotypical traits of the underlying protein "hardware". This bottleneck acts as a filtering process, permitting only those configurations that successfully balance informational stability with functional utility to persist. Such configurations must achieve system-level coherence between genotype and phenotype, satisfy multiple and competing selection pressures (including replication fidelity, catalytic efficiency, and resource constraints), and integrate spatial structuring, selective permeability for metabolite retention, and genetic regulation to suppress freeloading. Failure to evolve and stabilize this coupled set of mechanisms would likely prevent life from advancing beyond rudimentary and easily disrupted proto-cell communities. Fig. 5 illustrates saddle-type game-theoretic equilibria in multichannel signaling-coordination games where an evolutionary trajectory passes across two saddle landscapes.

From the perspective of algorithmic bottlenecks, the early emergence of widely shared genetic code on Earth does not contradict its difficulty. Rather, it suggests that the coding threshold is characterized by saddle-type equilibria that must be crossed early, before evolutionary dynamics settle into stable non-coding regimes. Trajectories that fail to traverse this bottleneck not only stabilize elsewhere, but may also block access to the bottleneck altogether. For example, early replicator systems may have achieved considerable complexity through catalytic RNA networks without evolving protein-based translation; increasing dependence on such networks could have reduced opportunities for the emergence of a translation apparatus. Thus, this bottleneck is difficult both to enter and to traverse. However, trajectories that successfully cross it may subsequently be drawn toward a stable coding regime, where selection rapidly reinforces and preserves the coding solution. Thus, the genetic code may be both rare across evolutionary possibilities and early in those rare histories where it succeeds.

### *3.2. Emergence of symbolic language*

It is well established that, while cooperation can exist without language, the emergence of language substantially enhances and restructures cooperative behavior [142]. Agents participate in collective actions and experience similar environmental feedback, successes and failures. Because agents experience themselves and others acting within the same loop of communication and cooperation, they can eventually form compressed, symbolic representations that model both their own behavior and that of the group. Ultimately, this generates a basis for differentiating between "self" and "non-self", and ultimately, for self-modeling: "the key criterion for modern human behavior is not the capacity for symbolic thought but the use of symbolism to mediate behaviour" [74]. These dynamics motivate an examination of the material, behavioral, and cognitive precursors that made such self-modeling and symbol-mediated regulation of behavior possible, and that ultimately culminated in the symbolic explosion.

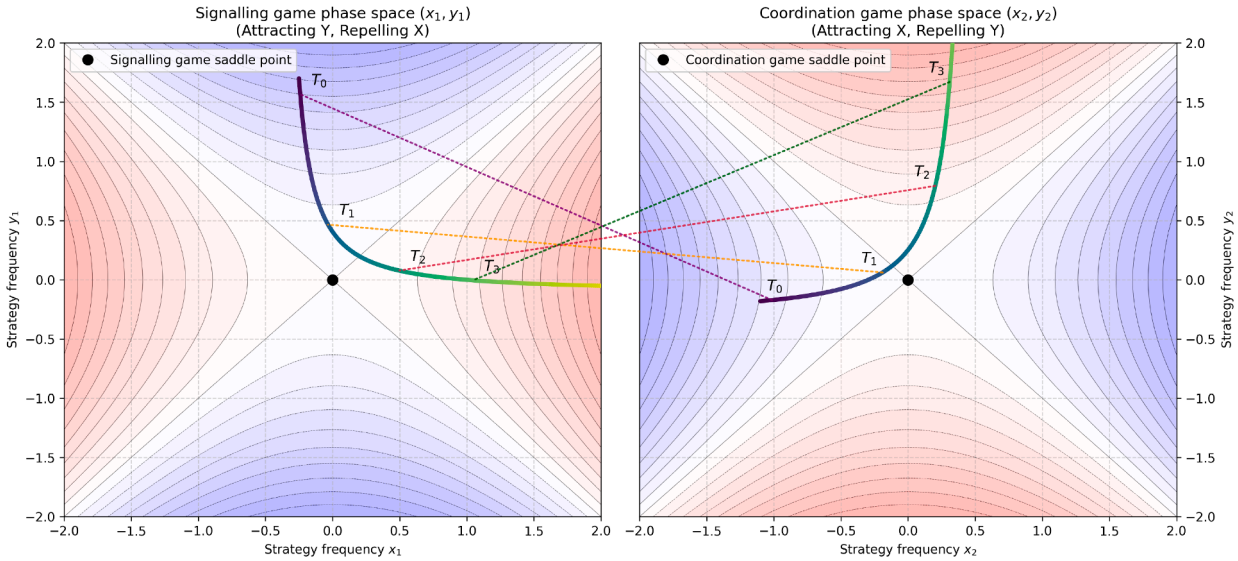



**Fig. 5.** Multichannel signaling-coordination games: synchronized evolutionary trajectory across two saddle landscapes. Left: The trajectory is pulled toward the equilibrium along the $y$-axis: $y$ is the stable (attracting) direction. The trajectory moves away from the saddle in the horizontal direction: $x$ is the unstable (repelling) direction. Signaling game trajectory initially moves toward the horizontal manifold, then escapes along the $x$-direction. Right: The trajectory is pulled toward the equilibrium along the $x$-axis: $x$ is the stable (attracting) direction. The trajectory moves away from the saddle in the vertical direction: $y$ is the unstable (repelling) direction. Coordination game trajectory initially moves toward the vertical manifold, then escapes along the $y$-direction.

### *3.2.1. Precursors of the symbolic explosion*

Humans stand apart from all other primates by the use of hard, durable tools, mastery of fire, deliberate burial practices, and creation of art [21,55,116,134]. In turn, these practices led to spatial patterning and shape creation, which formed the precursors of verbal language [55]:

> "The hard debris provided markers on the landscape for the location of human groups, a feature reinforced by the visibility of fires and the marking of the landscape with art. Human activity became entangled with these locational signals, creating an ever more complex milieu for our behaviour and putting great selective pressure on the analytic ability needed to exploit this information. Those humans who could perceive and remember the patterns of markers would have been advantaged in moving around the landscape, whether seeking to meet or avoid other humans."

In particular, once hominins started using color such as ochre circa 400–320,000 years ago [60], they could mark trails and territory. Crucially, hard objects, such as stone and bone debris scattered on occupation sites, not only left durable markers on the landscape but also reshaped and complicated our modes of interaction. Locational information became essential for guiding approach and avoidance behaviors, and the palimpsests of accumulated, layered debris that formed over time provided observers with cues about temporal sequence, activity order, and relational context (see Box 4).

This process can be naturally viewed in terms of niche construction [67,148]. Once humans began to modify their environment symbolically — through marks, tools, conventions, or shared representations — the resulting environment supported the groups that could interpret, manipulate, and extend those symbolic structures. In this way, symbolic practices reshaped the selective landscape, creating a positive feedback loop in which environmental modification and symbolic capacity mutually reinforced one another and led to stabilization of symbolic systems beyond the capacities of non-human animal communication [31,53,147].

In short, the symbolic explosion was preceded by several interrelated innovations: the use of durable objects as landscape markers and cues (i.e., spatial patterning), increasing sophistication in stone-tool production, the controlled use of fire (including fire-based signaling), a growing cognizance of space and time existing independently of the individual, and the advent of art-making practices such as shape creation and color-coding.

**Box 4: Expanding cognitive horizons: The emergence of symbolic reference.**

The increasing elaboration of stone-tool production demanded a higher cognitive capacity, with ever larger brains capable of remembering and executing progressively longer and more intricate action sequences [21]. The controlled use of fire generated spatial signals visible over substantial distances, creating reference points that extended far beyond the immediately accessible environment [134]. This capacity to determine the location of other hominins at long range — often without being detected — offered further selective advantages.

The emergence of abstract symbols may have been supported by an expansion of mental time horizons. The cognitive mechanisms underlying episodic memory and future projection are closely related, and increased capacities for remembering the distant past may have been accompanied by improved abilities to imagine distant futures [140]. As these temporal horizons expanded, hominins may increasingly have represented events extending beyond their own lifetimes, fostering more abstract conceptions of time analogous to already existing abstractions of space. Such an emerging awareness that the temporal world extends beyond the individual may have found expression in practices such as the deliberate burial of the dead, although the cognitive significance of mortuary behavior remains debated [116]. Together, these developments may have contributed to an emerging view of reality as extending beyond immediate personal experience [31], while also supporting richer representations of other agents as individuals with perspectives and trajectories independent of oneself.

We note that such capacities are not unique to humans: non-human animals are known to exhibit forms of spatial, temporal, and even social projection without possessing symbolic language. While virtual projection of space, time, and agency can exist without language [14,18,19,140], what appears to be distinctive is not abstraction itself [1,64], but the stabilization of abstract representations through shared, conventionalized, and recursively manipulable symbols [31,53] (see also next subsection).

Art depicting animals, humans, and composite beings that existed independently of the observer emerged around 70–50,000 years ago and became widespread during the Upper Paleolithic: this has often been interpreted as evidence for increasingly abstract representational capacities [20,34]. Such artistic expressions may have deeper roots in the intentional collection and processing of pigments, particularly ochre, documented in Africa during the Middle Pleistocene, dating back 300–400,000 years [9]. Significantly, art provided a means of storing information outside the brain, thus enabling increasingly sophisticated and structured forms of information processing [34]. Such external symbolic representations may also have encouraged increasingly structured relationships among signs, including shapes nested within other shapes, potentially foreshadowing some of the recursive properties characteristic of later symbolic systems [34]. These developments also encompassed the capacity to reimagine the familiar and to conjure entities never directly observed, for example, the lion-headed human (Löwenmensch) statuettes dated to 40–30,000 BP [20].

### *3.2.2. Least-effort linguistic game*

When individuals engage in collective actions, their repeated interactions create a common reference frame with aligned expectations and interpretations, allowing symbols or signals to reliably refer to the same things for different agents. However, this common reference frame comes at a cost.

A well-developed class of information-theoretic models [46,48,50,120,130] examines the evolution of language, conceived as a mapping between symbols and objects through the principle of least effort, which seeks to minimize the combined communicative burden of both speaker and listener. These models contrast, for instance, a degenerate system employing a single signal for all objects (thus, minimizing speaker effort) with a fully disambiguated system assigning a unique signal to each object (thus, minimizing listener effort). Different formulations introduce distinct cost functions. In spoken language, for example, the intrinsic cost of using a signal may be approximated by the time required to articulate the word, which imposes effort on both speaker and listener.

An early approach by Mandelbrot [100] treated language as a symbol-production process in which each word carries an inherent cost that increases logarithmically with its complexity. More recent formulations of the least-effort principle propose minimizing a composite cost function that balances the communication inefficiency (measured by the sum of speaker's effort of encoding objects and the listener's effort of decoding signals) with the intrinsic cost of each signal (word) [130]. Notably, models in this class often exhibit a phase transition separating degenerate languages that use a single signal and fully disambiguated languages with one-to-one mappings. This transition marks the critical coding threshold at which the competing information-theoretic costs of the speaker and listener come into balance.

Similarly, one may examine the emergence and evolution of language within a game-theoretic framework in which players act as speakers (encoding information, e.g., mapping signals to objects) and listeners (decoding information, interpreting signals). Players interact under payoff structures that reward reliable and efficient communication, reflected in more accurate belief updates, enhanced trust or group inclusion, and penalize ambiguity or excessive costs, such as increased cognitive load or delayed responses. For example, high payoff is achieved when sender and receiver converge on an unambiguous yet low-effort mapping, whereas low payoff results from extreme polysemy (a single signal referring to multiple objects) or high synonymy (multiple signals referring to the same object). Within such models, strategies may range from random signal-object assignments to more structured conventions that minimize miscommunication effort. Evolutionarily stable strategies may emerge when signaling systems settle into equilibria that balance efficiency with cost [22]. In this way, game-theoretic interactions can naturally give rise to increasingly efficient signaling conventions, providing plausible evolutionary precursors to the language threshold.

However, if signaling becomes too noisy or too costly, reliable and rapid recognition, and thus effective communication, breaks down. As with codon-to-amino acid mappings, the space of possible signal-object assignments is combinatorially vast, and interacting players may fail to agree on a common signaling scheme. This can produce fragmented or incompatible systems of communication that favor short-term survival strategies over long-term optimization of the language.

From a dynamical-systems perspective, such fragmented communication systems may correspond to saddle-type coexistence equilibria, in which partially established signaling conventions are stable to some perturbations but unstable to others. In these states, populations can be temporarily drawn toward a common signal-object mapping, creating the appearance of linguistic convention, yet small fluctuations, competing conventions, or stochastic errors may be amplified along unstable directions. Consequently, particular conventions may temporarily dominate, but remain susceptible to displacement by alternative mappings, preventing convergence to a universally shared language. Similar instability and equilibrium-selection phenomena have been studied in evolutionary signaling games and replicator-based models of language evolution [82]. The resulting dynamics are characterized by continual reorganization around unstable coexistence states, making the long-term stabilization of weakly aligned linguistic systems difficult to achieve.

Just as the early emergence of genetic coding reflects the traversal of a narrow algorithmic bottleneck rather than a long waiting time, the emergence of symbolic language can be understood as a rare arrangement of mutually dependent signaling and coordination requirements. Once a sufficiently robust and locally shared convention that supports communication despite ambiguity and variation is achieved, it rapidly stabilizes and expands. However, if missed, alternative lineages may be confined to stable yet non-symbolic communicative niches, e.g., rich in indexical and iconic signaling, but lacking recursion, abstraction, and collective control mechanisms such as norms.

Under our conjecture, coding thresholds can manifest as saddle-type bottlenecks whose stable manifolds occupy only a narrow region of state space. In this view, phase transitions remain one possible expression of evolutionary hardness, but not its only source. This perspective can be naturally connected to the optimization framework of Ferrer-i-Cancho [47]. Rather than requiring languages to remain finely tuned at a critical point, the language threshold may correspond to reaching a restricted class of near-optimal coding solutions. Hardness then reflects the difficulty of evolutionary search and coordination required to enter this class, with criticality emerging as a special case. Importantly, this yields a testable criterion: human languages appear to occupy such near-optimal coding regimes, whereas non-human communication systems do not, suggesting that the language threshold may be defined by entry into a distinctive class of coding-optimal systems [47].

This interpretation does not imply that other Earth species, e.g., non-human apes, are intrinsically incapable of language. Rather, it suggests that once a lineage becomes highly adapted around a particular communicative architecture, the evolutionary trajectory leading to a symbolic coding transition may occupy an increasingly narrow region of phase space. Recent discoveries of increasingly structured communication in birds, cetaceans, and other species [3,42,117,136] are consistent with the saddle-point bottleneck hypothesis: many evolutionary trajectories may approach the language threshold along stable directions, acquiring combinatorial and compositional features. Such trajectories may linger near the threshold and exhibit increasingly language-like properties. Nevertheless, they may fail to traverse the bottleneck, being diverted along unstable directions before achieving recursively closed symbolic language with cumulative symbolic inheritance. Thus, a species' capacity for language is distinct from the likelihood of traversing the language bottleneck through natural evolutionary search.

An interesting boundary case is the waggle dance of honey bees that perform a sophisticated movement to communicate displaced spatial location of food sources, water, or new nest sites to hive mates [57]. In particular, dancing bees are able to encode navigational information and translate waggle dances into a vector guiding a searching flight, essentially solving a computational, mapping problem, which in turn is decoded by follower bees [37]. However, the waggle dance is not recursively manipulable: it does not support embedding of signals within signals, composition of messages into higher-order structures, or operations such as negating or conditionalizing [53,71]. Moreover, while being symbolic in a minimal sense, the dance leaves no durable external record, and cannot be revisited, edited, or re-interpreted after the performance ends. Even though the waggle dance is socially shared, it does not enable cumulative modification of symbols or trans-generational symbolic accumulation [31]. In short, it can be argued that bees did not cross the symbolic language threshold.

While human spoken language does not generate a permanent material record, we note that it can nonetheless create relatively durable external representations through collective memory, ritualized repetition, teaching, storytelling, and so on. Many complex societies maintained extensive historical, legal and cosmological knowledge before writing [112,151]. Oral traditions, pedagogical lineages, and linguistic conventions function as forms of distributed external memory that can preserve information through repeated transmission across many generations, albeit with higher transmission error than written records, due to drift, loss, and reinterpretation. Writing externalizes symbolic content into a medium that substantially reduces dependence on memory and repeated reconstruction, and should be understood not as the origin of external symbolic storage, but as a dramatic enhancement of its persistence, fidelity, and scalability [63]. Thus, the relevant distinction is not durability *per se* but fidelity of preservation.

### *3.2.3. The coordination game*

During collective actions, each agent adjusts their behavior and internal models based not only on their own outcomes, but on patterns that are reinforced through responses of others, enabling coordination and prediction. In particular, the confrontational scavenging scenario suggests that proto-linguistic capacities in *Homo erectus* coevolved with early forms of cooperative behavior among individuals who were not necessarily closely related [142]. In this scenario, communication and cooperation must interleave together in a synergistic fashion, as otherwise, the resulting imbalance imposes a significant fitness cost.

In analyzing this cooperative scenario, Szathmáry [142] compared two well-known game-theoretic setups. In the $n$-person Prisoner's Dilemma, individuals may choose to cooperate or defect: cooperation entails paying a contribution cost, while defection does

not. All contributions are linearly aggregated, scaled by a reward factor $r$, and the resulting sum is evenly redistributed to all individuals, irrespective of their contribution. To defect is known to be the only evolutionary stable strategy (ESS) when $r < n$, which leads to complete absence of cooperation and a Pareto-inefficient equilibrium [2]. This is in stark contrast with the $n$-person public goods game, also known as the Teamwork Dilemma.

As argued by Szathmáry [142], the dynamics of cooperation in the confrontational scavenging scenario follows the Teamwork Dilemma where the collective benefit increases with the number of cooperators in a non-linear (sigmoid) fashion, and hence, attains a cooperative equilibrium. This dynamics is attributed to language which allows for a negotiated division of labor [142]:

> "This transition happened in early *Homo erectus*, who faced the problem of starvation due to the disappearance of fruits in that period. There was, however, plenty of meat around, including carcasses of the megafauna. …To use this resource, three crucial actions are needed: First, members of the group who cannot know about the carcass must be informed about its nature, location, and distance; second, they need to be recruited; and, third, execution of the task requires intense cooperation with limited opportunity to cheat. …It was this niche that allowed a wedge to penetrate the previous animal communication system by signals for displaced items. Given the fact that by then *erectus* had already had a large brain and was very likely equipped with Machiavellian social intelligence (13, 14), the process did not stop there, and protolanguage with increasing richness of symbolic reference started to evolve, to be followed by syntax that presumably emerged with the speciation of *Homo sapiens* (12)."

In general, advances in tool use increase the potential payoff of cooperation, while simultaneously raising the complexity of trust formation and signaling. The resulting tension between individual safety and collective gain can also be captured by the *Stag Hunt* game (i.e., the assurance game, trust dilemma, or common-interest game) [45]. In this framework, two or more agents choose between non-cooperative strategies of acting alone (safe but yielding low payoffs, analogous to hunting a hare), or cooperative strategies that are risky but potentially highly rewarding (analogous to hunting a stag). Although communal action generates higher payoffs when agents successfully coordinate and combine their skills, it crucially depends on mutual trust.

Although the payoff structure of the Stag Hunt game rewards mutual cooperation and penalizes unilateral risk-taking, achieving the cooperative equilibrium requires trust and mutual assurance. In the absence of trust and under uncertainty about others' actions, the non-cooperative equilibrium becomes *risk-dominant*, while the cooperative (payoff-dominant) equilibrium is not. As a result, cooperation is difficult to sustain (not robust) under uncertainty: players tend to favor the safer, lower-payoff option when facing the possibility of coordination failure. Consequently, the Stag strategy is not evolutionarily robust without dependable coordination mechanisms. Early intelligent agents may oscillate between cooperation and defection, producing easily destabilized collective arrangements. Without stabilizing proto-conventions or proto-norms, populations are likely to retreat to safer, though less collectively optimal, strategies.

From a dynamical-systems viewpoint, the sensitivity of outcomes to variations in trust, signaling and other perturbations can be understood through the presence of a saddle-type equilibrium that separates the basins of attraction of cooperation and non-cooperation. Which equilibrium is risk-dominant depends not on the existence of the saddle itself, but on the relative sizes of the corresponding basins of attraction. Under evolutionary dynamics such as replicator dynamics, mutual cooperation and mutual defection are typically stable attractors, while an intermediate mixed-strategy equilibrium acts as a saddle point. The saddle's stable manifold forms a basin boundary: populations that begin on one side converge toward the payoff-dominant cooperative equilibrium, whereas those on the other side converge toward the risk-dominant non-cooperative equilibrium. Trust, signaling, and emerging social conventions can therefore be viewed as mechanisms that shift populations across this boundary or enlarge the basin of attraction of cooperation [72,131,149]. In their absence, uncertainty and stochastic fluctuations may drive populations toward the non-cooperative attractor, rendering cooperative arrangements difficult to sustain.

#### *3.2.4. Algorithmic bottleneck in the emergence of the symbolic language*

The teamwork dilemma (formalized as a public-goods game or the Stag Hunt game) and the least-effort linguistic game can be modeled as coupled through a self-referential feedback loop, in which the output of each serves as an input to the other. The outcome of the linguistic game — the emergence of a shared symbolic code (language) — constrains the collective benefits achievable by teams of individuals, via coordination, trust, tool use, and the efficiency of cooperative action. In turn, the payoff structure and outcomes of cooperation feed back into the linguistic game by influencing which signaling conventions and norms are stabilized, refined, or abandoned.

This coupling would induce a multichannel dynamic analogous to that considered in the co-evolution of genetic code and metabolism. On the one hand, an effective and broadly shared language can stabilize cooperation, increase group-level payoffs, and expand the range of viable cooperative strategies and tool use. On the other hand, cooperative success creates selection pressure for signaling systems that are more expressive, reliable, and widely shared, thereby promoting the convergence of linguistic conventions. However, this same coupling can also destabilize otherwise well-behaved dynamics: failures of signaling can undermine trust and cooperation, while breakdowns in cooperation can prevent the stabilization of shared linguistic codes.

In effect, desirable properties of separate games are no longer guaranteed in the coupled system. Instead, the joint (multichannel) dynamics may either converge to mutually reinforcing equilibria, in which language and cooperation stabilize one another, or collapse into regimes characterized by defection, fragmentation of signaling conventions, and the loss of collective benefits. In this sense, the coupled signaling–cooperation system introduces another algorithmic bottleneck. Only those configurations in which symbolic communication ("software" or compressed functional information) and cooperative behaviors and tool use ("hardware") are jointly

compatible can persist and evolve further. Importantly, this loop can be interpreted as involving self-modeling of shared experiences: a body of interactions, perceptions, and outcomes that are jointly encountered by multiple agents and that become mutually recognizable and interpretable over time [16,76,96,147].

The self-referential nature of the "software"–"hardware" dynamics is highlighted in the reflection-capable computational architecture described by Barron et al. [10]:

> "Natural language also provides an efficient means for externalizing control flow by distributing complexity among groups of individuals, allowing groups to perform tasks too complex for any individual [92]. Dor [24] has proposed that language enables 'collaborative computation' between groups of individuals. This crucially involves treating utterances not merely as informative messages but as instructions which can in turn influence control flow. Collaborative work can enable the small-scale apprentice learning that allows for technological transfer and cooperative hunting [93]. On a broader scale, our ability to distribute computational labour allows for complex computations that would be beyond the grasp of any particular human. Reflection thus enables distinctively human forms of learning and cultural accumulation."

The emergence of symbolic artifacts, the capacity to store symbols externally, and the evolution of symbolically mediated behavior can be related to a cognitive change that enabled modern humans to represent different and even conflicting perspectives of objects: "the capacity to simulate how objects look from conflicting viewpoints" [74]. This also assumes a capacity to represent the perspective of others on objects (including oneself), for example, the practice of wearing personal ornaments like beads indicates an underlying sense of self-awareness and self-recognition [74]. Thus, it has been suggested that the cognitive capacities required to transform an object into a symbol are closely tied to the ability to adopt a self-reflective perspective [74].

Crucially, perspective-taking appears vital for crafting formal tools, as it allows the tool-maker to maintain awareness of the subordinate category of the tool during production. For example, the Middle Stone Age artefacts found in Blombos Cave (southern Africa) provided evidence for an early bone industry, with tools falling not only within the general category of "bone tools" but also within subordinate categories such as "bone projectile points" or "bone awls" [73,74].

The analysis of the precursors of symbolic language through the cognitive ability to imagine and mentally simulate how objects appear from multiple, potentially divergent viewpoints, offered by Henshilwood and Dubreuil [74], supports our thesis on the "algorithmic origin of language". Specifically, we can identify the "symbolic explosion" or the "language threshold" with the emergence of a tangled information hierarchy between language and symbolically mediated behavior. Importantly, the self-reflecting ability, enabled by this hierarchy, allowed for recursive syntax, with clauses embedded within clauses [71,85]: "the Rubicon of modern language probably lies in recursion", as without recursive syntax, one cannot express or integrate conflicting perspectives [74].

In summary, crossing the "language threshold" and associated coordination barrier — amplified by strategic uncertainty and environmental risks — poses a central challenge for cooperating tool-using agents. To exploit high-reward opportunities such as hunting large prey, building tools, or knowledge sharing, these agents must reliably communicate and coordinate. This coupling constitutes an algorithmic bottleneck that constrains the transition to symbolic language, whose successful scaling alongside tool use and cooperative behaviors is not guaranteed.

## 4. Discussion

In this article we explored the conjecture that the hardest steps on the path towards technological societies capable of interstellar travel can be interpreted as algorithmic bottlenecks, shaped by (nested) tangled information hierarchies. Symbolic encoding, interpretive machinery and recursive closure may be viewed as prerequisites for these bottlenecks, whereas durable external symbolic memory and recursive symbol manipulation are the informational capabilities that fully emerge once those prerequisites are established. We conjecture that the mechanisms generating these bottlenecks may correspond to saddle-type equilibria in evolutionary game-theoretic models, which reduce the probability of complex lifeforms and societies.

We point out that both considered evolutionary transitions — the emergence of genetic code and the emergence of symbolic language — combine sophisticated signaling and strategic coordination problems in competitive and noisy environments. These dynamics must not only stabilize themselves, but also establish new information architectures that unlock further evolutionary phase spaces, enabling novel ways to store, transmit and interpret information. Therefore, these evolutionary transitions are far from guaranteed. Their success depends on navigating biochemical and socio-biological landscapes in which signaling and coordination remain vulnerable to perturbations that push dynamics along unstable directions, and where defection, whether through reduced signaling effort or exploitation of shared resources, is tempting. Societies with advanced technologies may therefore be rare because at least one code-invention is very hard, path-dependent, or successful only along a narrow range of evolutionary trajectories. Thus, the Great Filter arising from algorithmic bottlenecks may leave only a small subset of trajectories capable of reaching long-term survival.

To reiterate, a hard algorithmic bottleneck need not be crossed late. It may instead be traversed early in those rare evolutionary histories where the requisite conditions arise. Furthermore, because the second bottleneck lies within the evolutionary landscape made accessible by the first, the two transitions should be regarded as sequentially dependent (nested). Thus, the fact that both the genetic-coding and symbolic-communication thresholds have been crossed on Earth does not imply that such transitions are generally easy. Rather, it suggests that our evolutionary trajectory was among the few that reached and traversed both bottlenecks before becoming trapped in alternative stable regimes. Furthermore, these thresholds should not be understood as instantaneous

inventions emerging without antecedents. Most likely, they have been preceded by periods of gradual evolutionary accumulation, leading to a discontinuous expansion of the accessible evolutionary phase space.

A recent study hypothesized the existence of a biological arrow of time underlying major evolutionary transitions, and formalized this conjecture through the lens of computation theory [121]. According to this interpretation, biological organisms are hierarchical dynamical systems that generate regularities in their phase-spaces through interactions with their environment. These emergent information patterns can be encoded within the organism (e.g., via the genotype-phenotype mapping), leading to "fundamental tensions (computational inconsistencies) between what is encodable at a particular evolutionary stage and what is potentially realizable in the environment" [121]. This study argued that open-ended evolution comprises a continual process in which a resolution of these tensions drives an evolutionary transition, expanding the problem-space and then generating new tensions in the expanded space.

Computation-theoretically, the problem-space expands when computational novelty (e.g., a new "code of life") is generated by self-modelling agents which gain access to meta-simulation [121], e.g., by Turing oracles or recursively generated logics [103,104,119,139,150], or receive the feedback from external environment, e.g., via natural selection [141]. Self-modeling, e.g., via genome or language, yields big advantages: robustness, division of labor, and cooperation at scale. Under this perspective, crossing the "coding threshold" may require agents to effectively "Jump Out Of The System" (JOOTS) [80], only to inevitably encounter yet another incomputable paradox within the expanded, higher-level system.

The two algorithmic bottlenecks examined in this work punctuate the path of open-ended evolution by necessitating the emergence and stabilization of a "software"–"hardware" separation under selection pressures. Crossing the associated coding thresholds requires the formation of self-referential, self-modeling loops that, in turn, enable novel inheritance systems and shared representational frameworks.

This perspective suggests that despite continually opening new evolutionary possibilities, this process may progressively narrow the evolutionary path available for recursively closed code–interpreter systems. Every successful coding transition expands the accessible phase space while simultaneously increasing its topological complexity. The resulting accumulation of nested information hierarchies may lead to a diminishing set of high-dimensional saddle-point pathways connecting viable regions, thereby reducing the fraction of traversable evolutionary trajectories. This may provide a mechanism by which open-ended evolution generates increasingly severe algorithmic bottlenecks that constrain subsequent innovations.

Contemporary evolutionary AI provides an instructive analogue of these coding thresholds. For example, neuroevolution studies have shown that strong attentional, memory, and information-processing bottlenecks can drive agents to develop compressed, robust, and partially self-interpretable internal representations [146]. Such results suggest that informational constraints need not merely limit adaptation; they may also catalyze the emergence of higher-level abstractions that improve coordination between representation and action. In this sense, artificial evolutionary systems offer a computational proof-of-concept for our conjecture that algorithmic bottlenecks can promote the formation of new information architectures rather than simply impede evolutionary progress.

As has been noted, "just as the evolution of powerful epigenetic inheritance systems allowed the evolution of complex multicellularity, natural language allowed the emergence of complex human societies" [84,142]. However, as the capacities for communication and shared representation scale into forms of collective intelligence, societies begin to confront analogous "coding thresholds" at the cultural and institutional level. These higher-order systems would need to stabilize cooperative codes, shared protocols, norms, and governance mechanisms in environments where strategic pressures can easily destabilize them. In this way, the challenges faced by evolving technological societies echo the earlier algorithmic bottlenecks of biological evolution.

The emergence of technological society can be modeled as an iterated or evolutionary $n$-player *Prisoner's Dilemma* (Commons Dilemma), in which cooperation must be maintained despite incentives to defect. As societies become increasingly interconnected and globalized, the *tragedy of the commons* extends across planetary boundaries, amplifying the difficulty of attaining cooperative equilibria. National governments and legal systems struggle to construct multilateral institutions that are robust enough to manage shared resources, regulate collective *externalities* (e.g., climate change, biodiversity loss, pollution), and coordinate actions across diverse actors whose interests are only partially aligned [157].

Another generic aspect of the challenges facing complex societies is *short-termism*: a decision-making bias prioritizing short-term rewards over long-term outcomes, which can manifest as a lack of inter-generational equity, e.g., with respect to the climate change crisis [95]. For example, there is a strong argument that our modern society (the Anthropocene) is operating at — or has surpassed — the Earth's carrying capacity in respect of planetary boundaries [126]. Without overarching constraints, short-termism is unable to capture complex responses required by long-term challenges [52].

Strategies that may be employed by technological societies in an attempt to resolve these challenges include innovation (in creating a new shared code of societal life). Crucially, innovation may go beyond the directed invention of a specific function, and instead recombine the accumulated material features and devices that were often generated for unrelated purposes. As emphasized by Fletcher [56], novel capacities frequently arise when existing assemblages of tools, practices, and infrastructures are brought into new configurations, thereby acquiring functions not anticipated at their origin[5]. Societies that encourage experimentation and variation are more likely to introduce new modes of information processing, coordination, and control, leading to qualitative shifts in collective behavior rather than incremental improvement. Other strategies, such as steady-state dynamics (e.g., steady state economy [25]), or degrowth/post-growth, are aimed at reducing production, consumption and environmental impacts, but these approaches have their own challenges [88,91]. Overall, technological societies that fail to build resilient institutions or internalize all key impacts of tech-

[5] For example, redesigning pendulum clocks into marine chronometers solved the "longitude problem" and revolutionized maritime navigation.

nological development, such as environmental or social costs, may decline or collapse before reaching interstellar/astroengineering capability.

Another emerging source of systemic risk arises from advanced forms of Artificial Intelligence (AI), such as agentic AI [43]. As has been observed, human society can be understood as a kind of "operating system": a multi-layered structure built from stories, norms, legal frameworks, monetary systems, and religious worldviews, all of which are mediated by language [66]. Harari [66] argued that generative AI systems have effectively "hacked" this societal operating system, intervening directly in the linguistic substrate on which collective understanding and decision-making depend. In other words, AI has gained the capacity to act upon our societal "code of life", influencing the symbols, narratives, and discourses through which large-scale cooperation is organized. The result is a destabilizing feedback loop: AI systems trained on human discourse begin to influence that same discourse, altering the informational ecology in ways that may erode cooperative norms or amplify collective vulnerabilities.

This "operating system" metaphor again underscores the algorithmic character of tangled societal hierarchy in which language-mediated information functions as system software coordinating both the social substrate (hardware) and its various informational resources (other software components). It is well-known that some of the most destructive systemic threats arise from exploiting vulnerabilities in a system's capacity for accurate self-description [105]. Examples include computer viruses that alter host code to invert the intended computation [102], and novel antigens that compromise the immune system's "self versus non-self" discrimination, precipitating autoimmune diseases [58]. By the same logic, the "hacking" of language-mediated societal functions represents an existential risk, because it targets the very substrate through which collective coordination is maintained. Thus, the formation of a stable tangled societal hierarchy may constitute yet another algorithmic bottleneck, susceptible to disruption by accelerating, and potentially runaway, AI-dominated dynamics. A precise delineation of the emerging "coding threshold" for societal intelligence [43, 93] remains unresolved, representing a future hard step that humanity has yet to navigate.

## 5. Conclusion

Informally, life, mind, and society are all information engines that extract order from disorder. However, early conditions constrain future possibilities, and Great Filters may correspond to algorithmic bottlenecks, manifested in many ways, as limits in error correction, e.g., mutation rates [39]; cognitive constraints, e.g., Dunbar's number [36,68]; social fragility, e.g., loss of knowledge in collapse [11,101], and so on. Under our conjecture, each algorithmic bottleneck, beginning with the emergence of genetic coding and followed by the emergence of symbolic language, moves evolution toward the next level of informational complexity. These bottlenecks are nested in the sense that each higher-level transition presupposes the successful stabilization of earlier coding system(s), constraining subsequent evolution to occur only within lineages that have already crossed preceding threshold(s). Specifically, agents capable of symbolic language evolve within lineages that have first stabilized genetic coding, rather than emerge along trajectories that bypass this transition, for example, following purely chemical pathways. We hypothesize that because these tangled information hierarchies are nested rather than independent, the sequence of coding thresholds forms a Great Algorithmic Filter, with generic fundamental properties.

A limitation of this perspective is the potential inflation of evolutionary filters through anthropocentric biases or generalization from the single known history of life on Earth. Features that appear uniquely important when viewed through the lens of human evolution may reflect path-dependent contingencies rather than general principles or universal evolutionary constraints. Consequently, some elements of the identified bottlenecks may prove less restrictive in alternative biological, cognitive, or ecological contexts, including those realized by other Earth-originating or extraterrestrial intelligences.

Our game-theoretic analysis is intended to shift the interpretation of "hard steps" from elusive improbability to equilibrium structure, arguing that algorithmic bottlenecks may correspond to saddle-type equilibria in multichannel signaling–coordination games where signaling, coordination, and control must tightly align. Importantly, this game-theoretic analysis is not intended to provide a detailed mechanistic simulation of the evolutionary dynamics leading up to these bottlenecks. Instead, it aims to diagnose the structural difficulty of these transitions and expose the equilibrium-level constraints that make their traversal inherently difficult. Rather than presuming that complexity inevitably accumulates through adaptive innovations, this hypothesis suggests that information-processing innovations may be vulnerable to strategic collapse due to multichannel failures, thus increasing the risk of systemic fragility and evolutionary dead ends.

## Author contributions statement

All authors participated in discussing this perspective, contributed to and reviewed the manuscript.

## Funding declaration

This research did not receive funding.

## Competing interests

Authors declare no competing interests.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No new data were created or analysed in this study.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During manuscript preparation, the authors used Microsoft Copilot to paraphrase text and refine figures. The authors confirm that where content was modified by generative AI, the resulting material was reviewed for possible errors, inaccuracies, and bias.

## Appendix A. Conceptual analogy between genetic and linguistic coding systems

**Table A.2**
Conceptual mapping between genetic and linguistic coding systems.

| Genetic system | Linguistic system | Generic role | References |
|---|---|---|---|
| Nucleotide bases | Graphemes, letters | Basic symbolic units | [5,86,114,138] |
| DNA sequences | Symbol strings | Abstract, discrete information carriers | [5,86,114,138] |
| Codons | Words, morphemes | Basic symbolic units combined to form higher-level structures | [5,86,114,138] |
| Genes | Sentences, lexical items | Minimal functional units carrying meaning or instruction | [5,86,114,138] |
| Genetic code (codon → amino acid mapping) | Lexicon, semantic conventions (symbol → meaning mapping) | Arbitrary but shared mapping between symbols and meanings | [5,86,114,138,166] |
| Translation machinery (ribosome, tRNA, enzymes) | Interpreter (parser plus semantic engine) | Mechanisms that interpret symbolic strings into functional outcomes | [5,32,114,138] |
| Ribosome | Parser applying grammar, syntax | Rule-based structure governing valid combinations | [5,32,114,138] |
| tRNA adapters | Semantic interpreters, conceptual schemas | Mediators linking symbols to functional meanings | [5,32,114,138] |
| Nested RNA secondary structures | Recursive syntax | Nested, recursive symbolic structures | [71,135] |
| Proteins | Communicative acts, structured thoughts and behaviors | Functional products of interpretation | [5,6,32,113,114] |
| Protein folding | Pragmatics, contextual interpretation | Context-dependent shaping of function or meaning | [5,32,79,113] |
| Gene regulation | Discourse rules, conversational norms | Control of when and how expressions are produced | [5,32,79,113] |
| Replication fidelity | Transmission fidelity (learning, teaching) | Accuracy of inheritance across generations | [13,29,143] |
| Mutations | Linguistic variation, semantic drift | Source of novelty and variation | [13,23,24,29,108,143] |
| Horizontal gene transfer | Language contact, borrowing | Exchange of information across lineages or groups | [23,35,62] |
| Genome | Book / Linguistic competence | Particular repository of symbolic information | [32,113,135] |
| Pangenome | Library, corpus / Language | Collective repository of symbolic information, durable external memory | [13,108,135] |

The conceptual analogy between genetic and linguistic coding systems is more than merely metaphorical, see Table A.2. In both cases, symbols are inert in the absence of an interpreting machinery capable of mapping them onto functional outcomes. The code itself is arbitrary but conventional — widely shared within a population — and therefore requires collective stabilization to persist. Importantly, functional consequences arise only through rule-governed translation, rather than via direct causal interaction between symbols and outcomes. This shared structure suggests that both transitions constitute genuine hard steps in an algorithmic sense, marked by the emergence of a new level of symbolic representation and interpretation within a recursively closed architecture.

This analogy does not imply a strict or unique correspondence, and alternative mappings are possible [44,135]. Moreover, some of the proposed correspondences remain debated. For example, an individual genome may be likened to a single text (e.g., book), whereas a pangenome may be compared to a corpus, text collection or library. At a more abstract level, however, genome may also be viewed as the repertoire from which potential texts, utterances, and actions can be generated. In this, generative sense, an

individual genome is analogous to linguistic competence [32,113], whereas a pangenome would be more closely comparable to language or discourse traditions [13,108]. Another non-unique correspondence involves recursive syntax, which has been compared both to nested RNA secondary structures [71,135], and to hierarchically organized gene regulatory networks [26].

In this article, we distinguish symbolic reference from syntax. Symbolic reference denotes the use of socially shared and conventionalized symbols whose meanings are not fixed by immediate context, resemblance, or physical causation. Syntax denotes the combinatorial principles governing how such symbols may be assembled into higher-order expressions, including recursively structured constructions. We remain agnostic regarding the precise historical order of their emergence. Some accounts propose increasingly rich symbolic reference preceding syntax, whereas others argue that the two capacities co-evolved. Our language threshold refers not to either capacity in isolation, but to the stabilization of a jointly interpretable symbolic system capable of supporting open-ended communication, cumulative inter-generational transmission, and shared control flow. An operational perspective treats symbolic reference as a non-degenerate mapping between signals and referents, while syntax emerges when combinatorial organization among signals becomes advantageous under communicative constraints. Under this view, symbolic reference and syntactic structure may arise jointly as consequences of pressures toward efficient communication [49].

## Appendix B. A glossary of terms

**Collectively catalytic system** is a system where multiple components work interdependently to accelerate processes. A collectively autocatalytic set is a collection of molecular entities in which each member's creation is catalyzed by another entity within the same set. As a whole, the set sustains and accelerates its own production [89].

**Evolutionary singularities** are defined by De Duve [30] as "events or properties that have the quality of singleness, uniqueness". In other words, evolutionary singularities appear to have occurred once (or extremely rarely) and have left no independent alternatives or parallel realizations. In this sense, an evolutionary singularity is a path-dependent turning point, constraining future possibilities, by establishing a new organizational principle or informational architecture that becomes foundational for all subsequent evolution. Canonical examples considered by De Duve [30] in the context of the Earth's evolutionary record, include the origin of life, the genetic code, and the transition to complex cellular organization, each of which appears universally shared across extant life and admits no known competing realizations.

**Externality**, in economic theory, refers to a cost or benefit that arises from an agent's activity and affects third parties who are not directly involved in the transaction [12]. For example, motor vehicle emissions is a negative externality, with the societal costs of air pollution — such as public health impacts and environmental degradation — are not fully borne by either the producers or consumers of motorized transport. It has been argued that since these third-party effects — whether negative (external costs) or positive (external benefits) — are not captured by market prices, they yield potential inefficiencies in resource allocation [99,132].

**Iconic references** are signals that resemble or mimic their referents through structural or perceptual similarity. Their meaning arises from likeness rather than arbitrary convention, e.g., gestures, or words that phonetically imitate, resemble, or suggest the sounds they describe (onomatopoeia).

**Indexical references** are linguistic expressions whose reference shifts depending on the context of the utterance, such as the identity, location, or time of the speaker. Examples include words such as "I", "here", "now", which do not have a fixed, stable referent across different situations. In other words, the interpretation of indexical references is fixed by context, not by convention.

**Iterated Prisoner's dilemma** is a repeated game where two players repeatedly choose to "cooperate" or "defect", allowing for strategies based on past moves, unlike the single-shot version where defecting is always best. In the iterated Prisoner's dilemma, developing trust and reciprocating actions (like "Tit for Tat": start by cooperating and then mirror the opponent's previous move, i.e., cooperate if they cooperated, defect if they defected) can lead to mutual cooperation and better overall outcomes than always defecting. However, if players know the game will end after a fixed, finite number of rounds, this logic unravels through backward induction. Players will defect in the last round, which means they should also defect in the second-to-last round, and so on, leading to defection from the start. Under these conditions, the cooperative outcome is less robust to uncertainty than the non-cooperative which makes the latter the *risk-dominant equilibrium*.

**Misincorporation** is an addition of a wrong amino acid into a growing polypeptide chain during translation, relative to the codon specified by the mRNA.

**Mixed-strategy Nash equilibrium** in game theory is a strategy profile where players randomly choose between their available pure strategies with specific probabilities (i.e., there is a probability distribution over different strategies, e.g., $1/3$ assigned to each strategy in Rock-Paper-Scissors game).

**Multichannel games** in game theory refer to a framework where multiple, distinct strategic interactions are interconnected, such that the actions, outcomes, or payoffs of one game influence the parameters or choices in another [33].

**Pareto efficiency** describes a resource allocation outcome without an alternative allocation that increases the payoff of any player without decreasing the payoff of at least one other player. A Nash equilibrium is payoff dominant if it is Pareto superior to all other *Nash equilibria* in the game.

**Risk-dominant equilibrium** in game theory is a stable Nash equilibrium that has the largest basin of attraction, making it less sensitive to uncertainty. That is, players are more likely to choose this profile when they are uncertain about their opponents' actions, favoring strategies that minimize potential losses rather than maximize potential gains (payoff dominance). For example, the Stag Hunt is an assurance game with two pure strategy Nash equilibria: a cooperative, payoff-dominant equilibrium (both hunt stag) and a safer, non-cooperative equilibrium (both hunt hare). While both equilibria are stable, the cooperative equilibrium is not risk-dominant: it is not robust to uncertainty. A lack of trust and the fear of the worst-case scenario makes players choose the safer,

less rewarding option, creating a dilemma between the high-reward, high-risk cooperative outcome and the low-reward, low-risk non-cooperative one.

**Self-regarding signals** in language theory are communicative acts whose reference is anchored to the signaler's own immediate state, perspective, or situation, rather than to shared, abstract, or displaced entities. They convey information about "me, here, now", and do not require a shared symbolic code.

**Short-termism** is a decision-making bias which gives priority to short-term profit or reward at the expense of long-term results and far-seeing action, thus motivating quickly executed projects [51,160]. Abstract reasoning about time and the effects of decisions and actions assumes the abilities to conceive, represent and imagine (or simulate) dynamic trajectories. These abilities are likely to have evolved in many extraterrestrial lifeforms via adaptations which improved the accuracy of prediction (forward simulation) and explanation (backward simulation), and thus, provided selective advantages. In turn, a partial inability to look beyond the short-term horizon is a bias that can be generalized to the extraterrestrial lifeforms which developed a concept of time, as well as a capacity to reason about past and future scenarios.

**Stable equilibrium**, in the context of replicator dynamics, is a population strategy distribution that persists under the evolutionary dynamics: following small perturbations, selection pressures drive the population back toward the equilibrium, rather than away from it, because any mutant or deviating strategy earns lower payoff than the resident mix. A stable equilibrium under replicator dynamics may not be risk-dominant.

**Technosignature-bearing society** refers to an extraterrestrial society that produces measurable technosignatures: observable properties or effects providing scientific evidence of past or present technology, e.g., radio emissions or large-scale astroengineering [98].

**Tragedy of the commons** describes how individuals, acting in their own self-interest, will overuse and deplete a shared, limited resource (a "common") even when it is not in anyone's longer-term best interest, leading to collective ruin, exemplified by overfishing or deforestation. This may be exacerbated by *short-termism*.

**Unstable equilibrium** in game theory is a strategy profile where any slight player deviation pushes the system away from that point, rather than back towards it. For example, in the driving ("chicken" or the Hawk-Dove game) game, with two drivers moving in opposite directions and choosing to swerve either to the left or to the right of the road, the mixed-strategy equilibrium (1/2 probability of each action) is unstable. If a player changes their probabilities, while maintaining the expected payoff under the assumption that the other player's mixed strategy is still (1/2, 1/2), then the other player immediately has a better pure strategy at either (0, 1) or (1, 0). In the Rock-Paper-Scissors game, the mixed Nash equilibrium (1/3,1/3,1/3) is a neutrally stable center under idealized replicator dynamics. However, because it is not an evolutionarily stable strategy and not an attractor, arbitrarily small perturbations prevent convergence, leading populations to continually adjust their strategies and cycle indefinitely rather than settle at the equilibrium.

## References


[1] Anderson, JR, Comparative thanatology, Curr Biol 26 (2016) R553–R556.
[2] Archetti, M, Scheuring, I, Game theory of public goods in one-shot social dilemmas without assortment, J Theor Biol 299 (2012) 9–20.
[3] Arnon, I, Kirby, S, Allen, JA, Garrigue, C, Carroll, EL, Garland, EC, Whale song shows language-like statistical structure, Science 387 (2025) 649–53.
[4] Axelrod, R, Hamilton, WD, The evolution of cooperation, Science 211 (1981) 1390–6.
[5] Barbieri, M, The organic codes: an introduction to semantic biology, Cambridge University Press, (2003).
[6] Barbieri, M, Biosemiotics: a new understanding of life, Naturwissenschaften 95 (2008a) 577–99.
[7] Barbieri, M, The mechanisms of evolution: natural selection and natural conventions, in: The codes of life: the rules of macroevolution, Springer, (2008b), pp. 15–35.
[8] Barbieri, M, A general model on the origin of biological codes, Biosystems 181 (2019) 11–19.
[9] Barham, L, Systematic pigment use in the middle pleistocene of south-central Africa, Curr Anthropol 43 (2002) 181–90.
[10] Barron, AB, Halina, M, Klein, C, Transitions in cognitive evolution, Proc R Soc B 290 (2023) 20230671.
[11] Bostrom, N, The vulnerable world hypothesis, Glob Policy 10 (2019) 455–76.
[12] Boudreaux, DJ, Meiners, R, Externality: origins and classifications, Nat Resour J 59 (2019) 1–34.
[13] Boyd, R, Richerson, PJ, Culture and the evolutionary process, University of Chicago press, (1988).
[14] Call, J, Tomasello, M, Does the chimpanzee have a theory of mind? 30 years later, Trends Cogn Sci 12 (2008) 187–92.
[15] Carter, B, The anthropic principle and its implications for biological evolution, Philos Trans R Soc Lond A: Math Phys Sci 310 (1983) 347–63.
[16] Clark, A, Surfing uncertainty: Prediction, action, and the embodied mind, Oxford University Press, (2015).
[17] Clark, A, Chalmers, D, The extended mind, Analysis 58 (1998) 7–19.
[18] Clayton, NS, Bussey, TJ, Dickinson, A, Can animals recall the past and plan for the future?, Nat Rev Neurosci 4 (2003) 685–91.
[19] Clayton, NS, Dickinson, A, Episodic-like memory during cache recovery by scrub jays, Nature 395 (1998) 272–4.
[20] Conard, NJ, Palaeolithic ivory sculptures from southwestern Germany and the origins of figurative art, Nature 426 (2003) 830–2.
[21] Coolidge, FL, Wynn, T, Working memory, its executive functions, and the emergence of modern thinking, Camb Archaeol J 15 (2005) 5–26.
[22] Crandall, JW, Oudah, M, Tennom, Ishowo-Oloko, F, Abdallah, S, Bonnefon, JF, Cebrian, M, Shariff, A, Goodrich, MA, Rahwan, I, Cooperating with machines, Nat Commun 9 (2018) 233.
[23] Croft, W, Explaining language change: An evolutionary approach, Pearson Education, (2000).
[24] Croft, W, Evolutionary linguistics, Annu Rev Anthropol 37 (2008) 219–34.
[25] Daly, H, Steady state economy, San Franc 545 (1977).
[26] Davidson, EH, The regulatory genome, 785, Elsevier, (2006).
[27] Davies, PCW, The fifth miracle: The search for the origin and meaning of life, Simon and Schuster, (1999).
[28] Davies, PCW, The eerie silence, Phys World 23 (2010) 28.
[29] Dawkins, R, The selfish gene, (1976).
[30] De Duve, C, Singularities: landmarks on the pathways of life, Cambridge University Press, (2005).
[31] Deacon, TW, The symbolic species: The co-evolution of language and the brain, WW Norton & Company, (1998).
[32] Deacon, TW, Incomplete nature: How mind emerged from matter, WW Norton & Company, (2011).
[33] Donahue, K, Hauser, OP, Nowak, MA, Hilbe, C, Evolving cooperation in multichannel games, Nat Commun 11 (2020) 3885.
[34] Donald, M, Origins of the modern mind: Three stages in the evolution of culture and cognition, Harvard University press, (1993).
[35] Doolittle, WF, Bapteste, E, Pattern pluralism and the tree of life hypothesis, Proc Natl Acad Sci 104 (2007) 2043–9.

[36] Dunbar, RI, Neocortex size as a constraint on group size in primates, J Hum Evol 22 (1992) 469–93.
[37] Dyer, FC, The biology of the dance language, Annu Rev Entomol 47 (2002) 917–49.
[38] Dyson, F, Origins of life, Cambridge University Press, (1985).
[39] Eigen, M, Selforganization of matter and the evolution of biological macromolecules, Naturwissenschaften 58 (1971) 465–523.
[40] Eigen, M, McCaskill, J, Schuster, P, The molecular quasi-species, Adv Chem Phys 75 (1989) 149–263.
[41] Ellis, GF, Top-down causation and emergence: some comments on mechanisms, Interface Focus 2 (2012) 126–40.
[42] Engesser, S, Ridley, AR, Watson, SK, Kita, S, Townsend, SW, Seeds of language-like generativity in bird call combinations, Proc R Soc B: Biol Sci 291 (2024) 20240922.
[43] Evans, J, Bratton, B, Agüera y Arcas, B, Agentic AI and the next intelligence explosion, (2026).
[44] Faltỳnek, D, Matlach, V, Lacková, L, Bases are not letters: on the analogy between the genetic code and natural language by sequence analysis, Biosemiotics 12 (2019) 289–304.
[45] Fang, C, Kimbrough, SO, Pace, S, Valluri, A, Zheng, Z, On adaptive emergence of trust behavior in the game of stag hunt, Group Decis Negot 11 (2002) 449–67.
[46] Ferrer-i-Cancho, R, The variation of Zipf's law in human language, Eur Phys J B-Condens Matter Complex Syst 44 (2005) 249–57.
[47] Ferrer-i-Cancho, R, On the class of coding optimality of human languages and the origins of Zipf's law, Europhys Lett 151 (2025) 62001.
[48] Ferrer-i-Cancho, R, Díaz-Guilera, A, The global minima of the communicative energy of natural communication systems, J Stat Mech: Theory Exp 2007 (2007) P06009.
[49] Ferrer-i-Cancho, R, Riordan, O, Bollobás, B, The consequences of Zipf's law for syntax and symbolic reference, Proc R Soc B: Biol Sci 272 (2005) 561.
[50] Ferrer-i-Cancho, R, Solé, RV, Least effort and the origins of scaling in human language, Proc Natl Acad Sci 100 (2003) 788–91.
[51] Fisher, R, The perils of short-termism: Civilisation's greatest threat, BBC Future, (2019), https://www.bbc.com/future/article/20190109-the-perils-of-short-termism-civilisations-greatest-threat, [Online; accessed 11-January-2026].
[52] Fisher, R, The long view: Why we need to transform how the world sees time, Hachette UK, (2023).
[53] Fitch, WT, The evolution of language, Cambridge University Press, (2010).
[54] Flack, JC, Coarse-graining as a downward causation mechanism, Philos Trans R Soc A 375 (2017) 20160338.
[55] Fletcher, RJ, The evolution of human behavior, in: Burenhult, G (Ed.), The first humans: human origins and history to 10,000 BCE, Harper, (1993), pp. 47–51.
[56] Fletcher, RJ, The limits of settlement growth: A theoretical outline, Cambridge University Press, (1995).
[57] Frisch, Kv, The dance language and orientation of bees, Harvard University Press, (1993).
[58] Geenen, V, The thymus and the science of self, Semin Immunopathol 43 (2021) 5–14.
[59] Gianni, E, Kwok, SLY, Wan, CJK, Goeij, K, Clifton, BE, Colizzi, ES, Attwater, J, Holliger, P, A small polymerase ribozyme that can synthesize itself and its complementary strand, Science 0 (2026) eadt2760, https://doi.org/10.1126/science.adt2760.
[60] Gibbons, A, Complex behavior arose at dawn of humans, Science 359 (2018) 1200–1.
[61] Gilbert, W, Origin of life: the RNA world, Nature 319 (1986) 618.
[62] Gogarten, JP, Doolittle, WF, Lawrence, JG, Prokaryotic evolution in light of gene transfer, Mol Biol Evol 19 (2002) 2226–38.
[63] Goody, J, The logic of writing and the organization of society, Cambridge University Press, (1986).
[64] Grush, R, The emulation theory of representation: motor control, imagery, and perception, Behav Brain Sci 27 (2004) 377–96.
[65] Hanson, R, The Great Filter — Are we almost past it?, (1998), https://mason.gmu.edu/~rhanson/greatfilter.HTML, [Online; accessed 15-Sept-2025].
[66] Harari, YN, Yuval Noah Harari argues that AI has hacked the operating system of human civilisation, Economist 28 (2023).
[67] Harré, MS, Drysdale, C, Ruiz-Serra, J, An AI theory of mind will enhance our collective intelligence, (2024). arXiv:2411.09168
[68] Harré, MS, Prokopenko, M, The social brain: scale-invariant layering of Erdős–Rényi networks in small-scale human societies, J R Soc Interface 13 (2016) 20160044.
[69] Hauert, C, Holmes, M, Doebeli, M, Evolutionary games and population dynamics: maintenance of cooperation in public goods games, Proc R Soc B: Biol Sci 273 (2006) 2565–71.
[70] Hauert, C, Wakano, JY, Doebeli, M, Ecological public goods games: cooperation and bifurcation, Theor Popul Biol 73 (2008) 257–63.
[71] Hauser, MD, Chomsky, N, Fitch, WT, The faculty of language: what is it, who has it, and how did it evolve?, Science 298 (2002) 1569–79.
[72] Heinsalu, E, Patriarca, M, Sanchez, D, Language dynamics: complex systems approaches, Int Encycl Lang Linguist 14 (2026) 241–7.
[73] Henshilwood, CS, d'Errico, F, Marean, CW, Milo, RG, Yates, R, An early bone tool industry from the Middle Stone Age at Blombos Cave, South Africa: implications for the origins of modern human behaviour, symbolism and language, J Hum Evol 41 (2001) 631–78.
[74] Henshilwood, CS, Dubreuil, B, Reading the artifacts: gleaning language skills from the Middle Stone Age in southern Africa, in: Botha, R, Knight, C (Eds.), The cradle of language, Oxford University Press, (2009), pp. 41–61.
[75] Hernández-Orozco, S, Hernández-Quiroz, F, Zenil, H, Undecidability and irreducibility conditions for open-ended evolution and emergence, Artif Life 24 (2018) 56–70.
[76] Heyes, C, Cognitive gadgets: The cultural evolution of thinking, Harvard University Press, (2018).
[77] Hoel, EP, When the map is better than the territory, Entropy 19 (2017) 188.
[78] Hoel, EP, Quantifying emergent complexity, Patterns 7 (2026).
[79] Hoffmeyer, J, Signs of Meaning in the Universe, Indiana University Press, (1997).
[80] Hofstadter, DR, Gödel, Escher, Bach: An Eternal Golden Braid, Harmondsworth, Penguin, (1980).
[81] Hopfield, JJ, Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity, Proc Natl Acad Sci 71 (1974) 4135–9.
[82] Huttegger, S, Skyrms, B, Tarres, P, Wagner, E, Some dynamics of signaling games, Proc Natl Acad Sci 111 (2014) 10873–80.
[83] Igamberdiev, AU, The drawbridge of nature: evolutionary complexification as a generation and novel interpretation of coding systems, Biosystems 207 (2021) 104454.
[84] Jablonka, E, Inheritance systems and the evolution of new levels of individuality, J Theor Biol 170 (1994) 301–9.
[85] Jackendoff, R, Possible stages in the evolution of the language capacity, Trends Cogn Sci 3 (1999) 272–9.
[86] Jakobson, R, Selected writings [of] Roman Jakobson: Word and language, Mouton, (1971).
[87] Joyce, GF, The antiquity of rna-based evolution, Nature 418 (2002) 214–21.
[88] Kallis, G, Hickel, J, O'Neill, DW, Jackson, T, Victor, PA, Raworth, K, Schor, JB, Steinberger, JK, Ürge-Vorsatz, D, Post-growth: the science of wellbeing within planetary boundaries, Lancet Planet Health 9 (2025) e62–e78.
[89] Kauffman, SA, Autocatalytic sets of proteins, J Theor Biol 119 (1986) 1–24.
[90] Kauffman, SA, The origins of order: Self-organization and selection in evolution, Oxford University Press, (1993).
[91] Keyßer, LT, Lenzen, M, 1.5 C degrowth scenarios suggest the need for new mitigation pathways, Nat Commun 12 (2021) 2676.
[92] Khersonsky, O, Roodveldt, C, Tawfik, DS, Enzyme promiscuity: evolutionary and mechanistic aspects, Curr Opin Chem Biol 10 (2006) 498–508.
[93] Kim, J, Lai, S, Scherrer, N, Evans, J, et al., Reasoning models generate societies of thought,(2026). arXiv:2601.10825
[94] Kun, Á, The major evolutionary transitions and codes of life, Biosystems 210 (2021) 104548.
[95] Lee, H, Calvin, K, Dasgupta, D, Krinner, G, Mukherji, A, Thorne, P, Trisos, C, Romero, J, Aldunce, P, Barret, K, et al., Climate change 2023: Synthesis report, summary for policymakers. Contribution of working groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core writing team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland., Technical Report, Intergovernmental Panel on Climate Change (IPCC), (2023).
[96] Lewis, D, Convention: A philosophical study, John Wiley & Sons, (2008).
[97] Lingam, M, Loeb, A, Role of stellar physics in regulating the critical steps for life, Int J Astrobiol 18 (2019) 527–46.
[98] Lingam, M, Loeb, A, Life in the cosmos: From biosignatures to technosignatures, Harvard University Press, (2021).

[99] Malik, A, Lenzen, M, Li, M, Mora, C, Carter, S, Giljum, S, Lutter, S, Gómez-Paredes, J, Polarizing and equalizing trends in international trade and sustainable development goals, Nat Sustain 7 (2024) 1359–70.
[100] Mandelbrot, B, An informational theory of the statistical structure of language, Commun Theory 84 (1953) 486–502.
[101] Manheim, D, The fragile world hypothesis: complexity, fragility, and systemic existential risk, Futures 122 (2020) 102570.
[102] Marion, JY, From turing machines to computer viruses, Philos Trans R Soc A: Math Phys Eng Sci 370 (2012) 3319–39.
[103] Markose, SM, Novelty in complex adaptive systems (CAS) dynamics: a computational theory of actor innovation, Phys A: Stat Mech Appl 344 (2004) 41–9.
[104] Markose, SM, Complex type 4 structure changing dynamics of digital agents: Nash equilibria of a game with arms race in innovations, J Dyn Games 4 (2017) 255–84.
[105] Markose, SM, Genomic intelligence as über bio-cybersecurity: the Gödel sentence in immuno-cognitive systems, Entropy 23 (2021) 405.
[106] Massey, SE, Mishra, B, Origin of biomolecular games: deception and molecular evolution, J R Soc Interface 15 (2018).
[107] Mills, DB, Macalady, JL, Frank, A, Wright, JT, A reassessment of the "hard-steps" model for the evolution of intelligent life, Sci Adv 11 (2025) eads5698.
[108] Mufwene, SS, Language evolution: Contact, competition and change, Bloomsbury Publishing, (2008).
[109] Myers, PVN, Ancient history, Ginn, (1904).
[110] National Human Genome Research Institute, Genetic code, (2026), https://www.genome.gov/genetics-glossary/Genetic-Code, [Online; accessed 5-Aug-2026].
[111] Obst, O, Polani, D, Prokopenko, M, Origins of scaling in genetic code, in: Advances in artificial life. Darwin meets von Neumann. Proceedings, 10th European conference, ECAL 2009, Budapest, Hungary, September 13–16, 2009, Springer, (2011), pp. 85–93.
[112] Ong, WJ, Orality and literacy, Routledge, (2003).
[113] Pattee, HH, The physics of symbols: bridging the epistemic cut, Biosystems 60 (2001) 5–21.
[114] Pattee, HH, How does a molecule become a message?, in: Laws, language and life: Howard Pattee's classic papers on the physics of symbols with contemporary commentary, Springer, (2012), pp. 55–67.
[115] Penrose, LS, Penrose, R, Impossible objects: a special type of visual illusion., Br J Psychol 49 (1958) 31–3.
[116] Pettitt, P, The Palaeolithic origins of human burial, Routledge, (2010).
[117] Phaniraj, N, Wierucka, K, Burkart, JM, Dynamic vocal learning in adult marmoset monkeys, Proc R Soc B: Biol Sci 293 (2026) 20260743.
[118] Pinotsis, DA, Miller, EK, Ephaptic coupling can explain variability in neural activity, Cereb Cortex 36 (2026) bhag098.
[119] Post, EL, Recursively enumerable sets of positive integers and their decision problems, Bull Am Math Soc 50 (1944) 284–316.
[120] Prokopenko, M, Ay, N, Obst, O, Polani, D, Phase transitions in least-effort communications, J Stat Mech: Theory Exp 2010 (2010) P11025.
[121] Prokopenko, M, Davies, PC, Harré, M, Heisler, MG, Kuncic, Z, Lewis, GF, Livson, O, Lizier, JT, Rosas, FE, Biological arrow of time: emergence of tangled information hierarchies and self-modelling dynamics, J Phys: Complex 6 (2025) 015006.
[122] Prokopenko, M, Harré, M, Lizier, J, Boschetti, F, Peppas, P, Kauffman, S, Self-referential basis of undecidable dynamics: from the liar paradox and the halting problem to the edge of chaos, Phys Life Rev 31 (2019) 134–56.
[123] Prokopenko, M, Polani, D, Chadwick, M, Stigmergic gene transfer and emergence of universal coding, HFSP J 3 (2009) 317–27.
[124] Rappaport, RA, Ritual and Religion in the Making of Humanity, 110, Cambridge University Press, (1999).
[125] Richerson, PJ, Boyd, R, The pleistocene and the origins of human culture: built for speed, Perspect Ethol 13 (2000) 1–45.
[126] Rockström, J, Donges, JF, Fetzer, I, Martin, MA, Wang-Erlandsson, L, Richardson, K, Planetary boundaries guide humanity's future on Earth, Nat Rev Earth Environ 5 (2024) 773–88.
[127] Rodin, S, Rodin, A, On the origin of the genetic code: signatures of its primordial complementarity in trnas and aminoacyl-trna synthetases, Heredity 100 (2008) 341–55.
[128] Rosen, R, Life itself: A comprehensive inquiry into the nature, origin, and fabrication of life, Columbia University Press, (1991).
[129] Sagan, C, Direct contact among galactic civilizations by relativistic interstellar spaceflight, Planet Space Sci 11 (1963) 485–98.
[130] Salge, C, Ay, N, Polani, D, Prokopenko, M, Zipf's law: balancing signal usage cost and communication efficiency, Plos One 10 (2015) e0139475.
[131] San Miguel, M, Toral, R, Individual based models of language competition, Int Encycl Lang Linguist 14 (2026) 248–57.
[132] Sandholm, WH, Negative externalities and evolutionary implementation, Rev Econ Stud 72 (2005) 885–915.
[133] Sayama, H, Construction theory, self-replication, and the halting problem, Complexity 13 (2008) 16–22.
[134] Scherjon, F, Bakels, C, MacDonald, K, Roebroeks, W, Burning the land: an ethnographic study of off-site fire use by current and historically documented foragers and implications for the interpretation of past fire practices in the landscape, Curr Anthropol 56 (2015) 299–326.
[135] Searls, DB, The language of genes, Nature 420 (2002) 211–17.
[136] Sharma, P, Gero, S, Payne, R, Gruber, DF, Rus, D, Torralba, A, Andreas, J, Contextual and combinatorial structure in sperm whale vocalisations, Nat Commun 15 (2024) 3617.
[137] Shklovskii, IS, Sagan, C, Intelligent life in the Universe, Holden-Day, (1966).
[138] Smith, JM, The concept of information in biology, Philos Sci 67 (2000) 177–94.
[139] Soare, RI, Turing oracle machines, online computing, and three displacements in computability theory, Ann Pure Appl Log 160 (2009) 368–99.
[140] Suddendorf, T, Corballis, MC, The evolution of foresight: what is mental time travel, and is it unique to humans?, Behav Brain Sci 30 (2007) 299–313.
[141] Svahn, AJ, Prokopenko, M, An ansatz for computational undecidability in RNA automata, Artif Life 29 (2023) 261–88.
[142] Szathmáry, E, Toward major evolutionary transitions theory 2.0, Proc Natl Acad Sci 112 (2015) 10104–11.
[143] Szathmáry, E, Maynard Smith, J, The major evolutionary transitions, Nature 374 (1995) 227–32.
[144] Szolnoki, A, Perc, M, Conditional strategies and the evolution of cooperation in spatial public goods games, Phys Rev E 85 (2012) 026104.
[145] Szostak, JW, Bartel, DP, Luisi, PL, Synthesizing life, Nature 409 (2001) 387–90.
[146] Tang, Y, Nguyen, D, Ha, D, Neuroevolution of self-interpretable agents, in: Proceedings of the 2020 genetic and evolutionary computation conference, (2020), pp. 414–24.
[147] Tomasello, M, Carpenter, M, Call, J, Behne, T, Moll, H, Understanding and sharing intentions: the origins of cultural cognition, Behav Brain Sci 28 (2005) 675–91.
[148] Trappes, R, Nematipour, B, Kaiser, MI, Krohs, U, Van Benthem, KJ, Ernst, UR, Gadau, J, Korsten, P, Kurtz, J, Schielzeth, H, et al., How individualized niches arise: defining mechanisms of niche construction, niche choice, and niche conformance, BioScience 72 (2022) 538–48.
[149] Tria, F, et al., Language games from a physics perspective, in: International encyclopedia of language and linguistics, Elsevier, (2026), pp. 1–7.
[150] Turing, AM, Systems of logic based on ordinals, Proc Lond Math Soc 2 45 (1939) 161–228.
[151] Vansina, JM, Oral tradition as history, University of Wisconsin Press, (1985).
[152] Vetsigian, K, Woese, C, Goldenfeld, N, Collective evolution and the genetic code, Proc Natl Acad Sci 103 (2006) 10696–701.
[153] Wächtershäuser, G, Before enzymes and templates: theory of surface metabolism, Microbiol Rev 52 (1988) 452–84.
[154] Wakano, JY, Nowak, MA, Hauert, C, Spatial dynamics of ecological public goods, Proc Natl Acad Sci 106 (2009) 7910–14.
[155] Walker, SI, Davies, PCW, The algorithmic origins of life, J R Soc Interface 10 (2013) 20120869.
[156] West, SA, Fisher, RM, Gardner, A, Kiers, ET, Major evolutionary transitions in individuality, Proc Natl Acad Sci 112 (2015) 10112–19.
[157] Wiedmann, T, Lenzen, M, Environmental and social footprints of international trade, Nat Geosci 11 (2018) 314–21.
[158] Wikipedia contributors, Drawing hands — Wikipedia, the free encyclopedia, (2024a), https://en.wikipedia.org/wiki/Drawing_Hands, [Online; accessed 21-June-2024].
[159] Wikipedia contributors, Penrose stairs — Wikipedia, the free encyclopedia, (2024b), https://en.wikipedia.org/wiki/Penrose_stairs, [Online; accessed 21-June-2024].
[160] Wikipedia contributors, Short-termism — Wikipedia, the free encyclopedia, (2026), https://en.wikipedia.org/wiki/Short-termism, [Online; accessed 11-January-2026].
[161] Woese, CR, On the evolution of the genetic code., Proc Natl Acad Sci 54 (1965) 1546–52.

[162] Woese, CR, The emergence of genetic organization, in: Ponnamperuma, C (Ed.), Exobiology, North-Holland Publishing Co., Amsterdam, (1972), pp. 301–41.
[163] Woese, CR, Interpreting the universal phylogenetic tree, Proc Natl Acad Sci 97 (2000) 8392–6.
[164] Woese, CR, A new biology for a new century, Microbiol Mol Biol Rev 68 (2004) 173–86.
[165] Yanofsky, NS, A universal approach to self-referential paradoxes, incompleteness and fixed points, Bull Symb Log 9 (2003) 362–86.
[166] Yockey, HP, Information theory and molecular biology, Cambridge University Press, (1992).
[167] Zahavi, A, Mate selection—a selection for a handicap, J Theor Biol 53 (1975) 205–14.