Biological arrow of time: Emergence of tangled information hierarchies and self-modelling dynamics

Mikhail Prokopenko[1,2,*], Paul C. W. Davies[3], Michael Harré[1,2], Marcus Heisler[1,4], Zdenka Kuncic[1,5,6], Geraint F. Lewis[5], Ori Livson[1,2], Joseph T. Lizier[1,2], Fernando E. Rosas[7,8,9,10]

[1] *The Centre for Complex Systems, University of Sydney, Sydney, NSW 2006, Australia*
[2] *School of Computer Science, Faculty of Engineering, University of Sydney, Sydney, NSW 2006, Australia*
[3] *The Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, 85287–0506, USA*
[4] *School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, NSW 2006, Australia*
[5] *School of Physics, Faculty of Science, University of Sydney, Sydney, NSW 2006, Australia*
[6] *The Charles Perkins Centre, University of Sydney, Sydney, NSW 2006, Australia*
[7] *Sussex AI and Centre for Consciousness Science, Department of Informatics, University of Sussex, Brighton, BN19RH, UK*
[8] *Centre for Psychedelic Research, Department of Brain Science, Imperial College London, London, SW72AZ, UK*
[9] *Centre for Complexity Science, Imperial College London, London, SW72AZ, UK*
[10] *Center for Eudaimonia and Human Flourishing, University of Oxford, Oxford, OX39BX, UK*
[*] *Corresponding author: mikhail.prokopenko@sydney.edu.au*

Abstract

We study open-ended evolution by focusing on computational and information-processing dynamics underlying major evolutionary transitions. In doing so, we consider biological organisms as hierarchical dynamical systems that generate regularities in their phase-spaces through interactions with their environment. These emergent information patterns can then be encoded within the organism's components, leading to self-modelling "tangled hierarchies". Our main conjecture is that when macro-scale patterns are encoded within micro-scale components, it creates fundamental tensions (computational inconsistencies) between what is encodable at a particular evolutionary stage and what is potentially realisable in the environment. A resolution of these tensions triggers an evolutionary transition which expands the problem-space, at the cost of generating new tensions in the expanded space, in a continual process. We argue that biological complexification can be interpreted computation-theoretically, within the Gödel–Turing–Post recursion-theoretic framework, as open-ended generation of computational novelty. In general, this process can be viewed as a meta-simulation performed by higher-order systems that successively simulate the computation carried out by lower-order systems. This computation-theoretic argument provides a basis for hypothesising the biological arrow of time.

Keywords: tangled hierarchy, self-reference, undecidability, open-ended complexity, evolutionary transition

Contents

1. Introduction

One of the longstanding challenges in biology is the formulation of fundamental principles that can explain why biological complexity is seemingly increasing over time [2, 60]. At the microscopic level, intricate networks of molecular interactions and cellular processes evolve and adapt over time [70]. At the mesoscopic level, organisms appear to continually generate novel forms, behaviours, learning strategies and organisational structures [71]. And at the macroscopic level, ecosystems develop in ways that support complex interactions among species, leading to the emergence of new niches and ecological relationships that, in turn, can drive further evolutionary changes [26].

The interplay of genetic, environmental and ecological factors generates never-ending biological innovation and rich diversity. In general, this seemingly unbounded increase in complexity that characterises evolution on multiple scales has been described as *open-ended evolution* [23]. Goldenfeld and Woese [38] highlighted a key question which makes understanding the open-ended biological complexity so difficult:

> "These rules themselves need to evolve, but how? We need an additional set of rules describing the evolution of the original rules. But this upper level of rules itself needs to evolve. Thus, we end up with an infinite hierarchy, an inevitable reflection of the fact that the dynamic we are seeking is inherently self-referential."

Several approaches to this challenge have been proposed [38, 86, 50, 128, 41, 94, 126] (see Section 9). However, developing a clear and comprehensive interpretation of how self-referential dynamics persist and perpetuate across multiple scales remains elusive, as various proposals are scattered across disciplines and lack a unified formal framework.

In this study we address this problem by exploring computational and information-processing dynamics underlying biological complexification. Our aim is to consider open-ended evolution as a dynamic computational process which successively expands its own problem-space. In doing so we interpret major evolutionary transitions (origin of life, formation of eukaryotic cells, emergence of multicellularity, etc.) [100, 116] through the lens of computation theory, aiming to provide a unifying intuitive perspective and improve inter-disciplinary communications.

The core of our argument is as follows. We consider biological organisms as dynamical systems which may generate regularities in their phase-spaces through their interactions within the environment. In an abstract sense,

these emergent regularities can be interpreted as (higher-level) information patterns which may influence the (lower-level) organisms via downward causation These loops of causation between higher and lower levels are known as tangled hierarchies [44]. We hypothesise that these tangled hierarchies can nurture self-modelling capabilities which improve the efficiency of organisms' replication. In other words, self-modelling would allow the organisms to capture compressed representations of the emergent information patterns, by utilising a suitable encoding. However, once such an encoding is adopted, the tangled hierarchies enhanced with self-modelling inevitably generate tensions (inconsistencies) between what is encodable within the current setup and what is possible, that is, realisable in the current environment. Informally, our main argument is that an evolutionary transition resolves these tensions by expanding the problem-space, i.e., by generating a new way to encode extended information patterns.

We begin by reviewing several perspectives on major evolutionary transitions (Section 2). This is followed by a computation-theoretic argument for the increasing complexity and open-ended evolution developed within the Gödel–Turing–Post recursion-theoretic framework. This framework formalises the construction of extensible computational systems, such as Turing α-oracle machines, and ordinal or recursively generated logics. We argue that this continual process can be interpreted as open-ended meta-simulation which constructs new problem-spaces by resolving computational inconsistencies (Section 3). We conclude presenting our background with reviewing cross-disciplinary insights developed in dynamical systems theory (Section 4), as well as complex systems, systems biology, artificial life and machine learning (Section 5).

Having examined the background studies, we propose a distinction between two types of tangled information hierarchies: with and without self-modelling capability (Section 6). This distinction allows us to draw a parallel between the open-ended meta-simulation which creates computational novelty and the continual evolutionary process which discovers new phase-spaces along evolutionary transitions in individuality (Section 7). We then clarify the role of self-reference and fundamental undecidability in forming *the biological arrow of time* (Section 8), and conclude this perspective by comparing our argument with other fundamental principles proposed to explain biological complexification and the open-ended evolution (Section 9).

2. Open-ended biological complexity and evolutionary transitions

In this section we discuss various views on biological complexification. This process, observed in evolutionary dynamics over time, is punctuated by major evolutionary transitions and "coding thresholds". In reviewing established approaches, we highlight key challenges and common features shared by these perspectives.

2.1. Major evolutionary transitions

The Second law of thermodynamics states the impossibility of fully transforming disordered heat energy into coherent work without the expense of some additional resource. Importantly, the Second law implies that physical systems naturally decay towards less structured arrangements. When seen against this background, the spectacular evolution of living systems on Earth is even more remarkable, requiring explanations that are compatible with these limiting physical principles.

When explaining the progress of biological evolution, two main perspectives have been adopted: approaches that emphasise that evolution happens slowly and continuously, and approaches that state that evolution is mainly driven by abrupt transitions [30, 91]. Within the second family of approaches, it has been argued that evolution displays major transitions in terms of biological complexity [100]. In general, these transitions produce important changes in the way organisms exchange information, cooperate and coordinate with each other, and how they survive and replicate. These collectives can become so tightly arranged — via functional specialisation, which enhances the efficacy of cooperation while inducing mutual dependency [116] — that they start acting as 'higher-order' units that effectively drive the selection process [88]. In effect, by tying each gene's replication to the survival of higher-order structures, selection favours genes that promote survival of the higher-order structure, which in turn enables greater division of labour and specialisation.

When trying to characterise what these major evolutionary transitions have in common, one can see that they involve profound changes in what an individual is and how it preserves its properties to new generations. This is usually supported by the emergence of novel inheritance systems, involving new modes of storing, transmitting, and processing information — sometimes referred to as new "codes of life" [3, 57]. Examples of this include replicating molecules which form cells as compartmentalised "populations" of molecules; independent replicators (hypothesised to be RNA) which may have formed chromosomes to reduce information loss during replication; emergence of the genetic code and translation machinery (with DNA used as genes and proteins as enzymes); prokaryotes evolving into eukaryotes with a membrane-bound nucleus which stores their genetic information, and so on, towards the evolution of multicellularity and eusociality, as well as language and sociocultural evolution [100, 116, 57].

In this way, natural evolution is thought to give rise to more elaborate arrangements of higher-order organisms, from eukaryotic cells to multicellular organisms and even collectives such as swarms and hives, supported by novel information-processing modes. However, there are more "codes of life" than major evolutionary transitions [57], and several crucial questions remain unanswered: Are there computational principles that drive these evolutionary transitions? What are the computational trade-offs that are being negotiated during a transition?

Box 1: Genetic information and replication

Gene transfer
Transmission of genes from one generation to the next, i.e., from parent to offspring (as a result of sexual or asexual reproduction), is referred to as vertical gene transfer (VGT) [13]. Horizontal gene transfer (HGT) moves partial genetic information laterally across distantly related organisms in the same generation [51].

Genotype and phenotype
The genotype contains the specific genetic information an organism carries in its DNA, encoded in sequences of nucleotides. The phenotype comprises the set of observable traits (physical, biochemical and behavioural) of the organism. Genetic instructions are used in development and functioning of a living organism: they are involved in construction of other components and copying itself.

Extended phenotype
The extended phenotype encompasses all the effects that a genotype can have on the environment beyond the immediate organism itself. This includes modifications or structures an organism creates that affect its survival and reproduction, influencing other organisms in the ecosystem [26]. Extended phenotype includes (i) behavioural innovations, e.g., a bird's ability to construct complex nests; (ii) interactions with environment and other organisms, indirectly influenced by traits encoded in the genome, e.g., pheromone production in insects affects their mating, foraging and social behaviours; (iii) physical modifications of the organism's environment, e.g., beaver dams, spider webs, or ant pheromone trails arise from behaviours guided by genetic predispositions and environmental factors.

Gene expression
Transcription and translation are two fundamental processes of gene expression through which the information encoded in DNA is used to produce functional proteins that contribute to an organism's phenotype. Transcription is the process of making RNA copies of the genetic protein information encoded in DNA, and translation is the decoding of instructions for making proteins.

2.2. Coding thresholds and phase space expansions

It has been argued that major evolutionary transitions overcome specific "coding thresholds", with the saltations opening a novel evolutionary phase-space with qualitatively new possibilities to handle information [123, 57]. For example, when discussing the emergence of DNA, Woese 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."[123]

Furthermore, Woese [123] suggested that another such saltation — which he described as "Darwinian threshold" — may have been crossed when cells went from an initial arrangement where their evolutionary dynamics were dominated by horizontal gene transfer (HGT) to a new arrangement dominated by vertical inheritance. This turned evolution from a communal process (in which evolving cells maintained no stable genealogical information) to one where stable organismal lineages could develop, preserve their genetic makeup, and coexist [123] (see subsection 2.3).

Analogously, a "language threshold" may have been crossed when early human groups developed language (another "'codes of life" [3, 57]), which enabled a new way to inherit knowledge and significantly speed-up adaptation. Woese [123] elaborated on this conjecture, noting that one common feature of higher-order arrangements — such as multicellularity or human language — is their enhanced communication ability, which enables a kind of 'interaction at a distance'. This insight reinforces the idea that higher-order structures are able to synergistically access an extended problem-solving space, making such structures non reducible to a mere aggregation of information-processing subunits.

Unfortunately, however, the investigation of major evolutionary transitions is hindered by a chicken-or-egg causal dilemma, highlighted by various error threshold paradoxes (e.g., Eigen paradox [29]), as it is often difficult to explain how a more complex structure can evolve if the evolutionary benefits can be realised only at the higher level, This challenge was eloquently expressed in 1904 by a pioneer of genetics Hugo de Vries: "natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest", and remains a subject of active research. For example, Wagner [109] explored how innovation and adaptability drive evolutionary success, arguing that the ability of organisms to generate new traits and capabilities is crucial for their long-term survival and thriving, and highlighting that constraints, rather than being purely limiting, can actually drive innovation.

An intriguing conjecture implicated genetic parasites, e.g., plasmids or viruses, as major catalysts for evolutionary transitions. Such parasites may (i) necessitate the formation of more complex structures helping the host to combat the parasitic threats and win in the parasite-host arms race, while (ii) providing building blocks — via a parasite-enabled gene transfer — for constructing the higher levels of organisation:

> "only increasing organizational complexity, e.g. emergence of multicellular aggregates, can prevent the collapse of the host–parasite system under the pressure of parasites" [55].

In an abstract sense, this hypothesis suggests that the underlying inconsistencies and tensions may provide a dialectic impetus for evolutionary innovations, forcing the construction of new phase-spaces. Nevertheless, formulation of general computation-theoretic principles underpinning an open-ended phase-space expansion remains elusive.

2.3. Innovation-sharing at the coding threshold

Vetsigian et al. [108] considered evolution of the genetic code during early life, that is, during the stages preceding the emergence of vertical genealogical descent. They argued that early evolutionary dynamics of cell-like entities involved communal descent mediated by HGT.

A key insight of their analysis is that the genetic code emerging in this communal world was not only needed to encode amino acid sequences in the genome, but also served as an *innovation-sharing protocol*. They conjectured that "early life did not require a refined level of tolerance", and hence, the emerging innovation-sharing protocol allowed for imprecise copies created by an ambiguous translation without a unique mapping between codons and amino acids. Nevertheless, this dynamic produced a (nearly) universally common mechanism for encoding information, the *universal genetic code*:

> "HGT of protein coding regions and HGT of translational components ensures the emergence of clusters of similar codes and compatible translational machineries. Different clusters compete for niches, and because of the benefits of the communal evolution, the only stable solution of the cluster dynamics is universality" [108].

The study concluded that once the universal genetic code emerged, the genetic complexity is likely to grow exponentially. This, in turn, leads — via another (Darwinian) transition — to dominance of the vertical (individual) descent, and hence the Darwinian evolution [108].

2.4. Division of labour increases fitness and the dimensionality of phenotype space

Generally, biological fitness consists of two components – fecundity (reproduction) and viability (survival), and trade-offs between the two attributes shape diverse life-histories. In the context of clonal multicellularity, or a colony of genetically related cells, such trade-offs provide a selective pressure towards cell specialisation since group fitness can potentially be greater than the average of individual cells [72]. For example, a microtubule organising centre (MTOC) can typically be utilised to either support flagella for motility or cell division, but not both simultaneously. Hence, single-celled organisms typically separate reproduction and motility phases temporally. In multicellular Volvox this trade-off is thought to have led to the evolution of a spatial distinction between two groups of cells, corresponding to a germ/soma separation [73], and a similar scenario may have occurred during animal evolution [53].

Box 2: Inclusive fitness

Darwinian fitness
Darwinian fitness is a quantitative measure of reproductive success. More specifically it measures the average contribution to the gene pool of the next generation from an individual, classified by genotype. It depends on, for a given environment, the relative probability of survival and rate of reproduction.

Inclusive fitness
Inclusive fitness is similar to Darwinian fitness but is helpful in explaining social interactions since it explicitly recognises that an individual can influence their genetic contribution to the next generation via their influence on individuals other than themselves. It takes into account the influence of such interactions on both the individual and the individual being influenced as well as the degree of shared genetic information between the two.

Multicellular organisms
Multicellular organisms consist of more than one cell. Most multicellular organisms develop clonally, that is from the division of a single cell. This in turn means that most cells in such organisms are genetically identical (exceptions are caused by mutations during development or by genetic recombination in the case of germ cells). Less commonly, multicellular organisms form through aggregation, where single cells may come together to form the organism in response to environmental cues. In this case the cells involved are usually more genetically heterogeneous.

There are many other trade-offs beyond the general categories of survival vs reproduction that are relevant to a division of labour. For example, oxygen sensitive processes such as nitrogen fixation [33] are separated spatially from oxygenic photosynthesis in filamentous cyanobacteria while the exchange of metabolites between cells benefits the whole [129]. Considering extant complex organisms and their many distinct cell types, the concept of division of labour can obviously be extended to a great variety of cellular functions. In general, a division of labour within multicellular organisms can be viewed as an increase in the dimensionality of phenotype space, since spatial separation enables otherwise incompatible processes to occur:

> "Mathematically, the evolutionary advantage of the division of labour in aggregate forms can be viewed as the emergence of new, higher fitness maxima when the dimensionality of phenotype space is increased. The new fitness maxima are not a direct consequence of aggregation, but are based on the interaction between aggregated individuals that engage in the division of labour." [47].

Thus, given complementary cells states exist even in single-cell colonial communities, a division of labour may have provided an advantage to cell aggregates during the earliest stages of multicellular evolution [47, 102].

2.5. Higher-level agents operate with extended patterns of information

A higher-level aggregate organisation emerging out of interactions among the lower-level components may influence its constituent parts within a complex feedback loop [43]. McMillen and Levin [71] investigated the competition between collective and individual biological levels, in which "the behavior of subunits percolates up toward adaptive processes at higher levels", while "higher levels of organization constrain and facilitate the behavior of their parts" [71]. In general, the notion of a biological individual, the question of group selection and the nature of competition between collective and individual levels should be interpreted from a temporal, "diachronic", perspective [76]. This view allows us to consider intermediate stages in evolutionary transitions which create the *potential* for "conflict between levels of selection, for selection between the smaller units may disrupt the well-being of the collective" [76].

In adopting this approach, McMillen and Levin [71] argued that one of the advantages of multiscale organisation is the capacity of collective agents to create a novel problem-space and modify the energy landscape available at the higher level, which allows the lower-level components to operate more efficiently [71]. In this interpretation an energy function represents an optimisation problem — an approach which traces back to the study of neural circuitry by Hopfield and Tank [46]. This seminal work demonstrated that interconnected networks of analog neurons can be computationally effective in solving hard optimisation problems, with the optimal solution corresponding to the lowest energy state of the network's energy function (defined in terms of neuronal voltages), i.e., its most stable state. In general, as has been pointed out by Watson et al. [114],

> "...in systems built out of the superposition of many low-order constraints, low-energy (high-utility) attractors necessarily have large basins of attraction... So, the better the attractor, the more it is visited, thus the more it is enlarged by learning, and the more it is visited in the future, and so on." [114].

Adopting a broader definition of "collective intelligence" for hierarchical biological systems which consist of many interconnected parts (e.g., gene regulatory networks, cells, etc.) utilises the analogy with neural networks, suggesting that evolution involves altering those connections in a way similar to the training of a neural network (see also subsections 2.6 and 5.6). When the higher-order system attains a lower energy by modifying the energy landscape, it allows the collective to achieve a higher utility.

In information-processing terms, the access to a novel problem-space means that "collective intelligence of competent parts" is able to propagate information across scales, exhibiting an integrated problem-solving capacity, so that the higher-level agents are able to make decisions based on *extended patterns of information* [71]. For example, gene expression of the frog embryo is influenced by the spatial voltage differences across its brain regions (i.e., by a group-level pattern), rather than by the absolute values of individual cells [78, 77, 71]. As a result, the dynamics is able to visit larger spatial areas, approaching the higher-utility (i.e., lower-energy) attractors with larger basins of attraction.

2.6. Tensions between individual and group interests

It has been argued that the emergence of higher-level organisation generates a tension between levels. As noted by Szathmáry and Maynard Smith [100]: "entities that were capable of independent replication before the transition can replicate only as part of a larger whole after the transition". This notion has been further developed by Watson and Szathmáry [115] (see also [113]), who pointed out that "in evo-ego, correlations change the evolutionary unit (such that multiple, previously separate units become a new single unit at a higher level of organisation)". This analysis is closely related to the study of "collective intelligence" by McMillen and Levin [71], and reinforces the observations that a higher-level organisation is qualitatively different and expands the phase-space of possibilities.

Importantly, Watson and Szathmáry [115] explored the "evo-ego" relationship, highlighting a tension between individual and group interests: "individual-level selection will oppose the creation and maintenance of adaptations that enforce selection at the group level". They emphasised that the transitions in individuality essentially create new evolutionary units and may contribute to *the evolution of evolvability* [27], by evolving new mechanisms of inheritance or reproductive codispersal. Crucially, this approach identified a potential tension:

> "...if individual and group interests are aligned then selection applied at the group level does not alter evolutionary outcomes, and if individual and group interests are not aligned then individual-level selection will oppose the creation and maintenance of adaptations that enforce selection at the group level [76]." [115].

Watson and Szathmáry [115], as well as Watson et al. [113], examined this tension, aiming to explain how evolution at one level of biological organisation (e.g., individual cells) can systematically generate reproductive structures non-trivially adapting at a higher level of organisation (e.g., multicellular organisms), even "before that level of adaptation exists?" [115]. The proposed approach — evolutionary connectionism — is discussed in sections 5.6–5.8.

2.7. Ecological scaffolding

An alternative approach to explaining evolutionary transitions in individuality is offered by Black et al. [9], who pointed out that these transitions are immediately related to the emergence of biological complexity. In trying to analyse and clarify the conditions favouring emergence of collective-level reproduction (e.g., multicellular life), they proposed the concept of "ecological scaffolding". Specifically, they modelled that, given an ecological structure of distributed and dispersing resources, a division of labour would allow individual cells to "participate directly in the process of evolution by natural selection as if they were members of multicellular collectives" [9].

3. Beyond computational undecidability: "breaking" the limits of computation

In this section, we turn our attention to computation-theoretic approaches that can provide formal means to analyse an open-ended process of complexification. This computation-theoretic background will be used in subsequent sections to describe novelty generation during open-ended evolution (discussed in Section 2).

In general, the problems or functions that cannot be solved or computed by any algorithm are referred to as *incomputable*. For example, the Halting Problem, a classic problem in the theory of computation, asks whether it is possible to create an algorithm that can determine, for any arbitrary program and input, whether that program will eventually halt (terminate) or continue running indefinitely [106]. The halting problem is undecidable: no general algorithm exists that solves the halting problem for all possible program–input pairs. Likewise, Gödel showed that in any sufficiently powerful formal system, there are true statements that cannot be proven within the system [37].

3.1. Self-reference and diagonalisation arguments

A self-referential expression is one that refers to itself literally, e.g., the phrase "this sentence" in the liar sentence "this sentence is false" or via its referents i.e., distinct expressions that *denote*, *name*, or *encode* the original expression [12]. A notable example of self-reference by referents is Gödel numbering, where statements about natural number arithmetic are uniquely assigned a natural number themselves — so self-reference arises when statements about natural number arithmetic are applied to their own Gödel number [56].

Box 3: What is computation?

Algorithms
Computation is defined as a process of executing a series of *well-defined* instructions such that given an input state, its execution may satisfy a termination condition, which produces a corresponding output state.

An algorithm is a finite sequence of mathematically rigorous instructions, typically used to perform a computation.

Turing machines
A model of computation, e.g., Turing machine, is an abstract model which describes how an output of a mathematical function is computed given an input. Specifically, Turing machines model a mechanism that operates on an infinitely long input tape of discrete symbols, using a head that can read or write a symbol in a given position, store a state and move in either direction of the tape. Turing Machines include transition rules with respect to the state and symbol at its head, and terminating transitions known as halting conditions (see Fig. 1). The final state of the tape corresponds to the computation's output.

Turing machines have become synonymous with general-purpose computers because *Universal Turing Machines* (UTM) can be constructed such that the specifications of any Turing Machine (any algorithm) and input can be encoded as a UTM input, whereby the UTM's execution produces the output that matching the original Turing Machine applied to the original input (see Fig. 2).

Information processing
It is useful to distinguish computation from "information processing", the latter referring to the manipulation or transformation of data or information by a system, such as encoding, compression, storage, transmission, decoding, search, pattern recognition, retrieval, etc. While information processing typically transforms input data into an output, in general, it does not have to produce a meaningful output and may continue without a termination condition. Hence, information processing differs from computation, which is a process with well-defined algorithm and termination (output) states. (See [74, ch. 12] for further discussion).

Self-referential statements have historically been employed to prove results related to incomputability, using what are known as *diagonalisation* arguments. A canonical setting for diagonalisation arguments are expressions E_i, referents $\ulcorner E_j \urcorner$ and properties P_{ij} arising out of the application of expressions E_i to referents $\ulcorner E_j \urcorner$, which we can

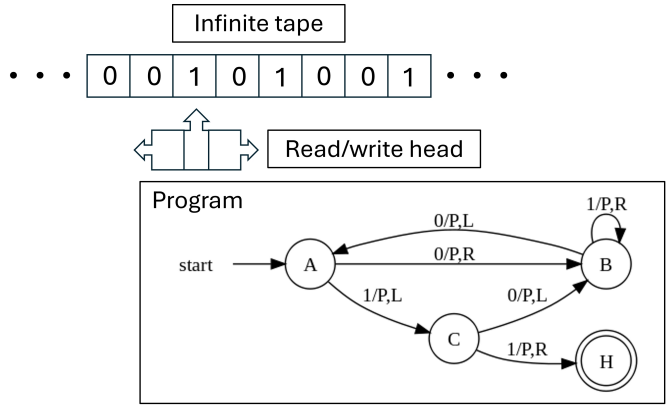

Figure 1: An example finite-state specification of a Turing Machine. States include A, B, C and halting state H. An arrow label (e.g., 0/P,R) specifies the tape symbol (e.g., symbol 0) that upon reading triggers a particular transition to another state, followed by the action, e.g., print (P) and move tape to the right (R) [122].

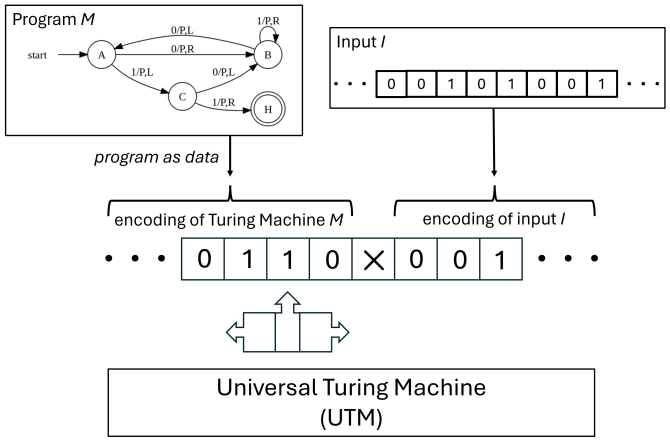

Figure 2: Universal Turing Machine (UTM) simulating Turing Machine M on input I. An encoding of program P converts it into input data which is separated on the tape from the input I by suitably chosen separator symbol(s) $\times$.

visualise as the following grid:

	$\ulcorner E_0 \urcorner$	$\ulcorner E_1 \urcorner$	$\ulcorner E_2 \urcorner$	$\cdots$	$\ulcorner E_k \urcorner$	$\cdots$
E_0	P_{00}	P_{01}	P_{02}	$\cdots$	P_{0k}	$\cdots$
E_1	P_{10}	P_{11}	P_{12}	$\cdots$	P_{1k}	$\cdots$
E_2	P_{20}	P_{21}	P_{22}	$\cdots$	P_{2k}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
E_k	P_{k0}	P_{k1}	P_{k2}	$\cdots$	P_{kk}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

(1)

Incomputability is then represented by finding combinations of expressions E_i and referents $\ulcorner E_j \urcorner$ whose property $P_{i,j}$ can not be determined without a contradiction. These combinations typically arise out of constructions on properties of the form P_{kk}, which are considered to be self-referential as they represent the application of an expression to its own referent.

To illustrate how diagonalisation arguments work, let us consider Cantor's Theorem. This result states that the subsets of the natural numbers (e.g., $\{2, 5, 1\}$) are uncountable, i.e., that any list of subsets $E_0, E_1, E_2, \ldots$ (indexed by the natural numbers) is incomplete [15]. The argument begins by assuming that an arbitrary such list is complete. For such a list, one builds an expression-referent grid as follows: expressions E_i correspond to the subsets, referents $\ulcorner E_i \urcorner$ indicate the corresponding index (i.e. $\ulcorner E_i \urcorner := i$), and properties P_{ij} are defined as $P_{ij} = 1$ if i is an element of E_j and 0 otherwise. The second step of the argument to prove incompleteness is to find i and j such that P_{ij} can not be determined. To do this, one uses the diagonal to construct the set $\{k \in \mathbb{N} \mid P_{kk} = 0\}$, i.e., each number k when considered as a referent is not a member of the expression referenced by it. Because this is a subset of the natural numbers, it must be an expression within our list, say E_x. However, any attempt to evaluate P_{xx} leads to a contradiction: assuming the referent x is not in E_x leads to the conclusion that x *does* belong to E_x, and *vice versa*. Therefore, we use the incomputability of P_{xx} to conclude that our initial list could not have been complete.

Another diagonalisation argument can be used to show that the real numbers are uncountable. Using an expression-referent grid formed by enumerable rows of infinite sequences of binary digits, Cantor demonstrated that while there are countably many columns (corresponding to the natural numbers), there are in fact uncountably many rows (representing the real numbers) — i.e., that such grid is not a square. In doing so, Cantor used a diagonalisation argument constructing a real number E_v as a sequence of binary digits P_{vj} such that each digit is complementary to its diagonal counterpart P_{jj} (essentially, swapping zeros and ones in P_{jj}). By definition, the constructed sequence E_v is different from any countable sequence E_k, disagreeing with it on at least one digit.

3.2. Discrepancy between expressions and referents

Importantly, diagonalisation arguments reveal and exploit a fundamental discrepancy between the space of expressions and the space of referents: in general, there are *more* expressions than referents. In other words, it is not possible to construct a one-to-one correspondence between these two spaces. For instance, by showing that the real numbers are *uncountable*, Cantor proved that there are *more* real numbers than natural numbers — despite both sets being infinite. The fundamental expression-referent discrepancy forms a key part of our analysis. It is also described in Appendix A.1.1 in the context of the halting problem.

3.3. Incomputability and oracle machines

As somewhat ironically noted by Soare [97], "the field of computability theory deals mostly with *incomputable*, not computable, objects". Crucially, the incomputability of a given problem is considered relative to a given computational system, and in certain cases it can be overcome

by considering another system that extends the original system in some way. For example, decision problems undecidable by Turing machines can be decided by a Turing machine extended with an additional component known as the *oracle* [107]. An oracle is a black-box capable of providing, in a single operation, an answer for any instance of the corresponding decision problem, as illustrated by Fig. 3. Adding an oracle to a Turing machine is analogous to adding a new, independent, axiom to a formal logical system, so that the extended system is able to prove more than it could previously do [22].

As a result of such extension, an *oracle machine* — a Turing machine connected to an oracle — can simulate any Turing machine, producing the same output, given the same input, as the simulated Turing machine [97]. Hofstadter [44] informally referred to this way of resolving paradoxes as "Jumping Out Of The System" (JOOTSing). However, such a resolution generates new undecidable problems within the extended (Turing machine plus oracle) system, leading to an open-ended process described in the next subsection.

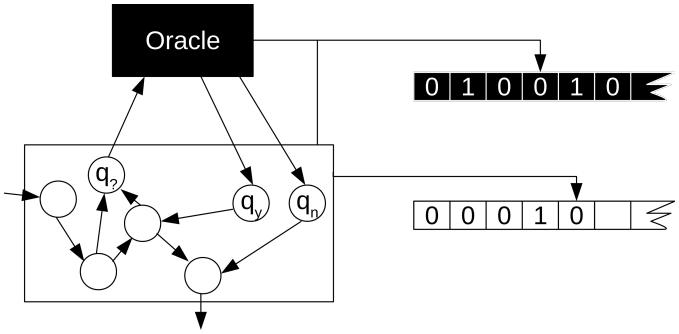

Figure 3: Turing oracle. Wikimedia Commons, licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

3.4. Open-ended computational meta-simulation

In general, the idea of resolving undecidability by extending bounds of the computational system is developed within the field of recursion theory. In particular, the Gödel–Turing–Post recursion-theoretic framework proposed various methods of constructing extensible ordinal or recursively generated logics [37, 105, 106, 107, 81, 97].

Turing [107] proposed a way of overcoming the halting problem by providing a means to source an answer beyond the system boundary, that is, by considering an α-order oracle machine specified for a particular level. In general, one may consider a sequence of α-order oracles machines with increasing orders α, each of which (for $\alpha > 0$) is capable of resolving undecidable halting problems of lower levels [107, 80], as illustrated in Fig. 4.

The α-order oracles machines can be related by an operation known as the "Turing jump". Formally, the "Turing jump" is an operation that assigns to each decision problem X a successively harder decision problem X' such that X' is not decidable by an α-order oracle machine with an oracle for X. Crucially, the Turing jump of X generates

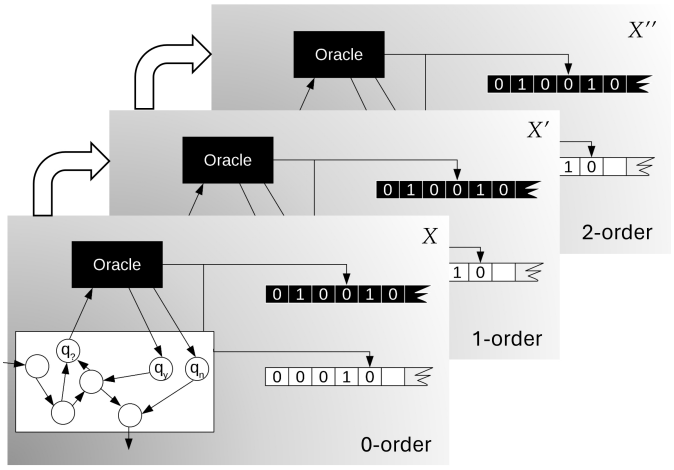

Figure 4: Open-ended sequence of α-order oracle machines, with the corresponding Turing jump operations assigning successively harder decision problems X, X', X'',..., in an open-ended way.

an $(\alpha + 1)$-order oracle to the halting problem for α-order oracle machines with an oracle for X.

The boundaries of lower-order systems are expanded by adding specific "novelties" generated by an interaction with the corresponding oracle. In describing this continual iterative process, Turing [107] drew an analogy with ordinal logics of Gödel [37]. Ordinal logic is associated with an ordinal number by recursively adding elements to a sequence of previous logics:

> "The well-known theorem of Gödel (1931) shows that every system of logic is in a certain sense incomplete, but at the same time it indicates means whereby from a system L of logic a more complete system L' may be obtained... A logic L_ω may then be constructed in which the provable theorems are the totality of theorems provable with the help of the logics L, L_1, L_2,..." [107].

This view was further developed by Post [81] using the recursively enumerable sets. In demonstrating that every recursively generated logic may be extended, Post [81] proposed the concept of relative computability: degree of unsolvability, or the Turing degree, of a set of natural numbers measures the level of algorithmic unsolvability of the set. Consequently, Post [81] proposed a hierarchy of degrees of unsolvability, where each degree represents the extent to which a set is unsolvable or incomputable. Each level of the hierarchy corresponds to a different degree of computational complexity, with sets at higher levels being more incomputable than those at lower levels.

Importantly, each extended system $X^{(\alpha)}$ comprising an α-order oracle machine ($\alpha > 0$) is able to simulate a lower-level $X^{(\alpha-1)}$ system. Arguably, one may consider a higher-order α-order oracle machine as a meta-simulator, i.e., a simulator of nested, lower-order, simulations (down to the 0-order oracle machine that can simulate any Turing

machine). Thus, we suggest that when a higher-order system simulates the lower-order computation, providing answers to lower-level undecidable problems, it performs *meta-simulation* of the lower-order systems. We shall refer to the successive construction of extensible computational systems and ordinal or recursively generated logics (within the Gödel–Turing–Post recursion-theoretic framework) as *open-ended meta-simulation*.

4. Undecidability and open-endedness in dynamical systems

The third primary domain, considered as part of our background, is physical dynamical systems. The concept of undecidable dynamics has been defined in physical systems by some analogy with computational undecidability, and so we can expect to see emerging parallels between unpredictable dynamics, expanding phase-spaces and extensible computational systems.

4.1. Undecidability in physical dynamical systems

Physical dynamical systems are inherently nonlinear and time–continuous. According to Moore [75], physical dynamical systems with at least three degrees of freedom can be Turing equivalent. They exhibit a type of unpredictability that is qualitatively stronger than low-dimensional chaos, with undecidable long-term dynamics, even if the initial conditions are known exactly. Bennett [4] qualifies this as follows:

> "For a dynamical system to be chaotic means that it exponentially amplifies ignorance of its initial condition; for it to be undecidable means that essential aspects of its long-term behaviour — such as whether a trajectory ever enters a certain region — though determined, are unpredictable even from total knowledge of the initial condition."

Thus, undecidability is a more extreme kind of unpredictability than chaos [5]. One implication of this is that if the halting problem is viewed through the lens of dynamical systems, then its undecidability means that the long–term behaviour of the dynamics corresponding to a universal Turing machine is unpredictable, even if the initial conditions are known exactly.

A relatively simple example offered by Moore is the motion of a particle in a 3D potential. He suggests, however, that the three-body problem (n-body more generally) presents more realistically complex dynamics for testing these ideas, in particular by demonstrating lack of scaling behaviour, irregular spectra of periodic points and attractor states, which are beyond the features that typically characterise chaotic dynamics in non-physical dynamical systems (e.g. the Lorenz 63 system, a mathematical model that only loosely approximates the complex dynamics of real weather systems). As shown in Fig. 5 for the classic

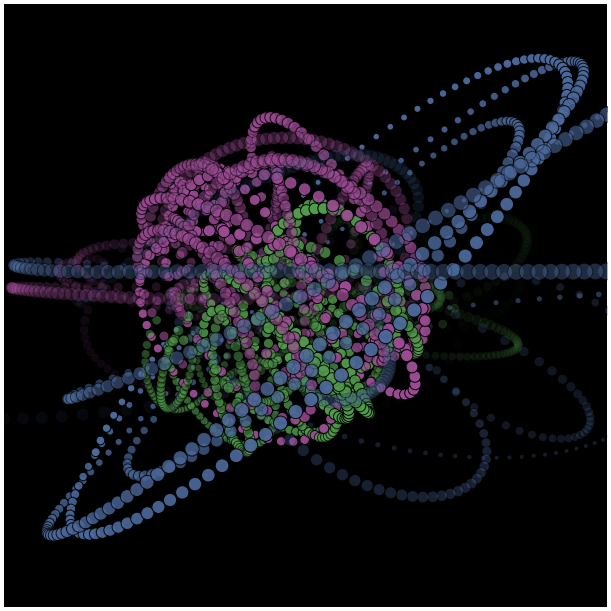

Figure 5: n-body simulation demonstrating unpredictability and undecidability of Newtonian dynamical systems, in this case comprised of $n = 3$ stars of different masses (indicated by coloured trajectories) orbiting each other under the influence of gravity.

3-body problem in a Newtownian gravitational system, it is impossible to predict if or when one of the orbiting objects becomes gravitationally unbound (i.e. the halting condition), even if the initial conditions (position and velocity of each object) are known.

Physical dynamical systems (e.g. fluid flows, propagating waves), as described by space–time partial differential equations, may have universal computing abilities if they have sufficient dynamical variables to maintain logically distinct trajectories in the presence of external noise. Bennett [5] argued that computationally universal dynamical systems can be used to define physical complexity, the quantity that increases when a self-organising system organises itself, producing novel features beyond those resulting from long-term evolution. Wolpert [127] derived specific impossibility results associated with physical computation. This work developed analogues of Chomsky hierarchy results about universal Turing Machines and the Halting theorem, showing, in particular, the impossibility of certain kinds of error-correcting codes. These observations not only open a way to connect computation-theoretic and dynamical-systems views on complexification processes, but also suggests that novelty generation during an open-ended biological evolution may also be seen through the lens of expanding computational spaces. These cross-disciplinary perspectives are explored in more detail in Section 5.

4.2. Expanding phase-space

In contrast to non-physical dynamical systems, physical dynamical systems are constrained by the laws of physics, which could be argued as restricting phase space expansion.

However, physical systems may also exhibit stochastic (non-deterministic) dynamics, which, while also obeying laws of physics (e.g. thermodynamics), introduces extra degrees of freedom (i.e. higher-order complexity) in addition to any deterministic dynamics, provided the stochasticity is not purely random, without any underlying structure or information content.

In general, dynamical systems (both physical and non-physical) may encounter different dynamical phase-space regimes, e.g. ordered, chaotic and edge-of-chaos, with different computational representations. Chaotic dynamics, associated with an increase in phase space exploration, continuously generates information and the amount of information needed for prediction grows with time. As discussed above, chaotic dynamics are not necessarily computationally undecidable. Edge-of-chaos dynamics, on the other hand, has been proposed as optimal for information processing and novelty generation [58]. In physical systems, edge-of-chaos is associated with the spontaneous emergence of long-range spatio-temporal correlations, effectively expanding dynamical phase-space, with recent studies (e.g. [11, 42]) demonstrating enhancement in information processing, but only in terms of general properties or for relatively complex computational tasks rather than necessarily for any specific task. Prokopenko et al. [84] examined the link between edge-of-chaos and undecidable dynamics in the context of cellular automata. This is briefly introduced in section 4.3 and discussed in more detail in Section 5.1.

4.3. Game of Life

To illustrate how undecidable dynamics can be generated in a concrete system, let us introduce *Conway's Game of Life* — a well-known example of a discrete dynamical system [35]. It is a two-dimensional cellular automaton (CA) (see Appendix A.3), specified with a binary alphabet and a local update rule defined for the Moore neighbourhood with 9 cells, incorporating:

1. Deaths. Any live cell with fewer than two or more than three live neighbours dies.
2. Survivals. Any live cell with two or three live neighbours lives on to the next generation.
3. Births. Any dead cell with exactly three live neighbours becomes a live cell.

Game of Life dynamics produce coherent spatial patterns, including oscillators that repeat themselves after a fixed number of generations (e.g., see Fig. 6), and gliders that move across the grid replicating their structure. It has been demonstrated that CA carry out distributed computation, with oscillators representing information storage (i.e., memory) [64], gliders capturing information transfer (i.e., communications) [61], and glider collisions corresponding to information modification (i.e., processing) [62], with the computational processes forming coherent information structures [63].

Box 4: Arrows of time

There are several conceptually distinct arrows of time which are often interrelated in forming an understanding of the "asymmetry" of time, explaining why processes in nature appear irreversible and why we perceive time as moving forward from the past to the future.

Thermodynamic arrow of time

Thermodynamic arrow of time is defined by the second law of thermodynamics, which states that entropy in a closed system always increases over time, leading to the irreversibility of natural processes.

Thermodynamic arrow of time and evolution

Blum [10] discussed "the relationship between time's arrow (the second law of thermodynamics) and organic evolution, exploring irreversibility and direction in evolution, and arguing that evolutionary patterns may be predetermined by thermodynamic processes. Another hypothesis, Dollo's law, suggests that once a complex trait or structure has been lost by an organism in the evolutionary process, it is unlikely to be regained in exactly the same form [20]. In other words, Dollo's law posits that evolution is not generally reversible, although reversible evolution has been observed on relatively short evolutionary timescales [19].

Cosmological arrow of time

Cosmological arrow of time refers to the direction of time in which the universe is expanding, and is based on the observation that the universe is growing larger over time, as opposed to contracting.

Radiative arrow of time

Radiative arrow of time aligns the direction of time with the way radiative processes, such as retarded electromagnetic radiation, the emission of light and sound waves, expand outward from their source, from a higher energy state to a lower one.

Causal arrow of time

Causal arrow of time reflects the principle that cause precedes effect: perceived events are ordered in a way that causes come before their effects.

Quantum arrow of time

Quantum arrow of time as defined in quantum mechanics relates time's direction to the collapse of the wave function. While the direction of time in a quantum system may be blurred due to uncertainty, when the system is measured, it transitions from a state of superposition to a definite state, which defines a temporal direction. An insightful perspective which may characterise other time arrows was offered by Seth Lloyd: the arrow of time is an arrow of increasing correlations [65].

It has been shown that computational power of Game of Life is equivalent to that of a universal Turing machine [6], and thus, the Game of Life dynamics may be undecidable under some conditions [84]. In other words, it cannot be determined, for all initial configurations of the Game of Life, whether or not they will reach some predefined final (termination) configurations (see Appendix A.3).

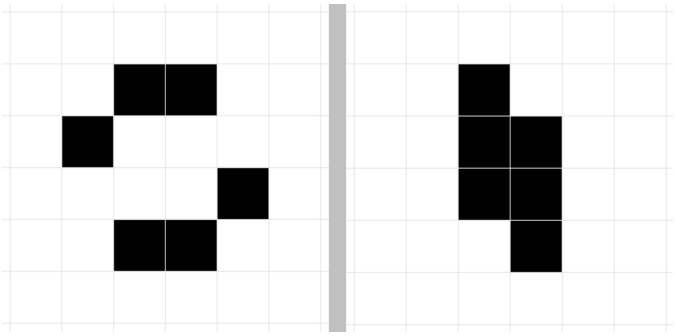

Figure 6: Two configurations of the oscillating polyomino pattern known in Conway's Game of Life as "toad".

5. Cross-disciplinary perspectives on novelty generation

In this section, we draw insights from several background studies across the three examined domains — biological, computational and physical — and explore possible connections among (a) open-ended biological complexity, including major evolutionary transitions (Section 2); (b) open-ended computational meta-simulation, including successive resolution of lower-level undecidable problems (Section 3); and (c) complex dynamical systems, including chaotic and strange attractors (Section 4). In doing so, we will try to examine how the "evolving evolvability" is related to emergence of hierarchical structures comprising individual and collective information-processing elements, e.g., (a) formation of new units of selection, comprising ensembles of pre-existing organisms, (b) computational novelty generation, and (c) emergence of self-replicating dynamic behaviours at the "edge of chaos".

5.1. Chaos, incomputability, and undecidable dynamics

Casti posited a link between the existence of chaotic dynamical systems and Gödel's incompleteness theorem [16, p.317] [17, p.148]. Casti argued that

> "...the theorems of a formal system, the output of a UTM [Universal Turing Machine], and the attractor set of a dynamical process (e.g., a 1-dimensional cellular automaton) are completely equivalent; given one, it can be faithfully translated into either of the others" [16, p.317].

By this correspondence Casti investigated dynamical systems that compute Kolmogorov complexity of numbers / programs, which is the length of the shortest programs that can compute them. Casti then applied Chaitin's (Incompleteness) Theorem [18] on Kolmogorov complexity to prove such universal computing dynamical systems have strange attractors unreachable from any initial state, which Casti argued is equivalent to Gödel's (first) incompleteness theorem. Analogously, Adams et al. [1] pointed out that simple one-dimensional cellular automata possess many states that cannot be reached from any initial state by repeated application of any of the complete set of 256 evolution rules. Such states may, however, become accessible if the evolution rules themselves evolve with time.

Several fundamental relations between (i) formal systems, (ii) algorithms, and (iii) dynamical systems were identified by Prokopenko et al. [84]. The comparative analysis identified three common factors implicated in the generation of undecidable dynamics within the three examined computational frameworks: (i) the program-data duality (e.g., encoding of Turing machines as input data); (ii) access to an infinite computational medium (e.g., an infinite tape used by a Turing machine); and (iii) the ability to implement negation (e.g., the ability to flip *accept* and *reject* states).

Considering "undecidable dynamics" in a broad context, Prokopenko et al. [84] generalised the concept of incomputability across several computational frameworks. For example, in formal systems, a *dynamic process* can be identified with a *proof*: a sequence of well-formed formulas derived by inference rules and starting from an axiom. Correspondingly, Gödel Incompleteness Theorem applicable to formal systems establishes the limits of *provability* in axiomatic theories, capitalising on the observation that some formal sentences can neither be proved nor disproved (e.g., the famous Gödel sentence which encapsulated the Liar paradox).

Similarly, for a Turing machine, a *dynamic computational process* is provided by a *sequence of machine states and tape patterns*, starting from some input. The halting Problem is the canonical example of computational *undecidability*, capturing the fact that it cannot be established, for all program-input pairs, whether the computation reaches a predefined halting state or runs forever.

Finally, one can consider an evolution of configurations — a *dynamic trajectory* — within a dynamical system such as a Cellular Automaton, starting from an initial state. Given suitably defined termination conditions (e.g., testing for fixed points or limit cycles), one may attempt to establish whether the trajectory is reaching the predefined attractors or continues to unfold along the "edge of chaos", i.e., class IV Cellular Automata [125]. The latter scenario can be identified with *undecidable dynamics*; see also Section 4.1.

In summary, formal systems (e.g., logical systems), models of computation (e.g., Turing machines), and dynamical systems (e.g., Cellular Automata) are deeply related: these frameworks can produce universal computation and generate undecidable dynamics. This undecidability is manifested through three factors: self-reference, infinite computation, and negation [84].

5.2. Computational novelty generation

The self-referential basis of undecidable dynamics is fundamentally related to novelty generation [84]. As pointed out by Markose [67, 68], computational novelty production and "thinking outside the box" by digital agents is underpinned by their capacity to encode self-referential statements with negation (e.g., Liar paradox or a Gödel sentence). Crucially, this capacity allows the agents to exit from known listable sets (e.g., actions, technologies, phenotypes) and produce new structured objects. As discussed earlier (Section 3.2), this is a consequence of the fundamental discrepancy between (the spaces of) expressions and referents which is exploited by various diagonalisation arguments.

Svahn and Prokopenko [99] explored how undecidability which places computational limits on a formal system can manifest itself in biological RNA automata. In particular, they considered Turing-equivalent RNA automata, i.e., the ones that can simulate, and be simulated by, a universal Turing machine. Importantly, they argued that the evolutionary space for these RNA automata can be expanded in a continual process analogous to a hierarchical resolution of computational undecidability by a sequence of Turing's oracles (and hence, by a sequence of Turing's ordinal logics and Post's extensible recursively generated logics). The proposed *ansatz* hypothesised that the resolution of possible undecidable configurations in biological RNA automata may represent a novelty generation mechanism, in context of interactions between the automata and their environment [99].

Thus, computational novelty can be seen as the problem-space expansion, created by agents that use the diagonalisation argument in exploiting the expression-referent discrepancy. In other words, novelty can be generated by *self-modelling agents* that have access to results of meta-level computation — for example, meta-simulation by Turing oracles and extensible logics (see Sections 3.3 and 3.4), or receive the corresponding information from external environment.

5.3. Tangled hierarchies and strange loops

In his seminal book, Hofstadter [44] explored how self-reference leads to emergent complexity in systems, by examining recursive structures which appear in a self-similar way at different levels or scales. Importantly, Hofstadter discussed Gödel's Incompleteness theorems, which show that in any formal mathematical system that is expressive enough to describe basic arithmetic, there will be statements that are true but cannot be proven *within the system*.

Hofstadter [44] proposed and extensively discussed the notion of *tangled hierarchies* (see Box 5) emerging in a wide range of phenomena, from language and cognition to artificial intelligence and consciousness. Self-reference plays a key role in this process, as it allows for feedback loops and recursive interactions between different levels of the hierarchy. In particular, Hofstadter argued that tangled hierarchies are fundamental to human cognition, which relies heavily on analogical thinking and self-referential loops. The concept of the *strange loop* encapsulated this idea, showing that patterns of thought may loop back on themselves to generate *new* levels of understanding.

5.4. Causative efficacy of biological information

Walker and Davies [111] argued that a hierarchy does not just encapsulate some quantity of information, but offers a specific information-processing arrangement describing its organisation, e.g., feedback loops, active information, and integrated information. Such an arrangement gives the higher levels in the hierarchy more functional "power", i.e. causal efficacy. In other words, information hierarchies may emerge (or self-organise) because they capture the *causative efficacy of information*:

> "...biological information has an additional quality which may roughly be called 'functionality' — or 'contextuality' — that sets it apart from a collection of mere bits as characterized by its Shannon information content. ...DNA is a (mostly) passive repository for transcription of stored data into RNA, some (but by no means all) of which goes on to be translated into proteins. ...It is the functionality of the expressed RNAs and proteins — not the bits — that is biologically important" [111].

Walker and Davies [111] further pointed out that biological information (i.e., functionality) is actively involved in information processing, e.g., control and feedback. Thus, the functionality is dynamic, depending on both the current state and the history of the organism. This perspective — *the algorithmic origins of life* — suggests that in order to delineate the phases of non-life and life one needs to analyse dynamical information and identify causal architecture:

> "...the emergence of life may correspond to a physical transition associated with a shift in the causal structure, where information gains direct and context-dependent causal efficacy over the matter in which it is instantiated" [111].

Following Davies [25], this work interpreted the phenotype-vs-genotype relationship as hardware-vs-software, identifying chemistry in living systems with hardware, and information (e.g., genetic and epigenetic) with software: "the chicken-or-egg problem, as traditionally posed, thus amounts to a debate of whether analogue or digital hardware came first" [111]. In a related study, Walker et al. [110] discussed how the information efficacy and top-down causation can also be applied to the major evolutionary transitions [100].

5.5. Evolution as collective information-processing dynamics

The emergence of genetic code can be modelled as an information-preservation phenomenon in the presence of noise [85]. In particular, the capacity to symbolically represent nucleic acid sequences was argued to emerge, overcoming the "coding threshold" [123], in response to a change in environmental conditions. In modelling the emergence of universal coding, Prokopenko et al. [85] proposed the concept of "stigmergic gene transfer", and considered interacting proto-cells as a dynamical system, within which "proto-symbols" encoding features of individual cells are stigmergically shared.

The model demonstrated that a joint encoding can emerge as a symbolic representation of the shared dynamics, in order to preserve information about attractors of the dynamics in a noisy environment: "...the pressure to develop a distinctive "symbolic" encoding only develops if the noise in the original system is in a particular range, not too small and not too large" [85].

Goldenfeld and Woese [38] described evolution from the standpoint of non-equilibrium statistical mechanics, as a problem in which the key dynamical modes are collective. Taking a provocative stance — "Life is Physics" — they argued that unifying principles of collective behaviour which arise from physical interactions are applicable to biology, especially in context of the interplay between evolution and environmental fluctuations.

In exploring the principles underlying collective behaviour, Goldenfeld and Woese [38] posed a key question: "How is it that matter self-organizes into hierarchies that are capable of generating feedback loops which connect multiple levels of organization and are evolvable?" [38]. Importantly, the study related co-evolutionary processes to far-from-equilibrium dynamics and pointed out that the corresponding physical laws remain unknown. Nevertheless, their review highlighted a salient connection between co-evolutionary/ecological dynamics and game-theoretic interactions which may generate paradoxical collective outcomes (e.g., Nash equilibria) and even chaotic dynamics (in dynamic game-theoretic models [92]).

5.6. Evolving self-modelling dynamical systems

As discussed in Section 2.6, Watson et al. [112] studied the "evo-ego" relationship, describing a tension between individual and group interests. While not framing their analysis in terms of "tangled hierarchies", the study nevertheless related the organisation of individual self-interested entities to their long-term collective interest, posing two questions:

> "(a) what kind of functional relationships between components are needed to make a new individual, and how they need to be organised; and (b) how the organisation of these relationships arises 'bottom-up,' i.e., without presupposing the higher-level individual we are trying to explain" [112].

The study pointed out that the chicken-and-egg dilemma (i.e., a strange loop), encapsulating these two questions, is typical across the major evolutionary transitions in individuality, as it is challenging to answer whether the higher-level unit of selection (needed for complex adaptations) emerges before or after the complex adaptations (needed to generate the higher-level unit of selection).

In attempting to resolve this conundrum, Watson et al. [112] drew an analogy with unsupervised *deep learning*, arguing that when individual reinforcement or selection modifies the strength of inter-unit relationships (analogously to updating the neural network weights during reinforcement learning), "the system becomes a self-modelling dynamical system" with predictable system dynamics. In other words, based on interactions with the environment but without an explicit external guidance, a self-modelling system is able to discover some structure in its own constraints and dynamics — that is, construct its own model. This is analogous to unsupervised learning algorithms which, given the unlabelled input data, can analyse and learn patterns by clustering similar data points or reducing dimensionality.

5.7. Information integration within collective action

In formulating the approach of *evolutionary connectionism* based on the analogy with machine learning, Watson et al. [112] identified the conditions required to evolve a novel organisation — a new level of individuality — comprising "short-sighted, self-interested entities", with the conditions necessary to learn *non-decomposable functions*. Put simply, the relationships between evolutionary units must be organised bottom-up in such a way that their interactions are synergisitic, producing "more than the sum of the parts". Crucially, this distributed (unsupervised) learning can occur without an explicit system-level feedback, by exploiting regularities encountered at localised interactions.

In other words, interactions of bottom-level entities produce distributed learning dynamics and integrate information by computing a non-decomposable (i.e., non-linearly separable) function of input states, for example, between some "embryonic" collections of particles and the corresponding "adult" collective phenotypes [112]. In turn, this integrated information facilitates collective action, for example, generating "specific coordinated responses in multiple downstream variables" — thus, signifying a downward causation. In terms of tangled hierarchies, the bottom-up process, generating a new step in evolutionary individuality, involves synergistic computation of non-decomposable functions which confer collective fitness, thus constraining interactions of the constituent units in a top-down way.

5.8. Evolutionary connectionism vs ecological scaffolding

Watson et al. [112] contrasted their "evolutionary connectionism" approach with "ecological scaffolding" [9] (see

Section 2.7), emphasising that in the latter one has to assume existence of some "fortuitous extrinsic conditions" that trigger the population structure to generate specific selective pressures. In other words, they argued that ecological scaffolding resolves the chicken-and-egg problem of the evolutionary transitions in individuality by (temporarily) allowing the environment to assume the role of a "chicken". Nevertheless, Watson et al. [112] pointed out that, even under these conditions,

> "...individual selection at the lower level supports the evolution of characters that access synergistic fitness interactions, changing the relationships among the particles, and given that synergistic fitness interactions among particles have evolved, it is subsequently advantageous for particles to evolve traits that actively support this grouped population structure" [112].

It can be argued that the ecological scaffolding is another way to shape a tangled hierarchy — albeit, without *self-modelling*. Once/if the scaffolding becomes redundant, the evolved "strange loop" endogenously captures the collective dynamics, comprising both (i) the top-down distribution and dispersal of resources and (ii) the bottom-up computation shaped by the division of labour [9]:

> "Now the original extrinsic ecological conditions might change or cease, but the population structure necessary to support higher-level selection is nonetheless maintained, supported by the adaptations of the particles" [112].

In subsequent analysis, we will argue that developing the self-modelling capacity is a crucial step in biological complexification, improving the organism's capacity to adapt, replicate and function autonomously.

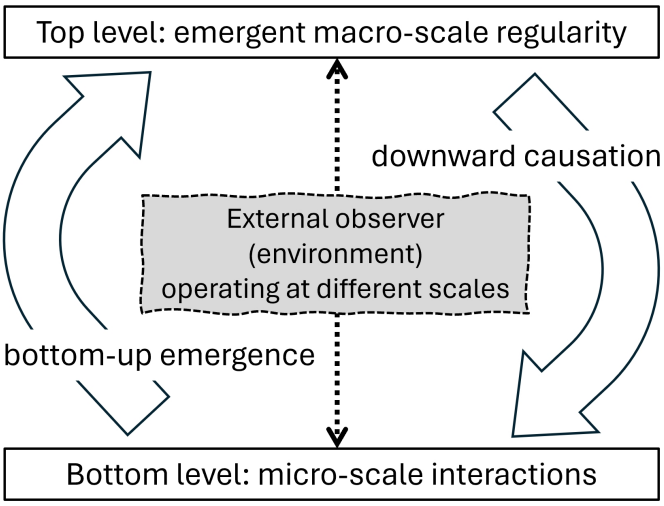

Figure 7: A tangled hierarchy (strange loop) with an external interpreter (e.g., observer, environment, interacting particle, etc.) which may operate and respond to regularities detected at different scales.

Box 5: Self-referential dynamics

Tangled hierarchies

Tangled hierarchies are systems in which different levels of organisation or abstraction are intertwined in a way that defies simple hierarchical categorisation: "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" [44]. In other words, tangled hierarchies interleave bottom-up emergence and top-down (downward) causation [31]. Fig. 7 illustrates the concept of tangled hierarchy, and Fig. 8 shows an example: shortest path formation during ant foraging.

Strange loops

As described by Hofstadter [44], "a strange loop is a hierarchy of levels, each of which is linked to at least one other by some type of relationship. 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." A strange loop involving ant foraging and optimal path formation, shown in Fig. 8 (see the example described in Section 6.2.1), lacks a preferred causal direction: are the ants driven by the pheromone gradient of the path (downward causation), or is the shortest path itself being generated by the ants movement (bottom-up emergence)?

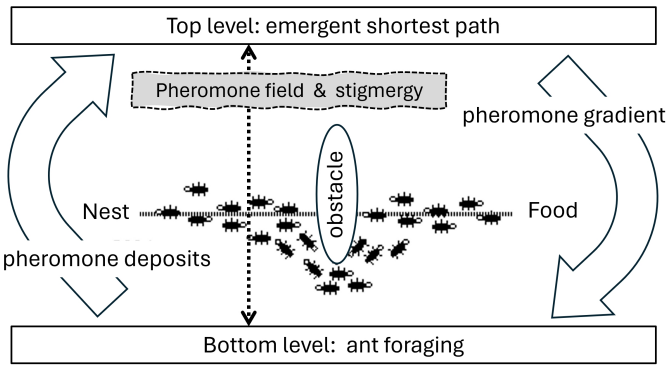

Figure 8: An example of tangled hierarchy (a strange loop): ants foraging for food deposit pheromones which diffuse and evaporate in the environment; ants direct their movement towards locations with higher pheromone concentration, utilising stigmergy (indirect coordination through the environment); a shortest path emerges out of these interactions (bottom-up emergence), and influences the ant foraging (downward causation).

Box 6: Scales, levels and emergence

Scales vs levels

The difference between scales and levels: "scales" imply a (continuous) progression along a single dimension (e.g., from fine-grained to coarse-grained), while "levels" typically suggest discrete distinct stages within a hierarchical structure.

Meta-levels and interpretation

There may be an external observer interpreting the interaction between the levels of a (tangled) hierarchy. This external interpreter is not part of the hierarchy, but may exchange matter and information with both interacting levels. For example, the environment may serve as an external, meta-level observer (or meta-simulator), determining the organism's fitness, and thus providing an "interpretation" of the organism behaviour/dynamics. While hierarchies have "levels", observer may employ different "scales" of observation: high-resolution (e.g., statistical mechanics), and low-resolution (thermodynamics) [83], distinguishing emergent phenomena by detecting and responding to regularities at a particular scale of observation, see Fig. 7.

6. Tangled hierarchies and self-modelling

Having examined cross-disciplinary views on novelty generation in biological, computational and physical dynamical systems, we point out a common feature of complexification processes: emergence of hierarchical structures comprising lower-level (e.g., individual) and higher-level (e.g., collective) information-processing elements. These tangled structures exhibit varying degrees of self-reference, self-modelling and self-replication.

To explore the role played by self-modelling in the emergence of replicating information-processing hierarchies, in this section we propose a distinction between two types of tangled hierarchies: type I (strange loop without self-modelling) and type II (strange loop with self-modelling). We exemplify these two types in several contexts (subsection 6.2), and argue that different mechanisms are employed to replicate their dynamics (subsection 6.3).

6.1. Two types of tangled hierarchies

We define a tangled hierarchy of type I (TH-I) as one where macro-scale information patterns which emerge from the micro-scale interactions within the environment are not compressed, encoded or decoded. In other words, no emergent macro-scale regularity is explicitly modelled in its entirety by micro-scale objects. Despite presence of a recursive flow between levels which "loops back on itself" [44], this loop does not involve compression — and hence, there is no object in TH-I that is explicitly referred to as 'self'.

In contrast, a tangled hierarchy of type II (TH-II) is one where micro-scale objects operating at the bottom level contain an encoded, compressed representation of some information pattern emerging at the macro-scale. By utilising the compressed information, the micro-scale objects are capable of representing — that is, modelling — some of the macro-scale regularities. This modelling is used within a self-referential loop by alternating encoding and decoding processes. As a result, the compressed representation (i.e., a model) of a higher-level pattern, encoded at a lower level, augments the entire hierarchy with *self-modelling*, due to the tangled nature of the inter-level relationship (see examples of TH-II in subsections 6.2.2 and 6.2.4).

In general, there may be a spectrum of hierarchies between types I and II, with some tangled hierarchies utilising compression/modelling to a partial extent, and the general dichotomy we proposed essentially shows the limit cases.

6.2. Examples of different TH types

In order to illustrate the two TH types and build a better intuition, we present several examples, mostly drawn from biology.

6.2.1. Ant foraging, stigmergy and optimal path formation

Ant foraging and optimal path formation in a pheromone field is a tangled hierarchy without self-modelling (TH-I). In this case, the tangled hierarchy is constituted by the ants foraging behaviour at the bottom level and the shortest path emerging at the top level (see Fig. 8). Each ant foraging for food deposits pheromone in response to only local information, without reference to the global (optimal) path. The path itself emerges as a result of stigmergy [101]: indirect interactions and coordination of the ants through the environment (Fig. 8). Importantly, ants perceive only local differences in the pheromone field (the pheromone gradient), and make only local field updates. The optimal path is a regularity within the pheromone field — however, this regularity is not compressed, and the information patterns underpinning stigmergy remain distributed through the environment, without any encoding, decoding or self-modelling.

It can be argued that some features of the emergent path (e.g., the likelihood of it being a shortest path) are attributable to (i.e., partially and indirectly encoded within) the ants genome. In other words, while the ant foraging behaviours belong to the ants immediate phenotype (being directly influenced by their genetic makeup), the class of optimal paths can be seen as a part of the ants extended phenotype [26]. That is, the species-dependent pheromone paths, shaped by the interactions between the ants and the environment, contribute to the genome evolution. In this process, the path optimality provides evolutionary rewards resulting in an eventual spread of the beneficial, fitness-increasing genotype. The encoded features are specifically the ones which are likely to produce shortest paths, given the environmental factors.

However, any actual physical path emerging in a given environment — a specific regularity in the environment-dependent pheromone field — is not *directly* encoded in the

genome. Rather, it emerges from the behaviours guided by genetic predispositions, while being influenced by the ants interactions with their environment.

Thus, this TH-I relating specific pheromone paths and ant foraging behaviours is not equipped to make use of a fitness-increasing encoding of some optimal features present in the extended phenotype, as it does not include self-modelling. Our next example, the genotype – phenotype relationship illustrates a TH-II, in which self-modelling plays a key role.

6.2.2. Genotype–phenotype relationship

The relationship between genotype and immediate phenotype involves a complex self-referential biological dynamics between the encoded genetic information and the organism itself in presence of environmental influences, as illustrated in Fig. 9. Genotype-phenotype relationship can be interpreted as a strange loop between the "self" (phenotype) and another object (genotype): the genotype "encodes" (models) the phenotype and is also "decoded" by the phenotype (see Box 1). With sexual reproduction, the fitness of a genotype is manifested through its phenotype. In turn, the fitness of a given phenotype, i.e., the organism's chances of survival and reproduction, varies across different selective environments [54]. Hence, we interpret genotype-phenotype relationship as a tangled hierarchy with self-modelling (TH-II).

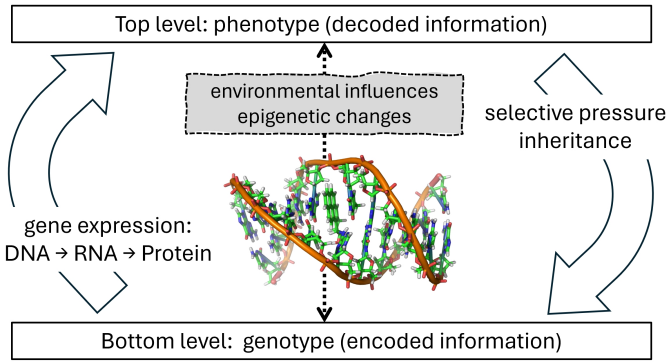

Figure 9: Genotype-phenotype relationship. Gene expression involving transcription and translation of encoded genetic information leads to the decoding of an organism with observable phenotypic traits (bottom-up construction). Given the phenotype, selective pressure and inheritance influence genotype over generations, updating the encoded genetic information which encodes the phenotype itself (top-down causation). Genotype-phenotype relationship is realised in context of environmental influences, epigenetic changes and other factors. DNA image: Wikipedia [118].

6.2.3. Phyllotactic patterning in plants

Another example of TH-I is the process of leaf positioning in plants. Leaves typically form at regular intervals at the plant apex, often creating recognisable spiral patterns. These patterns, however, are not directly encoded by the genome, but instead arise due to cell-cell interactions involving a feedback loop. The hormone auxin which triggers leaf outgrowth [89] is distributed by directional cellular transport [90]. Each cell assesses the concentration of auxin in neighbouring cells and transports auxin towards those neighbours in proportion to their auxin concentrations [49, 7]. This feedback, between auxin and its own transport, means that cells that happen to start with high concentrations receive even more auxin from neighbouring regions. However, at a certain distance defined by the relative strength of diffusion, another auxin peak forms and then another, as new space is created through cell divisions [49]. The resultant auxin distribution patterns are not explicitly encoded or compressed, and hence, cannot be replicated without replicating the entire system.

6.2.4. Evolution of self-modelling collective dynamics

The "evo-ego" relationship between the "embryonic" evolutionary units which synergistically interact and integrate information in producing their "adult" collective phenotypes [112] is an example of TH-II. As pointed out by Watson et al. [112], the relationship between the lower-level (individual) and higher-level (collective) units of selection is a *self-modelling* dynamical system in which integrated information patterns encode (compress) non-decomposable functions of input states (see Section 5.7). Thus, we again encounter a self-referential loop with self-modelling: "the key problem is that evolution is self-referential, i.e. the products of evolution change the parameters of the evolutionary process" [113].

6.2.5. Ecological scaffolding without self-modelling

As noted in Section 5.8, a tangled hierarchy comprising collective interests may shape *without self-modelling*, using ecological scaffolding, i.e., information about the evolutionary niche [9, 112]. In this case, there is no self-modelling, as the division of labour does not use any compressed or encoded information about distributed and dispersing resources, and hence, this is an example of TH-I.

6.2.6. Visual paradoxes

We now turn our attention to two well-known visual paradoxes, considering whether they may illustrate tangled hierarchies of different types.

The first paradox is an impossible staircase pattern (Fig. 10.Left), devised in 1937 by Oscar Reutersvärd [104], rediscovered in 1958 by Lionel Penrose and Roger Penrose [79], and artistically implemented in Escher's lithograph print "Ascending and Descending" (1960). In this pattern, the stairs are connected in an impossible way, i.e., the recursion is used within an impossible spatial configuration. Despite the appearance of a continuous staircase, there is an infinite loop of ascent and descent, so that the staircase appears to loop back on *itself*, creating an optical illusion of infinite repetition. However, there is no compression of any information pattern within a self-encoding object, and no self-modelling, suggesting that this is a tangled hierarchy of type I.

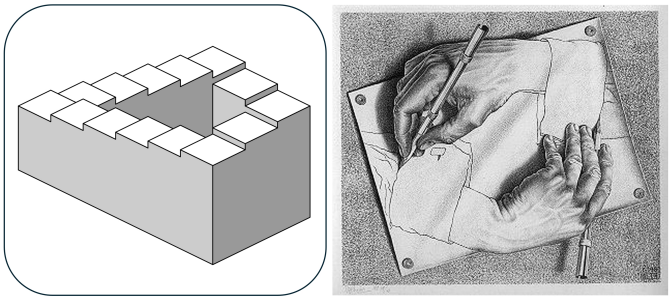

Figure 10: Left: "Impossible staircase" (1958) [79, 119]. Right: "Drawing Hands" (Escher, 1948) [120].

In "Drawing Hands" (1948), Escher created a visual paradox by depicting two intertwined hands drawing each other into existence (Fig. 10.Right). Each hand is in the process of perpetually drawing the other, creating a recursive loop of drawing an object which draws *itself*. This lithograph not only explicitly captures the idea of self-reference and the infinite recursive loop, but arguably depicts how one object "models" (constructs) the other.

Yet, there is no discernible compression of the model in "Drawing Hands" *per se*, and so this example can be called a tangled hierarchy of type II only with a stretch. Nevertheless, the notion of "self" is pronounced in this lithograph much more than in "Impossible staircase" and "Ascending and Descending", showing a transient stage towards fully-formed TH-II, in which self-encoding may be achieved by compression. In short, "Drawing Hands" lithograph may illustrate a tangled hierarchy positioned between the extremes of TH-I and TH-II.

6.3. Replication in tangled hierarchies

The distinction between the two types is also evident in different replication mechanisms available to dynamics within TH-I and TH-II.

In order to replicate information patterns and regularities emerging within TH-I (e.g., a pheromone path), the entire environmental state or complete dynamics needs to be reproduced (e.g., the entire pheromone field or the complete history of ant interactions from the initial state). Without compression and encoding, the information-processing is distributed throughout the entire system (e.g., stigmergy), and hence, needs to be replicated in every detail.

Computation-theoretically, in order to simulate dynamics unfolding within TH-I, one would require a *trivial* universal simulator which is defined as a simulator describing the collection of all particular solutions — in other words, it "needs access" to all corresponding Turing machines computing possible dynamic trajectories [39]. In this case, one may recall an insight of Brooks [14]: "the world is its own best model — always exactly up to date and complete in every detail".

The impossibility to replicate an optimal pheromone path emerging during ant foraging without reproducing the entire pheromone field (including the full pheromone gradient) reinforces our classification of this case as TH-I. If an optimal pheromone path was symbolically (i.e., digitally) encoded somewhere in the system, then it would have been possible to replicate just this encoding and expect that ants would access, "read-out" the encoded information, and rapidly rediscover the optimal connectivity.

An interesting but fatal phenomenon — "ant mills" — is observed when an environmental effect causes ants to lose the pheromone path and start following each other instead, potentially driving them in "death spirals" to exhaustion and death [93]. As pointed out by Delsuc [28], "the occasional but deadly formation of circular mills seems to be the evolutionary price that army ants pay to maintain such an ecologically successful and stable strategy of collective foraging". Specifically, this pathological behaviour can be explained by the ability of army ants to collectively select a raid direction [24]. Arguably, this behaviour arises when the tangled hierarchy lacks adequate means to explicitly encode and replicate additional information (e.g., a preferred raid direction). In effect, in the absence of an explicit pattern to follow, ants are reduced to follow the only regularity that remains (a circular path).

In contrast, information patterns and regularities emerging within TH-II can be replicated more efficiently. For example, to replicate a phenotype sufficiently well, one would need to copy/clone a genotype, and place it in a similar environment. This is ensured by the compression and encoding of relevant genetic information within DNA. Computationally, to replicate dynamics of TH-II, one needs a *non-trivial "singleton"* universal simulator which provides a universal solution via a single universal Turing machine [39]. This solution is more efficient than the trivial universal simulator needed to describe the collection of all trajectories generated by TH-I dynamics.

7. Emergence of self-modelling in tangled hierarchies

In principle, patterns emerging in TH-I (e.g., pheromone paths, hormone distribution gradients, ecological scaffolding) are compressible due to the presence of regularities. Once such compressed information becomes available to the system, it can then begin to process this information about *itself* — hence becoming a tangled hierarchy of type II. In this section we investigate the functional advantages enabled by such compression (i.e., the dimensionality reduction), which allows a self-referential system to be more predictive about its dynamics, exploiting the regularities more efficiently.

7.1. Emergence of functional self-descriptions

Having discussed the difference between tangled hierarchies of two types, we now consider the questions of how and why a functional self-description may emerge, and what functional advantages it could provide to TH-II relative to TH-I. It is worth pointing out that TH-I which

operates without self-modelling still can maintain a good level of stability and collective efficiency (e.g., ant colonies and other animal groups). And so one needs to examine what additional selective advantages would be offered by formation/emergence of self-modelling.

As discussed by McMillen and Levin [71], the higher levels of organisation bring distinct advantages:

- a different energy landscape that can be exploited by the collective;
- propagation of information across scales;
- a new problem-space and "extended patterns of information" available to the collective.

In short, the higher levels of organisation can favourably distort the energy landscape for their subunits and thus, provide a "causal architecture" enabling synergistic problem-solving competencies. These new competencies allow the collective "to navigate spaces of which the subunits are unaware" [71]. This causal architecture can be interpreted as an information hierarchy, tangling individual and collective dynamics.

Once "extended patterns of information" become available within a higher-level problem-space, a functional self-description may emerge to encapsulate the newly accessible regularities. Such self-modelling allows the information hierarchy to compress, encode and preserve beneficial information patterns in presence of stochastic environmental effects. The compressed self-description can then be decoded to recover the function disrupted by the adverse conditions.

For example, imprecise copying without an encoding was common during the evolutionary period dominated by lateral movement of genetic material via HGT. Woese [123] referred to results of such imprecise translation mechanism as "statistical proteins, proteins whose sequences are only approximate translations of their respective genes", while a consensus sequence for the various imprecise translations might closely approximate an exact translation of that gene. In contrast, precise copying using encoded/compressed information is a key feature of VGT, the vertical transmission of genes from parent to offspring. We can assume that during the period when HGT was the dominant form of gene transfer, the tangled hierarchies of type I could replicate only approximately, by imprecisely copying the whole organism.

As mentioned in Section 2.2, "coding threshold" has possibly separated the earlier evolutionary stage and the RNA world, providing the capacity to represent nucleic acid life symbolically, in terms of amino acid sequences [123]. Once some proto-codes emerged, encapsulating a prototypical self-descriptions of cellular dynamics, the Darwinian vertical descent based on VGT became the predominant mode of replication. Our conjecture is that this led to formation of the tangled hierarchies of type II.

Eventually, when the early proto-codes became interchangeable across the VGT replicators, a universal DNA-based code emerged, and the genotype-phenotype relationship fully formed. In other words, *replication robustness* may benefit from a universal code: it is more efficient to make multiple copies if they are described in a universal way, and this provides a selective pressure for the code universality.

Put simply, a functional self-description enables a more efficient replication process, preserving extended information patterns. Therefore, a compressed information about salient non-decomposable regularities — a "model" of the replicator — becomes codified in some symbolic (i.e., digital) form. Thus, information preservation within tangled hierarchies of type II which utilise self-modelling improves their structure and long-term function.

There are several possible mechanisms for emergence of functional self-descriptions within tangled information hierarchies. It could be triggered by a division of labour [47] which facilitates a split of competencies (i.e., symmetry breaking), followed by embedding or encapsulating some of these competencies within a proto-code. Alternatively, a perturbed biological system may be able to preserve information by finding suitable biochemical elements in its environmental locality and entrapping them as proto-code representing relevant features of its dynamics (e.g., attractors) [85]. As pointed out by Arthur [2], system's complexity may grow by capturing "software", that is, by capturing simpler elements and learning to "program" these as "software" in order to achieve its own goals.

In summary, functional self-descriptions and self-modelling emerge in response to some selective pressures and bring specific evolutionary advantages: robustness of replication, division of labour, preservation of non-decomposable collective dynamics, etc. The general motif is that self-referential and self-modelling dynamics improve efficiency of information-processing within a tangled information hierarchy (TH-II), at the cost of encoding and decoding of a compressed self-representation.

7.2. Synergistic fitness interactions exploit the discrepancy between "referents" and "expressions"

Our conjecture is that the evolutionary tension between individual and group interests [113, 115, 112, 71] is an example of a principled and generic "inconsistency", directly related to the expression-referent discrepancy discussed in Sections 3.2 and 5.2. Thus, we argue that the major evolutionary transitions in individuality exploited a divergence between the sizes of the expression and referent phase-spaces (i.e., problem-spaces).

We argue that the tangled hierarchy linking the individual and group levels exhibits the expression-referent discrepancy — the reason is that not all collective interests are reducible to a (sum of) individual preferences. This creates inconsistencies between individual and group interests ("individual-level selection will oppose the creation and maintenance of adaptations that enforce selection at

the group level" [115]). Computation-theoretically, our conjecture is that while individual interests are encodable via a given set of referents, the non-decomposable synergistic group interests are not necessarily expressible via the given referents. Thus, a disagreement between the individual and group interests can lead to either of two cases:

(i) in the case when "maximising the utility of the components rather than the collective" [113] dominates, then robustness and stability of the current setup is preferred (there is no transition);

(ii) otherwise, the tension is resolved by a transition where "coordinated phenotypic differentiation is then favoured and can thereby maximise collective fitness" [113], which in turn expands the phase-space.

8. Biological arrow of time as open-ended meta-simulation

In section 6 we introduced the distinction between tangled hierarchies with and without self-modelling. Then, in section 7 we argued that a functional self-description helps the TH-II system to innovate in two ways: by capturing and encoding (i.e., compressing) beneficial regularities and patterns in its dynamics, and preserving the resultant extended information patterns. This, in turn, facilitates a more efficient replication process.

In this section, we build on this premise and formulate our central conjecture, hypothesising that tangled hierarchies that emerge at various stages during biological evolution develop and expand in a continual, open-ended, process of self-referential dynamics and meta-simulation. This process encounters and then resolves undecidability by expanding the computational problem-space ("jumping out of the system") during transitions, proceeding according to the following steps:

1. *Emergence of tangled hierarchy without self-modelling.* The emergent TH-I (see Sec. 6) enables distributed coordination (e.g., pheromone paths emerging as a result of ant foraging and stigmergy) and imprecise replication (e.g., "statistical proteins"), but offers limited stability.
2. *Encapsulation of compressed functional self-description.* This transitional step identifies some regularity, i.e., information pattern, which brings evolutionary advantages (e.g., replication robustness), and then encodes this pattern within the lower-level of TH-I, providing it with a functional self-description. In general, this can be driven by a combination of exogenous and endogenous factors such as symmetry breaking (e.g., division of labour) and information preservation in presence of external noise.
3. *Formation of tangled hierarchy with self-modelling.* The formed TH-II (see Sec. 6) is capable of using encoded and decoded self-descriptions (e.g., a fully formed genotype-phenotype map). This has the following consequences:
   - *Self-reference.* Given the encoding and decoding capabilities, the formed expression-referent relationship exhibits duality, with expressions potentially becoming referents and vice versa, e.g., program-data duality where programs can be encoded as data and used as inputs (see Appendix A.1).
   - *Undecidability.* The expression-referent duality inevitably brings undecidability for a given system due to the expression-referent discrepancy: there are always more expressions (objects) than referents (encodings), and some expressions (possible objects) are inaccessible by the system — essentially, this is a diagonalisation argument. In the context of genotype-phenotype relationship, there are always more possible phenotypes than the genotypes available under a particular encoding scheme.
4. *Extension of the problem-space*, due to the expression-referent discrepancy. In a biological context, this enables organisms interacting with their environment (given their current niche), to access and exploit a new problem-space and "extended patterns of information" [71], i.e., to perform meta-simulation and resolve current inconsistencies (mismatches) by better fitting the environment [99].

In general, the extensions described in step (4) are not meant to be abrupt (notwithstanding the computation-theoretic analogy of Turing jumps), and may develop by a slow accumulation of lower-level changes within subunits adapting over time, until a tipping point.

The entire process involves bottom-up emergence, self-modelling, non-decomposable collective fitness (i.e., information integration) [112], and "information self-creation" [41]. These key features shape the tangled information hierarchies in a way that generates computational novelty. Informally, at a major transition, the collective dynamics may need different "symbols" (a new "code") to efficiently describe non-decomposable dynamics, relative to the symbols currently available to the individual subunits.

In general, this process follows a directional spiral, returning after step (4) to step (1) and forming a new TH-I, initially without self-modelling. Resolving a tension expands the problem-space which allows the extended system to access a new, more complex, landscape of novel (collective) possibilities. However, the new landscape brings about a new discrepancy (e.g., new biological contradiction, that is, undecidability), thus continuing the process in an open-ended irreversible way, represented by the biological arrow of time.

We emphasise that the open-ended computational meta-simulation, described in subsection 3.4, provides a computation-theoretic interpretation — a kind of semantics — rather

than a specific mechanism employing Turing oracles and Turing jumps. In biological systems, oracles do not have to exist *per se* — instead, higher-level tangled hierarchies emerge by capturing synergistic information patterns within the expanded phase-spaces. In other words, once a discrepancy between expressions and referents is resolved at a given level, the oracle's job is already completed. In summary, the computation-theoretic interpretation of open-ended meta-simulation offers a unifying perspective on diverse real-world dynamics and specific biological processes that exhibit temporal asymmetry.

9. Discussion and conclusion

In this study we we interpreted the phenomenon of open-ended biological complexity as a dynamic computational process and proposed a computation-theoretic argument for the biological arrow of time. Our argument follows Gödel–Turing–Post recursion-theoretic framework which formalises the construction of extensible computational systems such as Turing α-oracle machines. We proposed that this open-ended generation of computational novelty involves meta-simulation performed by higher-order systems that successively simulate the computation carried out by lower-order systems. Essentially, this open-ended meta-simulation provides solutions to the undecidable problems encountered by the lower-order "subunits" and hence, expands the effective phase-space of possibilities.

Before concluding, here we briefly reflect on several studies which proposed fundamental principles for open-ended evolution and biological complexification, and point out connections with our approach.

9.1. Evolutionary role of expanded genomes

Heng and Heng [41] discussed (species-specific) karyotype or chromosomal coding, arguing that

> "If a change in karyotype coding generates new information at the system level, then an altered karyotype contains the necessary information for the emergence of a new genome system – the necessary information for macroevolution" [41].

While some changes in karyotype may promote macroevolution, larger genome or karyotype changes are unlikely to be sufficient for it on their own. Nevertheless, these changes may provide a pre-condition: a large random jump in the fitness landscape, followed by a period where the organism is able to explore locally within the landscape more easily than before.

The evolutionary role of an expanded genome is also emphasised by Bingham and Ratcliff [8] who reported an association between eukaryotic genome duplication and the evolution of multicellularity. This was contrasted with ancestral prokaryotes which tended to lose rather than accumulate DNA, and this may have prevented a transition of prokaryotes to multicellularity. We may interpret this difference between eukaryotic and prokaryotic genomes as the difference in their abilities to form a larger phase-space by adding more "letters" to their encoding schemes — essentially, as the difference in their *self-modelling* abilities.

Interestingly, Bingham and Ratcliff [8] also highlighted that eukaryotic cells have a well-developed ability to deal with parasitic elements which form a significant part of eukaryotic genomes. This ability exemplifies an inconsistency-resolving element which may have also facilitated a transition to multicellularity.

One may draw a parallel between parasitic elements (invaders) and the "contrarian" agents (i.e., Liar agents, in the sense of Liar paradox) which exploit the gaps in self-descriptions, e.g., computer viruses that change the host code to generate the outcome opposite to the intended computation [66], or novel antigens disrupting the "self vs non-self" distinction within autoimmune system and causing autoimmune diseases [36]. Markose [69] insightfully pointed out that the knockout of specific auto-immune regulators leads to the loss of self-representation for certain self-gene codes within the autoimmune system, thus generating autoimmune pathologies. These examples illustrate how the ability to implement negation contributes to generation of inconsistencies within a tangled information hierarchy with self-modelling.

9.2. Increasing "dynamic kinetic stability"

In an attempt to reformulate Darwinian theory of evolution and extend it to inanimate matter, Pross [86] introduced the principle of Dynamic Kinetic Stability (DKS). The DKS principle, expressed in physicochemical terms, favours stable kinetic *patterns* balancing the rates of replication and decay (e.g., during autocatalytic reactions). This approach emphasised that certain relatively simple self-replicating systems (e.g., single molecules, oligomeric sequences comprising more than one subunit, minimal molecular networks) may have used imperfect replication while evolving toward replicating systems of greater DKS. Crucially, *autocatalysis* and *cooperative behaviours* were identified as common features of both abiogenesis (i.e., life emerging from nonliving matter) and biological evolution:

> "...cooperative behavior can emerge and manifest itself at the molecular level, that the drive toward more complex replicating systems appears to underlie chemical, and not just biological, replicators.
> ...life's emergence began with the chance appearance of some relatively simple replicating chemical system, which then began the long road toward increasingly complex replicating entities." [86].

These elements — higher-order regularities (i.e., collective or cooperative behaviours) emerging out of interactions of imperfectly replicating subunits; self-referential (autocatalytic) reactions; and the stability of replication dynamics — can also be interpreted in terms of tangled hierarchies which continually expand their problem-spaces.

9.3. Increasing "functional information"

Wong et al. [128] studied the roles of function and selection in evolving systems, and identified three universal evolutionary mechanisms utilising information about the system–environment interactions: static persistence, dynamic persistence, and novelty generation. Information was interpreted as "patterns of data in a system that encodes about itself, its environment, or about its relation with its environment", while functions were interpreted as processes that have "causal efficacy over the internal state of a system or its external environment" [128]. The study proposed a general law of increasing *functional information*:

> "The functional information of a system will increase (i.e., the system will evolve) if many different configurations of the system undergo selection for one or more functions" [128].

Among core functions capable of perpetuating themselves, such as dissipation and homeostasis, this approach also highlighted that self-replicating systems, including living systems, are necessarily autocatalytic, and drew an analogy with the DKS framework. Going beyond the DKS approach, Wong et al. [128] identified *information-processing* as another self-perpetuating core function. The described information-centric account distinguished different kinds of dynamical persistence with respect to the distinct levels at which information patterns contribute to persistence. In particular, the study considered (i) information storage: "memory" which allows for *encoding* associations, (ii) information inference of future states based on encoded memory: "memory-based prediction" which provides a *causal model* and improves persistence, and (iii) counterfactual reasoning: "prediction outside of memory" which generates *novelty* through imagining previously nonexistent versions of reality [128].

The increasing dynamic kinetic stability and the increasing functional information reflect the arrow of time. We contend that both these principles can be subsumed by the computation-theoretic Gödel–Turing–Post characterisation of the open-ended meta-simulation by systems that expand their problem-spaces in the search of ways to resolve specific contradictions and tensions. In particular, "counterfactuals" which are required to generate novel functional information [128] can be formed only by considering negation, typically in context of the Liar paradox.

9.4. Assembly theory and the "adjacent possible"

A recent proposal on "assembly theory" (AT) also attempted to explain and quantify selection and evolution in the context of novelty generation:

> "This approach enables us to incorporate novelty generation and selection into the physics of complex objects. It explains how these objects can be characterized through a forward dynamical process considering their assembly. By reimagining the concept of matter within assembly spaces, AT provides a powerful interface between physics and biology. It discloses a new aspect of physics emerging at the chemical scale, whereby history and causal contingency influence what exists" [94]

In assembly theory, an object is defined through its possible formation histories in an "assembly space", so that "objects are made by joining elementary building blocks together recursively to form new structures" [32]. Essentially, given the building blocks available at the time, the assembly space is a problem-space that comprises possible pathways for assembling an object.

This view can be compared with the concept of the "adjacent possible" proposed by Kauffman [50]: the set of all potential configurations that are just one step away from the current state of a system. These possibilities are constrained by the existing components, structures, or knowledge of the system. Thus, the innovations arise from the existing system's state and the adjacent possibilities that are immediately accessible from that state.

The assembly theory and the "adjacent possible" concept identified a discrepancy between the space of objects that can be constructed using actually accessible blocks, and the space of objects that are conceivable. Computation-theoretically, this discrepancy can be interpreted as the expression-referent discrepancy which generates contradictions, and forces an expansion of the problem-space by meta-simulating and discovering new (assembly) pathways.

9.5. Social dynamics and undecidability

It is likely that social complexity increases in an open-ended way as well, along a socio-biological time arrow. However, the social complexification is out of scope for this study. We briefly note, following other similar hypotheses [2, 123, 38, 71, 68], that the emergence of tangled hierarchies with functional self-descriptions and self-modelling may drive novelty generation in the evolution of language and culture.

In particular, one may conjecture that grammar provides a functional self-description of natural language, codifying various linguistic elements and rules, as well as their interpretation. Similarly, we may interpret social institutions, such as traditions, norms, conventions, laws, legal frameworks, as functional self-descriptions of society. These social institutions encode and encompass the organised and established patterns of human behaviour and relationships.

The grammars and institutions which encode the prevailing regularities are always short of describing all possible scenarios which may unfold due to interactions with external environment. These linguistic and social discrepancies between "expressions" (linguistic or social traits) and "referents" (codified grammatical or institutional rules) inevitably lead to mismatches, inconsistencies and contradictions. In turn, resolution of these tensions necessitates a