Biosemiotics
https://doi.org/10.1007/s12304-024-09554-1

RESEARCH

Efficiency in Organism-Environment Information Exchanges: A Semantic Hierarchy of Logical Types Based on the Trial-and-Error Strategy Behind the Emergence of Knowledge

Mattia Berera[1]

Received: 3 November 2023 / Accepted: 26 January 2024

Abstract
Based on Kolchinsky and Wolpert's work on the semantics of autonomous agents, I propose an application of Mathematical Logic and Probability to model cognitive processes. In this work, I will follow Bateson's insights on the hierarchy of learning in complex organisms and formalize his idea of applying Russell's Type Theory. Following Weaver's three levels for the communication problem, I link the Kolchinsky–Wolpert model to Bateson's insights, and I reach a semantic and conceptual hierarchy in living systems as an explicative model of some adaptive constraints. Due to the generality of Kolchinsky and Wolpert's hypotheses, I highlight some fundamental gaps between the results in current Artificial Intelligence and the semantic structures in human beings. In light of the consequences of my model, I conclude the paper by proposing a general definition of knowledge in probabilistic terms, overturning de Finetti's Subjectivist Definition of Probability.

Keywords Type theory · Communication theory · Semantic efficiency · Cognitive processes · Cybernetics

Introduction

One of the historical problems of philosophy concerns the emergence of knowledge, meaning to find out whether the link between sensible experience and abstraction exists and where it lies. Current and fruitful neurological approaches frequently fall victim to a materialistic reductionism that neglects the complexity of cognitive phenomena (for instance, see Velazquez, 2020). On the other hand, every metaphysical one needs to postulate some external cause solely justifiable by faith (for instance,

✉ Mattia Berera
matti.berera@gmail.com; mattia.berera@edu.unito.it

[1] Dipartimento di Matematica Giuseppe Peano, Università degli Studi di Torino, Torino, Italy

Published online: 12 March 2024

see Achella, 2022). Within a semiotic interpretation of life (see Barbieri, 2008), I propose an understanding of the problem by lying within the coordinates of Darwin's teaching. That is, within a dialectic of nature that seeks the causes of phenomena in the relationships among phenomena themselves (for instance, see Bishop, 2022). I choose to base the argument on what I believe to be the most recent and complete version of such an understanding of nature, namely on those interpretations of Physics according to which the fundamental process shaping reality is nothing but a continuous and plural exchange of information between physical systems; in particular, I embrace the interpretation provided by Rovelli (1995, 2015) through his Relational Quantum Mechanics (RQM). For the soundness of such a starting hypothesis, it is good to clarify a notion of information broad enough to model all kinds of interactions between any physical systems. Shannon's classical 'measure of information' (1948) provides a solution for this purpose, according to which any information exchange – i.e., any communication defined by its constraints – is more informative as the greater the number of possible messages it could convey. Thus, Shannon describes the amount of information of a random source X with probability distribution p as the entropy in Statistical Mechanics:

$$\mathcal{H}(X) = -\sum_{x} p(x) \log p(x) \,. \tag{1}$$

In this framework, the question arises of when and how one can define some information exchanges as cognitive processes.

According to common sense, a **cognitive process** is any conscious and unconscious processes by which knowledge is accumulated, such as perceiving, recognizing, conceiving, and reasoning (for instance, see https://www.britannica.com/topic/cognition-thought-process). Before giving a formal definition of 'knowledge,' I can start by stating a general point. Any knowing subject is a physical system in the first place, and according to RQM, the exchange of information between systems and their environment is an inevitable constant process. Thus, any notion of knowledge requires some distinction to be definable in the totality of exchanged information to identify the only information that should be 'treated as knowledge,' regardless of the sense of such an expression. More specifically, for any notion of knowing not to be empty, the information conveying knowledge must possess meaning in some sense, which is to say, it is necessary to achieve some definition of semantics. A property of information should exist relative to each knowing subject, by which it can distinguish 'meaningful' from 'not-meaningful' exchanges. From a gnoseological point of view, such a characteristic should be endogenous to the system and independent of its history to avert studying the results of some systems' behavior a posteriori. Identifying intrinsic, non-etiological semantics is crucial in investigating the foundations of cognitive processes without arguing in favor or against some interpretation of them.

In the following, I will treat knowing subjects as biological systems; however, some results of the reasoning could provide a framework for better understanding some gaps between Artificial Intelligence and humans (see the end of "A Hierarchy of Semantic Efficiency"). One can thus frame the work I present in the context

of Cybernetics, the discipline inaugurated by Norbert Wiener in the 1940s, which aims to mathematically study the "essential unity of the set of problems centering about communication, control, and statistical mechanics, whether in the machine or in living tissue." (Wiener, 1965, 11) The reader may find a biosemiotic framing of Cybernetics in Brier (1999, 2013), and a proposal for a synthesis of Biosemiotics and Cybernetics in Sharov (2010). The key feature of such a framework concerns the attention given to the self-regulating mechanisms of systems, called **feedback loops**; mechanisms whereby if

> we desire a motion to follow a given pattern the difference between this pattern and the actually performed motion is used as a new input to cause the part regulated to move in such a way as to bring its motion closer to that given by the pattern. (Wiener, 1965, 6)

As physical systems, organisms possess some peculiar characteristics that one should consider in the following reflections. By introduction, I give an excerpt from the pages in which Erwin Shrödinger paves the way for the Thermodynamic Theory of Living Systems. The Austrian physicist asks himself:

> What is the characteristic feature of life? When is a piece of matter said to be alive? When it goes on doing 'something', moving, exchanging material with its environment, and so forth, and that for a much longer period than we would expect an inanimate piece of matter to 'keep going' under similar circumstances. When a system that is not alive is isolated or placed in a uniform environment, all motion usually comes to a standstill very soon as a result of various kinds of friction. [...] The physicist calls this the state of thermodynamical equilibrium, or of 'maximum entropy'. [...] It is by avoiding the rapid decay into the inert state of 'equilibrium' that an organism appears so enigmatic [...]. How does the living organism avoid decay? The obvious answer is: By eating, drinking, breathing and (in the case of plants) assimilating. The technical term is metabolism. The Greek word [...] means change or exchange. Exchange of what? (Schrödinger, 1944, 69-70)

Due to the Second Law of Thermodynamics, like any other physical system not in equilibrium,

> a living organism continually increases its entropy - or, as you may say, produces positive entropy - and thus tends to approach the dangerous state of maximum entropy, which is death. It can only keep aloof from it, i.e. alive, by continually drawing from its environment negative entropy [...] What an organism feeds upon is negative entropy. Or, to put it less paradoxically, the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive. (Schrödinger, 1944, 71)

I recall that, in Statistical Mechanics, the Second Law of Thermodynamics is a theorem stating that in an isolated system, entropy is a non-decreasing function of time and that it follows from the ergodic hypothesis that we cannot a priori exclude any possible state from the probable states in which the system may be.

Although there is no agreement on whether or not Schrödinger's approach is comprehensive in defining life, it undoubtedly represents a necessary framework. So an organism is some open thermodynamic system – i.e., one that exchanges energy and matter with the environment – that possesses mechanisms keeping it away from thermodynamic equilibrium, that is, inner devices for constantly lowering its entropy. By the Second Law of Thermodynamics, processes triggered by these mechanisms always involve a thermodynamic cost to the system.

This article is organized as follows. "Problems Concerning Viability" explains the starting semantic model, according to Kolchinsky and Wolpert, and identifies the relationships between semantic and syntactic aspects of information exchanges between organisms and the environment. "Semantic Efficiency and Learning" is devoted to linking the semantic problem to that of learning and arguing the need for a hierarchical model of types in the formulation of the concept of learning in complex organisms, following Bateson's insight. In "A Hierarchy of Semantic Efficiency", I propose a hierarchical model to represent concept production in a complex organism and show how such a model contradicts current research in Artificial Intelligence. In the fifth and final section, I suggest a probabilistic model for knowledge and argue in favor of the latter by explaining the connection with the semantic hierarchy model constructed in the previous section. In the Appendices, I offer a schematic description of the two major mathematical theories employed throughout the paper; in A, respectively, I discuss Kolchinsky and Wolpert's model, and in B, Russell's Type Theory in the Wiener-Kuratowski proposal.

Problems Concerning Viability

Since in Shannon's Communication Theory, one can identify the amount of information exchanged by a physical system with its entropy, the thermodynamic equilibrium state – i.e., of maximum entropy – is thus representable in information-theoretic terms as the existence of many more ways of being in a dead state than otherwise (for instance, see Summers 2023). Therefore, I choose to follow the idea of Kolchinsky and Wolpert (2018) of defining the meaningfulness of an exchange of information between the system and the environment as the outcome it has in keeping the system 'alive.' One can find fruitful developments from their approach in the fields of thermodynamic interpretation of life's evolution (Jeffery et al., 2019), robotics and artificial life (Roli–Kauffman, 2020), biosemiotics applied to economics (Herrmann-Pillath, 2021); a similar proposal grounded on Friston's Free Energy Principle is in Kiverstein et al. (2022).

The two authors introduce a non-negative real-valued function of time describing the system's 'degree of survival' and study how such environmental information contributes to keeping this function around a given value. From Schrödinger's considerations, it naturally follows that a good choice for such a function is the negative of entropy, i.e., the opposite of Shannon's measure of information.

If one considers a joint system organism-environment (X, E), whose evolution is associated with some probability distribution, one can define the **viability function** of X at time τ as the quantity

$$\mathcal{V}(X_\tau) := \sum_{x_\tau} p(x_\tau) \log p(x_\tau),$$

where p is the marginal distribution of X, and x_t is a particular outcome of random variable X_t representing the system X at time t (see Kolchinsky and Wolpert, 2018, 6). From now on, I identify a physical system with its state space; the continuous case follows, as Shannon shows in Part III of his fundamental article (1948).

A further probabilistic reason for such a choice of the $\mathcal{V}$ function is provided by Kolchinsky and Wolpert themselves: in fact,

> entropy provides an upper bound on the amount of probability that can be concentrated in any small subset of the state space X [...] this is relevant because there is often a naturally defined 'viability set' [...], which is the set of states in which the system X can continue to perform self-maintenance functions. Typically, the viability set will be a very small subset of the overall state space [...] If the entropy of system X is large and the viability set is small, then the probability that the system state is within the viability set must be small [...] Thus, maintaining low entropy is a necessary condition for remaining within the viability set (Kolchinsky and Wolpert, 2018, 6).

I schematically discuss in Appendix A the Kolchinsky and Wolpert model for its parts needed in the following reasoning. From these, one can define the **initial viability value** of mutual information $\mathcal{I}(X;E) = \mathcal{H}(X) - \mathcal{H}(X|E)$ between X and E as follows:

$$\Delta\mathcal{V}(X_\tau) = \mathcal{V}(X_\tau) - \mathcal{V}(X_\tau^F), \tag{2}$$

where X_τ^F is the system X at time $t = \tau$ if X evolved with distribution independent from the distribution of E.

> The difference [$\Delta\mathcal{V}$] can also be negative, which means that the [...] information decreases the system's ability to exist. This occurs if the system behaves 'pathologically', i.e. it takes the wrong actions given available information (Kolchinsky and Wolpert, 2018, 3).

The definition of initial viability value only provides a measure of the 'significance' of the information for time $t = 0$ but still does not allow for discerning the meaningful information within it. By some interventions summarized in the appendix, it is possible to identify an optimal joint distribution $p^{opt}_{X,E}$ of the system (X, E) with which to give the following two definitions. One can call the **amount of semantic information** that X has about E at time $t = 0$ the mutual information at time $t = 0$ between E and X if the joint system evolve with distribution $p^{opt}_{X_0,E_0}$, i.e.

$$\mathcal{S}(X_0;E_0) := \mathcal{I}(X_0^{opt};E_0);$$

and **pointwise semantic information** of a state x_0 of X over one state e_0 of E at time $t = 0$ the quantity

$$\mathcal{S}(x_0;e_0) := \log \frac{p^{opt}_{X_0,E_0}(x_0, e_0)}{p^{opt}_{X_0}(x_0)p^{opt}_{E_0}(e_0)} .$$

By extending the reasoning so far, it is possible to measure the semantic information acquired by the system X in the dynamic interaction with the environment at a given time interval (see Kolchinsky and Wolpert, 2018, 10-11). The latter situation models what we can interpret as a stochastic behavior 'by trial and error' of a system in its environment. An example is exhibited by

> a chemotactic bacterium, which makes ongoing measurements of the direction of food in its environment, and then uses this information to move towards food. (Kolchinsky and Wolpert, 2018, 7)

Given these initial ingredients, it is worth paying attention to the relationships within a communication between the pure syntactic correctness of the message and its semantic understanding. In defending Shannon's Theory, Warren Weaver – one of its co-authors – preliminarily distinguishes the communications problem into three sub-problems:

Level A How accurately can the symbols of communication be transmitted? (The technical problem.)

Level B How precisely do the transmitted symbols convey the *desired* meaning? (The semantic problem.)

Level C How effectively does the received meaning affect conduct in the *desired* way? (The effectiveness problem.)

(Weaver, 1949, 262, italics added)

As well-known, the mathematical Theory of Communications only deals with the technical aspects concerning Level A; however, Weaver himself states that not only do the limitations found in the solutions to the first problem have effects on the solutions to the other two, but that most likely a theory that accounts for the technical-syntactic problem is a sufficient theory to address the semantic problem and the subsequent one that concerns effectiveness as well (see Weaver, 1949, 277 et seq).

Kolchinsky–Wolpert model – along with the classical semantic models developed by game theory and economic informatics (for instance, see Morgenstern and von Neumann, 1944, and Polani et al., 2001) – argue in favor of Weaver's conjecture. Care must be taken, however, with the term 'desired' used by Weaver; in Kolchinsky and Wolpert's framework, it is not permissible to assume the existence of a shared meaning between the living system and the environment but only of semantics

intrinsic to each organism. So the possibility arises of overlooking what Weaver (1949, 279) calls the **semantic decoding problem**:

> the *semantic problems* [that] are concerned with the identity, or satisfactorily close approximation, in the interpretation of meaning by the receiver, as compared with the intended meaning of the sender. (Weaver, 1949, 262)

Those problems are to be understood in this context as problems concerning the receiver's free interpretation according to some of its intrinsic discriminants, free except for some phylogenetic constraints. These constraints depend within the model on choosing the $\mathcal{V}$ viability function that discriminates semantics: Weaver's semantic problem is thus expressible as follows.

Level B′. How precisely do the transmitted symbols convey *semantic content*?

In more formal terms, the semantic problem is equivalent to measuring the ratio of semantic information over total mutual information, which Kolchinsky and Wolpert (2018, 8) call **semantic efficiency**:

$$\eta := \frac{\mathcal{S}}{\mathcal{I}} \in [0, 1]\,. \tag{3}$$

At this point, I can consider the effectiveness problem in light of the proposed solution to the semantic problem. According to Weaver again,

> the *effectiveness problems* are concerned with the success with which the meaning conveyed to the receiver leads to the desired conduct on his part. It may seem at first glance undesirably narrow to imply that the purpose of all communication is to influence the conduct of the receiver. But with any reasonably broad definition of conduct, it is clear that communication either affects conduct or is without any discernible and probable effect at all. (Weaver, 1949, 263)

Yet, within the present model, the adjective 'desired' cannot be loosely associated with the 'conducts' of the receiver either, since one cannot assume a conscious desiring agency by the environment on the living system. Therefore, one should intend as desired those behaviors of the system that ensure its viability and must reformulate Weaver's effectiveness problem as follows.

Level C′. How effectively does the *semantic content* affect conduct in a way *that ensures viability*?

This problem is not trivial in our framework since one of the distinctive features of semantic information is that it

> should be able to be 'mistaken', i.e. to 'misrepresent' the world. This emerges naturally in our framework whenever information has a negative viability value (i.e. when the system uses information in a way that actually hurts its ability to maintain its own existence). (Kolchinsky and Wolpert, 2018, 12)

Here, I am assuming something that may appear unwarranted: the idea that the semantic we are considering is the only one that stimulates the organism's interactions with the environment. Namely, I am deliberately neglecting situations in which

the organism attaches meaning to information in ways not directly related to maximizing its survival, for instance, when it sacrifices itself to ensure others' survival. Such constraint at this level should be regarded as necessary though non-sufficient to represent the knowing subjects and their many possible responses in the face of the information they exchange with the environment.

Back for a moment to the Second Law of Thermodynamics: every interaction between the system and the environment – i.e., every process of exchanging information – is not for free from the thermodynamic point of view, i.e., obtaining information about the environment requires some system's work.

> If this work were not spent acquiring the initial mutual information, it could have been used at time τ to decrease the entropy of the system, and thereby increase its viability, by $[\mathcal{I}(X_0;E_0)]$ [...] The benefit of the mutual information is quantified by the viability value $[\Delta\mathcal{V}]$, which reflects the difference in entropy at time $t = \tau$ when the system is started in its actual initial distribution $[p_{X_0,E_0}]$ versus the fully scrambled initial distribution $[p^F_{X_0,E_0}]$. (Kolchinsky and Wolpert, 2018, 9)

Therefore, it makes sense to define the **benefit/cost ratio** of the mutual information

$$\kappa := \frac{\Delta\mathcal{V}}{\mathcal{I}} = \eta\frac{\Delta\mathcal{V}}{\mathcal{S}}\,, \tag{4}$$

which measures a system's ability to use information about the environment to maintain its viability. From the second equation in Eq. 4, in cases where the initial viability value of the mutual information is positive – that is, if the organism behaves in a way that maximizes its viability – one gets that having a low semantic efficiency η means to have a low benefit/cost ratio κ.

In an intuitively understandable way, the amount of information a living system exchanges with its environment is directly proportional to its complexity, definable as its "number of factors which are interrelated into an organic whole" (Weaver, 1948, 539). But the same cannot be said of its meaningful part. One can expect that the amount of meaningful information will give an efficiency $\eta \ll 1$ for large amounts of information exchanged. For a complex living system, trial-and-error behavior is inefficient at maximizing its viability function. It involves a very high exchange of information from which the organism gets a low yield of meaningfulness. Insofar as mentioned above, this also means that the semantic information embodied in the information exchanged will have little effect on the behaviors that ensure the system's viability.

Semantic Efficiency and Learning

Starting from this point, I shall leave aside quantitative modeling to focus on a qualitative study of the adaptive mechanisms of semantic efficiency. To the current state of my knowledge, no approach in literature has dealt with the hierarchy I will construct in the following with quantitative methods. The reader may

find suggestions to this effect in Rovelli (2018) and Surov (2022). Exploring such potential developments remains for future works.

Understanding how the behaviors of an organism are affected by interactions with the environment, i.e., studying the Level C′ problem, falls under the investigation of those processes referred to in the literature as 'learning.' For common sense, a **learning process** is any change in a living system's behavior due to an exchange of information with its environment (for instance, see https://www.britannica.com/science/learning). From Eq. 4, one finds that the effectiveness of information exchange in preserving the viability of an organism is related to how that organism renders more efficient attribution of meaning to the information it exchanges with the outside world. Hence, studying an organism's learning processes should be related to studying its strategies for making its exchanges of information semantically efficient. Following the thoughts of cyberneticist Gregory Bateson, Evolutionary Biology tells us that

> all biological systems (organisms and social or ecological organizations of organisms) are capable of adaptive change. [...] Whatever the system, adaptive change depends upon *feedback loops*, be it those provided by natural selection or those of individual reinforcement. In all cases, then, there must be a process of *trial and error* and a mechanism of *comparison*. (Bateson, 1969, 278)

As argued above, a trial-and-error strategy is all but efficient beyond a certain threshold of organism complexity; therefore, according to Bateson, in a complex organism,

> there is needed not only that first-order change [in its behavior] which suits the immediate environmental (or physiological) demand but also second-order changes which will reduce the amount of trial and error needed to achieve the first-order change. And so on. By superposing and interconnecting many feedback loops, we (and all other biological systems) not only solve particular problems but also form *habits* which we apply to the solution of *classes* of problems. (Bateson, 1969, 279)

Bateson's idea is that it is possible to construct a type hierarchy among the different ways of learning to explain the strategies that organisms have adaptively matured to optimize their environmental interaction processes. Yet, for modeling of ethological problems through Russell's Type Theory (see Whitehead and Russell, 1963) to be permissible, it is a matter of testing "whether the distinction between a *class* and its *members* is an ordering principle in the behavioral phenomena which we study" (Bateson, 1968, 287). Russell says that

> the division of objects into types is necessitated by the vicious-circle fallacies which otherwise arise. These fallacies show that there must be no totalities which, if legitimate, would contain members defined in terms of themselves. Hence any expression containing an apparent variable must not be in the range of that variable, *i.e.* must belong to a different type. (Whitehead and Russell, 1963, 161)

As Bateson states (see Bateson, 1968 and 1969), ethological evidence shows that, in the case of some organisms, one can not use predicative propositions that describe behavioral modules to explain the categories one uses to group those behavioral modules. Namely, a property can be simultaneously verified by a behavior of the organism and contradicted by a class of its behaviors to which that behavior belongs.

For example, an organism X that meets an unknown object Y will start exploring it through a trial-and-error method. According to Kolchinsky–Wolpert model, one can say that interaction is successful if and only if the viability value $\Delta\mathcal{V}$ of the information X obtains about Y is positive, where $\Delta\mathcal{V} > 0$ characterizes what is called in ethology a **positive reinforcement**. It implies that if interaction with Y negatively affects $\mathcal{V}(X)$, the organism is supposed to treat the interaction as a failure. But this description is not adequate in the case of some organisms; for example, mammals learn from their mistakes, and therefore, if X were to face an object $\tilde{Y}$ similar to Y in the future, it would know how to keep away from it without having to expend the thermodynamic cost of interaction. Yet, X will not refrain from behaving again by trial and error when faced with new unknown objects, namely, when faced with objects that it will categorize as different from Y.

This means that a failure at the level of the interactions results in success at the level of exploration – meaning learning – and that such success does not change the general pattern of future behavior. Thus, if we intend to model with a formal system the learning modes of X, we must consider that given an interaction (x_t, e_t) between X and the environment E, it must be possible to consider some function v that when interpreted yields

$$v := \text{‘to be successful’}$$

and such that $v\big((x_t, e_t)\big)$ is simultaneously true, false, and undecidable. However, we want our system to be consistent, so we need to guard against the possibility that such functions might exist. So, we need to hierarchically distinguish propositions referring to interactions from propositions referring to learning that results from those interactions, precisely what allows us to model the problem via Russell's Type Theory.

In Appendix B, I define a formal system that I call Type Theory (**TT**), grounded on the following two schemes of axioms:

T1. is the Principle of Abstraction – stating that for every property, there exists a set whose elements are precisely the elements for which such a property holds – constrained to apply, for every type, only for sets whose elements are tokens of the preceding type;

T2. is the Extensionality Principle – stating that two sets are identical if and only if the same elements fall under their domain – constrained to hold only within any given type.

Any interpretation of **TT** is called **type hierarchy**: it is a recursive collection $\mathcal{T}(D_0)$ generated by a non empty set D_0, consisting of:

- the elements of D_0, which are said to be of **type 0**;
- the elements of $D_{n+1} := \mathscr{P}(D_n) = \mathscr{P}^{n+1}(D_0)$, which are said to be of **type $n+1$**, for all $n \in \mathbb{N}$;

where $\mathscr{P}$ is the power-set operator (see Hatcher, 1982, Section 4.6). It is easy to prove (see Hatcher, 1982, 107) that a type hierarchy is uniquely determined by the cardinality of D_0, regardless of the nature of its elements, and that all items of $\mathcal{T}(D_0)$ are relations that can be constructed on the elements of D_0. Yet, only those relations whose arguments are constrained to take values among the tokens of a given type belong to $\mathcal{T}(D_0)$.

Back to the exploration-in-mammals example, Bateson applies Russell's theory as follows:

(a) Changes in frequency of items of mammalian behavior can be described and predicted in terms of various 'laws' of reinforcement.
(b) 'Exploration' [...] is a category, or class, of mammalian behavior. [...]
(c) [...] If, as asserted in (b), 'exploration' is not an *item* of mammalian behavior but is a *category* of such items, then no descriptive statement which is true of *items* of behavior can be true of 'exploration.' If, however, descriptive statements which are true of items of behavior are also true of 'exploration,' then 'exploration' is an item and not a category of items.

(Bateson, 1968, 286-287)

In general, we can assume that an organism's behavior is modeled as a trajectory γ from time $t = 0$ to some time $t = \tau$ in the metric space X representing its state space:

$$\begin{aligned} \gamma \,:\, [0,\tau] &\longrightarrow X \\ t &\longmapsto x_t \end{aligned} .$$

Here, I am interested in formalizing the common notion of learning, that is, changes in X behavior due to interactions with the environment E; a behavior change is thus a situation in which

$$p_{X_\tau}(x_\tau) \neq p_{X_\tau|X_0,E_0}(x_\tau|x_0,e_0)\,.$$

In particular, I first must represent the most elementary changes that occur in X behavior after an interaction with E, that is, the immediate reactions of an organism when it receives a significant stimulus.

Definition 1 Given an organism X at time t_1, an instant $t_0 = 0$ in its past, and two possible behaviours γ_1, γ_2 of X such that $\gamma_1(t_0) = \gamma_2(t_0)$, and $\gamma_1(t_1) \neq \gamma_2(t_1)$, I call **0-learning**, or a **simple change** in the behavior of X, the distance

$$\ell := \operatorname{dist}\big(\gamma_1(t_1), \gamma_2(t_1)\big).$$

In the continuous case, one can consider infinitesimal time intervals and identify a neighborhood of x_0 with the tangent vector space $T_{x_0}X$, that is, with the set of 'directions' of trajectories based on x_0. Therefore, for τ sufficiently small, one can identify behaviors by the directions of their derivative at $t = 0$. At this point, a simple change has to be defined as an angle ℓ such that

$$\cos \ell = \frac{\langle \dot{\gamma}_1 | \dot{\gamma}_2 \rangle}{\|\dot{\gamma}_1\| \|\dot{\gamma}_2\|},$$

for $t = 0$ and for any two behaviours γ_1, γ_2 of X such that $\gamma_1(0) = \gamma_2(0)$.

Definition 2 I call **learning hierarchy** the type hierarchy $\mathcal{T}(\mathfrak{L})$ based on the set $\mathfrak{L}$ of simple changes in the behavior of an organism X.

I then follow Bateson (1968) to make some points about what this hierarchy makes us understand about the phenomena it models.

Type 0 It contains simple changes in behavior, i.e., immediate responses to environmental stimuli: its elements model the event that an organism being stimulated reacts with some behavior, regardless of the different trajectories its future history will follow.

Type 1 It includes the relationships between simple changes in behavior, i.e., all classes that group reactions to stimuli. These model an organism's ability to change its response to certain stimuli it classifies as interchangeable, i.e., equal. An organism that carries out type 1 learning will give at time $t = \tau > 0$ an answer different from the one it gave at time $t = 0$ to an *identical* stimulus. The mammal X we met before shows evidence of 1-learning when avoiding interaction with $\tilde{Y}$.

Type 2 It contains the classes of type 1 elements; for instance, its elements model a change in the organism's response to a class of stimuli or a shift in the classification of stimuli itself. An organism experiences several contexts and classifies them as similar; if, when confronted with a subsequent context of the same kind, its response is more efficient, then the organism is showing an instance of 2-learning. The same can be said of an organism that begins to implement different behavioral modules for contexts to which it previously responded in the same way. The mutual impermeability of types also enables us to model the phenomenon whereby a single negative reinforcement ($\Delta\mathcal{V} < 0$) in an interaction with the environment does not easily undermine type 2 classes, i.e., such occurrences are treated by the organism as 'exceptions that prove the rule.' Such a model property accounts for what has been said about exploration in mammals, which

persist in a stochastic strategy (type 2) even after suffering a failure (type 0). From a cognitivist point of view, we could say that the 'ego' in humans is the aggregate of specific outcomes of 2-learning, meaning that personality consists of certain learned patterns of how to relate to the outside world.

Type 3 Thus, type 3 learning "is likely to be difficult and rare even in human beings" (Bateson, 1968, 307). Items of this type are classes of 2-learnings – for example, they model a change in shifts within the classifications of stimuli. Roughly speaking, we can call 'habits' the elements of type 1 and 'habit formations' the elements of type 2. Thus, type 3 learning should be about a change in how one forms its habits. Any individual ontogeny is an instance of type 3 in the learning hierarchy; if humans were to learn how to juggle among classes of type 2 consciously, they would experience a kind of irrelevance of their 'ego.'

Type 4 4-learning "probably does not occur in any adult living organism on this earth" (Bateson, 1968, 298). But since human ontogeny is an example of 3-learning, the phylogenetic process that led to the homo sapiens species, combined with a specific individual's ontogeny, is an element of type 4. One could ascribe unconditional reflexes to level 4 of the individual learning hierarchy since they are the outcome of the organism's specific phylogeny. It is crucial to highlight that the learning hierarchy is not a chronological order in the organism's history, so it does not suggest that the individual gradually refines an initial trial-and-error strategy. The division into types has only explanatory value for the evolutionary emergence of abstract thinking. Bateson proposes the example of respiration, which can clarify the allocation between unconditional reflex and conscious act: "For human beings it is rather constantly true that air is present around the nose; the reflexes which control respiration can therefore be hard-programmed in the medulla. For the porpoise, the proposition 'air around the blowhole' is only intermittently true, and therefore respiration must be controlled in a more flexible manner from some higher center." (1968, 279) Concepts such as 'unconditional reflexes' or 'instinct,' rather than having explanatory value, seem to represent black boxes. Instead, including the outcome of the organism's specific phylogeny within its characteristic probability distribution provides possibilities for deeper analysis.

It is crucial that at type 1, we made use of the idea that two stimuli at two different instants of time can be considered equal, i.e., given a trajectory in the product space $X \times E$ relative to the organism and the environment, for some $\tau \neq 0$ there must hold

$$(x_0, e_0) = (x_\tau, e_\tau)\,.$$

Yet Biology is not an experimental science for the very reason that it does not generally admit repeatability. That is, because of the complexity of the phenomena of

interaction between organisms and the environment, one cannot find two identical stimuli in the history of an organism. Therefore, one could model the organism's strategy as the experimental scientist's when confronted with some experiments, ultimately different, of which the scientist chooses to neglect some aspects instead of others. In this way, the study of the class of phenomena is reduced to the study of its quotient modulo some equivalence relation, which cannot eventually be the identity.

For such a reason, Bateson introduces the concept of 'context' and refers to the phenomenon I just described as the hypothesis of its **repeatability**. For a framing within the biosemiotic field of the concept of 'context,' the reader may refer to Ongstad (2022) and Gabora and Kitto (2013).

> *Without* the assumption of repeatable context [...], it would follow that all 'learning' would be of one type: namely, all would be zero learning. [...] What previously we called 'learning' we would now describe as 'discrimination' between the events of Time 1 and the events of Time 1 *plus* Time 2. It would then follow logically that all questions of the type, "Is this behavior 'learned' or 'innate'?" should be answered in favor of genetics. We would argue that without the assumption of repeatable context, our thesis falls to the ground, together with the whole general concept of 'learning.' (Bateson, 1968, 293)

Yet, a notion of learning is necessary since much of one cannot explain behavioral phenomena in deterministic terms based on genetics. This leads one to broad support for the validity of the $\mathcal{T}(\mathfrak{L})$ model.

A Hierarchy of Semantic Efficiency

I now consider the set of possible meaningful interactions between a system X and its environment E at any time t

$$\mathcal{B} = \{ (x_t, e_t) \mid \mathcal{S}(x_t;e_t) \neq 0 \} .$$

I again refer to the ergodic hypothesis mentioned in the Introduction. Namely, I assume that from the observer's point of view, all the possible organism's behaviors are equiprobable. The observer here possesses no knowledge about the organism. For Eq. 4, increasing the viability effectiveness in information exchanges implies increasing semantic efficiency, so every type of learning – i.e., every level in a hierarchy of increasing effectiveness – is connected with a level in a hierarchy of increasing semantic efficiency. Semantic efficiency η is the ratio of semantic information overall mutual information; therefore – being the latter constrained by organism complexity – Eq. 3 states that the only way an organism can improve its semantic efficiency is by growing the semantic information extracted for the same amount of mutual information.

The theory thus points out that each type of learning must correspond to a specific semantics, i.e., a concrete way in which the organism ascribes meaning to its interactions with the environment, and that these semantics are hierarchically

arranged among themselves through abstraction relationships, with a semantics corresponding to simple changes in behavior at the base.

Definition 3 Suppose one considers the type hierarchy $\mathcal{T}(\mathcal{B})$, based on the set $\mathcal{B}$ of possible meaningful interactions between a complex organism X and its environment E at some time t. In that case, one can call **semantic hierarchy** the type hierarchy $\mathcal{T}(\mathcal{S}_{\mathcal{B}})$ based on the set $\mathcal{S}_{\mathcal{B}} := \{\, \mathcal{S}(b) \mid b \in \mathcal{B} \,\}$ of pointwise semantic informations of X over E.

Type 0 For any interaction $b = (x_t, e_t)$ between the organism X and the environment E, the **0-meaning** of b is pointwise ascribed to it by Kolchinsky–Wolpert semantics considering $t = 0$:

$$\mathcal{S}^{\mathbf{0}}(b) := \mathcal{S}(x_t;e_t)\,;$$

i.e., it is the attribution of meaning to that information about the environment that ensures the maintenance of the value of the viability function.

Type 1 These are the attributions of meaning the organism implements by linking the quantities of 0-meaning contained in interactions with the outside world. As is well known, the partitions of any set Z – i.e., the elements of $\mathscr{P}^2(Z)$ – are in a one-to-one correspondence with the equivalence relations on it. So, for each equivalence relation σ on the set $\mathcal{B}$, the **1-meaning** of the class of an interaction b is the set of 0-meanings of the interactions contained in $\{\, b_i \mid \sigma(b_i, b),\ i \in I \,\}$, namely

$$\mathcal{S}^{\mathbf{1}}([b]_\sigma) := \{\, \mathcal{S}^{\mathbf{0}}(b_i) \mid \sigma(b_i, b) \,\}.$$

One may quantify the 1-meaning of a class $[b]_\sigma$ by the cardinality of I, i.e., by the number of interactions – after b – the organism will avoid experiencing, albeit behaving in a viability-consistent way.

Type 2 These are the meanings the organism makes by grouping instances of type 1; that is, they are meanings assigned to equivalence relations on $\mathcal{B}$. The **2-meaning** of an equivalence relation σ on $\mathcal{B}$ is

$$\mathcal{S}^{\mathbf{2}}(\sigma) := \{\, \mathcal{S}^{\mathbf{1}}([b]_\sigma) \mid b \in \mathcal{B} \,\}.$$

One can say that an equivalence relation σ has the more 2-meaning, the greater the cardinality of the partition it induces on $\mathcal{B}$.

The hierarchy is, of course, not limited to type 2; indeed, one might expect the meanings of mathematical entities to stand at levels much higher than 2, but it is at this level that a gap arises between viability-consistent meanings and types of meanings that may 'forget the initial goal' of preserving $\mathcal{V}$ values. At level 2, the notation also loses reference to the level 0 interaction that generated the

2-meaning, and it is possible to interpret the gap that arises at this level as the emergence of what we can call 'abstract meanings.'

If the transition from type 0 to type 1 implies an increase in meaning on a like for thermodynamic cost, in the next step, one should note the following contradiction, which one should expect at each rank in the hierarchy.

Contradiction 1 For $n > 1$, the amount of n-meaning of an equivalence relation conflicts with the amount of $(n-1)$-meaning of its classes since the cardinality of the induced partition is inversely proportional to the cardinality of the classes.

At this point, it is worth formalizing the concept of 'concept.' The following definition is an explication of the common understanding of the term.

Definition 4 I call **conceptual hierarchy** the type hierarchy $\mathcal{T}(\mathcal{B})$, and **concept** the interpretation in $\mathcal{T}(\mathcal{B})$ of any constant of any type $n > 0$ of **TT**; an ***n*-concept** is a concept of type n.

The generality of the hypotheses allows me to propose the conjecture that any concept built by a person will be subject to the constraints of the model; in particular, any n-concept will be the result of n abstractions – in the sense of **TT** – from the set $\mathcal{B}$ of interactions with 0-semantic content – i.e., meaningful in the sense of Kolchinsky and Wolpert. It means that for each person and each concept C thought by that person, there always exists a tree graph rooted in some past interaction with the environment, having C as a leaf. Asserting the existence of such a root does not, of course, mean giving a method for finding it; this is usually one of the tasks set by psychological disciplines.

I stress that $\mathcal{T}(\mathcal{B})$ and $\mathcal{T}(\mathcal{S}_{\mathcal{B}})$ are no more than explicative models for the adaptive genesis of concepts and by no means provide a theory of semantics in linguistic communication. We are obviously to expect that communication between human beings violates the constraints of **TT**, e.g., that receiving sentences of a natural language could relate concepts of a person's $\mathcal{T}(\mathcal{B})$ in ways forbidden by the rules of **TT** (see Appendix B). The model developed so far does not allow us to describe two human beings' shared semantics, and thus, we cannot strictly speak about linguistic contexts.

Nonetheless, $\mathcal{T}(\mathcal{S}_{\mathcal{B}})$ enables us to intuitively account for situations in which a word is 'of little significance,' i.e. when it does not express "a difference which makes a difference[, that] is an idea" (Bateson, 1969, 279). Contradiction 1 models the case in which a person understands a linguistic expression as the correlation of a concept encompassing many instances with concepts of an inferior type; in such cases, the former tends to be not very expressive. Moreover, even metaphors and rhetorical figures are violations of Russell's vicious circle principle (see Appendix B); one understands them as finitary relations over distant types of $\mathcal{T}(\mathcal{B})$ which do not belong to $\mathcal{T}(\mathcal{B})$. Yet, if $\mathcal{T}(\mathcal{S}_{\mathcal{B}})$ cannot represent the sense of the sentence

$$P := \text{‘it’s raining cats and dogs,’}$$

nonetheless, it allows us to say that the sense of ‘it’s raining a lot’ can always be traced back to the viability-consistent meaning of past environmental interactions in the history of any human involved in the communication of P. I use ‘sense’ to refer to the semantic content conveyed by language and ‘meaning’ to speak about the elements of $\mathcal{T}(\mathcal{S}_{\mathcal{B}})$. I implicitly assume a gap between them, according to Wittgenstein (1963), that one can mathematically justify following Harris (1991).

According to Weaver (see “Problems Concerning Viability”), any possible interaction b is meaningful if and only if it has an effect, i.e., if it generates a simple change ℓ in the organism’s behavior. Thus, I can state that $\mathcal{B}$ and $\mathfrak{L}$ have the same cardinality – whether continuous or countable as appropriate – that is, there is a bijection between the set of possible meaningful interactions and the set of 0-learnings. Therefore, for what I stated before, we have that

$$\mathcal{T}(\mathfrak{L}) \equiv \mathcal{T}(\mathcal{B}),$$

i.e., the argument so far concerns two different instances of the same hierarchy. In particular, at level 1, we have a one-to-one correspondence on the set of interpreted 1-constants:

$$\text{habit} \longleftrightarrow \text{1-concept};$$

thus, the necessity of the hypothesis of repeatability of contexts mentioned above implies that building a 1-concept requires the organism to overlook certain features of the generating interaction. Higher levels cannot overcome such omissions; each level yields new ones. Bueno (2022) has recently argued in favor of the context dependence of mathematical theorems, which are linguistic representations of high-level concepts in the individual hierarchy. However, we must point out a contradiction between the two hierarchies.

Contradiction 2 If the learning hierarchy $\mathcal{T}(\mathfrak{L})$ is kept below type 5 in nature, building concepts in humans has no constraint on the degree of levels of abstraction; indeed, mathematics and philosophy are fields of thought in which one tends to ascend the types of $\mathcal{T}(\mathcal{B})$ by a great deal.

From the semantic model outlined so far, it is possible to draw some reflections on current Artificial Intelligence (AI) developments. Indeed, the foundation of conceptual and semantic type hierarchies is provided by the work of Kolchinsky and Wolpert, who model the situation of a general autonomous agent, i.e.,

> a far-from-equilibrium system which actively maintains its own existence within some environment [...]. A prototypical example of an autonomous agent is an organism, but in principle, the notion can also be applied to robots [...] and other non-living systems. (Kolchinsky and Wolpert, 2018, 2)

While current research has lost the focus it had in the middle of the last century, namely, to understand the functioning of processes carried out by a human mind

to reproduce them artificially, current technologies still have as their stated purpose the aim of replacing human activity with that of machines. Hence, we are facing a sharp discord between ambition and practice in AI research, which the model developed so far can help clarify.

In general, current research in AI aims to construct trainable architectures for recognizing the affiliation of specific objects to given concepts. The definition of such a recognition ability is the cornerstone around which the various algorithmic proposals have evolved over the past twenty years. According to the seminal work of Bengio (2009), the current success of the Deep Learning (DL) method lies in the following considerations, which at first glance appear to closely resemble the discourse presented in the previous section of this work.

> Lower level abstractions are more directly tied to particular percepts, whereas higher level ones are what we call "more abstract" because their connection to actual percepts is more remote [...] The focus of deep architecture learning is to automatically discover such abstractions, from the lowest level features to the highest level concepts. Ideally, we would like learning algorithms that enable this discovery with as little human effort as possible, i.e., without having to manually define all necessary abstractions or having to provide a huge set of relevant hand-labeled examples. (Bengio, 2009, 5)

In particular, "depth of architecture refers to the number of levels of composition of non-linear operations in the function learned" (Bengio, 2009, 6), where one must recognize that "the mammal brain is organized in a deep architecture [...] with a given input percept represented at multiple levels of abstraction" (Ibid). Philosophical advocacy is provided, for instance, by Niiniluoto (2022). Yet, I believe that the notion of 'abstraction' used in the definition of the DL method cannot properly model mammal thought's capacity for abstraction, better represented by **TT**. If one attempts to recursively represent a hierarchy $\mathcal{T}(\mathcal{B})$, one should not compose functions but ascend in dimension. Let's consider, for example, the function

$$f(x) = x * \sin(a * x + b)$$

which Bengio provides on page 14; if we represent it as a graph of compositions between the functions from the set $\{*, +, \sin\}$, we find that $f(x)$ requires an architecture of depth 4 to be learned. However, from the perspective of **TT**, if x, a, b are elements of $\mathcal{B}$ then $f(x)$ is an element of $\mathscr{P}(\mathcal{B})$ just like $\sin(a * x + b)$ and $a * x$. In other words, no matter how many non-linear functions we compose, we cannot go beyond level 1 of the hierarchy $\mathcal{T}(\mathcal{B})$.

What we need to represent, instead, is a logical leap between types. Let's consider $x \in \mathcal{B}$; an element $g_x^1 \in \mathscr{P}(\mathcal{B})$ is a set of elements from $\mathcal{B}$ that are in some relation to x. Any element of $\mathscr{P}^2(\mathcal{B})$ is an equivalence relation on $\mathcal{B}$, which is not dependent on x but on an element of $\mathscr{P}(\mathcal{B})$, such as g_x^1. Thus, a computable hierarchical representation should be in the form

$$
\begin{aligned}
x \mapsto g^1_x \quad &= \{\, y \in \mathcal{B} \mid g^1(x,y) \,\} \\
g^1_x \mapsto g^2_{g^1_x} \quad &= \{\, \bar{y} \in \mathscr{P}(\mathcal{B}) \mid g^2\left(g^1_x, \bar{y}\right) \,\} \\
g^2_{g^1_x} \mapsto g^3_{g^2_{g^1_x}} \quad &= \{\, \sigma \in \mathscr{P}^2(\mathcal{B}) \mid g^3\left(g^2_{g^1_x}, \sigma\right) \,\} \\
&\vdots
\end{aligned}
$$

where the notation indicates that the n^{th} recursive function directly depends on the previous one, being a class represented by the latter.

From the viewpoint of **TT**, the layers of DL are relationships among elements of $\mathcal{B}$, meaning they are all 1-concepts. The semiotic analysis recently conducted by Muşat and Andonie (2020) on Convolutional Neural Networks may help to understand how the aggregation of signs into supersigns – through what they call type I superization – stops at the first level of the conceptual hierarchy. The data sets with which an intelligent agent (IA) is fed represent sets of meaningful interactions or subsets of $\mathcal{B}$. A model such as $\mathcal{T}(\mathcal{B})$ should thus be explicative for why AI today requires ever larger data sets, incomparably bigger than those needed by any mammal, to construct concepts and make decisions sufficiently coherent to ensure survival. Remaining at level 1 of the conceptual hierarchy, we can say that the current AI is not a good model of human intelligence but rather an attempt to build an artificial intelligence that mimics the responses, for example, that an ant's mind might give to certain human stimuli. The massive extension of data sets is a highly thermodynamically inefficient response to complexity (see "Problems Concerning Viability").

The fundamental gap is that an IA cannot invent the criteria by which to construct the classes of $\mathscr{P}(\mathcal{B})$, namely the 2-concepts belonging to $\mathscr{P}^2(\mathcal{B})$. As a result, the IA will not even know how to proceed with the subsequent steps of a $\mathcal{T}(\mathcal{B})$, where I recall that the leap from type 1 to type 2 is where the transition to abstract thinking occurs. An IA built upon a deep architecture can recognize whether an object x is similar to those other objects $x_1, \ldots, x_n$ with which it has been trained and had categorized in some way $X = \{\, x_i \mid i \leq n \,\}$ unknown to the human user. However, it cannot generate equivalence relations on the set of x using abstraction rule 4 of Definition 9 in the Appendix B. For instance, a deep architecture that can consistently identify a new chair as resembling the chairs seen before – assuming it could hypothetically do so an enumerable number of times – still would not possess the ability to employ the concept of 'chair' to construct the concept of 'sitting,' which involves the 2-concept

$$\text{`}\forall x, x \text{ is a chair.'}$$

I have already highlighted how language consistently violates the Principle of Vicious Circle. Individual thought also frequently engages in such infractions; e.g., we commonly establish relationships between concepts of different types, as in the case of the concern that an earthquake (any instance of the 'earthquake' concept) could destroy my house (this specific house here, not just any house). However, the fact that **TT** represents solely the conceptualization of an ideal individual rather than that of real human beings does not mean that one can construct a sound model of

thought without taking it into account: violating a regulatory principle is not equivalent to not possessing it at all.

Knowledge and Semantics

Thanks to the models constructed through **TT**, I can better justify a notion of 'knowledge' in humans as an outcome of learning processes and as mental coordination of classes of phenomena. However, I want to formalize that there are several senses with which the verb 'knowing' is used (for instance, see https://www.dictionary.cambridge.org/dictionary/english/knowledge). We commonly refer to knowledge in terms of abstract thinking as getting a concept of a phenomenon – such as a theory, a person, a city, etcetera; otherwise, we use the same word to refer to perceptual experiences of physical reality. Far from confusing, one can easily model such a semantic dichotomy from the hierarchies I have just constructed.

However, Contradiction 2 claims that if one wants to model concept understanding, one must dispense with the idea that beyond level 4, knowledge can link to actual learning processes, that is, to n-type changes in the behavior of the knowing subject, for $n > 4$. By this, of course, I do not mean that the study of a mathematical theory does not affect the student's behavior, but that at the very best, one can expect that the deepening of, say, Topology will change the way the student forms the habits of relating to the world, and not changing how the student intervenes in the mode of shaping those habits (see the argument in "Semantic Efficiency and Learning").

Therefore, one must tie a definition of knowledge to $\mathcal{T}(\mathcal{B})$ and not to $\mathcal{T}(\mathfrak{L})$. I mentioned that only a succession of omissions allows the individual to construct an element of $\mathcal{T}(\mathcal{B})$ of type higher than 0. From a gnoseological perspective, one must interpret such omissions as an 'absence of certainty' in the individual's representation of phenomena. So cognition can aspire to be sure only if it does not involve any concept, i.e., whether it is the perception of a meaningful interaction $b \in \mathcal{B}$ in bi-univocal correspondence with a 0-learning $\ell \in \mathfrak{L}$.

Such considerations recall the well-known Laplace's epistemological approach, who, in the introduction to his *A philosophical essay on Probabilities*, states:

> Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it an intelligence sufficiently vast to submit these data to analysis [...] for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers [...] a feeble idea of this intelligence. [...] All these efforts in the search for truth tend to lead it back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed. (Laplace, 1814, 4)

The idea one can recover from Laplace's philosophy is that even for deterministic phenomena – those for which trajectories that evolve from states close to the initial time always stand at a finite distance on the state space – one must necessarily express human knowledge of them in probabilistic terms. Of course, the ontology of Quantum Mechanics even obliterated the theoretical possibility of the so-called

'Laplace's demon,' capable of knowing the entire history of the universe from knowledge of its conditions at a given time. Nevertheless, as for deterministic phenomena, current science still relies on probability theory in purely epistemic rather than ontological terms. Regardless of the deterministic or non-deterministic nature of the phenomena under knowledge, I am only interested in underlining the cognitive processes of both classes as having a common probabilistic nature.

The approach with which de Finetti founds Probability Theory easily accounts for the necessities of what I am saying here. According to his subjectivist proposal,

> notion of probability is relative: the fact that two cases appear equally probable depends on which circumstances are known or unknown to us. (de Finetti, 1929, 12, translation by the author)

Namely, he identifies probability attributions to phenomena with subjective expectations of their occurrence. Reversing the terms of this identity, we get that the things we know depend on the probability that certain events may occur. In taking this approach, it is possible to define knowledge, which, of course, already assumes that one knows what probability is and that a theory of it does not rest on subjectivist grounds. In what follows, I will implicitly assume Kolmogorov's formal definition as the theoretical foundation of my argument.

It is now worthwhile to give a precise statement of the notion of **phenomenon** as a string of events $(Y_1 \doteq y_{1_j}), \ldots, (Y_n \doteq y_{n_\ell})$, where $(Y_i \doteq y_{i_j})$ indicates that a physical system Y_i is at some time in its state y_{i_j}, where $i \in I$ and $j \in J$ for some index sets I, J. Based on the latter, I can give the following definition.

Definition 5 I call **knowledge** about a physical system the assignment of a probability distribution to its evolution, and knowledge of a phenomenon the assignment of a probability value to its occurrence.

I stated above that a good definition should depend on the elements of $\mathcal{T}(\mathcal{B})$; this is guaranteed by the Joint Probability Theorem, which I quote from de Finetti's words:

> What is the probability of event E after experience led me to know the set of circumstances A? If, before I knew the outcome of the complex of circumstances A, it was p the probability that event E and circumstances A would occur, and q the probability that circumstances A would occur, the probability of E conditional on the occurrence of circumstances A is p/q. (de Finetti, 1929, 35-36, translation by the author)

That is to say: $P(E|A) = P(E \cap A)/P(A)$. In particular, we can interpret the set of circumstances A as a subset of elements of $\mathcal{T}(\mathcal{B})$. Indeed, the occurrence of the circumstances A implies that a knowing subject X perceives some determinations of A and includes these perceptions among the instances of some of its particular concepts. Thus, such a definition of knowledge is not only bound to a system object of knowledge Y but also presupposes that one has fixed a system-subject X and its conceptual hierarchy. I emphasize that the definition is not concerned

with whether the future behavior of Y will verify X's expectations; in fact, we also wish to expect cases in which X's knowledge of Y turns out to be 'fallacious,' i.e., cases in which the actual behavior of Y has received low or null probability from X. What matters in this context is to state that X has nonzero knowledge of Y if it 'coherently' expects certain behaviors of Y to occur with different probability values between them. Following de Finetti, I consider X to be consistent in its evaluations if it behaves according to the formal rules of the calculus of probabilities.

> The calculus of probability is the logic of the probable. Just as formal logic teaches how to deduce the truth or falsity of some consequences from the truth or falsity of some premises, the calculus of probability teaches how to deduce the greater or lesser likelihood or probability of some consequences from the greater or lesser likelihood or probability of certain premises. [...] one can show that the well-known theorems of the calculus of probability are necessary and sufficient conditions for the opinions of a given individual not to be inherently contradictory. (de Finetti, 1930a, 2, translation by the author)

When considered backward, de Finetti's foundational proposal renders a genuine model for actual individuals' decisions:

> From this point of view, the calculus of probabilities is truly assimilable to experimental science. [...] In the experimental sciences, one replaces the world of sensations with a fictitious world in which magnitudes have an exactly determinable value; in the calculus of probabilities, I replace my vague and elusive state of mind with that of a fictitious individual who knew no uncertainty in judging the degrees of its confidence. (de Finetti, 1929, 41, translation by the author)

A quantitative measure of knowledge should thus vary between a minimum value of 0 if X associates to Y uniform probability distribution and a maximum value if the distribution is degenerate or Dirac, centered in some element of the space of states of Y. Taking back the information-theoretic framework, such a request brings up the notion of redundancy, for which I can state the following definition, where X plays both roles of member of the observed joint system and observer.

Definition 6 I define the **amount of knowledge** that a complex organism X has about a system Y as the redundancy of information exchanged between X and Y in dependence on the probability distribution p_Y attributed by X to the evolution of Y; i.e.

$$\varrho(p_Y) = 1 - \frac{\mathcal{H}_{p_{X,Y}}(X,Y)}{\mathcal{H}_{p_X}(X) + \mathcal{H}_{p_Y}(Y)},$$

where $p_{X,Y}$ and p_X are the joint distribution and the marginal distribution of X respectively, and $\mathcal{H}_{p_Z}$ is the entropy of the system Z with distribution p_Z (see Eq. 1).

I verify that the definition is consistent with the above. I first consider the minimal knowledge case, that is, the case in which the marginal distribution of Y is uniform. In this case, X and Y are completely independent, i.e., the joint distribution $p_{X,Y}$ is the product of the marginals and thus

$$\varrho(p_Y) = 1 - \frac{\mathcal{H}_{p_X}(X) + \mathcal{H}_{p_Y}(Y)}{\mathcal{H}_{p_X}(X) + \mathcal{H}_{p_Y}(Y)} = 0\,.$$

Instead, in the case where Y receives from X a probability distribution with support only around a single event y_0, we obtain that $\mathcal{H}_{p_Y}(Y) = 0$, and the maximum redundancy

$$\varrho(p_Y) = 1 - \frac{\mathcal{H}_{p_{X|Y}}(X|Y)}{\mathcal{H}_{p_X}(X)}\,.$$

I emphasize that the amount of knowledge X has about Y reasonably increases when the uncertainty in the behavior of X decreases if the behavior of Y is known and behaves inversely for freedom of choice in the behavior of X. In the degenerate case, only these quantities affect the knowledge that X has about Y.

At this point, I have a definition of the knowledge that X has about Y as the attribution of the probability distribution p_Y to the latter's behavior. I have defined the concepts of X as abstractions belonging to $\mathcal{T}(\mathcal{B}_X)$, built up from meaningful interactions. The two models – cognitive and semantic – tie in if one considers that from X's perspective, for every physical system Y, there exists a type 1 instance of its conceptual hierarchy; namely, Y is identifiable with the abstraction that X makes from some meaningful interactions $b = (x, y)$ for some $y \in Y$. From X's perspective, Y is a class of meaningful interactions; that is, it is a set of possible meaningful interactions that X considers to have in common the characteristic of 'being interactions with the same system Y.' Therefore, within the framework of my semantic model, I can give the following definition.

Definition 7 Given a physical system Y and a knowing subject X, I call **representation of Y for X** the 1-concept $\left[(x, y)\right]_{\sigma_Y} \in \mathcal{T}(\mathcal{B}_X)$, where (x, y) is any meaningful interaction of X with Y, and σ_Y is the equivalence relation that discriminates interactions with Y within $\mathcal{B}_X$.

In this formulation, I have limited myself to drawing general consequences from the hierarchy $\mathcal{T}(\mathcal{B})$; I expect future works to give a better hierarchical construction of knowledge, too. However, the above is sufficient for my purposes here, for I can model the understanding of an abstract phenomenon as the assignment of a shared probability distribution to all instances of the concept that the subject constructs in its own $\mathcal{T}(\mathcal{B})$. Indeed, the subjectivist approach in Probability Theory allows me to account for the informal assumption of 'repeatability of contexts' noted by Bateson (see the discussion above), as the notion of 'repeatable contexts' can be formalized through de Finetti's notion of **exchangeable events**.

> One says events E_i ($i = 1, 2, \ldots$; in finite or infinite numbers) to be exchangeable if [...] the probability in a problem concerning n of them does not vary no matter how they are chosen and permuted. (de Finetti, 1969, 56, translation by the author)

With such a definition, de Finetti can express an alternative notion to 'independent events' and, through his Representation Theorem (for a mathematical discussion, see de Finetti, 1930b), he can build a bridge from the subjectivist approach toward the formal results of classical frequentist setup. In my model, this theorem accounts for the fact that whatever probability we initially associate with a phenomenon,

> after a large number of trials, our state of mind is almost entirely determined by frequency. Of the two factors, initial opinion and experience, the second influences with increasing weight and generally becomes entirely preponderant with increasing numbers of trials. (de Finetti, 1929, 38, translation by the author)

I.e., the more interactions between X and Y, the more knowledge X has of Y and the more instances X places below the domain of its representation of Y.

In the framework constructed so far, I can better clarify the common-sense definition given in the Introduction by the hierarchy of learnings due to environmental interactions, i.e., what I model here as a shift in the attribution of probability distribution to systems.

Definition 8 A **cognitive process** is one in which a knowing subject switches from attributing some probability distribution to a system to attributing a second one characterized by a higher amount of knowledge. I.e., a process that increases the number of instances below the individual's representation of that system.

Conclusions

Within the Thermodynamics of living systems framework, I followed Kolchinsky and Wolpert's proposal to identify basic semantics shared by all organisms in their necessary information exchanges with the environment. Through a study of the communication problems posed by Weaver, I came to model the phylogenetic problem of increasing the efficiency of environmental interactions in complex organisms, that is, the problem of optimizing the effect that information exchange has on organism behavior.

A change in behavior due to an interaction with the outside world is what we call learning, i.e., the historically achieved way in which the phylogeny of complex species has addressed Weaver's effectiveness problem. Following Bateson's insights, I built a hierarchy of modes of learning modeled after Russell's Type Theory. An explication of adaptive history resulted, which allows complex organisms to deal effectively with different classes of phenomena by constructing concepts. Thus, I proposed a hierarchy of semantic types, which models all possible forms of meaning as classes of meanings. Such a hierarchy has no explicative value for any specific concept, but it provides a necessary condition; thus, it highlights some profound theoretical gaps between the

current approach followed by AI research and the functioning of the mind in mammals. My thesis is that we need a definition of 'abstraction' akin to the one proposed by Type Theory and that not every tree-like structure is a good representation of cognitive processes. Current AIs can learn how to abstract from items to 1-concepts but are not able to mime the way knowing subjects decide to miss classes of details, classes of classes of details, and so on.

Therefore, a good definition of 'knowledge' must account for the increasing loss of certainty as the degree of abstraction of its objects increases. Hence, as soon as it ceases to be perceptual knowledge and becomes knowledge of concepts, one must define knowledge in probabilistic terms following Laplace's approach. So, I proposed using Probability Theory as a Gnoseological theory, that is, as an abstract model of actual human phenomenon. By this model, following de Finetti's subjectivist approach, one studies the behaviors of an imaginary individual who attempts to make consistent predictions about the world of phenomena. Thus, an ideal individual assigning probabilities to the systems' behavior represents the human effort of knowing.

To conclude, one must define a cognitive process as the shift from one probability attribution to another, which raises the value of some measure of knowledge. I proposed a measure to be redundancy in information exchange between the knowing subject and object of knowledge. Thus, any cognitive process cannot be objective but has to be a subjective attribution of probability to classes of phenomena of reality, where the subject groups such phenomena by quotienting the set of interactions with the environment modulo equivalence relations bound to various types of learning.

Appendix A The Kolchinsky–Wolpert Model

As I stated, the viability function of X at time τ is the quantity $\mathcal{V}(X_\tau) = \sum_{x_\tau} p(x_\tau) \log p(x_\tau)$, where p is the marginal distribution of X, and x_t is a particular outcome of random variable X_t representing the system X at time t (see Kolchinsky and Wolpert, 2018, 6). I recall that if a joint system (X_1, X_2) has joint distribution p_{X_1,X_2}, then the two marginal distributions are $p_{X_i}(x_i) = \sum_{x_j} p_{X_1,X_2}(x_1, x_2)$, for $i, j = 1, 2$ and $i \neq j$.

Kolchinsky and Wolpert call **actual distribution** the joint distribution p_{X_t,E_t} of trajectories of the joint system (X, E) over time $t = 0$ to $t = \tau$ (2018, 7), and intervene in the X part of the joint system with a counterfactual method to measure the effects of changes on the rest of the system. Namely, they study modifications on the joint distribution function, which shuffle away the mutual information between X and E, to identify a threshold below which the viability of the scrambled system is lower than that of the actual system. Mutual information $\mathcal{I}(X;E)$ between X and E admits minimum 0 value if X and E are independent random variables, that is when $p_{X,E} = p_X \cdot p_E$. So, one can assume that an intervention such as

$$F : p_{X,E} \longmapsto p_X \cdot p_E$$

is the intervention on the distribution of (X, E), which maximizes mutual information destruction. Therefore, the initial viability value $\Delta\mathcal{V}(X_\tau)$ of mutual information between X and E defined by Eq. 2, is the difference at time τ between the viability of X if the distribution of (X, E) at time $t = 0$ is the actual ditribution p_{X_0,E_0}, and the viability of X if the distribution of (X, E) at time $t = 0$ is $F(p_{X_0,E_0}) = p_{X_0} \cdot p_{E_0}$ (Kolchinsky and Wolpert, 2018, 7).

To discern the meaningful information carrying $\Delta\mathcal{V}(X_\tau)$, the two authors introduce the set of deterministic endofunctions on the states of E. The procedure they use is said of coarse-graining on the conditional distribution $p_{X_0|E_0}$; roughly speaking, it consists of considering all functions of the possible outcomes of E – depending only on those outcomes and not, for example, on time – that act by exchanging or identifying their inputs. Given a deterministic function $\varphi : E \to E$, one can define the **intervened distribution induced by φ** as the joint distribution

$$p^{\varphi}_{X_0,E_0} := p_{X_0|\varphi(E_0)} \cdot p_{E_0},$$

where

$$p_{X_0|\varphi(E_0)}\big(x_0|\varphi(e_0)\big) = \frac{\sum_{e_0' : \varphi(e_0')=\varphi(e_0)} p_{X_0,E_0}(x_0, e_0)}{\sum_{e_0' : \varphi(e_0')=\varphi(e_0)} p_{E_0}(e_0)}.$$

As done above, one can rename X^{φ}_{τ} the system X at time $t = \tau$ if (X, E) evolved with initial joint distribution $p^{\varphi}_{X_0,E_0}$. I point out that $\big(X_0|\varphi(E_0)\big)$ is independent of E_0, and from the point of view of X two states e_0 and e_0' such that $\varphi(e_0) = \varphi(e_0')$ are indistinguishable. That is, X_0 has only information about $\varphi(E_0)$ and not about E_0.

Therefore one can define the **optimal intervention** $p^{opt}_{X_0,E_0}$ as the intervened distribution that holds the following conditions:

1. $p^{opt}_{X_0,E_0} \in \left\{ p^{\varphi}_{X_0,E_0} \,\middle|\, \mathcal{I}(X^{\varphi}_0;E_0) = \min_{\psi\in\Phi} \mathcal{I}(X^{\psi}_0;E_0) \right\}$,
2. $\mathcal{V}(X^{opt}_{\tau}) = \mathcal{V}(X_\tau)$,

where X^{opt}_t is the system X at time t if (X, E) evolved with initial joint distribution $p^{opt}_{X_0,E_0}$ (see Kolchinsky and Wolpert, 2018, 8). By such a definition, any further intervention on $p^{opt}_{X_0,E_0}$ would change the output of $\mathcal{V}$, i.e., all the mutual information contained in $p^{opt}_{X_0,E_0}$ causally contributes to X's viability at time $t = \tau$.

Appendix B The Russell's Type Theory

In this appendix, I will present the fundamental theory with which I formalized Bateson's insights. In particular, I will refer directly to the exposition that philosopher and mathematician William Hatcher (1982) gives of such a theory and its history. Contemporary Mathematical Logic has somewhat forgotten Russell's Type Theory in favor of the set theory approach inaugurated by Zermelo and Fraenkel. Both theories have historically developed to establish consistent logical foundations for the mathematical

structure, that is, to correct the approach of Frege's "Grundgesetze der Arithmetik" (1893), which hides the contradiction known as Russell's antinomy. One can formulate the latter as follows: if one considers the set y defined by the property $x \notin x$, y should be the set of all sets that are not elements of themselves. One may ask if y does belong to itself. By the law of excluded middle, either it does or not. If it does, then $y \in \{ x \mid x \notin x \}$ and so y must satisfy the defining property of the set y; i.e., it does not belong to itself. On the other hand, if y does not, then y satisfies the defining property of y and is thus an element of itself (see Whitehead and Russell, 1963, 60).

One of Frege's fundamental insights is recognizing that when we create concepts – or properties – we might want to express predicates about them. For instance, we construct the property of 'being a chair,' and we want to say that 'there is something that has the property of being a chair,' meaning 'there is a chair.' In such sentences, we are not predicating the property but objectifying it, i.e., nominalizing the predicate (see Frege 1892 and 1893). In overcoming Russell's antinomy, the Type Theory approach retains Frege's aim of formalizing that part of abstract reasoning, which Imre Lakatos calls 'quasi-experience' (1978). According to the latter, as the experimental sciences, Mathematics grows by explicating phenomena of thought in a quasi-empirical way. He argues that behind the definition of a mathematical concept lies an accidental choice due to unformalized thinking referring to a set of non-mathematical objects. On the other hand,

> Zermelo's system is more directly concerned with mathematics and the needs of mathematical structures [...]: Mathematics is (we believe) consistent. Thus, if we give a precise account of the intuitive use of sets as mathematicians use them, we shall have an adequate and correct foundation. [...] we observe that mathematicians do not normally use such sets as 'the set of all sets' or the 'set of all sets not elements of themselves'. We might contend that these contradictory notions are not really valid mathematical objects at all. (Hatcher, 1982, 135)

I am not concerned with the foundations of mathematics here; therefore, Russell's approach, although more uncomfortable and in some ways a failure concerning meta-mathematical purposes, better addresses what I need. As mentioned, the idea behind Type Theory is to build an axiomatic theory that prohibits antinomies due to self-reference while maintaining Frege's Law of Courses of Value (Frege, 1893). One can state the latter as follows: given any property P, there exists a set y such that for all x, x is in y if and only if x satisfies the condition P; i.e.,

$$\exists y \forall x \big(x \in y \iff P(x) \big) .$$

Therefore, it is to impose the following constraints:

> 'Whatever involves *all* of a collection must not be one of the collection;' or, conversely: 'If, provided a certain collection had a total, it would have members only definable in terms of that total, then the said collection has no total.' We shall call this the 'vicious-circle principle,' because it enables us to avoid the vicious circles involved in the assumption of illegitimate totalities. (Whitehead and Russell, 1963, 37)

Hereafter, I refer to the more recent formulation due to Wiener and Kuratowski in the version reported by Hatcher (1982), which can be expressed through the notation of Set Theory, closer to the contemporary reader's taste.

Definition 9 I call **Type Theory** (**TT**) the formal system in which

- the **language** is that of set theory plus the sets of symbols $\{x_i^n | i, n \in \mathbb{N}\}$, $\{a_i^n | i, n \in \mathbb{N}\}$ for variables and constants, respectively;
- the **well-formed formulas** (wffs) and **terms** are define as follows:

  1. $\{x_i^n\}_{i\in\mathbb{N}}, \{a_i^n\}_{i\in\mathbb{N}}$ are terms said to be of **type *n***;
  2. '$x_i^n \in x_j^{n+1}$' is a wff;
  3. if P, P' are wffs, then $\neg P$ and $P \vee P'$ are wffs;
  4. if P is any wff and x is any variable, then '$\forall x P(x)$' and '$\exists x P(x)$' are wffs;
  5. if $P(x_i^n)$ is a wff containing x_i^n free, then '$\{x_i^n | P(x_i^n)\}$' is a term of type $n + 1$.

- the **axioms** are the following schemes:

  T1. $\exists x_i^n \forall x_j^{n-1} \left(x_j^{n-1} \in x_i^n \iff P(x_j^{n-1}) \right)$,
  where x_i^n does not occur in the wff $P(x_j^{n-1})$, which contains the variable x_j^{n-1} free;
  T2. $\forall x_i^n \left(\left(x_i^n \in x_j^{n+1} \iff x_i^n \in x_\ell^{n+1} \right) \implies \left(x_i^{n+1} = x_\ell^{n+1} \right) \right)$;

- the **rules of inference** are the natural deduction rules (for instance, see Hatcher, 1982, 43-44), with the constraint that only variables and terms of a given type can be substituted.

I emphasize that the theory as defined cannot prove arithmetic because its axioms do not allow the existence of $\mathbb{N}$ to be established; to do so, an axiom of infinity must be added to **TT**. As mentioned, the present exposition has no foundational purposes, and I state a sufficient theory for my argument.

Remark I make some observations on the last definition.

1. The scheme *T2* constrains the Extensionality Principle – stating that two sets are identical if and only if the same elements fall under their domain – to hold only within any given type.
2. The operation described by rule 4 states a principle of abstraction; i.e., it is a method for constructing concepts.
3. The wffs formalize the intuitive Frege's notion of 'property;' in particular, the terms defined by rule 5, and whose existence is guaranteed by *T1*, model Frege's operation of 'objectivization' of 'concepts,' that one needs to 'speak about concepts.'

With Axiom *T1*, we thus have a new version of the Law of Courses of Value that only holds within any given type.

Principle (Abstraction) Given a type, for any property P, there exists a set *y* of this type such that for all x of the preceding type, x is in y if and only if x satisfies P.

Acknowledgements I want to acknowledge the philosopher Abdullah Öcalan, the highest guide in the quest for truth, who contributed more than anyone else to inspire this research.

Author Contributions M.B. wrote the entire manuscript.

Funding Not applicable.

Availability of data and materials Not applicable.

Declarations

Ethical standard Not applicable.

Competing interests The authors declare no competing interests.

References

Achella, S. (2022). Idealism and Science of Life: An Intersection Between Philosophy and Biology. In N. Rezaei, & A. Saghazadeh (Eds.), *Thinking. Integrated Science*, (vol. 7, pp. 111–131). Springer.

Barbieri, M. (2008). Biosemiotics: a new understanding of life. *Naturwissenschaften, 95*, 577–599.

Bateson, G. (1968). The Logical Categories of Learning and Communication, and the Acquisition of World Views. Extended in *Steps to an Ecology of Mind*. Jason Aronson Inc (1987), 284–314.

Bateson, G. (1969). Double Bind, 1969. In *Steps to an Ecology of Mind*. Jason Aronson Inc (1987), (pp. 276–283).

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning, 2*, 1–55.

Bishop, R.C. (2022). Contextual Emergence: Constituents, Context and Meaning. In S. Wuppuluri, & I. Stewart (Eds.), *From Electrons to Elephants and Elections.* The Frontiers Collection, (pp. 243–256), Springer.

Brier, S. (1999). Biosemiotics and the foundation of cybersemiotics. Reconceptualizing the insights of ethology, second order cybernetics and Peirce's semiotics in biosemiotics to create a non-Cartesian information science. *Semiotica*, 127(1/4), 169–198.

Brier, S. (2013). Cybersemiotics: A New Foundation for a Transdisciplinary Theory of Consciousness, Cognition, Meaning and Communication. In L. Swan (Ed.), *Origins of Mind. Biosemiotics*, (vol. 8, pp. 97–126). Springer.

Bueno, O. (2022). Content, Context, and Naturalism in Mathematics. In S. Wuppuluri, & I. Stewart (Eds.), *From Electrons to Elephants and Elections.* The Frontiers Collection, (pp. 287–306). Springer.

de Finetti, B. (1929). Probabilismo. Saggio critico sulla teoria della probabilità e sul valore della scienza. Biblioteca di Filosofia. *Editrice F. Perrella, 1931*, 1–57.

de Finetti, B. (1930). Fondamenti logici del ragionamento probabilistico. *Bollettino dell'Unione matematica italiana, 9*(1930), 1–3.

de Finetti, B. (1930). Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Accademia Nazionale dei Lincei*, S. **6th**, vol. 4, Fasc. 5 (1930), 251–299.

de Finetti, B. (1969). Sulla proseguibilità di processi aleatori scambiabili. *Rendiconti dell'Istituto di Matematica dell'Università di Trieste, 1*, 53–67.

Frege, G. (1892). Über Begriff und Gegenstand. Translated in P.T. Geach, On Concept and Object, *Mind* (1951) 60(238), 168–180.

Frege, G. (1893). Grundgesetze der Arithmetik. Selected, translated, and edited in M. Furth *The basic laws of Arithmetic*, University of California Press (1964).

Gabora, L., & Kitto, K. (2013). Concept combination and the origins of complex cognition. In L. Swan (Ed.), *Origins of Mind. Biosemiotics*, (vol. 8, pp. 361–381). Springer.

Harris, Z. S. (1991). *A Theory of Language and Information*. A Mathematical Approach: Oxford University Press.

Hatcher, W. S. (1982). *The Logical Foundations of Mathematics*. Pergamon Press.
Herrmann-Pillath, C. (2021). The Natural Philosophy of Economic Information: Autonomous Agents and Physiosemiosis. *Entropy, 23*(3), 277.
Jeffery, K., Pollack, R., & Rovelli, C. (2019). On the Statistical Mechanics of Life: Schrödinger Revisited. *Entropy, 21*(12), 1211.
Kiverstein, J., Kirchhoff, M. D., & Froese, T. (2022). The Problem of Meaning: The Free Energy Principle and Artificial Agency. *Frontiers in Neurorobotics*, 16.
Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *The Royal Society Publishing*, 8(6).
Lakatos, I. (1961). What does a mathematical proof prove? In J. Worrall, & G. Currie (Eds.), *Mathematics, Science and Epistemology* (1978, pp. 61–69). Cambridge University Press.
Laplace, P. S. (1814). *Essai philosophique sur les probabilités*. Translated in F. W. Truscott, & F. L. Emory (Eds.), *A philosophical essay on Probabilities* (from the 6th ed., 1902). John Wiley & Sons.
Morgenstern, O., & von Neumann, J. (1944). *Theory of games and economics behavior*. Princeton University Press.
Muşat, B., & Andonie, R. (2020). Semiotic Aggregation in Deep Learning. *Entropy, 22*(12), 1365.
Niiniluoto, I. (2022). Concepts, Experts, and Deep Learning. In S. Wuppuluri, & I. Stewart (Eds.), *From Electrons to Elephants and Elections.* The Frontiers Collection, (pp. 577–586). Springer.
Ongstad, S. (2022). Perceptions of Context. Epistemological and Methodological Implications for Meta-Studying Zoo-Communication. *Biosemiotics, 15*, 497–518.
Polani, D., Martinetz, T., & Kim, J. (2001). An Information-Theoretic Approach for the Quantification of Relevance. In J. Kelemen, & P. Sosík (Eds.) Advances in Artificial Life. *Lecture Notes in Computer Science*, vol. 2159. Springer.
Roli, A., & Kauffman, S. A. (2022). Emergence of Organisms. *Entropy, 22*(10), 1163.
Rovelli, C. (1995). Relational Quantum Mechanics. *International Journal of Theoretical Physics, 35*(8), 1637–1678.
Rovelli, C. (2015). Relative information at the foundation of physics. In A. Aguirre, B. Foster, & Z. Merali (Eds.), *It from Bit or Bit from It? On Physics and information* (pp. 79–86). Springer.
Rovelli, C. (2018). Meaning and Intentionality = Information + Evolution. In A. Aguirre, B. Foster, & Z. Merali (Eds.), *Wandering Towards a Goal* (pp. 17–27). Springer.
Schrödinger, E. (1944). *What is life? (1992)*. Cambridge University Press.
Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*(379–423), 623–656.
Sharov, A. A. (2010). Functional Information: Towards Synthesis of Biosemiotics and Cybernetics. *Entropy, 12*(5), 1050–1070.
Summers, R. L. (2023). Lyapunov Stability as a Metric for Meaning in Biological Systems. *Biosemiotics, 16*, 153–166.
Surov, I. A. (2022). Natural Code of Subjective Experience. *Biosemiotics, 15*, 109–139.
Velazquez, J. L. P. (2020). On the emergence of cognition: from catalytic closure to neuroglial closure. *Journal of Biological Physics, 46*, 95–119.
Weaver, W. (1948). Science and Complexity. *American Scientist, 36*(4), 536–544.
Weaver, W. (1949). Recent Contributions to the Mathematical Theory of Communication. *A Review of General Semantics*, 10(4), 261–281. Special issue on Information Theory (Summer 1953),
Whitehead, A. N., & Russell, B. (1963). *Principia Mathematica* (2nd ed.). Cambridge University Press.
Wiener, N. (1965). *Cybernetics, or control and communication in the animal and the machine* (2nd ed.), The MIT Press.
Wittgenstein, L. (1963). *Philosophical Investigations* (2nd ed.), Basil Blackwell Ltd.