Learning and the Evolution of Conscious Agents

Eva Jablonka[1,2*] and Simona Ginsburg[3]

[1]The Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University, Israel; Centre for Philosophy of Natural and Social Science; [2]London School of Economics and Political Science, UK; [3]Natural Science Department, The Open University of Israel, 1 University Road, POB 808, Raanana 4353701, Israel

*Corresponding author: Eva Jablonka, jablonka@tauex.tau.ac.il

Text: 13,244 words (including table and legends, but without title, abstract and bibliography).

Abstract

The scientific study of consciousness or subjective experiencing is a rapidly expanding research program engaging philosophers of mind, psychologists, cognitive scientists, neurobiologists, evolutionary biologists and bio-semioticians. Here we outline an evolutionary approach that we have developed over the last two decades, focusing on the evolutionary transition from non-conscious to minimally conscious, subjectively experiencing organisms. We propose that the evolution of subjective experiencing was driven by the evolution of learning and we identify an open-ended, representational, generative and recursive form of associative learning, which we call Unlimited Associative Learning (UAL), as an evolutionary transition marker of minimal consciousness. This evolutionary marker provides evidence that the evolutionary transition to consciousness has gone to completion and allows reverse-engineering from this learning capacity to the system that enables it – making possible the construction of a toy model of UAL. The model allows us to identify some of the key processes and structures that constitute minimal consciousness, points its taxonomic distribution and the ecological context in which it first emerged, highlights its function and suggests a framework for exploring developmental and evolutionary modifications of consciousness. We point to ways of experimentally testing the relationship between UAL and consciousness in human and in non-human animals and discuss the theoretical and ethical implications of our approach. The framework we offer allows the exploration of the evolutionary changes in agency, value systems, selective processes and goals that were involved in the transition to subjective experiencing from a perspective that resonates with the approaches of bio-semioticians.

Keywords: Agency; Evolutionary Transition; Goal-directed behavior (GDB); Imagination; Unlimited Associative Learning (UAL); Vivaciousness

1. Introduction

According to the journal *Biosemiotics*, "biosemiotics is dedicated to building a bridge between biology, philosophy, linguistics, and the communication sciences. Biosemiotic research is concerned with the study of *signs* and meaning in living organisms and systems. Its main challenge is to naturalize biological meaning and information by building on the belief that signs are fundamental, constitutive components of the living world." (https://www.springer.com/journal/12304). However, the frequent use of terms such as meaning, goal and agency is often regarded with suspicion by biologists, who feel that this use of language introduces notions which have been developed for and tailored to the symbolic-linguistic human world and are inappropriate in other contexts. One of the problems is that the bridge between the symbolic-linguistic human world of meaning and the biological world of cells and bacteria cannot be effectively constructed without going through the evolutionary processes that have led to non-symbolic consciousness in non-human animals. Although bio-semioticians engage with evolution (e.g., Pagni & Simanke, 2021; Sharov & Tønnessen, 2021), the origins and evolution of different varieties and levels of consciousness, a research project which has been neglected for much of the 20th century, have not received sufficient theoretical attention, so important parts of the bridge are missing.

Our own approach to the study of consciousness, which addresses many of the questions that are central to the biosemiotic approach, is evolutionary. We use the terms *consciousness* and *subjective experiencing* as synonyms and qualify consciousness when we want to highlight different levels of consciousness such as imaginative consciousness or reflective/symbolic (human) consciousness. Our focus in this paper is on the evolutionary transition between non-conscious and minimally conscious modes of being, but we also (very briefly) discuss the evolutionary transitions to imaginative and to reflective-symbolic modes of conscious life. Our conceptual framework overlaps with that of bio-semioticians in that the notions of goal and agency are central to our approach (Ginsburg & Jablonka, 2019, 2020, Birch et al., 2020a, 2021; Zacks et al., 2022). However, in this article we restrict our notion of *agent* to organisms, we differentiate between *goal* and *function*, and we employ the notion of *functional information* (Jablonka, 2002a, Fresco et al., 2020), which is seen as problematic by some bio-semioticians (central concepts are italicized and defined in Box 1).

Our approach to the study subjective experiencing is comprehensively discussed in our 2019 book *The Evolution of the Sensitive Soul* (henceforth referred to as TESS). We see subjective experiencing as constituted by cognitive-neurological functional operations such as integration within and between modalities, sensory-motor mappings, memory and evaluation processes, and regard the functions of these constitutive processes as some of the specific functions of subjective experiencing (TESS, Ginsburg & Jablonka, 2020a). However, before we present our position, we discuss the central explanatory frameworks for studying living organisms, which, like our evolutionary framework, recognize the goal-directed nature of biological activities (section 2). We then describe our constitutive-evolutionary approach to consciousness (section 3), which focuses on what we call the teleological transitions in evolution, present our methodology for studying the transition to the conscious mode of being[1], and identify the coupled functional processes and structures that constitute its dynamic organization. Section 4 focuses on the evolution of minimal consciousness, which, we suggest, was driven by the evolution of *learning* and examine its ecological context and its evolutionary effects, and in section 5 we follow the evolutionary sophistication of consciousness in birds and mammals, and the emergence of symbolic consciousness in the *Homo* genus. We return to our general conceptual framework in the summary section (6) where we compare our model to some current theories of consciousness and discuss some of its implications and the future research directions it opens up.

2. Frameworks of explanation

An explicit framework for the study of nature was suggested by Aristotle and has served as a foundation for thinking about nature for millennia, although the specific terms he used, especially the notion of the final cause, received theological or mystical non-Aristotelian interpretations by later thinkers. Aristotle identified four types of causal accounts that are all required for a comprehensive explanation of natural phenomena:

---

[1] We use the term teleology and teleological as general terms covering all goal-directed behavior (GDB). These include behaviors that do not depend on conscious will or preconceived design (teleonomic GDB) as well as GDB that is driven by mental intention, desires or reflectively and rationally guided goal-directed behavior. The term "mode of being" is used in this article within the Aristotelian teleological framework. For a discussion of an ecological notion of a mode of being (e.g. terrestrial aquatic, aerial) see Ginsburg and Jablonka 2020a.

"There are four causes: first, the final cause, that for the sake of which; secondly, the definition of essence (and these two we may regard pretty much as one and the same); thirdly, the material; and fourthly, that from which the source of movements comes." (Aristotle, Generation of Animals, 715a 407).

Let us take an example of which Aristotle could not have been aware, but which fits his scheme just as well as those he was familiar with: a natural object, a bacterium that we want to study and understand. The molecular building blocks such as proteins, nucleic acids,sugars and lipids are the bacterium's "material cause"; the chemical reactions between them are the "efficient cause" (the source of action and re-action), and the dynamic architecture of the bacterial cell as a whole, which leads to its persistence over time is the "formal cause". The formal cause is also, in this case, the intrinsic "final cause" because it is the organizational dynamics of the organism that leads to the nutrition of the bacterium and to its reproduction (i.e., to its long-term persistence). A satisfactory explanation of all living and non-living products of nature must give an account in terms of all these four causes.

The living products of nature, living organisms, were of special importance for Aristotle, and he called the intrinsic, dynamic organization that ensures an organism's (e.g., a plant's) persistence over time as an individual and as a type, "soul". He defined the soul in terms of the final, formal and efficient causes: "The soul is the cause or source of the living body. The terms cause and source have many senses. But the soul is the cause of its body alike in all three senses which we explicitly recognize. It is (a) the source or origin of movement, it is (b) the end, it is (c) the essence of the whole living body." (On the Soul 415b 9-13). Aristotle did not include matter in the definition of the soul because although every mortal soul is embodied, the soul of different types of organism requires different matter (the relevant matter for Aristotle is the *parts* of the organism; different organisms, such as oaks and dogs have different parts). He provided a general definition that can be applied to all mortal living beings including the problematic case of rational (human) mortals.

Two and a half millennia later, Nikolaas Tinbergen suggested a general framework of explanation that also consists of four causes, and that, like the Aristotelian scheme, can be applied to all living organisms (Tinbergen, 1951, 1963). Tinbergen's causes (also called "questions") include phylogenetic "ultimate" causes, functional causes which provide explanation in terms of current utility, developmental causes that give an account in terms of the ontogenetic construction of the trait of interest, and immediate causes – an account in terms of the current underlying mechanisms. Tinbergen regarded this explanatory framework as both necessary and sufficient for the comprehensive scientific study of all living organisms (Bateson and Laland, 2013). For example, when investigating a bacterial biofilm, biologists need to provide an explanation that includes an account of the mechanisms involved in producing a shared matrix in which the bacteria are embedded, in terms of the ontogeny of the biofilm over time, in terms of the current function of the bacterial biofilm, and in terms of the evolutionary history of bacterial biofilm formation. It is quite clear that higher-level goals and corresponding *values* such as mental motivations (passions and desires), or a plan of action based on deliberate logical reasoning serve no explanatory causal role in this case. The similarities and differences between Aristotle's and Tinbergen's "causes" are shown in figure 1.

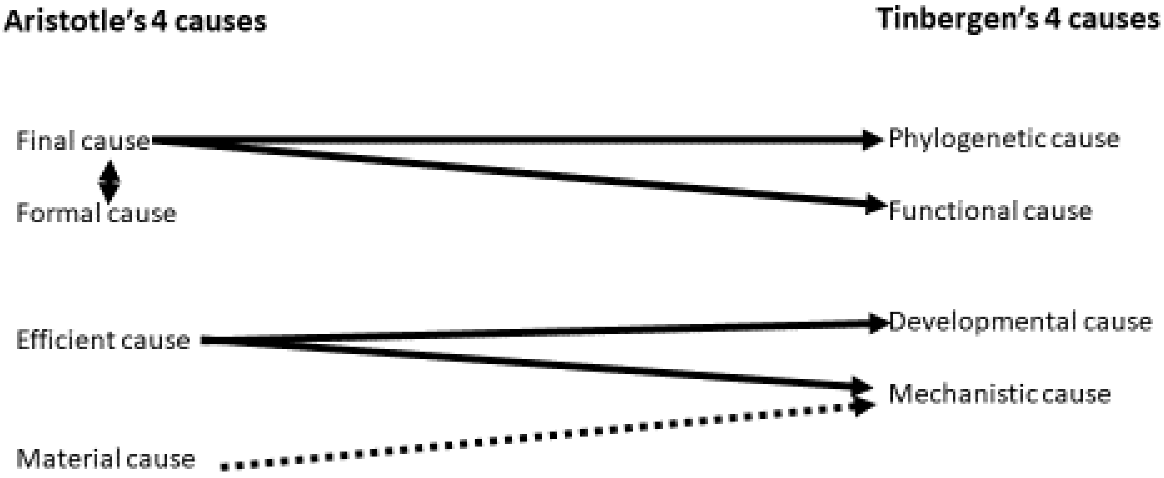

Figure 1: Correspondences between Aristotle's and Tinbergen's 4 causes.

There are clear correspondences between most of Aristotle's and Tinbergen causes with the exception of Aristotle's material cause, which has no obvious parallel with causes in Tinbergen's scheme. However, "material" in Aristotle's scheme corresponds to the parts of the system, and the parts are components of what we call mechanism today. There is therefore, some, albeit not very clear relation, between Aristotle's material cause and Tinbergen's mechanistic, immediate cause (hence the dashed arrow). Note that the Aristotelian formal and final cause are one and the same – this is the case when thinking of living dynamics in general. The exception is the case of human-made artefacts, where a distinction between the formal and the final cause can be made.

What about about explanations in terms of subjective *mental states*, in terms of desire, passion, imagination, that seem warranted in the case of animals such as corvids and apes? Tinbergen's main interest was animal behavior and he assumed that some animals may have mental states. However, he completely avoided mental causes because he believed that even in cases where mental causes do exist, as for example when a feeling of hunger impels a dog to seek food, such subjective feelings cannot be studied: "The ethologist does not want to deny the possible existence of subjective phenomena in animals, he claims it is futile to present them as causes, since they cannot be observed by scientific methods" (Tinbergen, 1951, p. 5).

Mental causation did, however, play a central role in Aristotle's framework. He distinguished between three soul levels: "plant, beast, and man" (On the Soul., 431b2–4) although as a good biologist he recognized some gray areas. The most basic and non-mental is the "nutritive/reproductive" soul of plants, the sole goal of which is self-maintenance (through nutrition at the individual level and through reproduction at the lineage level). The second soul is the "sensitive" soul of animals, where goals are driven, in addition, by mental states such as passions, desires and in some animals by imagination, and its goal is to satisfy felt needs, while the third is the "rational" soul of humans, the goal of which is the satisfaction of abstract symbolic values like "the good" or "the true" (see TESS chapter 1 and 10, and Ginsburg & Jablonka, 2020a for discussions). These soul-levels are hierarchically nested and constrained: the sensitive soul is nested within the nutritive/reproductive soul, and the rational soul is nested within the sensitive.

A similar carving of biological reality, though one that avoids the explicit mentioning of mental states, was suggested by Daniel Dennett (1995). Dennett based his categorization on a nested hierarchy of targets of selection rather that values and goals, and distinguished between four types of organisms and four types of *selection*: Darwinian, such as bacteria and plants, where goals are set by natural selection during phylogeny and where genetic selection alone needs to be considered; Skinnerian organisms, the behaviors of which are selected as well as their genes; Popperian organisms, in which selection also occurs among imagined actions, and Gregorian symbolizing animals (humans), who can select, in addition, between symbolically-represented possibilities. Figure 2 describes the relations between the Aristotelian soul levels and Dennett's generate-and-test tower, which is his metaphor for types of selection and adaptation operating at different organizational biological levels.

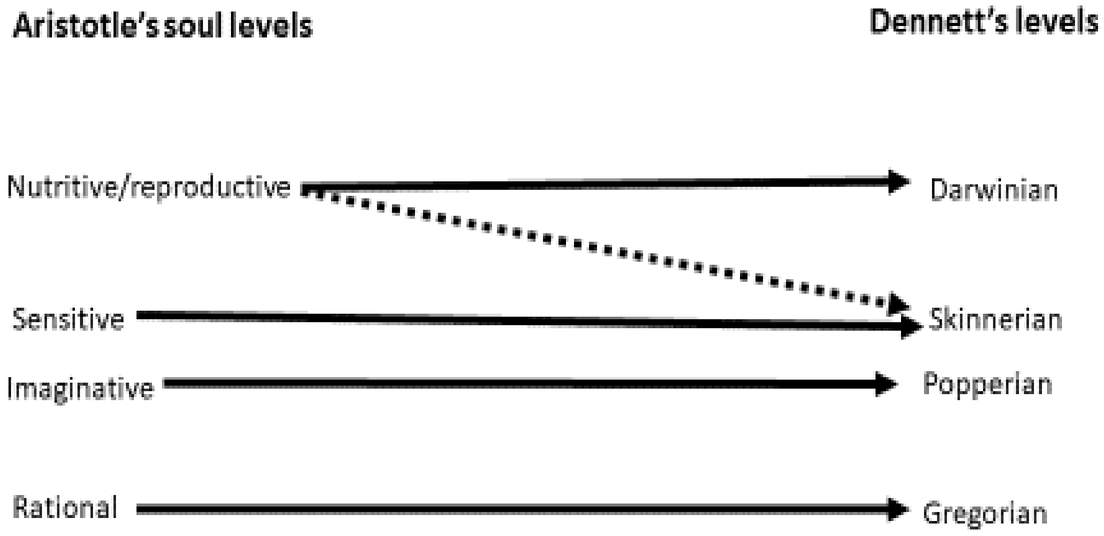

Figure 2: Correspondence between Aristotle's soul levels and the levels in Dennett's generate-and-test tower. Aristotle considered animals with imaginative souls, and distinguished between animals that can move and relocate and those that cannot, but his general scheme is based on the "plant, beast, man" trio, which categorizes living organisms according to their value systems and goals (survival and reproduction; satisfaction of desires and felt needs; and satisfaction of abstract values like truth and justice). Dennett's characterization is based on types of selection (genetic, behavioral, representational, representational-symbolic) as a classifying criterion. The dashed arrow indicates that some Skinnerian organisms may have only a nutritive/reproductive soul.

2.1 An Aristotelian-evolutionary approach and the centrality of goals

Our own approach is an evolutionary reframing of the Aristotelian scheme. However, we also use the selectionist framework suggested by Dennett to explore the transition to imaginative consciousness, and adopt the evolutionary-transition approach developed by Maynard-Smith and Szathmáry (1995) to examine the transition to imaginative consciousness (discussed in section 5). We reframed the Aristotelian approach to the soul in terms of evolutionary teleological transitions: from non-living to living systems, from living non-conscious to living conscious systems and from the latter to symbolic-rational ones, focusing on the transition to the sensitive soul, the transition to animal consciousness (TESS chapters 6-8; Ginsburg & Jablonka, 2020a; 2021). Distinct sets of *value* system delineate new types of goals for each of these three teleological modes of being, which can be construed as an open-ended evolutionary processes, driven by different value systems. According to this approach, consciousness is not a universal property of all matter. It is thought to be constituted by a specific type of dynamic organization that can be found only is *some* groups of living organisms[2].

We reasoned that if we can identify the evolutionary transition from a non-conscious to a conscious mode of being and describe this transition in terms of the changes in the system's functional organization, we would be able to characterize the mechanisms and dynamics that constitute a minimal conscious system without being misled by later evolved neural and behavioral associations and dissociations.

Our view that the representation of goals and the striving for goals are central to the process of subjective experiencing and can be regarded as it overall function was influenced by William James' approach to consciousness (James, 1890). James introduced his view by addressing a dilemma: "A low brain does few things, and in doing them perfectly forfeits all other use. The performances of a high brain are like dice thrown forever on a table. Unless they be loaded, what chance is there that the highest number will turn up oftener than the lowest?" (James, 1890, volume I, p. 139). He suggested that consciousness "loads the dice" of the noisy neural activity of a complex brain, and this is its function, it is "a *fighter for ends*, of which many, but for its presence, would not be ends at all" (James, 1890, James's emphasis). In other words, James suggested that the function of consciousness is to open a new, hitherto inaccessible, realm of goals, which are consciously perceived and desired.

2.2 Agency and plasticity

A notion that is related to the general concepts of function and goal in both non-sentient and sentient organisms and that requires discussion from an evolutionary perspective is the notion of biological *agents*. Biological agents are defined as dynamic systems (organisms are the paradigmatic example) that display unified, adaptive, goal-directed, plastic (flexible) behaviors

---

[2] We are aware that our evolutionary approach is not universally shared – there are panspsychists who believe that all matter is conscious, dualists who separate mind and body, and bio-psychists according to whom living entails sentience, so all living organisms are considered sentient. The discussion of these different approaches their merits and problems is beyond the scope of this article.

and adaptive internal organization (Okasha, 2018). Since there is evolutionary continuity between different biological agents (bacteria, dogs, humans), how is the agency of non-sentient organisms like bacteria related to that of sentient and reflective ones like humans?

The unified, goal-directed, adaptive-functional aspects of biological agency have been discussed earlier, but what about agential *plasticity*? There is something about living organization that cannot be fully captured by most formal models, and this "something" is the fundamental, inherent plasticity of all living organisms, the internal endless flux of material and energy which is the precondition for their amazing adaptive flexibility. All living organisms are spontaneously, inherently proactive, not just reactive (Bertalanffy, 1952; Brembs, 2011; Longo et al., 2015), and spontaneous exploratory activity occurs at all levels of biological organization. Examples are random and semi-random genetic mutations and epimutations; "noise" in biochemical and neural networks, default-network activity in the brain, behavioral-locomotory explorations in moving organisms, and cultural variations. The selective effects of most variations (including spontaneous mutations and epimutations in biochemical networks) are developmentally shaped though processes of differential stabilization involving silencing, elimination, activation or network reorganization that either return the system to a previous state of homeostasis, or shift development onto alternative developmental trajectories that lead to a new homeostatic state. The processes of exploration-stabilization that underlie these processes are the foundations of adaptive developmental and evolutionary canalization and plasticity (West-Eberhard, 2003). We call the biochemical, neural, and cultural networks that are the preconditions for any developmental adjustments *plasticity default networks.*

It is, we believe, the spontaneous activity, the internal flux, which is inherent to all living systems, that has led and is still leading biopsychists to assume that there must be something more to life than the functional and structural coupling of adaptive biochemical processes, something that only a recognition of turbulent inwardness, intuitively related to what we call subjectivity, can capture (Jonas, 1966; Thompson, 2007; Bray, 2009, Reber, 2019). It is not surprising that the machine metaphor is seen as problematic (Nicholson, 2013). Machines are usually not seen to be endowed with such restless inwardness (but see Riskin, 2016).

Although we disagree with biopsychists' assumption that all living organisms, in virtue of being alive, are also phenomenally conscious (Reber, 2019), we are in sympathy with their demand for the recognition of the inner, restless, turbulent state that is the condition for all modes of living, and that can be described in terms of a non-conscious, dynamic internal nascent "ego". All living beings can be described as dynamic networks with a default inner turbulent, water-based materiality, which is necessary for their *open-ended plasticity* that allows selection-evolution at different levels (genetic and epigenetic in all organisms, behavioral in some, and symbolic-cultural in humans). We suggest the term *vivaciousness* to describe this inner, dynamic default state of the living, water-based "wetware" of living beings (Bray 2009), which is necessary for their self-maintenance during ontogeny and which enables their re-production. Vivaciousness should not be confused with the old notion of vital force used by vitalists. It is the sum-total of all the internal physical dynamic processes of living beings, and although it is a biological primitive it has been further honed by natural selection.

In neural organisms, vivaciousness includes the dynamics of the nervous system. In some neural organisms, such as cnidarians, ctenophores, acoels and other phyla, neural dynamics is part of the vivaciousness and open-ended plasticity of these organisms which has not yet acquired an additional intrinsic type of value. In non-conscious neural organisms, the network activity can be described as "white-noise", an incessant activity that is the basis of exploration-stabilization processes that are involved in on-line responses, in the modulation of reflex reaction (sensitization and habituation) and in very limited associative learning (LAL; TESS chapters 6 and 7). In sentient organisms these neural dynamics have evolved to become the dynamics of consciousness, which is identified with the activity of the default network (Raichle et al., 2001; Buckner & DiNicola, 2019; although the default network was identified only in humans and a few mammals, we believe a default network will be found in all sentient organisms). In humans, we call the internal vivacious default state of the symbolic mode of being *reflectiveness*; it is required for shared, communicable and veto-able representations of norms (if and how this default network differs from the default network of a subjectively experiencing rat is a question we cannot at present satisfactorily answer). All three inner states – vivaciousness, consciousness and reflectiveness – can be described as having what Kant called "purposefulness without purpose" (he used the phrase to explain the notion of aesthetic judgment) and all are maintained and modulated by selection because all are forms of adaptive plasticity which is necessary for the adaptive responsiveness of the system.

How is the open-ended plasticity, which takes different forms in different types of organisms, related to the specific functions of the system? We suggest that as with the case of life, where functions are attributed to metabolism, membrane assembly, replication, etc., functions should be attributed to the parts and processes that constitute consciousness, including the default activity of the nervous system. However, since goals that are perceived, felt and driven by desires are possible only when consciousness is in place – we can generally say that the overall function of consciousness is to open up a new, open-ended, realm of goals. The question is how such a system evolved and how its simplest possible manifestations are instantiated.

3. Methodology: An evolutionary transition approach to consciousness

Our approach to the study of the transition to subjective experiencing was inspired by the methodology employed by one of the founders of systems chemistry, the Hungarian chemist Tibor Gánti, who developed it for the investigation of the transition to life from non-living complex chemical systems (Gánti, 1987, 2003). Gánti began by compiling a list of capacities that most biologists considered to be jointly sufficient for evolutionarily persistent life (we call such a list a consensus list). He identified eight such jointly sufficient capacities or criteria: maintenance of a boundary, metabolism, stability, information storage, regulation of the internal milieu, growth, reproduction, and irreversible disintegration (death). On the basis of this broad consensus, he constructed a system of coupled mechanisms and processes that implement these capacities, and built a simple model. Finally, he identified an experimentally tractable marker of a minimal living teleonomic system. Gánti suggested that *unlimited heredity* – the capacity to form lineages that vary in open-ended ways from the initial system, so the number of possible different variants is vast – is the marker for minimal life, so that any system endowed with open-ended heredity must show the consensus list of characterizing life-capacities (Gánti, 2003; Maynard Smith and Szathmáry, 1995 further sharpened and explained this concept). If we find a system with the capacity for unlimited heredity anywhere in the universe, we should be able to re-construct or reverse-engineer on its basis the simplest teleonomic living system of which it is part.

We applied Gánti's methodology to evolutionary transitions to other modes of being – the conscious and the rational modes, and we generalized his notion of a diagnostic transition marker (Ginsburg & Jablonka, 2015; Bronfman et al., 2016a,b; TESS chapter 1; Ginsburg & Jablonka

2020a; Birch et al., 2020a, 2021a). We called a diagnostic capacity that requires that all the properties attributed to a particular *teleological mode of being* are in place, an evolutionary transition marker (ETM), and suggested an ETM for the transition to minimal consciousness. Like Gánti, we started by compiling a list of characteristics of minimal consciousness that can be characterized in neural, cognitive, behavioral and phenomenological terms and that most scholars would regarded as jointly sufficient for the simplest conceivable agent to be deemed subjectively experiencing.

3.1 characterizing minimal consciousness

Below we present our consensus list of characteristics:

- Unification and differentiation: the capacity to perceive objects and processes as integrated wholes (an apple, a dance, a scene), and at the same time recognize that they are made of parts, so different wholes (different apples, different faces, different dances) can be discriminated, and a composite whole can be reconstructed from a partial combination of their parts. Many animals can discriminate between patterns and activities that predict danger and those that predict safety. The females of many bird species compare the patterns or songs or artefacts that males build, and choose the most impressive ones.
- Global accessibility and broadcast: This is the capacity to link and integrate information from perception, memory and evaluative systems and broadcast the output back to input and executive systems that lead to relevant actions. These networks of back and forth interactions construct maps of predictive relations between stimuli and their reinforcing outcomes, between actions and their predicted sensory outcomes, and between outcomes and their predicted value. These representations are formed in a common neural space that contextualizes and updates incoming inputs, enabling comparison, discrimination, generalization and prioritization of evaluations, all of which inform decision-making.
- Temporal depth: The capacity to hold-on to incoming information, to have a “working memory”, so the present has some duration.
- Flexible value attribution: the capacity to alter and update the rewarding or punishing values of actions and sensory stimuli. Since many inputs with different valences impinge

on the organism, the value system must allow the ranking of concurrently encountered reinforcing stimuli, enabling trade-offs and context-sensitive re-evaluations. Different types of physiological responses to internal and external signs and actions are perceived as positive internal states (caring, lusting, joyful states) or negative (painful, fearful, hungry states), guide different types of action and lead to different overall evaluative (affective) states.

- Exploration-stabilization (variation/selection) processes in the nervous system: the capacity for vigilance and for selective attention through excluding and amplifying mechanisms that alter the general alertness of the animal and can render some stimuli and actions more salient than others according to predictive evaluations based on present and past experience.
- Intentionality (aboutness): the capacity to map, not merely integrate, inputs from the world, body and their relations.
- Agency and *goal-directed behavior*: organisms have bodies enabling object-oriented spatial cognition requiring freedom of movement. Such organisms infer that their specific actions have specific consequences (e.g., some lead to damage to themselves and such consequences are to be avoided). They learn to choose the positively valued consequences of their actions and avoid the negative ones in a flexible way suggesting voluntary behavior.
- Self–other distinction from a point of view (a sense of self): the capacity to construct models of the world and body and to respond to them from a stable perspective. The organism is able to distinguish between a stimulus that is the result of its own action and an identical stimulus that is independent of its action (e.g., being tickled by someone else and self-tickling lead to very different reactions). This is true not just for the outcome of reflex actions, but for outcomes of *learned* action outcomes. The ability for instrumental self-learning and its distinction from (yet connection with) world-learning, requires the construction of flexible world and self-models.

We present the neurological, behavioral and phenomenological manifestations of these partially overlapping characteristics in Table 1 (this is an extended version of the table in Birch et al., 2020a).

Table 1: characteristics of minimally conscious organisms and their neurophysiological, behavioral and phenomenological correlates

	Neurophysiological and cognitive mechanisms	**Behavioral attributes**	**Phenomenological Manifestations**
Unification and differentiation	Integration of information through synchronous and sequential binding mechanisms; interaction between the capacity for feature separation and gestalt perception[1,2,3]	Learning to discriminate between composite complex patterns[4]	Different features of an object are perceived as bound together into a single percept (e.g., we experience an apple as round, red, fragrant and smooth). Thus, there are perceptual gestalts. Yet different apples can be discriminated[5]
Global accessibility and broadcast	Multimodal integration of inputs from sensory, evaluative and memory systems that inform the formation of adaptive action patterns[6].	Multimodal discrimination learning[7,8].	Unified experience of events that include sights, smells, sounds, emotions and memories, all together, leading to voluntary motor behavior[5].
Temporal depth	Working memory[9]	Capacity for trace-conditioning; delayed match-to-sample learning; ability to learn from video sequences [10,11]	A sense of a 'specious present'; a feeling that the present has duration[12]
Flexible value attribution	Integrative systems for valuing and revaluing different stimuli and for weighing different needs against each other[13, 14]	Capacity for reversal-learning and second-order conditioning; flexible decision-making in situations of conflict[15,16]	Changing feelings of pleasure and displeasure which depend on context; changing desires, emotions and moods, following world and body changes[5, 14,17]
Exploration-stabilization; selective attention	Attentional networks[18]	Habit- formation and autopilot behavior, that can be switched to vigilant attention. Degradation of learning under distracting conditions[6]	The focusing of attention and vigilance[5];vigilance attention to detail[19]
Intentionality (aboutness)	Hierarchical mapping of body and world[20]	Goal-directed behavior based on goal representation [21-24]	Things are perceived and felt as being about the world, about the body[17]

Agency and goal directed behaviour	Spontaneous brain activity and formation of flexible attractors for flexible learning[25]	Exploration guided by motor-sensory-motor (MSM) loops[26]; sensitivity to action outcome and outcome value re-evaluation[22,26,27]; suppression of action-control leads to depression[28]	A feeling of efficacy; exploration-joy[29]; in humans a feeling of free will
Self/other registration	Interaction of neural models of self, body and motivated action, generating egocentric representations of the moving animal in space[30,31]	Damage to self-model (e.g. following stroke) leads to feelings of disowning one's body parts[32]	The feeling of ownership of one's experiences; the structure of experience as a "point of view" on the world[17,33,34.]

[1,2] Baars 2005,a,b; [3]Dehaene & Changeux, 2011; [4]Couvillon & Bitterman, 1988; [5]Searle, 2004; [6]Dehaene, 2014; [7]Mansur et al., 2018; [8]Telles et al., 2017; [9]Baddeley, 1986; [10]Lucas et al., 1981; [11]Bangasser et al., 2006; [12]James, 1890; [13]Morsella, 2005; [14]Solms, 2021; [15]Hadar & Menzel, 2010; [16]Gewirtz & Davies, 2000; [17]Metzinger, 2003; [18]Petersen& Posner, 2012; [19]McGilChrist, 2010; [20]Feinberg & Mallatt 2016; [21]Dickinson, 2012a; [22]Dickinson & Balleine, 1994; [23,24]Balleine & Dickinson, 1998a,b; [25]Freeman, 2003; [26]Ahissar & Assa, 2016; [27]Dickinson & Balleine, 2000; [28]Macmillan, 2021; [29]Panksepp, 2005; [30]von Holst & Mittelstadt, 1950;. [31]Merker, 2005; [32]Vallar & Ronchi, 2009; [33]Merker, 2007; [34]Williford et al., 2018.

The listed characteristics are not an ad-hoc collection. As we show in the next sections they are partially overlapping and are functionally and causally related, constructing a unified complex dynamic system.

2.2 Unlimited associative learning is the evolutionary transition marker (ETM) for minimal consciousness.

After compiling the list, we identified an ETM for the transition to consciousness. As we noted earlier, an ETM is a capacity such that once we have evidence of it, we have evidence that all the capacities in the consensus list are in place. Once an ETM is identified, the corresponding mode of being (subjectively experiencing in our case) can be inferred (figure 3). Moreover, it is possible to reverse engineer or reconstruct on the basis of the ETM the minimal system that enables it (a protocell in the case of life, a cognitive neural system in the case of consciousness).

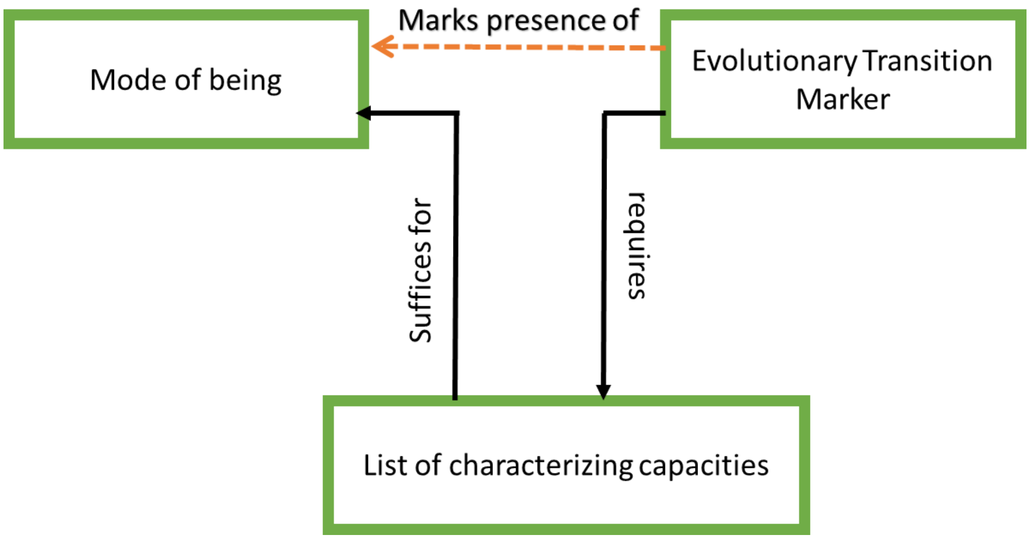

Figure 3: The general idea of an ETM (based on Birch et al., 2020a). An ETM is a diagnostic capacity that requires that all the consensus properties that are jointly sufficient to attribute a particular mode of being to an entity, are in place. It therefore marks the mode of being of interest.

We suggested that the ETM for subjective experiencing is a domain-general, open-ended form of associative learning, which we called unlimited associative learning (UAL), the ontogenetic equivalent of Gánti's unlimited heredity. This type of learning requires that all the consensus consciousness features we listed earlier are in place. UAL can be operationalized by the following (all testable) learning capacities:

(i) Discrimination learning: learning to discriminate among *differently organized, novel, multi-featured* patterns of sensory stimuli, and between novel, composite action patterns (e.g., Couvillon & Bitterman, 1988; Mansur et al., 2018; Telles et al., 2017). A capacity for such learning requires representations of predictive relations among world-stimuli, actions and outcomes.

(ii) Trace-conditioning: the capacity to learn about a predictive, novel, compound neutral stimulus or an action-pattern even when there is a time gap between the presentation of the compound stimulus or action and its reinforcement (e.g., Lucas et al., 1981; Dickinson et al., 1992; Bangasser et al., 2006; Moyer et al, 2015; Rodríguez-Expósito et al., 2017). This capacity requires temporal integration of information.

(iii) Learn to flexibly alter the evaluation of predictive stimuli and action patterns and show goal-directed behavior, which enables the animal to make motivational tradeoffs, prioritizing different outcomes in a context-sensitive manner (Solms, 2021). The animal can alter the valence attributed to patterns of sensory stimuli and motor actions when conditions change. Since animals have to *learn* about the desirability of commodities (incentive learning), if desirability is changed (as in "outcome devaluation" Holland & Rescorla, 1975; Adams & Dickinson, 1981; Mizunami, 2021), the animal can re-evaluate the outcome by direct contact with the commodity in the new state. This requires a flexible, integrating evaluative system, and allows goal directed behavior (Dickinson a& Balleine, 1994, 2000)

(iv) Second-order conditioning: learning about the predictive value of new stimuli or action on the basis on previously learned stimuli and actions. This leads to the formation of chains of actions (e.g., Holland & Rescorla, 1975; Hussaini et al., 2007) and to categorizations and transfers (e.g. Benard et al., 2006) and requires evaluative flexibility.

If one accepts that the list of capacities is a plausible characterization of minimal biological consciousness (consciousness as we currently know it), UAL can be considered as a good ETM of minimal consciousness. As Table 1 shows, the behavioral/learning attributes correspond to the characteristics of the consensus list. Unification and differentiation are needed for discrimination learning among patterns and updating the value of composite new stimuli; global accessibility enables multimodal discrimination-learning and the assignment of changed priorities to the same stimuli or actions according to internal and external conditions; trace-conditioning points to integration over time, to working memory and temporal depth; the capacity for reverse-learning and to the revaluation of goals points to a flexible evaluation systems enabling trade-offs and second-order conditioning; the ability to flexibility shift attention, to voluntarily maintain attention and to pick relevant stimuli out from the background or become vigilant, indicates that selective processes of exclusion and amplification are in place; mapping of patterns of stimuli, actions and their relations, as well as the storage of associative links is a manifestation of intentionality; goal-directed behavior requires agency, and self-world registration is needed so that the organism will distinguish between current and learned own-action-dependent outcomes (and the stimuli that predict them) and outcomes that are independent of its own actions.

2.3. The functional architecture of UAL and some testable predictions regarding the relation between UAL and consciousness

UAL is a system property. It is a domain-general, generative, recursive, and representational type of associative learning and requires an organization that support these features. First, since it is instantiated in a finite biological system it requires hierarchical, recurrent associations between world, body and prospective action-program representations. Second, memory cannot be local: a dedicated memory sub-system that stores event-representations which are perceived as composites, is required. Third, a dedicated evaluation sub-system that can assign valence to any compound input configuration and that enables context-sensitive prioritization needs to be present. Fourth, the motor sub-system must be based on body mapping allowing the representation of prospective actions. Fifth, these sub-systems must come together within a common neural space and the outputs must inform decision-making. Predictive processing, with top-down inputs generating “hypotheses” about the state of the world, the body, and their relations, which are updated by bottom up inputs, is an integral part of this picture (Solms & Friston, 2018; Seth, 2021; TESS, chapter 8). We present a toy model of UAL (figure 4) which portrays, in a very simplified and schematic manner, some central features of the functional architecture of the UAL system.

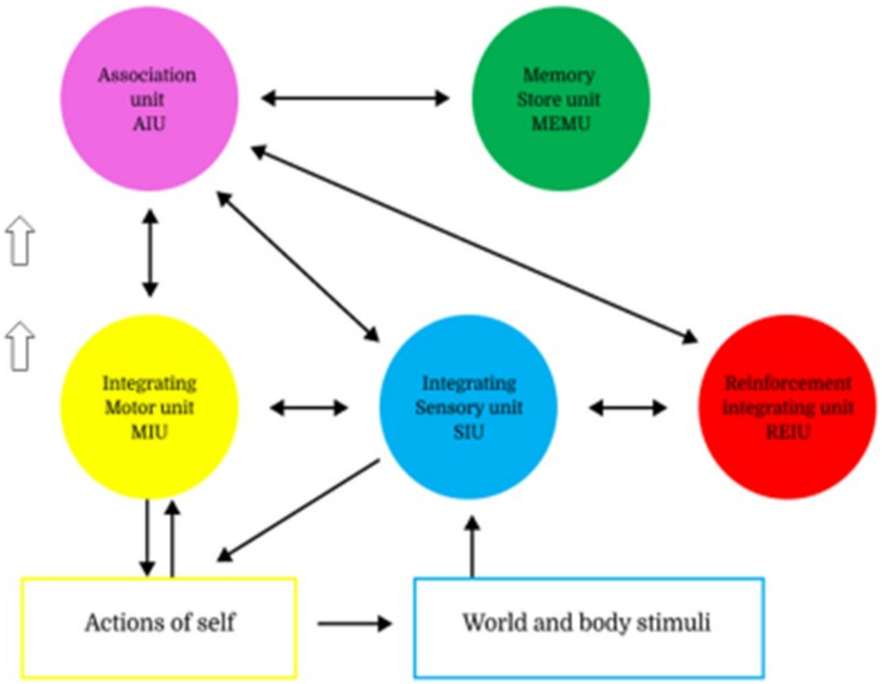

Figure 4. A highly simplified scheme of the functional architecture of UAL. UAL depends on reentrant (back and forth) connections (depicted by double-headed arrows) between sensory (SIU), motor (MIU), reinforcement (value, REIU) and memory (MEMU) integrating processors. There is a central association unit (AIU) at the core of the network (although it is possible that AIU is distributed between MEMU and REIU). Hierarchical intervening levels are indicated by empty arrows on the left. We did not show here the interactions between self and world monitoring, the direct interactions of the sensory and motor units with the memory unit, the mappings of SIU-REIU and MIU-REIU relations and their interactions with MEMU at intervening hierarchical levels, and the different levels and types of memory involved in this dynamics (see TESS chapter 8 and Bronfman et al., 2016 for more details).

We suggested that the dynamics of UAL architecture in living organisms give rise to minimal consciousness through reentrant interactions among the system’s representations. We called these active integrating systemic states “categorizing sensory states” (CSSs), because they represent, evaluate and categorize through their dynamics, input, action and outcome. If the organism is food-deprived, its internal sensory state interacts with both innate scaffolds and memory traces associated with the attainment of a desired outcome (e.g., food), as well as with represented predictors of past evaluated action-outcomes. We argued that the dynamic, overall sensory state of the organism constituted by these dynamics is as close as we can get to a third-person depiction of the architecture underlying private, mental states.

The UAL model we presented is descriptive – a computational model of UAL has not yet been developed, so quantitative predictions are not, as yet, possible. However, the proposal that UAL is an ETM of minimal consciousness does lead to some testable predictions (discussed in detail in Birch et al., 2020a).

First, the claim that UAL is a system suggests that the different elements group together in UAL animals, and lead to the predictions that:

(a) The elements of UAL are expected to be ontogenetically correlated. Developmental studies are expected to show that the development of one element facilitates or enables the development of one or more of the other elements.

(b) The elements of UAL are expected to be phylogenetically correlated. Finding that a species has evolved one of the elements of UAL (e.g., discrimination learning) increases the probability that the species has evolved the whole UAL package.

(c) The elements of UAL are expected to be medically correlated. Brain injuries that affect one element will not leave the other elements completely unaffected, but may leave more limited forms of learning unaffected.

Second, the claim that UAL is an ETM of conscious awareness suggests that:

(d) Experimental protocols such as backward masking that selectively switch off conscious perception in humans, leaving unconscious perception in place, are expected to selectively switch off or substantially degrade UAL, while leaving more limited forms of learning in place in both humans and animals. Hence, we expect that humans and non-human animals will perform poorly in UAL tasks such as spatial learning, discrimination learning, trace conditioning and reverse learning when the predictive perceptual cue is masked (subliminal). These UAL tasks require that the subject is aware of a goal to be reached on the basis of predictive relations among cues, controllable actions, and outcomes, so it will be sensitive to learned devaluation of action-outcomes and outcome re-evaluations. So far, experiments on human subjects (and few experiments on monkeys) support these predictions, although only some UAL tasks or their proxies have been tested (e.g., Öhman & Soares, 1993; Clark & Squire, 1998, 1999; Skora et al., 2021; Ben Haim et al., 2201).

(e) We expect that blindsighted humans and animals, which show degradation of specific aspects of consciousness, will be unable to perform UAL on stimuli presented in the blind region of the visual field but will be capable of more limited forms of learning.

(f) The neural signatures of subjective experiencing in humans and other animals, whatever they turn out to be, are expected to be correlated with UAL.

These predictions are empirically testable and are important constraints on the construction of any future UAL computational model. Additional insights into the architectural constraints and affordances of UAL require investigations of its evolutionary history.

4. The evolution of UAL

There are four aspects of the evolutionary history of UAL and minimal consciousness that we address in this section: (i) the evolutionary precursors of UAL, which were transformed during the evolution of learning; (ii) its distribution in the living world; (iii) the ecological context in which UAL first evolved and the evolutionary dynamics involved, and (iv) the evolutionary effects of UAL and minimal consciousness. Most of these aspects have been extensively discussed in TESS (chapters 6-9), so our exposition here is inevitably brief and simplified.

4.1 From limited learning to UAL

We suggested that the evolutionary transition to UAL was based on the elaboration of simple associative learning, which we called limited associative learning (LAL). LAL includes classical and operant conditioning of elemental predictors of positively and negatively valued stimuli and actions. Such limited learning, which is enormously adaptive (compared to learning by sensitization and habituation alone) is apparent in flat worms, sea slugs and nematode worms as well as most other animals. LAL *does no*t enable complex spatial or discrimination learning, trace-conditioning and flexible reverse-learning, and is expected not to be sensitive to action-outcome and outcome-value degradation. At the anatomical-functional level LAL does not require a dedicated memory sub-systems that supports basic declarative memory, nor does it require a dedicated integrating system for prioritizing actions and directing attention (the dominance of some values over others can be determined on the basis of locally determined salience, and local mechanisms of overshadowing and blocking). Surprisingly, LAL requires a centralized nervous system. This points to the need for central integration (of associatively learned sensory predictors and actions) in multicellular animals with sense organs and muscle sheets (TESS chapter 7).

We suggested (TESS chapters 7,8) that increase in size and the development of body parts with large ganglia drove (i) the evolution of the action-modeling motor integrating unit (MIU), enabling flexible control of the movement of body parts; (ii) the evolution of sense organs, such as eyes and olfactory organs; this occurred initially by increase in general size, and then drove the evolution of sensory integrating units (SIUs) that model the sensed world and the sensed body; (iii) the evolution of a declarative memory system that stores the mappings of world, body, and action; (iv) the evolution of a value system that can prioritize integrated needs, and (v) the evolution of integrating region/s (AIU), which enabled the interactions of these sub-systems (although the AIU may have been initially distributed among the integrating units and became distinct during later evolution). The construction of new dedicated, integrating memory, evaluation and association units and a new hierarchy of sensory and motor units is shown in figure 5.

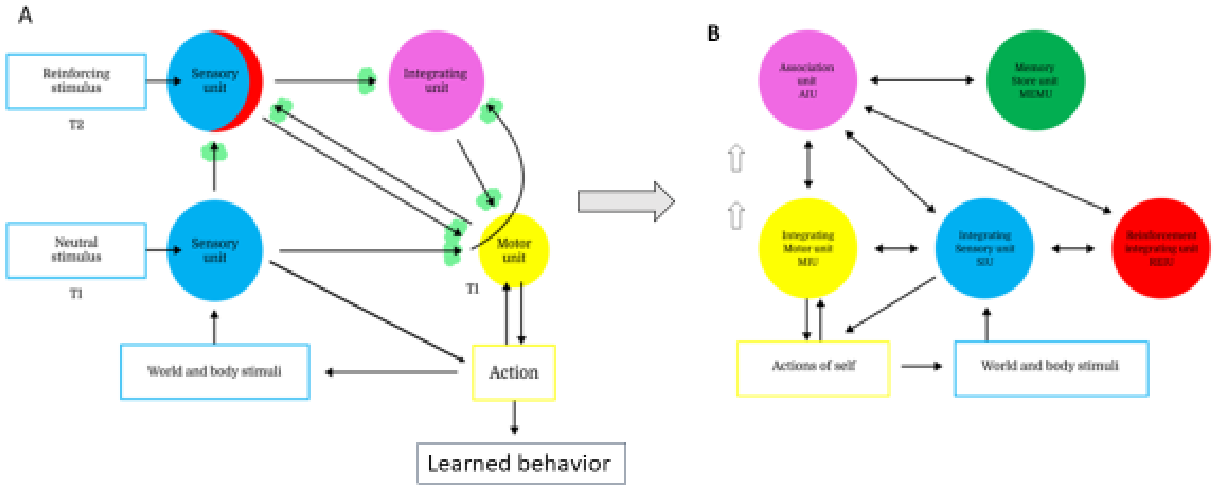

Figure 5: From Limited, subliminal associative learning (A, top) to UAL (B, bottom). Note that memory in (A) is local (green spheres at locations of association; there is no dedicated declarative memory system); reinforcement is local too (red crescent, representing the

reinforcement value of the representation of the unconditional stimuli; there is no dedicated system that integrates and prioritizes all evaluations); In (B) which reproduces figure 4, there are new hierarchical levels and new dedicated structures (reinforcement and memory) that implement UAL. The UAL functional architecture can be seen as an extension of the limited associative learning architecture in (A).

4.2 The distribution of UAL and minimal consciousness

The capacity for UAL, and by implication of minimal consciousness has been found in three animal phyla: most vertebrates, some arthropods and one group of mollusks, the coleoid cephalopods (the squid, the cuttlefish and the octopus) (see TESS table 8.1).

Although the animals belonging to these three phyla have very different brain structures, the functional organization of their brains is similar. There is, especially, striking overall similarity between the neural architecture and functional organization of the mammalian and insect brains. The mushroom bodies and the central complex, situated in the insect protocerebrum, are believed to be either homologous or analogous to the hippocampus and basal ganglia, respectively. There is also striking similarity between the cerebellum and insect mushroom body and central complex in motor learning, and between the vertebrate tectum and the insect fan-shaped body. Similarly, the hemiellipsoid bodies of crustaceans exhibit functional homologies to the hippocampus.

There are less detailed studies comparing the cephalopod molluscs and vertebrates or arthropods. However, in the octopus and cuttlefish, two main structures, the vertical lobe and the superior frontal lobe of the brain form complex networks that, together, are analogs of the vertebrate hippocampus. These integrating brain structures underlie cephalopod learning and memory (see table 8.2 in TESS).

Did UAL (and minimal consciousness) evolve independently in vertebrates, arthropods and coleoid cephalopods, did it emerge twice, or did the three groups have a single common UAL ancestor? The time of origin and the marked difference in brain organization of cephalopods, which evolved UAL ~250 million after the vertebrates and arthropods, suggest that UAL is likely to have evolved independently in coleoid cephalopods. The question then is if UAL in vertebrates and arthropods has single common origin or if UAL evolved in parallel in the two phyla during the Cambrian. The remarkable functional similarity between insect and vertebrate brains and their shared molecular kit supports the suggestion that UAL in arthropods and vertebrates had a common origin. However, if, limited associative learning evolved earlier and independently in the two groups, parallel evolution of UAL based on analogous sensory-motor biases enabled by the modular ganglia and brain organization in the two groups, may have led to parallel evolution. At present this question is not resolved, although we regard the latter alternative as highly likely (TESS, chapters 7,9).

We would like to stress that our conclusions about the distribution of UAL and minimal consciousness are very tentative. Although several different scholars came to the same distribution of consciousness on the basis of different criteria (Feinberg & Mallatt, 2016; Godfrey-Smith, 2020; Barron & Klein, 2016), our suggestion is open to extensions and modifications, since the information on the learning capacities of animals in other taxa is scant, patchy and in some cases completely lacking.

4.3 The ecological context and the evolutionary dynamics of UAL

Fossil evidence suggests that almost all animal phyla appeared during the Cambrian era, including many arthropod groups and one vertebrate group (fish), which had the brain structures that support UAL. In addition to fossil traces of the complex tripartite brains of arthropods and fish, there is also fossil evidence for the evolution of predation and escape from predation.

What was special about the Cambrian and what drove the enormous diversification of animals during this geologically-short era? There are many different hypotheses (reviewed in TESS chapter 9, table 9.1) but there is general agreement that the pre-conditions for the Cambrian explosion included: biologically significant increases in oxygen concentration; pulses of global warming, the result of methane release associated with polar movements, which led to increased nutrient cycles and productivity; changes in sea level that led to the flooding of continental margins, which greatly increased the range of habitable shallow-water areas. These led to changes in the chemical constitution of the oceans, including an increase in calcium and phosphate concentrations (the permissive conditions for the bio-mineralization, which animals exploited). It was in these permissive conditions that animals could grow in size and engage in productive burrowing and swimming. Movement and coordination of movement became important and the larger nervous systems and muscle sheets of the larger Cambrian animals enabled this. Predation exerted intense continuous selection for sensory organs and movement coordination.

Animals that could learn and predict had an upper hand in this interactive, competitive Cambrian world. We suggested that the evolution of learning – especially of UAL which appeared during this era – was one of the factors that drove the great Cambrian explosion (Ginsburg & Jablonka, 2010). We argued that the ability to fuse signs into composites and distinguishable percepts and to generate and distinguish between different action patterns that can be "analyzed", discriminated and evaluated, drove the evolution of UAL. We conjectured that arthropods first evolved this advanced learning capacity, and became the most efficient predators of that era and the most important drivers of the evolution of all other groups, both in their own phylum and in other phyla. Their superior learning abilities led to co-evolutionary arms races in interacting species, and to intraspecific diversification that led to speciation. It drove the evolution of hard parts, of camouflage and toxicity, and of improved locomotion and more sophisticated sensors such as eyes in prey and competing predators. One of the important effects of this arms race was the generation of *learning arms races*: a better learning ability of predators selected for the prey's improved ability to learn. The evolution of UAL in the grazing fish of the Cambrian was, we believe, an evolutionary response to the strong selection imposed by their UAL arthropod predators. The evolution of UAL in coleoid cephalopods, around 250 million years after the Cambrian, have been associated with the internalization of shells that increased the cephalopods mobility, and seem to have resulted from greatly increased competition for targeted prey with fish predators (Kröger et al., 2011). It is likely that fish's superior learning ability drove the evolution of maneuverability and self-learning in cephalopods and led to UAL in these mollusks.

Behavioral innovations, often the results of learning, are considered a major evolutionary engine of adaptation and diversification. Hardy (1965) suggested that the dramatic radiations of reptiles, mammals, and birds were driven by their enhanced learning ability, leading to the invasion of new niches and to new selection regimes. Alan Wilson and his coworkers called this guiding effect of behavior on evolution and adaptive radiations "behavioral drive" (Wyles et al., 1983), and Bateson (2006) called behavior "the adaptability driver". According to all these suggestions, evolutionary change starts with ontogenetic learning adaptations. Genetic variations supporting the adaptive ontogenetic learned adjustments follow. Such "phenotype-first" evolution is central to West-Eberhard's argument that plasticity, including the plasticity afforded by learning, has played a major role in evolution, with genes being, usually, "followers not leaders in evolution" (West-Eberhard, 2003, p. 20).

4.4 The evolutionary effects of minimal consciousness: suffering, active forgetting and selection through choice

There are many advantages to UAL and minimal consciousness. The ability to discriminate between composite percepts and acts, the ability for pattern-completion that induces composite-memory retrieval, the ability for plastic self-learning that enables goal-directed behavior driven by flexibly prioritized physiological needs, and the ability to make cumulative improvements to one's actions and build up skills, are probably the most obvious. Animals with these capacities can flexibly adjust to a huge range of conditions and events during their own lifetime, through ontogenetic selection processes.

But UAL and consciousness have also less obvious and less intuitive evolutionary effects. Like any great and complex adaptation, consciousness incurred costs and led to new selection regimes. Suffering is the most obvious, subjective and private, price of consciousness, but though costly at the personal level it is selectively advantageous. Feelings like pain, anxiety and fear, are unpleasant precisely because they are evaluations of actual or potential harm to oneself and usually lead to adaptive responses protecting the individual from greater harm: individuals that do not feel pain show no self-protecting behavior and often suffer injuries and early death (Schon et al., 2018/2020), and individuals who are not anxious may be easy prey. However, individual suffering exceeds its optimal adaptive effects. Since with UAL partial cues may serve (through pattern-completion) as predictors of more than one composite percept, each of which has a different valence and leads to a different response, reactions to partial cues of adversity may often lead to "false positives", overreactions that are inappropriate, overly fearful or overly aggressive. Randolph Nesse (2001) called the principle underlying such overreactions to adversity (he focused on flight and anxiety reactions) the "smoke detector principle". When the price of encountering "fire" (predator, foe, etc.) is very high, and the costs of flight upon encountering the predicting “smoke” are lower, such reaction are selectively advantageous. However, some (inevitably imperfect) countermeasure to the frequent neuro-hormonal subjectively felt stress and its resulting ill-health could and did evolve. The flexible nervous system underlying UAL requires that both memory and forgetting are regulated, and the chronic stress due to learning-induced overreactions (over-learning) led to further regulation – to the co-evolution UAL with a more efficient stress response, self-control, and active forgetting. While early Cambrian sentient animals may have been overly anxious and neurotic, mechanisms at the cellular, neuro-hormonal, and immunological level that restricted the duration and extent of memory, that promoted active forgetting, and that controlled and limited arousal must have been selected, rendering the post-Cambrian animals more mentally and physically healthy (for an extended discussion of the evolution of forgetting see TESS chapter 9).

Just as consciousness and UAL explain the origins of suffering, so do they explain the origins of joy. We (TESS, chapter 4, 5 and 8), followed the proposal of Panksepp (2005) that the emotion he called SEEKING (associated with spontaneous exploration) per-se is intrinsically positively valued. The spontaneous, perceptual and motor exploratory activity of animals with UAL is, we suggested, inherently pleasurable, because it enables learning and leads to knowledge, which is adaptive. The link between pleasure and knowledge was (not surprisingly) made by Aristotle: “All men by nature desire to know. An indication of this is the delight we take in our senses; for even apart from their usefulness they are loved for themselves; and above all others the sense of sight. […] The reason is that this, most of all the senses, makes us know and brings to light many differences between things.” (Aristotle, Metaphysics 1 1-6; 1984b). Perceptual and motor explorations in an animal with UAL enable world- and self-learning and entail the pleasurable feeling of the animal’s active, exploring, living body, making the animal desire to care for its survival (Humphrey, 2011). Animals show their joy of perceiving, freely acting and making decisions when, after suffering from sensory and motor deprivation, they are relieved of it and can exercise their agency. One touching example is that of pit-ponies who were imprisoned for years in the darkness of coal mines, which “…when brought to the surface, the mules tremble at the earth radiant in the sunshine. Later, they go almost mad with fantastic joy.” (Crane, 1894). We believe that the feeling of agency, the ability to explore, exercise curiosity and engage in goal-directed behavior, is intrinsically joyful and that alongside suffering, its inevitable complement, joy and suffering were the first basic feelings of a sentient animal.

There are many varieties of suffering and pleasure, and the evolution of the richness and diversity of felt emotions and drives was honed during the evolution of learning, with the emotional repertoire differing in animals inhabiting different niches (Birch et al., 2020b). Importantly, variations among perceived sensory and motor patterns and variations in the evaluations of such patterns determine the ways in which selection operates on interacting conspecifics as well as on individuals from other interacting species (Jablonka, 2021). Sexual selection is a case in point: the complex patterns on the peacock's tail could evolve only if peahens could discriminate among variant patterns and assign value to them; the song of the male nightingale evolved because females discriminate among and prefer complex and varied songs, and the evolution of the visual and olfactory patterns of flowers could evolve because of the ability of insects and birds to discriminate among visual and olfactory patterns. It is no coincidence that unlike animal-pollinated flowers, wind-pollinated flowers have no complex visual patterns and smells (Prum, 2017). Darwin argued that sexual selection of compound perceptual and action patterns by mate preference is an indicator of mentality or consciousness, and underlies animals' and humans' sense of beauty (Darwin, 1871).

Sexual selection through mate choice is a special case of selection through subjective choice. When communication signs are exchanged between prey and predator, for example between insects and their bird predators, elaborate and precise camouflage patterns such as those seen in stick insects could evolve because best camouflaged insects have a survival advantage. In all cases of such selection by choice or "intentional selection" the *receiver* of the sign has to be conscious (the term intentional selection was suggested by Noble, 2021)[3]. The sender need not be conscious: the evaluation of signs by conscious receivers, such as insects or birds, led to complex patterns of color and smell in the non-conscious flowers they feed on. Without such sentient receiver-based choices, the patterns in the living world would not have the sensory

---

[3] The term "semiotic selection" suggested by Maran and Kleisner 2010, is a broader concept than the intentional selection concept of Noble, which explicitly assumes consciousness. The qualification and elaboration of the concept of selection within the extended evolutionary synthesis (EES) is discussed by Jablonka and Lamb 2014; Jablonka 2021, and Kull 2021 discuss the term in their commentaries on Noble 2021 target article.

richness and intricacy we observe. If we find intricate sensory communication signs between beings on other planets, this may serve as an indicator that the receivers of these signs are sentient. Closer to earth, much of the behavioral richness and the morphological and perceptual diversity that we observe on our planet would not be possible without consciousness.

5. Beyond UAL: The evolution of imaginative and symbolic consciousness

Minimal consciousness, driven and constructed through the evolution of UAL, was only the first stage in the evolution of consciousness. Since we maintain that the evolution of the cognitive abilities that are based on UAL determines the contents and levels of consciousness, we can identify gradations and stages in the evolution of cognition and of consciousness through the study of their behavioral and neural facets. Focusing our attention on qualitative changes in the evolution of consciousness, we identify imaginative cognition and planning as a qualitative change not only in cognition but also in consciousness, since imaginative animals have subjective experiences not just of the directly impinging present but also of the remembered past and the planned future (Zacks et al., 2022 provide extended discussion of this topic). The teleological transition to the human symbolic, language-dominated mode of being was another, complex major evolutionary transition, leading to the ability not only to imagine but to share memories and plans by instructing the imagination of interlocutors (Dor, 2015).

We defined fully-formed imagination that allows planning for the future as the offline recombination, transformation, evaluation, and selection of simulated episodes that are based on past experiences. Behavioral studies of planning and episodic memory – memory of events and episodes, of what, where and when things happened – point to members of some vertebrate lineages (the only phylum where comparative studies of imagination are available) as endowed with episodic memory and imaginative cognition, and to the hippocampus as a central hub of episodic memory processes. This kind of imaginative capacity enables the animals to subjectively experience the events in the virtual worlds of the past and the future.

We argued that the transition to imagination, though not involving new value systems and hence not considered a teleological transition, was a major evolutionary transition sensu Maynard Smith and Szathmáry (1995). Such major evolutionary transitions entail the addition of (i) a new level of selection, (ii) a new level or type of individuality or “self”, (iii) an additional layer of hierarchical control, and (iv) an increased division of labor. The evolution of imaginative consciousness in some animal lineages (e.g. great apes and corvids) satisfies all these conditions: it necessitates selection among neural representations and imagined scenarios – "letting our hypotheses die in our stead" (Popper, 1972); it constructs a new kind of narrative self and enhanced self-monitoring; it is linked to the addition of layers, subdivisions and specializations in the hippocampus, the declarative memory hub, where episodic memory is encoded and reconstructed, as well as the elaboration of regulatory connections with executive and reward systems; and it involves new levels of top-down (cortical and neo-cortical) control. Animals with imaginative consciousness can defer gratification, "think" before acting, inhibiting not just reflexive responses but also, to a considerable extent, the feelings that elicit them, so actions can be voluntary and controlled. The comparative evidence suggests that the evolution of imaginative consciousness in vertebrates was gradual, and seems to have evolved several times to different degrees and along different consciousness dimensions in different taxa (Zacks et al., 2022).

The evolutionary transition to symbolic language, is, according to our criteria, a teleological transition, because it involves a new category of goals and values – symbolic values, which guide humans moral communicative and aesthetic activities (TESS chapter 10; Ginsburg & Jablonka, 2020a). Dor (2015) describes the function of language as the instruction of imagination: it allow individuals to intentionally and systematically instruct their interlocutors in the process of imagining the intended content delivered by the speaker without actually experiencing it. The chains of words the speaker emits instruct their listeners in the process of imagining the meaning she intended to convey. The process involves the paring-down of the mental representation by the speaker, transforming them to chains of ordered words – culturally-evolved signs that stand for concepts and the conventions of their usage. The listener uses these chains of ordered signs as scaffolds to construct her own mental representation of the intended message.

The evolution of symbolic language may have involved, as Dor (in press) argues, an important prior evolutionary transition. Pointing, facial and bodily intentional gestures, pantomime and ritual – a suite of communication devices, that Donald (1991) calls mimesis, plausibly preceded the evolution of symbolic language. The ability to mime and the use of gestures – in order to ask questions, describe things, express agreement or disagreement, perplexity or certainty, send feedback and seal arguments – is uniquely human and appears in pre-linguistic children. According to Dor, this mode of communication allows more than the use of the body as a representational device: it requires that the message is delivered with a set of norms of communication that guide its interpretation, allowing teaching and rich, cumulative cultural evolution. Although still anchored in the here-and-now of the collectively experienced situation, mimetic communication allows a new kind of mental collaborative computation, extending the cognitive possibilities of the collaboratively communicating person and molding her identity as part of the normative collective, in which she partakes and which she internalizes. The social emotions of shame, guilt, embarrassment and pride, which are expressed in the uniquely-human blush (Darwin, 1872), express the powerful internalization of social norms in our genus (TESS chapter 10).

Symbolic language evolved on these foundations through the culturally-guided assimilation of the genetic variations that facilitated the process of encoding, decoding and storing of linguistic signs (Dor & Jablonka, 2010). It enabled communication about a sharable virtual realm leading to a huge expansion of the number of messages that can be exchanged, making sense of the ubiquity of lying and the notions of truth, falsity and an “objective” world (Tomasello, 2014). Importantly, from our value-focused perspective, it led to a new set of collective symbolic, culturally-evolved and culturally-specific values and goals – the notions of virtue and of a worthwhile life, of good and evil, of justice and freedom, that guide human life.

6. Discussion

Our approach has implications for a large range of topics, including the question of plant consciousness, the possibility of robot consciousness, and the many aspects of animal welfare. Before we very briefly touch upon these subjects, we would like to compare our model to some current models of consciousness and point to some conceptual and theoretical issues that are central to our approach.

6.1 Comparing the UAL model to other theories of consciousness

Although the UAL model which attempts to describe minimal consciousness has been based on the evolutionary history of animals’ nervous system and their modes of learning, it is reassuringly compatible with current consciousness models that have been mostly based on work on humans' cognition and neuroanatomy. Most current models emphasize specific aspects consciousness characteristics, such as information integration (Tononi et al., 2016), hierarchical recurrent interactions among neural maps (Lamme, 2020; Feinberg & Mallatt, 2016), the construction of an ego-center (Merker, 2007; Williford et al., 2018; Seth, 2021), and evaluative emotions (Damasio, 2021; Panksepp, 2011; Solms, 2021). The UAL model brings these components together within a cognitive-learning context.

The UAL model is closest to two models of consciousness, the first developed from a cognitive-neurophysiological perspective and the second from a behavioral perspective. The first is the global neural network (GNW) model developed by Changeux, Dehaene and their colleagues. According to the GNW model the dynamics that give rise to conscious mental states are based on recurrent interactions among sensory, motor, memory, value and attentional processors. The inputs from these systems become integrated in a common workspace, forming unified, coherent representations of the world, that are broadcasted back to the input systems and onwards to executive motor systems (Dehaene & Changeux, 2011; Mashour et al., 2020). We see our UAL model as a minimal version of a GNW, which does not require a single dedicated attentional network nor the neural structures and processes supporting metacognitive tasks. We see the question of the evolution of UAL and the evolution of a minimal GNW as overlapping questions (Zacks et al., work in progress).

The second model, the Hedonic Interface Theory (HIT) of Dickinson and Balleine is the only model other than ours, which suggests that learning and consciousness are evolutionarily linked and that goal-directed learning (GDB) is the function of consciousness (Balleine & Dickinson, 1998a; Dickinson & Balleine, 2000, 2010; Dickinson, 2012a). Dickinson and Balleine defined GDB in terms of modifiable action-outcome and outcome-value representations: "…an action is goal directed if its performance is mediated by the interaction of two representations: (1) a representation of the instrumental contingency between the action and the outcome, and (2) a representation of the outcome as a goal for the agent." (Dickinson & Balleine, 1994, p. 1). They showed that in rats, desirability is learned by manipulating the value of a learned outcome of an action (reviewed in Dickinson & Balleine, 1994, 2000, 2010), and that rats' decisions and human causal judgment show similar illusions under manipulations of the action-outcome contingency (Dickinson & Balleine, 2000).

An animal that can learn to distinguish between the consequences of its own actions and identical consequences that are independent of them, can choose to act, modify its action, or decide not to act. However, if its decisions are to be adaptive they must be anchored in its physiological needs. The access to internal physiological states, occurs, according to Dickinson and Balleine, through feeling the hedonic value of the outcome. This, they argue, is the function of feelings, of sentience. They suggest a dual-psychology model consisting of a primitive learning system that is based on stimulus-response (S-R, a reflex machine), which enables animals to learn through simple Pavlovian or instrumental habit learning and a second, more complex, later-evolved psychological system (cognitive system), enabling the formation of action-outcome (causal) representations that increase control of action, and an outcome-value association that enables learning about outcome desirability. The interaction between the two systems leads to access to internal states through feelings (for details, see Dickinson, 2012b).

We agree with Dickinson and Balleine that the function of consciousness is to control behavior through subjective feelings – to inform the animal about the desirability of a goal and motivate it to reach it, and we also agree that only conscious animals can exhibit GDB (as they define it). The functional, Jamesian characterization of consciousness as opening up a new, intrinsic realm of GDB has been one of the pillars of our UAL theory, and is in broad agreement with the HIT model. However, we believe that HIT neglects major aspects of consciousness such as perceptual consciousness and the sense of self. Another problem is the assumption that desire is a representation of a belief in the value of a goal, which does not have a built-in affect. We disagree, and believe, that desire/wanting in UAL animals has an inherent, intrinsic affective value that is related to what Panksepp called the basic emotion of seeking (Panksepp, 2011), as shown by the distress expressed by animals that are deprived of exploration and the possibility to control the outcomes of their actions (McMillan, 2020). We also believe that a unitary scheme such as that of UAL, according to which the breadth of learning had continuously increased through the evolutionary elaboration of simple Pavlovian and instrumental conditioning, is more plausible from an evolutionary perspective than a dual psychology model, and addresses aspects of consciousness that are not address by HIT (for a more extensive discussion of the merits of HIT and UAL see Jablonka & Ginsburg, in press).

6.2 Theoretical and conceptual issues

The relationship between values, actions, perceptions and high-level associative processes such as generalization and categorization is central to our view of consciousness. As we stressed throughout, *consciousness is a system property which emerges from the dynamics of the relation between sensory, motor, value and memory systems.* The attempt to find a single Archimedean point among these interacting capacities – value (Solms, 2021), motor action (Keijzer et al., 2013), perception (Feinberg and Mallatt, 2016), is, we believe, misguided. It was the evolutionary boot-strapping of of the *relations among these capacities* which constructed both the varied perceptual and evaluative aspects of consciousness as the animals evolved increasingly complex forms of learning (TESS chapter 6-8). These relations, which constitute the conscious perception of the world and the body and their relations, are more akin to the relations between the three spatial dimensions that bring a 3D object into being than to the relation between parts of a composite whole such as the interaction between hydrogen and oxygen atoms that generate water. As we argued, the construction of percepts and feelings through these relational dynamics enable organisms to attain goals that would otherwise be inaccessible to them, and it is at this level of analysis that we pitch our account of the function of consciousness as a whole (as a fighter for *desired ends*); the more specific functions of the parts, processes and relations that constitute consciousness are many and varied, and include all the functions of UAL (TESS chapter 4).

As we see it, the idea that feelings and what is popularly called “thinking” are separate and even contradictory, is due to our idiosyncratic human-symbolic cognition. Metacognitive symbolic processes such as those underlying symbolic categorizations and logical inferences are phenomenally experienced as unrelated to feelings, as “free” from feelings. We argued that this phenomenal experience reflects the strong inhibitory effects of metacognitive processes on the emotional system, which are the result of the evolution of imagination and especially of the instruction of imagination through language in humans. In other mammals, the inhibitory effects of metacognitive process are less potent. The unprecedented level of emotional control in humans, accounts, we believe, for to humans' equally unprecedented feats of creativity, as well as to large-scale collaborative acts of unspeakable (forgive the pun) cruelty and destruction (TESS chapter 10; Ginsburg & Jablonka, 2022).

There is another aspect of the relationship between feelings and cognition which is noteworthy. Learning is often regarded as a cognitive process which is more complex, and evolutionarily later than feelings (see, for an example, Panksepp, 2011). As we have argued, the opposite is actually the case: learning is extremely ancient, apparent in all forms of life. More generally, basal cognition in the simplest life forms is the basis for the sophisticated forms of cognition to which the term is usually applied[4]. However, although feelings emerged *with* UAL during phylogeny (feelings are constituted by UAL dynamics), UAL develops *after* the development of feelings during ontogeny: human babies and other young animals, which cannot manifest UAL, express feelings and basic emotions, because they are born with the neural architecture that supports and enables the integration of values that lead to feelings. The manifestation of UAL requires time-consuming learning which is therefore displayed at later developmental stages. As we have stressed, ontogeny does not recapitulate phylogeny in this case.

As noted earlier, our focus on functional explanations, on agency and on goals strongly resonates with the ideas of bio-semioticians. However, we attribute agency and goals in the full sense only to organisms, not to parts of organisms such as ribosomes, DNA, neurons, brains, hands or machines (we regard machines as the extended phenotypes of humans). The goals of living organisms – survival and reproduction, felt needs and symbolically-valued goals – cannot be attributed to their parts. There are, however, and inevitably (because of the process of evolution), what we call "gray areas", cases where our definitions and characterizations do not capture the nature of the entity we study. There is no definitive answer from our perspective to questions such as: are self-replicating molecules in a complex chemical broth, alive? Is a termite colony an organism?

---

[4] The two issues of The Philosophical Transactions of the Royal Society 2020 and 2021 https://royalsocietypublishing.org/toc/rstb/2021/376/1821 discuss basal cognition at depth and from multiple perspectives; we have discussed major transitions in cognition from the learning perspective in Ginsburg & Jablonka, 2021.

This is a good opportunity to highlight the biosemiotics research perspective that centers on the conceptualization and study of signs, which we did not sufficiently discuss in previous publications. We use the term sign to denote a “carrier” of functional information: a predictive, designating or representing input (e.g., predictive sensory cue such as a black cloud signaling rain, an alarm call, a welcoming gesture, a word, etc.) that requires a process of interpretation that guides the interpreter’s actions and re-actions. As the bio-semioticians, building on Peirce’s theory of signs stress, signs are usually used to denote carries or “vehicles” of functional information (Sharov & Tønnessen, 2021). Since living organisms process or interpret incoming inputs, and since interactions that involve learning are ubiquitous, the bio-semioticians’ focus on signs complements and enriches other ways of studying biology. Signs that are emitted by living organisms are of particular interest because biotic interactions are not only ubiquitous but also especially dynamic and challenging. As we see it, functionally significant systematic interactions among is a biological primitive, since reproduction, which is evolutionarily necessary for sustainable life, entails that products of reproduction interact, so even if encounters are just fleeing, organisms are never completely solitary and functional biotic interactions are inevitable. Since biotic interactions are typically flexible, recurring and demanding, we believe that it is biotic interactions, which have been the main driver of the evolution of signaling and of learning.

6.3 Implications and future directions

As mentioned earlier, our approach is relevant to current discussions of plant consciousness, robot consciousness and animal welfare. According to our model, which suggests a positive criterion of consciousness (the capacity for UAL), plants, which have limited learning ability, cannot be said to be conscious although their adaptive plasticity is formidable (for an extensive discussion of why we do not deem plants conscious, see Ginsburg & Jablonka, 2020b). However, like biopsychists we regard the machine metaphor, which seems to disregard vivaciousness, as inadequate. The hiatus between vivacious living organisms and non-vivacious machines is as large as that between conscious and non-conscious living organisms.

Vivaciousness may also be relevant for discussions of robot consciousness (Man & Damasio, 2019). UAL, which is a domain-general, generative, recursive and representational type of associative learning, may require material dynamics that are equivalent to those exhibited by biological materials, demanding that a UAL robot is built of soft, pliable and sensitive materials. A computational model of UAL and the construction of a robot realizing UAL, would go some way to elucidate the necessity to fulfill these requirements. Although the consciousness of a robot able to engage in full blown domain-general UAL would be difficult to establish, the ethical cautionary principle requires that we treat such a being with consideration (see Bronfman et al., 2021, for a discussion of robot consciousness from a UAL perspective).

The ethical considerations that follow from our approach with regards to animals with UAL are simple: if members of many animal species that were deemed non-conscious are in fact sentient beings according to our criteria, then they are moral patients and require that we treat them accordingly, considering their specific felt needs. Indeed, the multidisciplinary research on animal sentience is already leading to the extension of the scope of welfare laws, with cephalopods and some crustaceans added to vertebrates as species requiring welfare considerations (https://www.lse.ac.uk/news/news-assets/pdfs/2021/sentience-in-cephalopod-molluscs-and-decapod-crustaceans-final-report-november-2021.pdf). An extended ethical view of animal welfare has the potential to broaden our view of the web of interactions of which we and other animals are part, and counteract the planetary-scale devastation that our collective behaviors cause.

There are many research directions that our approach opens up. In addition to the construction of computational and robotic models of UAL and domain general intelligence, much comparative work needs to be done on the neural structures that instantiate UAL in different animal lineages, their evolutionary and developmental transformation, and their brain activities in different states, for example, as animals gradually emerge from anesthesia, at different stages of sleep, under the influence of drugs, or during relaxed states that engage the default mode network (or its precursors). Extending the range, methodological rigor and comparability of such studies, and relating UAL and imaginative cognition in multiple groups to neuro-physiological and cognitive mechanisms are both badly needed, since the existing animal studies, at all levels, are extremely partial and patchy and the behavior of many animal groups have not been studied. We believe that such studies, within the evolutionary approach that we advocate, can dissolve the "hard" problem without undermining the magic of consciousness.

Declaration

We declare no conflicts of interest.

Acknowledgment: We are grateful to the referees of this paper and to the editors for their helpful suggestions. The paper is dedicated to the memory of Marion J. Lamb.

Box 1: Definitions of Terms.

Agent: a dynamic system displaying unified, adaptive, *goal-directed,* plastic (flexible) internal organization and behaviors.

Consciousness/subjective experiencing: an inner and private, dynamic system activity constituted, according to our theory, by the evolved architecture of *open-ended associative learning* processes; these processes construct perception (like experiencing red), cognition (like remembering) and feelings (like thirst, joy, pain). Subjective experiencing can be attributed, currently, only to some living organisms.

Function: attributed to a trait (structure, process, relation) that systemically contributes to the *goal-directed behavior* of the encompassing system.

Functional information: any difference that makes a systematic difference to the *goal-directed behavior* of an *agent*. Functional information implies an interpretation process and an interpreter, so there is no functional information without interpretation. A *sign* is an input that carries functional information.

Goal: attributed to an object, process or state that satisfies an intrinsic *value* that guides a system's behavior.

Goal-directed behavior (general, teleonomic): behavior that lead to the attainment of *goal/s*. Goal-directed behavior that is based on intentions and beliefs requires that there is a representation of the instrumental contingency between the action and the outcome and a representation of the outcome as a goal for the agent. The perceptual predictors of valued outcomes usually guide intentional behavior.

Learning: a process leading to an experience-dependent behavioural response of a system. It requires that: (i) A sensory stimulus that originates either from the activities of the system or from the external world leads to a change in the internal state of the system. (ii) A memory trace of this change is stored; storage requires active stabilization and involves valence mechanisms of positive or negative reinforcement. (iii) Future interactions with the stimulus or associated stimuli lead to a change in the threshold of the behavioral response.

Mental states: internal dynamic states that are based on ontogenetic learning, generating perceptual and affective qualia. The goals to be satisfied are constructed by the organism (and in this sense are *teleonomic*), but the objects of the felt needs (e.g., food, air, mates, body integrity) are perceived as goals to be reached.

Plasticity: The ability of the same substance or material to assume different forms; in biology it is usually used to refer to the capacity of a single genotype giving rise to multiple phenotypes, in response to internal and/or external changed conditions. Open-ended plasticity is based on generative mechanisms such as those underlying genetic recombination and trial-and-error learning that can generate a vast number of variations from limited components.

Plasticity default networks: dynamic networks that form a default state which is necessary for open-ended plasticity and selection-evolution in living organisms. These open-ended selection-evolution dynamics have purposefulness without purpose. The dynamic inner states of these networks give rise to *vivaciousness*, *consciousness* and *reflectiveness,* which correspond to the states of living, subjectively experiencing and symbolic reflectivity. Vivaciousness is the basic default state of all living organisms, necessary for their consciousness and reflectiveness; consciousness in living organisms is necessary for reflectiveness.

Reflectiveness: an inner dynamic state of a biological agent endowed with a rational-symbolic value system. Such a value system is necessary for the open-ended plasticity that is the basis of shared, communicable and veto-able representations of norms.

Selection: the process of picking out a subset from a set guided by criteria or values that can happen at different levels of biological organization. Selection may involve multiplication and reproduction (Darwinian selection) or may involve differential stabilization that does not involve multiplication (sample selection). Darwinian selection can be natural, sexual and artificial, and, depending on the selective conditions, can be directional, disruptive, frequency dependent, stabilizing, canalizing, etc.

Sign: A sign refers to, denotes, designates, implies, points to or represents, something that Charles Morris calls “designatum” (an object, a process, a relation, an absence). Hence there is no sign (i) without a desginatum, (ii) without a process of interpretation and (iii) without an interpreter. The interpreter is an *agent*. A sign thus “carries” *functional information*. We use sign in this paper in the way it used in the biosemiotics literature and in everyday speech, as the designating or representing "vehicle" (e.g., predictive sensory cue such as a black cloud signaling rain, an alarm call, a gesture, a word, etc.).

Teleonomic behavior: goal-directed behavior that does not depend on conscious will or preconceived design.

Teleological Modes of Being: denote the kinds of dynamic organization that are characterized by the possession of distinct value systems (intrinsic reinforcement systems) that guide attainment or non-attainment of goals. "Teleological mode of being" refers to the living-non-sentient mode of being, the sentient mode of being, and the rational-symbolic (human) mode of being. Different teleological modes of being are characterized by the possession of distinct, hierarchically nested *goals* and *value* systems.

Value: intrinsic reinforcement that guides *goal* attainment or nonattainment through a process of *selection*. Values can be phylogenetic/ultimate (maintenance of homeostasis and homeorhesis supporting survival and reproduction); ontogenetic-mental (affective states like pain and pleasure); symbolic (abstract concepts like truth). In biological organisms the ultimate value constrains all others over evolutionary time.

Vivaciousness: the inner, plastic, default dynamic state of a living system which is necessary for sustained active living.

References

Adams C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer evaluation. *The Quarterly Journal of Experimental Psychology 33B*(2),109–121. https://doi.org/10.1080/14640748108400816

Ahissar, E., & Assa, E. (2016). A closed loop hypothesis of perception. *eLife 5*, e12830. doi:10.7554/eLife.12830 xxx

Aristotle. (1984a). *Generation of animals* (J, Barnes, Ed., W. D. Ross, Trans.). In *The complete works of Aristotle*. The revised Oxford translation, vol. 2. Princeton University Press.

Aristotle. (1984b). *Metaphysics, Book 1.* (J, Barnes, Ed., W. D. Ross Trans.) In *The complete works of Aristotle*. The revised Oxford translation, vol. 2. Princeton University Press.

Aristotle. (1984b). *On the soul* (J. Barnes, Ed., J. A. Smith, Trans.). *The Complete Works of Aristotle.* The revised Oxford translation, vol. 1. Princeton University Press.

Baars, B. J. (2005a). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research, 150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

Baars, B. J. (2005b). Subjective experience is probably not limited to humans: the evidence from neurobiology and behavior. *Consciousness and Cognition*, *14*, 7–21. https://doi.org/10.1016/j.concog.2004.11.002

Baddeley, A. D. (1986). *Working memory*. Clarendon Press.

Balleine, B. W. & Dickinson, A. (1998a). Consciousness – the interface between affect and cognition. In J. Cornwell (Ed), *Consciousness and human identity*, pp. 57–85. Oxford University Press.

Balleine, B. W. & Dickinson, A. (1998b). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology 37*, 407–419. doi: 10.1016/S0028-3908(98)00033-1

Bangasser, D. A,. Waxler, D. E., Santollo, J., & Shors, T. J. (2006). Trace conditioning and the hippocampus: the importance of contiguity. *Journal of Neuroscience 26,* 8702–8706. https://doi.org/10.1523/JNEUROSCI.1742-06.2006

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the Natural Academy of Science USA, 113*, 4900-4908.

Bertalanffy L von. 1952. *Problems of life—an evaluation of modern biological thought.* New York: Wiley.

Bateson, P. (2006). The adaptability driver: Links between behavior and evolution. *Biological Theory 1*(4), 342–345. doi: 10.1162/biot.2006.1.4.342

Bateson, P., & Laland, K. N. (2013). Tinbergen's four questions: An appreciation and an update. *Trends in Ecology and Evolution*, *28*(12), 712–718. doi: 10.1016/j.tree.2013.09.013

Benard, J., Stach, S., & Giurfa, M. (2006). Categorization of visual stimuli in the honeybee Apis mellifera. Animal Cognition, 9(4), 257–270. https://doi.org/10.1007/s10071-006-0032-9

Ben-Haim, M. S., Dal Monte, O., Fagan, N. A., Dunham, Y. & Hassin, R. R. (2021). Disentangling perceptual awareness from nonconscious processing in rhesus monkeys (Macaca mulatta), *Proceedings of the National Academy of Sciences, 118,* 15. doi: 10.1073/pnas.2017543118

Birch, J., Ginsburg, S., & Jablonka, E. (2020a). Unlimited associative learning and the origins of consciousness: A primer and some predictions. *Biology and Philosophy 35*, 56. doi.org/10.1007/s10539-020-09772-0

Birch, J., Ginsburg, S., & Jablonka, E. (2021). The learning-consciousness connection. *Biology and Philosophy 36*, 49. https://doi.org/10.1007/s10539-021-09802-5

Birch, J., Schnell, A.K. & Clayton, N.S. (2020b) Dimensions of animal consciousness. *Trends in Cognitive Sciences, 24*(10), 789–801. doi: 10.1016/j.tics.2020.07.007

Bray, D (2009). *Wetware: A computer in every living cell*. Yale University Press

Brembs, B. (2011). Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proceedings of the Royal Society B*, 278, 930–939. doi:10.1098/rspb.2010.2325

Bronfman, Z., Ginsburg, S., & Jablonka, E. (2016a). The evolutionary origins of consciousness: suggesting a transition marker. *Journal of Consciousness Studies 23*(9/10), 7–34.

Bronfman, Z. Ginsburg, S., & Jablonka, E. (2016b). The transition to minimal consciousness through the evolution of associative learning. *Frontiers in Psychology 7*, 1954. doi:10.3389/fpsyg.2016.01954

Bronfman, Z.Z., Ginsburg, S. and Jablonka, E. (2021). When will robots be sentient? *Journal of Artificial Intelligence and Consciousness, 8*(2),1–21. https://doi.org/10.1142/S2705078521500168

Buckner, R. L. & DiNicola, L. M. (2019). The brain’s default network: updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience, 20*, 593–608. https://doi.org/10.1038/s41583-019-0212-7

Clark, R. E. & Squire, L. R. (1998). Classical conditioning and brain systems: the role of awareness. *Science, 280*, 77-81. doi: 10.1126/science.280.5360.77

Clark, R. E., & Squire, L. R. (1999). Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies. *Psychological Science, 10,* 14-18. https://doi.org/10.1111/1467-9280.00099

Couvillon, P.A., & Bitterman, M. E. (1988). Compound-component and conditional discrimination of colors and odors by honeybees: further tests of a continuity model. *Animal Learning and Behavior 16*, 67–74. https://doi.org/10.3758/BF03209045

Crane, S. (1894). In the depth of a coal mine. *McClure's Magazine III*(3). https://ehistory.osu.edu/exhibitions/gildedage/content/CraneDepths

Damasio, A. (2021). *Feeling and knowing: Making minds conscious*. Pantheon.

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. 1st edition. London, John Murray.

Darwin, C. (1872). *The expression of emotions in man and animals*. John Murray.

Dehaene, S. (2014). *Consciousness and the brain: deciphering how the brain codes our thoughts.* Penguin Books.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron 70,* 201–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dennett, D. C. (1995). *Darwin's dangerous idea*. Simon and Schuster.

Dickinson, A. (2012a). Associative learning and animal cognition. *Philosophical Transactions of the Royal Society of London B*, 367, 2733–2742. doi: 10.1098/rstb.2012.0220

Dickinson, A. (2012b). Why a rat is not a beast machine. *Frontiers of Consciousness. Chichele Lectures*. doi: 10.1093/acprof:oso/9780199233151.003.0010

Dickinson, S., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning and Behavior 22*(1), 1–18. https://doi.org/10.3758/BF03199951

Dickinson, A., & Balleine, B.W. (2000). Causal cognition and goal-directed action. In C. Heyes and L. Huber (Eds.), *Vienna series in theoretical biology: The evolution of cognition*, pp. 185–204. MIT Press.

Dickinson, A. & Balleine, B. (2010). Hedonics – the cognitive-motivational interface. In M. L. Kringelbach, & A. C. Berridge (Eds.), *Pleasures of the brain*, pp. 74–84. Oxford University Press.

Dickinson, A., Watt, A., & Griffiths, W. J. H. (1992). Free-operant acquisition with delayed reinforcement, *The Quarterly Journal of Experimental Psychology Section B, 45*(3), 241–258. http://dx.doi.org/10.1080/14640749208401019

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Dor, D. (2015). *The instruction of imagination: Language as a social communication technology.* Oxford University Press.

Dor, D., (in press). Communication for collaborative computation: Two major transitions in human evolution. *Philosophical Transaction of the Royal Society B.*

Dor, D., & Jablonka, E. (2010). Canalization and plasticity in the evolution of linguistic communication. In R. K. Larson, V. Deprez & H. Yamakido (Eds.), *The evolution of human language,* pp. 135–147. Cambridge University Press,.

Feinberg,, T.E. & Mallatt, J. (2016). *The ancient origins of consciousness: How the brain created experience.* MIT Press.

Freeman, W. T. (2003). Neurodynamic models of brain in Psychiatry. *Neuropsychopharmacology 28*, S54–S63. https://doi.org/10.1038/sj.npp.1300147

Fresco, N., Ginsburg S., & Jablonka E. (2020). Functional information: a graded taxonomy of difference makers. *Review of Philosophy and Psychology*, 11, 547–567. https://doi.org/10.1007/s13164-018-0410-7

Gánti, T. (1987). *The principle of life*. (L. Vekerd, Trans.). Budapest, Omikk.

Gánti, T. (2003). *The principles of life, with a commentary by James Griesemer and Eörs Szathmáry*. Oxford University Press.

Gewirtz, J. C., & Davis, M. (2000). Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. *Learning and Mem 7*, 257–266. https://doi.org/10.1101/lm.35200

Ginsburg, S. & Jablonka, E. (2010). The evolution of associative learning: A factor in the Cambrian explosion. *Journal of Theoretical Biology 266*, 11–20. doi:10.1016/j.jtbi.2010.06.017

Ginsburg, S., & Jablonka, E. (2015). The teleological transitions in evolution: A Gántian view. *Journal of Theoretical Biology, 381*, 55–60. doi:10.1016/j.jtbi.2015.04.007

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.

Ginsburg, S., & Jablonka, E. (2020a). Consciousness as a mode of being. *Journal of Consciousness Studies 27*(9-10), 148–162.

Ginsburg S., & Jablonka, E (2020b). Sentience in plants: A green red herring? *Journal of Consciousness Studies, 28*(1-2), 17–34.

Ginsburg, S., & Jablonka, E. (2021). Evolutionary transitions in learning and cognition. *Philosophical Transactions of the Royal Society B*, 376, 20190766. https://doi.org/10.1098/rstb.2019.0766

Ginsburg, S., & Jablonka, E. (2022). *Picturing the mind: consciousness through the lens of evolution*. MIT Press.

Godfrey-Smith, P. (2020). *Metazoa: Animal life and the birth of the mind*. Collins

Hadar, R., & Menzel, R. (2010). Memory formation in reversal learning of the honeybee. *Frontiers in Behavioral Neuroscience 4*, 186. https://doi.org/10.3389/fnbeh.2010.00186

Hardy, A. (1965). *The living stream: A restatement of evolution theory and Its relation to the spirit of man*. Collins.

Holland, P. C., & Rescorla, R. A. (1975). Second-order conditioning with food unconditioned stimulus. *Journal of Comparative and Physiological Psychology 88*, 459–467. https://doi.org/10.1037/h0076219

Humphrey, N. (2011), *Soul dust: The magic of consciousness*. Princeton University Press.

Hussaini, S. A,. Komischke, B., Menzel, R., & Lachnit, H. (2007). Forward and backward second-order Pavlovian conditioning in honeybees. *Learning and Memory 14*, 678–683. https://doi.org/10.1101/lm.471307

Jablonka, E. (2002). Information: its interpretation, its inheritance and its sharing. *Philosophy of Science*, 69, 578–605. doi:10.1086/344621

Jablonka, E, (2021). Signs of consciousness? *Biosemiotics 14*(1), 25-29. doi: 10.1007/s12304-021-09419-x

Jablonka, E., & Ginsburg, S. (in press). From teleonomy to mentally-driven goal-directed behavior: Evolutionary considerations.

James, W. (1890). *The principles of psychology*. Dover.

Jonas, H. (1966). *The phenomenon of life: toward a philosophical biology.* Harper and Row.

Keijzer, F., Van Duijn, M., & Lyon, P. (2013). What nervous systems do: early evolution, input–output, and the skin brain thesis. *Adaptive Behavior 21*(2), 67–85. doi: 10.1177/1059712312465330

Kröger, B., Vinther, J., & Fuchs, D. (2011). Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules. *Bioessays, 33,* 602–613. https://doi.org/10.1002/bies.201100001

Kull, K. (2021). Natural Selection and Self-Organization Do Not Make Meaning, while the Agent's Choice Does. *Biosemiotics* **14,** 49–53. https://doi.org/10.1007/s12304-021-09422-2

Lamme, V. (2020). Visual functions generating conscious Seeing. *Frontiers in psychology, 11,* 83. https://doi.org/10.3389/fpsyg.2020.00083

Longo, G., Montévil, M., Sonnenschein, C. & Soto, A. M. (2015). In search of principles for a theory of organisms. *Journal of Biosciences, 40*(5), 955–968. doi: 10.1007/s12038-015-9574-9

Lucas, G. A., Deich, J. D., & Wasserman, E. A. (1981). Trace autoshaping: acquisition, maintenance, and path dependence at long trace intervals. *Journal of the Experimental Analysis of Behavior 36*, 61–74. doi: 10.1901/jeab.1981.36-61

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence 1*, 446–452. doi: 10.1038/s42256-019-0103-7

Mansur, B. E., Rodrigues, J. R. V., & Mota, T. (2018). Bimodal patterning discrimination in harnessed honey bees. *Frontiers in Psychology, 9*, 1529. https://doi.org/10.3389/fpsyg.2018.01529

Maran, T., Kleisner, K. (2010). Towards an Evolutionary Biosemiotics: Semiotic Selection and Semiotic Co-option. *Biosemiotics* **3,** 189–200. https://doi.org/10.1007/s12304-010-9087-8

Mashour, G. A., Roelfsema, P, Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798. doi: 10.1016/j.neuron.2020.01.026

Maynard Smith, J., & Szathmáry, E. (1995). *The major transitions in evolution*. Oxford University Press.

McGilchrist, I. (2010). *The master and his emissary: the divided brain and the making of the Western world*. Yale University Press.

McMillan, F. D. (2020). The mental health and well-being benefits of personal control in animals. In F. D. McMillan (Ed.), *Mental health and well-being in animals*, 2nd edition. CABI.

Merker, B. (2005). The liabilities of mobility: a selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition 14*, 89–114. doi:10.1016/S1053-8100(03)00002-3

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences 30*, 63–134. doi:10.1017/S0140525X07000891.

Metzinger, T. (2003). *Being no one: the self-model theory of subjectivity*. MIT Press.

Mizunami, M. (2021). What Is learned in Pavlovian conditioning in crickets? Revisiting the S-S and S-R learning theories. *Frontiers in behavioral neuroscience, 15,* 661225. https://doi.org/10.3389/fnbeh.2021.661225

Morsella, E. (2005). The function of phenomenal states: supramodular interaction theory. *Psychological Review 112*, 1000–1021. https://doi.org/10.1037/0033-295X.112.4.1000

Moyer, J. R., Deyo, R. A., & Disterhoft, J. F. (2015). Hippocampectomy disrupts trace eye blink conditioning in rabbits. *Behavioral Neuroscience 129*, 523–532. https://doi.org/10.1037/bne0000079

Nesse, R. M. (2001). The smoke detector principle. *Annals of the New York Academy of Science 935*, 75–85. doi:10.1111/j.1749-6632.2001.tb03472.x

Nicholson, D. (2013). Organisms ≠ machines. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44*, 669–678. doi: 10.1016/j.shpsc.2013.05.014

Noble, D. (2021) The illusions of the modern synthesis. *Biosemiotics 14,* 5–24. https://doi.org/10.1007/s12304-021-09405-3

Öhman, A. & Soares, J. J. (1993). On the automatic nature of phobic fear: conditioned electrodermal responses to masked fear-relevant stimuli. *Journal of Abnormal Psychology 102*, 121–132. doi:10.1037/0021-843X.102.1.121

Okasha, S. (2018). *Agents and goals in evolution*. Oxford University Press.

Pagni, E., & Simanke, R. T. (2021). *Biosemiotics and evolution: The natural foundations of meaning and symbolism*. Springer.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness & Cognition, 14*, 30–80

Panksepp, J. (2011). Cross-species affective neuroscience decoding of the primal affective experiences of humans and related animals. *PLoS ONE 6*(9), e21236. doi:10.1371/journal.pone.0021236

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual review of neuroscience, 35*, 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

Popper, K. (1972) *Objective knowledge: An evolutionary approach.* Clarendon Press.

Prum, R. O. (2017). *The evolution of beauty: How Darwin's forgotten theory of mate choice shapes the animal world and us*. Anchor Books.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). Inaugural article: A default mode of brain function. *Proceedings of the National Academy of Sciences. 98*(2), 676–682. doi: 10.1073/pnas.98.2.676

Reber, A. S. (2019). *The first minds: caterpillars, 'karyotes and consciousness*. Oxford University Press.

Riskin, J. (2016) *The restless clock: A history of the centuries-long argument over what makes living things tick*. University of Chicago Press.

Rodríguez-Expósito, B., Gómez, A., Martín-Monzón. I., Reiriz, M., Rodríguez, F., & Salas, C. (2017). Goldfish hippocampal pallium is essential to associate temporally discontiguous events. Neurobiology of Learning and Memory 139, 128–134. https://doi.org/10.1016/j.nlm.2017.01.002

Schon, K. R., Parker, A. P. J., & Woods, C. G (2018, updated 2020). Congenital insensitivity to pain overview. In M. P. Adam., G. M. Mirzaa, R. A. Pagon, et al., (Eds.). GeneReviews®. University of Washington, Seattle. https://www.ncbi.nlm.nih.gov/books/NBK481553/

Searle, J. (2004). *Mind: a brief introduction*. Oxford University Press.

Seth, A. (2021). *Being you: A new science of consciousness*. Faber & Faber

Sharov, A., & Tønnessen, M. (2021). *Semiotic agency: Science beyond mechanism*. Springer.

Skora, L. I, Yeomans M. R., Crombag H. S., & Scott R. B. (2021) Evidence that instrumental conditioning requires conscious awareness in humans. *Cognition*. Mar;208:104546. doi: 10.1016/j.cognition.2020.104546.

Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. Profile Books

Solms, M. & Friston, K. (2018). How and why consciousness arises: Some considerations from physics and physiology. *Journal of Consciousness Studies*, *25*(5-6), 202–238.

Telles, F. J., Corcobado, G., Trillo, A., & Rodríguez-Gironés, M. A. (2017). Multimodal cues provide redundant information for bumblebees when the stimulus is visually salient, but facilitate red target detection in a naturalistic background. *PLoS ONE 12*(9), e0184760. https://doi.org/10.1371/journal.pone.0184760

Thompson, E. (2007). Mind in life: Biology, phenomenology, and the science of mMind. Harvard University Press.

Tinbergen, N. (1951). *The study of instinct.* Clarendon Press.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift fürTierpsychologie 20*, 410–433.

Tomasello, M. (2014) *A Natural history of human thinking.* Harvard University Press.

Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience 17*, 450–461. https://doi.org/10.1038/nrn.2016.44

Vallar, G., & Ronchi, R. (2009). Somatoparaphrenia: a body delusion. A review of the neuropsychological literature. *Experimental Brain Research 192*, 533–551. https://doi.org/10.1007/s00221-008-1562-y

von Holst, E., & Mittelstaedt, H. (1950). The reaference principle. Interaction between the central nervous system and the periphery. In: Selected papers of Erich von Holst: the behavioural physiology of animals and man (R. Martin, Trans.), pp. 39–73. Methuen,

West-Eberhard, M. J. (2003). Developmental plasticity and evolution. Oxford University Press.

Williford, K., Bennequin, D., Friston, K., & Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Frontiers in psychology, 9*, 2571. https://doi.org/10.3389/fpsyg.2018.02571

Wyles, J. S., Kunkel, J. G., & Wilson, A. C. (1983). Birds, behavior, and anatomical evolution. *Proceedings of the National Academy of Sciences USA 8*0, 4394–4397. doi: 10.1073/pnas.80.14.4394

Zacks, O., Ginsburg, S., & Jablonka, E (2022). The futures of the past. *Journal of Consciousness Studies, 29* (3-4), 29–61. doi: 10.53765/20512201.29.3.029

Zacks, O., Ginsburg, S., & Jablonka, E (in preparation).