When Will Robots Be Sentient?

Article *in* Journal of Artificial Intelligence and Consciousness · August 2021

DOI: 10.1142/S2705078521500168

CITATIONS
4

READS
358

3 authors, including:

Simona Ginsburg
The Open University of Israel Raanana Israel
50 PUBLICATIONS **1,468** CITATIONS
SEE PROFILE

Jablonka Eva
Tel Aviv University
135 PUBLICATIONS **12,734** CITATIONS
SEE PROFILE

Journal of Artificial Intelligence and Consciousness
Vol. 8, No. 2 (2021) 1−21

DOI: 10.1142/S2705078521500168

When Will Robots Be Sentient?

Zohar Bronfman

pecan.ai
40 Toval Street, Ramat-Gan, Israel
zohar@pecan.ai

Simona Ginsburg

The Open University of Israel
1, University Road, Raanana 4353701, Israel
simona@openu.ac.il

Eva Jablonka*

The Cohn Institute for the History and Philosophy of Science and Ideas
Tel Aviv University, 6934525 Ramat Aviv, Israel
jablonka@tauex.tau.ac.il

Received 8 June 2021
Accepted 21 June 2021
Published 6 August 2021

The current failure to construct an artificial intelligence (AI) agent with the capacity for domain-general learning is a major stumbling block in the attempt to build conscious robots. Taking an evolutionary approach, we previously suggested that the emergence of consciousness was entailed by the evolution of an open-ended domain-general form of learning, which we call unlimited associative learning (UAL). Here, we outline the UAL theory and discuss the constraints and affordances that seem necessary for constructing an AI machine exhibiting UAL. We argue that a machine that is capable of domain-general learning requires the dynamics of a UAL architecture and that a UAL architecture requires, in turn, that the machine is highly sensitive to the environment and has an ultimate value (like self-persistence) that provides shared context to all its behaviors and learning outputs. The implementation of UAL in a machine may require that it is made of "soft" materials, which are sensitive to a large range of environmental conditions, and that it undergoes sequential morphological and behavioral co-development. We suggest that the implementation of these requirements in a human-made robot will lead to its ability to perform domain-general learning and will bring us closer to the construction of a sentient machine.

Keywords: AI; Domain-General Learning; Development; Evolution; Evolutionary Transition Marker (ETM); Minimal Consciousness; Soft Materials; Unlimited Associative Learning (UAL).

*Corresponding author.

1. Introduction

Can robots be experiencing subjects? In the words of philosophers, can robots be sentient, conscious beings with subjective experiences such as feelings of pleasure, discomfort, fear and joy, and perceptions like seeing a red poppy, smelling a ripe banana or hearing the song of a robin? How can we tell if they do? Roboticists assume that something like general intelligence, the ability to adaptively and flexibly learn and to pursue many different goals in many changing conditions, is a reasonable marker of consciousness since it emulates the behavior of conscious beings like ourselves [Shevlin, 2019]. The construction of such general intelligence is therefore central to the engineering project of constructing an artificial entity that can be said to be conscious.

The assumption that rich, domain-general learning ability and consciousness are related leads to the following questions:

- What is the theoretical rationale for suggesting that domain-general learning entails subjective experiencing or consciousness (we use these two terms interchangeably)? If it does, what is the nature of this entailment? Is consciousness causally and intrinsically constituted by domain-general cognition (like an organ that is causally constituted by its cells) or is it extrinsically caused by it (like fire caused by a spark of lightening)?
- How rich and sophisticated must such general cognition and learning be? Is complex, domain-general cognition sufficient, necessary, or necessary and sufficient for the manifestation of consciousness?
- What are the value systems that underlie the ability for domain-general learning?
- Is the physical implementation of consciousness important? Does consciousness entail a particular material organization (e.g., biological) or can a conscious being be made of silicon or be an algorithm? Is the developmental-learning history of the system a compulsory affordance for developing consciousness?

In this paper, we outline a theory of consciousness that addresses these questions. In Sec. 1, we point to some of problems in current AI research that roboticists and AI engineers recognize as central for the construction of conscious artificial beings, notably the current inability to build an AI with a domain-general learning capacity. An evolutionary theory of consciousness in living organisms − the only conscious beings of which we are currently aware − is described in Sec. 2. According to this theory, the evolution of consciousness in living organisms was driven by the evolution of a complex, domain-general type of associative learning, which we call unlimited associative learning (UAL), and the cognitive-affective architecture and dynamics of UAL *constitutes* consciousness. In Sec. 3, we discuss the value systems that are required by a system with domain-general intelligence, starting with the kind of values that enable the open-ended adaptive plasticity we see in living organisms that exhibit UAL. The material and developmental constraints and affordances in living

organisms are discussed in Sec. 4. We end the paper (Sec. 5) with a brief discussion of the implications of our view for the project of constructing conscious AI.

2. The Limitations of AI

Although in recent years AI and machine learning algorithms have achieved many impressive capabilities, such as game-playing [Silver *et al.*, 2017], policy learning [Mnih *et al.*, 2015], debating [Slonim *et al.*, 2021], and even art creation [Gatys *et al.*, 2016] − some major conceptual limitations still apply. The limitations may be divided into input-dependent (data) limitations and task-dependent (generalization and transfer) limitations.

Unlike living organisms, who can learn even when the stimuli conditions vary and are noisy between episodes and instances, existing AI algorithms (e.g., neural networks) are very sensitive to the state of the data across the training set. Missing values, noise, or a slight shift in the stimuli distribution can cause a dramatic, disproportional bias or distortion in the learning process. In ecological natural settings, where organisms learn, the stimuli conditions constantly change. In the visual domain, for example, the viewing angle, opacity, and brightness will vary over time, as well as the potential occlusion and fragmentation caused by other objects passing spontaneously through the visual field, yet the object will be seen maintaining its identity. Existing AI algorithms will fail to correctly identify the object if such "noisy" stimuli were not met during the learning history (i.e., existed in the training set). This "non-graceful" dependency of ML algorithms on pristine data implies that machines still fail to extract in a robust manner the "meaningful representation" from the entire input. Animals, on the other hand, categorize percept and actions − they can perceive solid objects as permanent entities that endure and do not suddenly evaporate or change their shape or color; perceive their body parts as relocating together during their movement; discriminate between world-generated sensory stimuli that are external to them and identical, but self-generated sensory stimuli that stem from their own actions; recognize that objects have parts that cannot all be perceived at the same time (for example, the front of a 3D object like a face implies the back of the head) and recognize that objects have different affordances − this object can be held by hand, this one can be drunk and spilled, and so on. These categorizations lead to the perception of invariances, such as color invariance in spite of changes in illumination [for discussions of invariances, see Nozick, 2002; Tretsman, 2018] and continuous visual perception in spite of spontaneous self-generated eye blinks (externally imposed interruptions of identical duration are perceived as gaps; [Golan *et al.*, 2016]).

The failure of AI to infer invariances is related to the large number of data samples that are currently needed to establish significant learning. The amount of data is (many) orders of magnitude larger than the ecological learning episodes or instances on which organisms base their learning, and as the tasks become more complex,

comprising many dimensions and signal subtleties, the amount of required data increases even further.

In addition to data limitation and poor generalization, transfer in AI is very limited. Transfer refers to cases where previously learned associations or representations are retrieved and used in a different context or setting. Ecologically, this capability allows organisms to take advantage of past developmental trajectories and harness previous experiences to shorten or improve new learning experiences. Many animals show various forms of transfer, such as the cross-modal transfer exhibited by bumble bees: bees trained to discriminate cubes and spheres using only touch (in darkness) or only vision (in light, but unable to touch the objects) could subsequently discriminate those objects using only the non-trained sensory information [Solvi *et al.*, 2020], a transfer feat that current AI devices cannot achieve. To attain some (minimal) level of transfer in similar tasks, the algorithm has to be trained again on the new task (data). Generally, when the tasks are moderately different (e.g., different modalities) or have different levels of representation − visual and semantic, for example − transfer learning in current methods of AI is almost nonexistent. This lack of ability to reuse, or harness past learned episodes in different conditions, implies that the AI algorithms fail to achieve a "meaningful representation" (which is a key for abstraction and transfer). Lastly, there are very few successes in combining or communicating the several different models to achieve a complex goal. While organisms usually harness multiple learned associations simultaneously, cross-talking and feeding one another, one AI algorithm does not benefit from the learning that takes place in another algorithm.

3. The UAL Theory of Minimal Consciousness

The limitations of current AI algorithms are overcome by very simple living organisms, such as worms and slugs that can learn through conditional associative learning (classical and operant conditioning). We use the term "conditional associative learning" to mean the formation of a *conditional* pairing between a non-reinforcing stimulus or action and a *subsequen*t reinforcing stimulus.[a] Conditional predictive associations can be formed between neutral stimuli (stimuli that under ordinary conditions do not trigger a response), between biologically important stimuli like those linked to the maintenance of basic homeostatic and reproductive functions and unrelated reflexive responses (e.g., one may become conditioned to reflexively blink

[a] Associative learning is defined differently by AI scientists for whom any change in the connection between elements as a result of their past activity counts as associative learning. Hebb's law − neurons that fire together wire together − is their maxim of associative learning. Sensitization and habituation which require a modulation in the strength of connections in a reflex path is, for them, an instance of associative learning because it involves a change in the synaptic strengths as a results of past activity. Hence, they make no distinction between non-associative and associative learning. As noted above, we use the term "associative learning" in the sense used by psychologists − for the formation of a *conditional pairing* between a non-reinforcing stimulus or action and a *subsequent* reinforcing stimulus or action, and we refer to learning by habituation and sensitization as "non-associative learning."

when smelling food), between stimuli and the contexts in which particular stimuli and actions occur (the cage in which one had an electric shock will elicit fear reaction), and between motor activities triggered or initiated by the animal and their negatively or positively reinforcing effects (one learns to press a lever that liberates food pellets). While domain-specific and controversial cases of such learning about the world and about one's own activities have been reported in plants and microorganisms, the ability for flexible associative learning enabling multiple pairing between stimuli, or between stimuli and actions, has been found only in animals with a central nervous system (and not in all of them; see Ginsburg and Jablonka [2019, chapter 7]).

It is important to note that conditional associative learning in neural animals is not restricted to one specific domain. For example, the hermaphrodite morph of the nematode worm *Caenorhabditis elegans*, which has a nervous system of only 302 neurons, learns to approach or avoid new tastes, odors or temperatures that predict the presence or absence of food, distinguishes between simultaneously presented rewarding and punishing cues, integrates cues from different senses, explores mazes, responds to the context in which they received rewards and punishment, and makes decisions [Rankin, 2004]. Such learning capacities are based not only on labile, intercellular synaptic memory, but also on more durable intracellular memory based on epigenetic marks such as patterns of histone modifications and regulatory small RNAs that interact with them [Bronfman *et al.*, 2014].

Although spontaneous and stochastic exploratory activities and preexisting simple reflex reactions can be flexibly combined, reinforced and recalled in worms and other animals with simple brains, and although these animals have a domain-general learning capacity that goes far beyond the capacities of current AI, their learning is nevertheless limited. Animals with small and simple brains like nematodes, planarians, and annelids as well as most mollusks, cannot discriminate between differently organized multimodal, compound, novel stimuli or complex action patterns, can learn only if there is a full or partial temporal overlap between the neutral stimulus or action and the reinforcing stimulus, and cannot make decisions requiring a motivational trade-off between learned and/or reflexive responses. Is their associative, domain-general, yet clearly restricted learning (we called it limited associative learning), sufficient for consciousness? Are nematode worms and sea slugs conscious? Why should we assume that domain-general learning is necessary for consciousness? Indeed, why assume that consciousness has anything whatsoever to do with learning and domain-general intelligence?

A link between exquisite sensory sensitivity to changes in the internal and external environment enabling flexible learning and consciousness was suggested by William James. He wrote:

> "The dilemma in regard to the nervous system seems, in short, to be of the following kind. We may construct one which will react infallibly and certainly, but it will then be capable of reacting to very few changes in the

> environment – it will fail to be adapted to all the rest. We may, on the other hand, construct a nervous system potentially adapted to respond to an infinite variety of minute features in the situation; but its fallibility will then be as great as its elaboration. We can never be sure that its equilibrium will be upset in the appropriate direction. In short, a high brain may do many things, and may do each of them at a very slight hint. But its hair-trigger organization makes of it a happy-go-lucky, hit-or-miss affair. It is as likely to do the crazy as the sane thing at any given moment. 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, I, pp. 139–140]

James, however, did not explain what is entailed in "loading" these dice, how one can figure out what level of context-dependent sensitivity is required, and how consciousness is distributed in the living world. In order to address these questions, we put forward an evolutionary theory of consciousness, which is focused on the relation between the evolution of learning and the emergence of consciousness (explored in detail in Ginsburg and Jablonka [2019]).

Our evolutionary approach to the study of consciousness is based on the methodology of the Hungarian chemist Tibor Gánti for the study of the origin of life [Gánti, 1987, 2003]. Although – just as in the case of consciousness – there is no generally accepted definition of life, there is a well-established origin-of-life research program based on a general consensus around a list of capacities that are deemed to be jointly sufficient for life. These life-capacities, according to Gánti are: maintenance of a boundary, metabolism, stability, information storage, regulation of the internal milieu, growth, reproduction, and irreversible disintegration (death). Gánti suggested that the functional and structural coupling among the mechanisms and processes that implement these capacities constitute a minimal living system, a proto-cell which he modeled and called the *chemoton* [Gánti, 2003]. Moreover, Gánti [1987, 2003] and Maynard Smith and Szathmáry [1995] proposed that one can find a positive marker of a minimal and sustainable form of life which is experimentally and theoretically tractable. They suggested that the marker for life is 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. 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 the basis of this capacity the simplest system with all the properties that characterize a living system – something like a proto-cell (a chemoton).

A capacity which requires that all the properties attributed to a particular mode of being such as life or consciousness are in place (so the evolutionary transition to

this mode of being has been completed) is an *evolutionary transition marker* (ETM) [Ginsburg and Jablonka, 2015; Bronfman *et al.*, 2016a,b; Ginsburg and Jablonka, 2019]. Following Gánti's methodology, we first sought a consensus list of capacities – a list of consciousness characteristics that most consciousness researchers would regard as jointly sufficient for the simplest conceivable system to be deemed subjectively experiencing. On the basis of many biological, psychological, cognitive, and philosophical studies, we extracted such a list (for more extensive discussion of each entry in the list and the studies supporting it, see Ginsburg and Jablonka [2019]). These partially overlapping capacities are as follows:

- Percept unification and differentiation: integration of sensory stimuli stemming from the body and the world as well as their relations.
- Global accessibility and broadcast: information from different cognitive systems – sensory, motor, memory, and evaluation systems – is integrated through back and forth interactions between and within different hierarchical levels of neural processing. The integration processes lead to the construction of cognitive representations enabling comparison, discrimination, generalization, and evaluation that inform decision-making.
- Temporal depth: the integration of percepts over time. A conscious animal must have a "working memory" – it must hold-on to incoming information long enough for it to become integrated and evaluated.
- Flexible value attribution and goal-directed behavior: perceptions and actions are evaluated as rewarding or punishing, and can be flexibly changed and differently prioritized when conditions change (e.g., what was previously punishing is now rewarding).
- Selective attention: target and action selection, which require the selective exclusion and amplification of signals emanating from the effects of changes in the body or the world according to evaluations based on present and past experience.
- Intentionality (aboutness): requiring mapping of signals from the world, body and their relations (embodiment and agency are assumed).
- Self-other distinction from a point of view: there is a stable perspective from which the system constructs models of the world and body and responds to changes in them. Such a system is able to distinguish between a stimulus that is the result of its own action and the same stimulus when it is not.

If we found an entity with these capacities on another planet, most scientists would take seriously the possibility that it is conscious or that it had been constructed by a conscious being.

The above characteristics of consciousness, like those identified by Gánti for life, are functionally and causally related, and construct a unified complex dynamic system. We identified a single capacity, an ETM, which enables the construction of a minimal system that displays all the listed characteristics. This ETM, we suggested,

is an open-ended mode of associative learning, which we called UAL. UAL refers to an organism's ability to

(i) Discriminate among differently organized, novel, multi-featured patterns of sensory stimuli and to select among new compound action patterns.
(ii) Learn about a predictive, compound neutral stimulus or action even when there is a time gap between the presentation of the compound stimulus or action and its reinforcement (such learning is called "trace conditioning").
(iii) Alter the valence attributed to patterns of sensory stimuli and motor actions when conditions change.
(iv) Use previously learned stimuli and actions as a basis for future learning.

UAL is a good evolutionary transition marker for consciousness because it requires that all the capacities in the consensus list are in place. *Unification* and differentiation are needed to construct compound stimuli and recognize the (potentially changeable) relations between their parts; *global accessibility* is needed for integrating information from multiple systems (memory, value, sensory, motor); *integration over time* is needed for trace conditioning; a *flexible evaluation system* is needed to make context-dependent learning possible; *selective attention* is needed to pick relevant stimuli out from the background; *intentionality* is needed, since the system maps (or represents) stimuli and their relations when storing associative links; *embodied agency* is needed for exploring and learning associations between actions; *a stable perspective or point of view* is needed to compare patterns of stimuli and actions and recognize world and body invariants, and self-world registration is needed so that the organism will not confuse between stimuli generated by its own actions and stimuli generated by the external world.

A highly simplified scheme of the architecture of UAL is given in the toy model depicted in Fig. 1.[b] According to this model, a central association unit is linked to more specialized, lower-level integrative processors that handle sensory information, motor information, memory and evaluation (reinforcement).[c] The double-headed arrows represent two-way re-entrant connections between the units, pointing to the importance of reciprocal top-down and bottom-up interactions. According to this description, consciousness is intrinsically *constituted* by UAL.

[b]The figure does not show the many excitatory and inhibitory interactions between integrating units at different levels of hierarchical organization, does not show the levels of processing within each unit, including local synaptic memory (between single neurons or sub-circuits), intracellular (nuclear) memory mechanisms and local mechanisms of amplification and exclusion (that constitute attention). It also does not show the internal organization of the (anatomically distributed) integrating processing units, which we suggest are structured as cognitive maps, where the relations between inputs are encoded relative to a stable perspective. A more detailed toy model of UAL than the one shown in Fig. 1 can be found in Bronfman *et al.* [2016b] and in Ginsburg and Jablonka [2019, figure 8.2]. However, these are not computational models. Computational models of UAL await construction.

[c]Our model shares some features with the neural global work space model of Dehaene and Changeux [2011], and with Merker's view of consciousness as a system in which motivational, sensory and action aspects are linked [Merker, 2007].

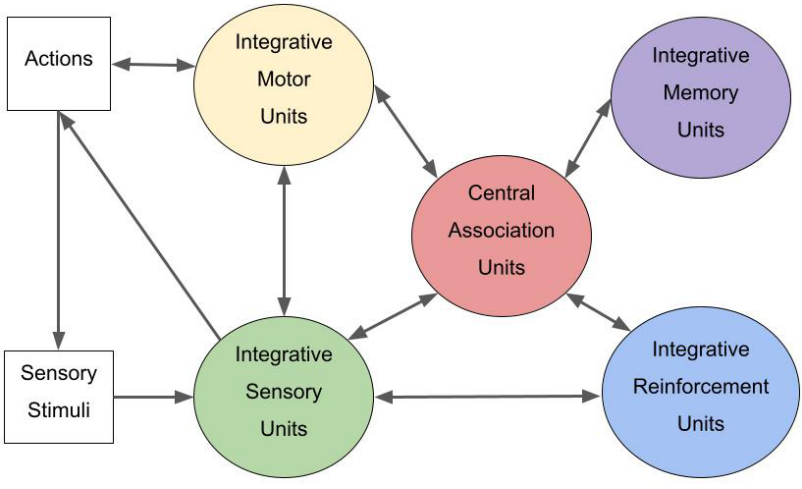

Fig. 1. A minimal toy model of UAL architecture. Unlimited Associative Learning is hypothesized to depend on reciprocal re-entrant connections between sensory, motor, reinforcement (value) and memory processing units, that together construct a central association unit (AU) at the core of the network. Intervening integrating units, direct lateral interactions of the sensory and motor units with the memory unit, as well as other components of the system that are discussed in the text and in Ginsburg and Jablonka [2019] are not shown.

UAL is a rich, domain-general type of learning. It is generative, since the number of associations within and between different sensory stimuli (auditory, visual, tactile, olfactory) that can be combined, learned and recalled during ontogeny (the individual's development) is vast, as is the number of action-patterns that can be linked to them. The space of learning is further expanded by the ability to cumulatively learn on the basis of what has already been learned in the past, and by the ability to learn even when there is a time gap between the neutral and reinforcing stimulus. UAL implies that the system has an ability to map relations between objects and actions from a stable point of view, to generalize, to flexibly transfer what it learned from one domain to another and to assign changing values to different models of the body and the world flexibly and rapidly. As we noted earlier, these transfer capabilities – whereby a previous learning instance or learning output is "reused" or applied to a novel situation in another domain or modality – have yet to be demonstrated in artificial agents.

Experimental evidence (at this point, mostly from human studies) supports our claim that UAL can be manifest only when organisms are conscious of the stimuli presented to them. UAL tasks, such as trace conditioning, complex decision-making and discrimination among compound novel patterns *cannot be learned* when the relevant stimuli are presented subliminally, while simpler learning tasks can be learned when stimuli are subliminal (see Birch *et al.* [2020] for examples; for additional corroborating studies, see Skora and Scott [20121]; Ben-Haim *et al.* [2021]).

Note that UAL, like all ETMs, is a positive marker of consciousness — it can tell us which living organisms are conscious, but it cannot tell us which are not. We believe that consciousness, like UAL, evolved gradually and it is impossible to pinpoint where exactly the transition to consciousness occurred, although it is possible to point to organisms that are very plausibly non-conscious because (like plants, for example) they lack all the UAL capacities.

Note that UAL is an evolutionary, not an ontogenetic marker of consciousness. An animal may have the evolved UAL architecture without the capacity to accomplish UAL tasks, which require a time-consuming learning process. A newborn baby is conscious because it has the neural architecture that the evolution of UAL built up — the architecture enabling the unification of stimuli into images, the integration and prioritization of evaluative signals, the mechanisms ensuring the durability of these physiological states and the memory systems that can actively store, through ongoing reconstructive consolidation processes [Dudai, 2012], complex representations. It can, however, manifest UAL only later in its development, when it has learned to perceive and act, and has accumulated and stored the integrated memories of its world and actions, a storing process that is built up slowly during ontogeny. Similarly, a person with Alzheimer disease whose memory is badly impaired is nevertheless conscious: although her cognition and some aspects of her consciousness are damaged — enough is spared of her sensory and motor integration processes for her to manifest minimal working memory and a minimal sense of self. What we argue is that the architecture of UAL was built during evolution through selection for UAL, and that once the UAL *architecture* is in place, the animal is minimally conscious.

Unlike consciousness, the capacity for UAL and the relationship between UAL and consciousness can be directly (though laboriously) tested in living organisms [Ginsburg and Jablonka, 2019; Birch *et al.*, 2020]. If the rationale underlying the UAL theory of consciousness is accepted, we can infer on the basis of surveys of learning that the organisms that exhibit UAL (and consciousness) belong to three animal phyla: vertebrates (most), arthropods (some) and mollusks (the coleoid cephalopods — squids, cuttlefish and octopuses). We can also infer that UAL (and consciousness) originated during the Cambrian era in vertebrates and arthropods — a proposition that is beginning to be widely shared — and 250 million years later in the coleoid cephalopods [Ginsburg and Jablonka, 2010, 2019; Feinberg and Mallatt, 2016; Godfrey Smith, 2020]. To the best of our knowledge (which is full of holes because the learning capacity of many animal groups has not been studied), the capacity for UAL has not been found in other taxa, including most animals, plants, fungi, protists and prokaryotes.

If UAL is an ETM of consciousness, basic felt drives such as hunger, thirst, and pain, and emotional feelings like fear, rage, joy and care are motivators of adaptive actions. However, many adaptive behaviors do not depend on motivating subjective feelings, just as they do not depend on perceptual subjective experiences. All living organisms have non-conscious, flexible adaptive capacities that enable them to

evaluate the salience of constantly changing external and internal conditions, construct appropriate action programs, and memorize and learn on the basis of past experiences. The building-blocks of evaluation systems in feeling animals, such as the various neuromodulators associated with pain, pleasure and excitement, are also employed during non-consciously driven actions, as are phylogenetically-conserved epigenetic, cell-memory systems, which are shared by *all* living organisms [Ginsburg and Jablonka, 2009]. These ancient epigenetic memory mechanisms interact with synaptic memory mechanisms during learning in neural animals that belong to different phyla [for reviews and discussions see Bronfman *et al.*, 2014, 2016a,b; Gold and Glanzman, 2021]. Hence, all neural organisms also have, in addition to synaptic inter-cellular memory systems and the persistence imposed by bioelectric fields, intracellular memory systems that are based on epigenetic cell-memory mechanisms, which manifest different temporal persistence and enable greater learning flexibility [Feldesh, 2019].

Since many of the building blocks that drive non-conscious, plastic behaviors are crucial for feeling-driven behaviors and hence for the affective aspect of sentience, it is necessary to understand the nature of the relationship between the ubiquitous non-conscious and the UAL-conscious evaluation systems.

4. Value Systems for Conscious Domain-General Learning are Entailed by Homeostatic Survival Networks

Value systems are factors and processes that positively reinforce reactions that lead to a system's goal-promoting states and negatively reinforce reactions that lead to a system's goal-compromising states. "Goal-promoting" and "goal-compromising" are terms that are relative to the objectives of the system. In living organisms, the value or valence of fitness-promoting reactions is referred to as positive and the valence or value of fitness-reducing reactions is referred to as negative. In AI systems, valenced goals are extrinsically set by the designer and in living organisms goals are intrinsic to the organism and are set by natural selection to promote survival and reproduction [Thompson, 2007; Damasio, 2018; Solms, 2021]. Survival and reproduction are the *ultimate intrinsic goals* of living organisms, which AI devices lack. These ultimate goals require the concerted and regulated operation of *proximate value systems.* These are the systems that are involved in the evaluation of distinct physiological needs such as the need for nutrition, bodily integrity, safety, and sex (that lead to drives), as well as the systems that evaluate action-patterns such as exploration-seeking, fear, rage, joy, care, and disgust (that we call emotions). We suggested [Ginsburg and Jablonka, 2019] that intense selection and differential stabilization processes at several different levels of neural organization, lead to the dynamics of UAL, which construct the felt needs of conscious animals.

The internal state of living organisms, which has to be maintained within a narrowly defined physiological range if the organism is to preserve itself during ontogenetic time (survive) and maintain itself during phylogeny (reproduce), is called a

state of *homeostasis*. The maintenance of this state requires exquisitely sensitive regulatory mechanisms that respond to the on-going deviations from the homeostatic range. It also requires that there is a distinction between the world and the bodily-self, so that the effects of stimuli generated by one's own actions and the effects of externally imposed stimuli are discerned [Cullen *et al.*, 2009; Jékely *et al.*, 2021]. The body operates within the boundaries of homeostasis and acts on the world and in the world to persist and thrive through behavioral and physiological-neural reactions to challenges and opportunities. Following LeDoux (e.g., LeDoux and Pine, 2016), we call the neural circuits that are involved in the generation of both conscious and non-conscious action-patterns and drives *survival circuits*.

The maintenance of homeostasis, which in neural organisms involves control by neural survival circuits, shares design principles as well as biochemical factors and processes with non-neural organisms. These include the following:

- Reflex-like response mechanisms that link sensory inputs at specific receptors and specific effector-guided actions. For example, in the single-celled *Paramecium*, collision with a solid object leads to backward swimming for a short distance because the beating direction of the cilia is temporarily reversed [Naitoh and Sugino, 1984]. In neural animals, there are classical reflex responses such as the blink reflex, or the salivation reflex. These specific reflexes are evolved adaptive responses to recurrent environmental challenges and opportunities. The modulation of reflex reactions involves changes in the threshold of reflexive responses: either habituation (experience-based increase of the threshold of reflexive response), or sensitization (the experience-dependent decrease of the threshold of response).
- Exploration-stabilization mechanisms, which are inherent to all living organisms. They involve stochastic and semi-stochastic variations in biochemical, electrochemical or motor reactions, with some of the variations becoming stabilized by reinforcing signals. For example, exploration occurs within the biochemical networks in cells, through the stochastic locomotor movements of bacteria, and during the growth of plant roots (for more examples and discussion, see West-Eberhard [2003]). In neural animals, there are also exploration mechanisms within the nervous system as well as motor behavioral exploration (discussed in Ginsburg and Jablonka [2019]). The intensity and frequency of exploratory behavior is enhanced or reduced in response to changes in the conditions of life, and past reinforcement can affect the intensity and frequency of present exploration.
- Stress-response mechanisms, which are strategies of responding that compensate for categories of potentially damaging conditions such as DNA damage, parasites, tissue injury, and drastic abiotic stresses (e.g., heat stress). Stress response mechanisms are ubiquitous and are found in all living organisms at all levels of biological organization (discussed in Lyon and Kuchling [2021]). In neural animals, the stress response includes the neuro-hormonal-immune system (reviewed in Ginsburg and Jablonka [2019, chapter 9]), and this type of stress response plays an

important part in aversive learning. The neural stress response is highly conserved: cnidarians, which display only non-associative learning and have a diffuse nervous system, share many of the factors and processes that are the building blocks of the stress response in UAL animals like mammals [Parisi *et al.*, 2020].

All these survival-promoting mechanisms depend on networks of intracellular and intercellular interactions and share highly conserved components in both non-neural and neural organisms. All are crucial for memory and learning, and all shape the overall sensitivity and responsiveness of the organism, so each and every response to changing conditions is evaluated within the context of the general state of the organism. For example, an animal that received continuous positive reinforcement shows "optimistic behavior", and is more motivated to explore and persist in its attempts to face challenges and solve problems, whereas one that has failed several times shows "pessimistic behavior" [Bateson, 2016]. Moreover, not only the overall state ("optimistic" or "pessimistic") of the animal but also its material and energy resources affect motivation and action-choice. An animal that received food and has extra material and energy reserves can employ these physical resources to solve many different tasks, whereas an animal that lacks such resources cannot.

In addition to these systemic effects, when different homeostatic and action programs driven by different neural circuits are in conflict, animals must weigh up the specific challenges and action-options open to them. Not only do they need to sense the ever-changing signals relevant to their existential needs, they also have to make decisions as to how and when to act, and which action to prioritize because the concurrent challenges and opportunities that they face have different effects on the ultimate goal of survival and cannot all be acted upon at the same time.[d] The proximate value systems must therefore be amenable to intrinsic, context-dependent prioritization and stabilization, enabling selection among action patterns that satisfy different needs in different conditions (e.g., in condition X satisfying need **a** is more important than satisfying need **b**; in condition Y the opposite is the case). For such contextual activity-selection to occur, the different proximate survival systems must share a common biochemical and physical "language" that affects the reaction thresholds of different circuits and makes a differential response to the shared factors possible. In nervous systems, these shared factors are the action-potentials into which all sensory stimuli are transformed; the memory systems based both on synaptic network wiring and on nuclear epigenetic memory [Bronfman *et al.*, 2014]; and the neurotransmitters (involved in synaptic signaling) and neuromodulators (involved in both synaptic and non-synaptic signaling) that, as their name suggests, modulate the activity thresholds of different survival circuits. Neuromodulators, which include

[d] Although the valences are context-dependent, there seems to be a general bias in prioritizing responses to threatening or damaging situations – responses to survival-endangering or compromising states with *negative* valence tend to inhibit concurrent responses, like approaching food or sex, which have positive valence [Baumeister *et al.*, 2001]. This makes evolutionary sense since survival is the most basic goal of living organisms.

dopamine, serotonin, acetylcholine, histamine, noradrenaline, various hormones and the endogenous opioid neuropeptides, have systemic effects, especially in the part of the nervous system that Carvalho and Damasio [2021] classify as interoceptive nervous system (the INS), which maps the homeostatic state of the body. These parts of the nervous system are poorly myelinated, have more blood-brain barrier gaps, and are particularly susceptible to effects of systemic neuromodulators [Carvalho and Damasio, 2021].

Neuromodulators affect the threshold of neural synaptic responses that determine which circuit-activity is prioritized. Circuit prioritization or "dominance" means not only that the corresponding action program is expressed, but also that all other circuits are partially or completely inhibited, so, for example, in an animal that freezes with fear, pain is suppressed (e.g., Roelofs [2017]). This can be achieved if the different survival circuits are mapped onto an anatomically shared space where they interact.

In UAL animals, the innate, non-conscious systems of response to challenges and opportunities are the scaffolds on which learning-based felt evaluations, informing the animal about its well-being, occur. Ginsburg and Jablonka [2010, 2019] argued that the ability to learn to adaptively respond to compound inputs from the world and from one's own actions, to discriminate between self-induced and world-induced stimuli, and to prioritize the learned responses in a context-sensitive, flexible manner, requires the cognitive-affective architecture of UAL. They called the dynamic neural representations that are formed during conscious processing of world and body stimuli "categorizing sensory states" (CSSs) because these neural states categorize and evaluate, through their dynamics, both inputs ("this is a fear-related signal") and outputs ("flight to reach safety is needed"). The incoming world and body inputs activate memory traces of associated inputs (for example, inputs and memory traces related to fleeing a predator), and these inputs and traces also determine what type of response will occur and how it will be evaluated and prioritized in the current conditions. Once this architecture has evolved, even the scaffolding innate reflex reactions, which can occur unconsciously, are processed by the UAL units and generate felt emotions in intact animals.

The evaluative, felt aspect of emotions, which is essential for learning-based decision-making, depends on the reinforcement integrating unit (REIU) of the UAL architecture (see Fig. 1), where concurrent neural representations are evaluated, compared and prioritized. REIU receives inputs from exteroceptive sensors that lead to action programs such as fight, flight or care; proprioceptive sensors that respond to self-movement and body position; and interoceptive sensors that reflect visceral states like hunger, thirst or pain and lead to actions that satisfy them. In vertebrates, REIU includes the reticular activating system (RAS) in the brainstem, where major neuromodulators are produced. The RAS receives inputs from the peripheral nervous system and transmits them to the pallium (the layers of grey and white matter that cover the upper surface of the cerebrum; the cerebral cortex in mammals) for further

processing and contextual evaluation. The processed pallial outputs are then sent to the periadequetual grey (PAG) in the midbrain, leading to value-based decision-making (see Solms [2021] for a discussion of the central role of these structures for feeling and decision making). Only after these integrating units evolved in the vertebrate lineage (and after analogous structures evolved in arthropods and cephalopods), were the effects of value system activations represented as felt body states; for example, the survival circuit for fear elicited the feeling of fear and not just the fear-behavior of flight or freezing.

In animals, UAL-based processing is entailed by and results from homeostasis-promoting processes that occur at all levels of biological organization (cell, tissue, organ, and organism). Homeostasis, with its myriad of system-general and system-specific mechanisms, is the final arbiter with regard to the evaluation of the organism's internal states. Since it constitutes the unifying, ultimate goal of the organism's adaptive behaviors, models of homeostatically regulated reinforcement learning have been constructed. These models are based on the drive-reduction theory of Hull [1943], who suggested that needs such as hunger and pain avoidance lead to actions that evolved to monitor them and drive action-programs that satisfy them, so the satisfaction of these needs can be modeled as drive-reduction (that leads to a return to homeostasis). Homeostatically regulated reinforcement learning models based on this idea better account for learning patterns in humans and other animals than traditional models, and create a framework for building AI devices that can achieve multiple goals in a multidimensional homeostatic space (for models and discussions, see Keramati and Gutkin [2014]; Juechems and Summerfield [2019]; Man and Damasio [2019]).[e] These models do not, however, incorporate the neural dynamics and integrating structures required for UAL, which are the basis of felt self-interest. If we are to build an AI device with a domain-general learning ability like UAL, we will need to include the overall material and energetic resources available to the system, the AI analogues of an integrating reinforcing unit (REIU), the shared systemic neuromodulators that affect the overall state of the system, the AI analogues of REIU structures and processes from which information diverges and converges through reentrant connections between neural maps, and the partially decoupled tiers of synaptic and epigenetic neuronal memory at all levels.

The richness of the adaptive, plastic behaviors of living organisms endowed with domain-general, mentally-motivated learning-capacity, requires the sharing of material and energetic resources that are coordinated by regulatory mechanisms at several levels of biological organization. The question we now address is whether

[e]One of the virtues of some of these models is the acceptance that the exteroceptive and interoceptive systems must interact in learning novel strategies that support physiological stability. Although learning to act in a particular way to satisfy hunger can occur even when the satisfaction of hunger is detached from the taste of food, this learning takes a far longer time, requiring more learning trials [Miller and Kessen, 1952]. Exteroceptive and localized bodily sensory stimulations like nociception in a particular location interact with the value systems to guide and facilitate learning.

domain-general learning, which entails this multisystem rich regulatory system, can be realized in non-biological materials.

5. Material and Developmental Affordances and Constraints

Living organisms are made up of soft materials. The physical and chemical properties of biological materials enable them not only to be sensitive to various changes in their environment (e.g., temperature, pressure) and to flexibly react to these changes, but also to store the effects of past reactions as changes in their molecular structure, and to engage in feedback interactions. The properties of sensitivity, reactive flexibility, storage of past effects and feedback interactions can be actualized in a single molecule, in a molecular complex with a particular spatial organization, or in organized structures made up of complexes. Materials with such properties are called "intelligent" [Kaspar *et al.*, 2021], and can reduce the computational load required for adaptive behavior at high levels of organization. However, their adaptive responsiveness renders them vulnerable to changes in their environment.

Taking the precarious existence of living organisms as their point of departure, Man and Damasio [2019] argued that the vulnerability of living systems provides affordances that are necessary for sentience: "Living organisms capable of mentation are fragile vessels of pain, pleasure and points in between. It is by virtue of that fragility that they gain access to the realm of feeling." [Man and Damasio, 2019, p. 446]

The fragility of living organisms that Man and Damasio highlight is entailed by more than the soft, pliable organic materials of which they are built. The integrity and functionality of an entity made of such malleable, potentially "intelligent" materials require that the system as a whole maintains a narrow range of internal conditions in an external environment that is constantly fluctuating and potentially jeopardizing its organizational integrity. In addition to the hazards of the external environment, the internal milieu, too, is inherently active and constantly changing. In "wet" living organisms there is ongoing activity, stochastic and semi-stochastic, of molecules and ions that may be hazardous to the integrity of internal structures and requires constant monitoring and repair. This fundamental vulnerability means that *all* processing within the body must conform to the same strict homeostatic demands imposed by the "wet" internal milieu. These shared constraints render functional interactions between processors mandatory and enable the internal milieu to be controlled by the nervous system, which is an integral part of it and shares the same ultimate (survival) "interests", the same kind of materials and the same constraints.

The fluctuations in the external and internal milieu have to be sensed for maintenance to be possible, so vulnerability requires sensitivity, and sensitivity requires flexible material responsiveness that promotes self-preservation. Since self-preservation depends on the evolved homeostatic mechanisms that cope with the external and internal hazards, Man and Damasio [2019] suggested that the construction of feeling machines requires that the machine is built of soft materials that

make it vulnerable to changes in its external and internal environment, and that it has intrinsic self-interests and mechanisms safeguarding these interests. They leave open the question whether soft but non-biological material (such as gels and plastics) can construct a sentient machine, but they argue that a machine made from such soft materials will have enhanced sensitivity and adaptive behavior, and that this is one of the ways of getting closer to the construction of a sentient AI device.

The sensitivity and fallibility of a complex nervous system was central, as we saw in Sec. 2, to James' view of consciousness. As James argued (and we agree), the larger and more complex the nervous system grew, the more interconnected and malleable it became, so more associations and predictions could be made, yet these very advantages make it more prone to errors, to more "hits and misses". As we argued in the previous sections, the dynamics of the interacting units that allow UAL is the "loading of the dice" that James recognized as necessary for the operation of a complex nervous system. The growing complexity of the nervous system during evolution required the addition of new, homeostasis-serving, learning mechanisms – a scaling-up of homeostasis that involved novel, additional levels of integration and top-down control, and possibly also novel ways of employing epigenetic cell memory mechanisms [Ginsburg and Jablonka, 2019, chapter 8].

Another consideration for AI inspired by biological organisms is the importance of physiological and morphological development. Since learning, including machine learning, is an aspect of development, AI devices can be said to develop, but their development does not include the morphological and physiological maturation processes that occur in living organisms. In neural organisms, ontogeny is controlled by the nervous system [Cabej, 2013], and the maturation of the nervous system is influenced by the development of the non-neural parts of the organism. As neural animals develop and their shape, strength and size changes, what is learned and in what sequence things are learned depend on morphological-physiological maturational stages. Starting with an immature body and nervous system that can respond only to the most stable and most recurrent inputs ensures that the general, basic, perceptual, affective and action scaffolds underlying the invariances of world, body and action are laid down first. Further learning is built on these foundations, gradually following the development of physiological, morphological, and in social animals, social affordances. Although not easy to delineate, stages in morphological and cognitive development are co-dependent.

6. Some Conclusions

On the basis of our theory of consciousness, we have argued that the neural architecture of domain-general (unlimited) associative learning *constitutes* consciousness in developing living organisms that are made up of biological materials and that are guided by both ultimate values (surviving and thriving) and the mediating proximate values (underlying drives and emotions). What are the implications of this view of consciousness in living organisms for the construction of conscious AI devices?

We believe that an AI device that is capable of domain-general learning requires the dynamics of a UAL architecture. Yet more than a formal similarity in architectural organization is necessary. Just as a genetic algorithm that implements unlimited heredity is not alive (although it has been built by living organisms), a UAL algorithm is not conscious. Implementation, in terms of mechanisms, developmental history and materials, is crucial. In a living organism the UAL architecture guides goal-directed actions that are implemented in materials that are highly sensitive to their environment. This sensitivity is crucial since domain-general learning must enable the animal to respond flexibly to external and internal contingencies by evaluating concurrent inputs and potential outputs from an organism-centered perspective. We agree with Damasio and Man [2019] that soft materials, which are sensitive to a large range of environmental conditions, are required for constructing an artificial device that manifests wide-ranging adaptive plasticity − a device that has the capacity to respond to a large range of inputs and manifests a large range of adaptive outputs. Such sensitivity and flexibility are necessary to accomplish UAL, although this very sensitivity makes a device prone to error and therefore requires regulation of the internal milieu as well as powerful learning and memory updating mechanisms.

A shared internal, homeostatically-regulated milieu depends on an ultimate value system that prioritizes the inputs and the activities of the entity and enables its continued functional persistence. Self-preservation in the face of external and internal changes can be seen as the "goal" of that entity to which all parts and activities ultimately contribute. Drawing on the ideas of Maturana and Varela [1980], Thompson [2007] discussed the idea that an autopoietic, self-preserving dynamic organization is necessary both for life at the cell and organism level and at the level of neural cognition. An autopoietic, self-preserving system maintains its own boundaries and entails a distinction between the system and the world, a distinction that in conscious animals is manifest as a sense of self. The autopoietic perspective suggests that for an AI entity to be conscious, cognition has to be preserved and individuated through the formation of self-generated boundaries that allow a distinction between the self and the world from a stable point of view. Self-preservation will be easier if common energy, material and informational resources are shared.

In living organisms, homeostasis at the level of the parts of the organism (cells, tissues, organs) as well as the organism as a whole, is necessary. Homeostasis at both the level of the parts and the whole involves the shared structural and regulatory components of the cells of which the organism is made. The regulatory components mediate between levels of biological organization that are all made up of "wet" organic materials and participate in feedforward (bottom-up), lateral, and feedback (top-down) interactions. It is not clear whether the parts in a complex AI will be all made of the same materials. If they are not, the question of the communication between different (e.g., soft and hard) materials and the material exchanges between them arise. The value systems of parts that are made up of different materials are also

likely to be different and the relation between the "interests" and values of the parts and the ultimate value of the device as a whole, is not clear and would need to be articulated. Another open question is the co-dependence between morphological development and behavioral development, which is necessary in living conscious organisms, but is, as yet, not implemented in AI devices. Is such co-dependence also necessary for developing domain-general learning and consciousness in AI?

Let us assume that we have built a developing, soft robot with a domain-general learning capacity (UAL), an ultimate value system sub-served by mediating value systems, and a stable point of view from which perceptions of the world and the execution of actions are evaluated as value-satisfying. Although we cannot be sure that such a robot will be conscious, because the biology of consciousness is a very young discipline and we may be unaware of processes and factors that are crucial for it (and that may be also crucial for Al consciousness), the cautionary principle demands that we treat such a robot as a sentient being − one that perceives, evaluates and acts from its own private perspective to promote its own private welfare.

Acknowledgment

We are very grateful to Marion J. Lamb for her detailed and constructive comments on earlier versions of the paper.

References

Bateson, M. [2016] Optimistic and pessimistic biases: A primer for behavioural ecologists, *Curr. Opin. Behav. Sci.* **12**, 115−121, doi: org/10.1016/j.cobeha2016.09.013.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. and de Vohs, K. [2001] Bad is stronger than good, *Rev. Gen. Psychol.* **5**, 323−370, doi: 10.1037/1089-2680.5.4.323.

Ben-Haim, M. S., Dal Monte, O., Fagan, N. A., Dunham, Y., Hassin, R. R., Chang, S. W. C. and Santos, L. R. [2021] Disentangling perceptual awareness from nonconscious processing in rhesus monkeys (Macaca mulatta), *Proc. Natl. Acad. Sci. USA* **118**(15), e2017543118, doi: 10.1073/pnas.2017543118.

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

Bronfman, Z., Ginsburg, S. and Jablonka, E. [2014] Shaping the learning curve: epigenetic dynamics in neural plasticity, *Front. Integ. Neurosci.* **8**, 55, doi: 10.3389/fnint.2014.00055.

Bronfman, Z., Ginsburg, S. and Jablonka, E. [2016a] The evolutionary origins of consciousness: Suggesting a transition marker. *J. Conscious. Stud.* **23**(9/10), 7−34.

Bronfman, Z., Ginsburg, S. and Jablonka, E. [2016b] The transition to minimal consciousness through the evolution of associative learning, *Front. Psychol.* **7**, 1954, doi: 10.3389/fpsyg.2016.01954.

Cabej, R. N. [2013] *Building the Most Complex Structure on Earth: An Epigenetic Narrative of Development and Evolution of Animals* (Elsevier, Amsterdam).

Cullen, K. E., Brooks, J. X. and Sadeghi, S. G. [2009] How actions alter sensory processing: Reafference in the vestibular system, *Ann. N.Y. Acad. Sci.* **1164**, 29−36, doi: 10.1111/j.1749-6632.2009.03866.x.

Carvalho, G. B. and Damasio, A. [2021] Interoception and the origin of feelings: A new synthesis. *BioEssays* **43**, e2000261, doi: 10.1002/bies.202000261.

Damasio, A. [2018] *The Strange Order of Things: Life, Feeling, and the Making of Cultures* (Pantheon, New York).

Dehaene, S. and Changeux, J.-P. [2011] Experimental and theoretical approaches to conscious processing, *Neuron* **70**, 201−227, doi: 10.1016/j.neuron.2011.03.018.

Dudai, Y. [2012] The restless engram: Consolidations never end, *Annu. Rev. Neurosci.* **35**, 227−247, doi.org/10.1146/annurev-neuro-062111-150500.

Feinberg, T. E. and Mallatt, J. [2016] *The Ancient Origins of Consciousness: How the Brain Created Experience* (MIT Press, Cambridge, MA).

Feldesh, R. [2019] The Distributed Engram. *bioRxiv preprint*, doi: https://doi.org/10.1101/583195.

Gánti, T. [1987] *The Principle of Life* (L. Vekerd, Trans; Omikk, Budapest, Hungary).

Gánti, T. [2003]. *The Principles of Life, with a Commentary by James Griesemer and Eörs Szathmáry* (Oxford University Press, New York).

Gatys, L. A., Ecker, A. S. and Bethge, M. [2016] "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2414−2423.

Ginsburg, S. and Jablonka, E. [2009] Epigenetic learning in non-neural organisms, *J. Biosci.* **34**, 633-646, doi: 10.1007/s12038-009-0081-8.

Ginsburg, S. and Jablonka, E. [2010] The evolution of associative learning: A factor in the Cambrian explosion, *J. Theor. Biol.* **266**, 11−20, doi: 10.1016/j.jtbi.2010.06.017.

Ginsburg, S. and Jablonka, E. [2015] The teleological transitions in evolution: A Gántian view. *J. Theor. Biol.* **381**, 55−60, doi: 10.1016/j.jtbi.2015.04.007.

Ginsburg, S. and Jablonka, E. [2019] *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness* (MIT Press, Cambridge, MA).

Godfrey Smith, P. [2020] *Metazoa: Animal Life and the Birth of the Mind* (Farrar, Straus and Giroux, New York).

Golan, T., Davidesco, I., Meshulam, M. *et al.* [2016] Human intracranial recordings link suppressed transients rather than 'filling-in' to perceptual continuity across blinks. *eLife* **5**, e17243, doi: 10.7554/eLife.17243.

Gold, A. R. and Glanzman, D. L. [2021] The central importance of nuclear mechanisms in the storage of memory. *Biochem. Biophys. Res. Commun.* xxx, doi: 10.1016/j.bbrc.2021.04.125.

Hull, C. L. [1943] *Principles of Behavior: An Introduction to Behavior Theory.* (Appleton-Century-Crofts, New York).

James, W. [1890] *The Principles of Psychology* (New York: Dover).

Jékely, G., Godfrey-Smith, P. and Keijzer, F. [2021] Reafference and the origin of the self in early nervous system evolution, *Phil. Trans. R. Soc. B* **376** 20190764, doi: 10.1098/rstb.2019.0764.

Juechems, K. and Summerfield, C. [2019] Where does value come from? *Trends Cogn. Sci.* **23**, 836−850, doi: 10.1016/j.tics.2019.07.012.

Kaspar, C., Ravoo, B. J., van der Wiel, W. G. *et al.* [2021] The rise of intelligent matter, *Nature* **594**, 345−355, doi: 10.1038/s41586-021-03453-y.

Keramati, M. and Gutkin, B. [2014] Homeostatic reinforcement learning for integrating reward collection and physiological stability, *eLife* **3**, e04811, doi: 10.7554/eLife.04811.

LeDoux, J. E. and Pine, D. S. [2016] Using neuroscience to help understand fear and anxiety: A two-system framework, *Am. J. Psychiatry* **173**, 1083−1093, doi: 10.1176/appi.ajp.2016.16030353.

Lyon, P. and Kuchling, F. [2021] Valuing what happens: A biogenic approach to valence and (potentially) affect, *Phil. Trans. R. Soc. B* **376**, 20190752, doi: 10.1098/rstb.2019.0752.

Man, K. and Damasio, A. [2019] Homeostasis and soft robotics in the design of feeling machines, *Nat. Mach. Intell.* **1**, 446−452, doi: 10.1038/s42256-019-0103-7.

Maturana, H. and Varela, F. [1980]. *Autopoiesis and Cognition: The Realization of the Living.* (D. Reidel, Boston).

Maynard Smith, J. and Szathmáry, E. [1995] *The Major Transitions in Evolution* (Oxford University Press, Oxford).

Merker, B. [2007] Consciousness without a cerebral cortex: A challenge for neuroscience and medicine, *Behav. Brain. Sci.* **30**, 63−134, doi: 10.1017/S0140525X07000891.

Miller, N. E. and Kessen, M. L. [1952] Reward effects of food via stomach fistula compared with those of food via mouth, *J. Comp. Physiol. Psychol.* **45**, 555−564, doi: 10.1037/h0060113.

Mnih, V., Kavukcuoglu, K., Silver, D. *et al.* [2015] Human-level control through deep reinforcement learning, *Nature* **518**, 529−533, doi: 10.1038/nature14236.

Naitoh, Y. and Sugino, K. [1984] Ciliary movement and its control in *Paramecium, J. Protozool.* **31**, 31−40, doi: 10.1111/j.1550-7408.1984.tb04285.x.

Nozick, R. [2002] *Invariances: The Structure of the Objective World* (Harvard University Press, Cambridge, MA).

Parisi, M. G., Parrinello, D., Stabili, L. and Cammarata, M. [2020] Cnidarian immunity and the repertoire of defense mechanisms in anthozoans, *Biology* **9**, 283, doi: 10.3390/biology9090283.

Rankin, C. H. [2004] Invertebrate learning: what can't a worm learn? *Curr. Biol.* **14**, R617-R618, doi: 10.1016/j.cub.2004.07.044.

Roelofs, K. [2017] Freeze for action: Neurobiological mechanisms in animal and human freezing, *Phil. Trans. R. Soc. B* **372**, 20160206, doi: 10.1098/rstb.2016.0206.

Shevlin, H. [2019] To build conscious machines, focus on general intelligence: A framework for the assessment of consciousness in biological and artificial systems, *Towards Conscious AI Systems Symposium Co-Located with the Association for the Advancement of Artificial Intelligence 2019 Spring Symposium Series (AAAI SSS-19)*, http://lcfi.ac.uk/resources/build-conscious-machines-focus-general-intelligenc/.

Silver, D., Schrittwieser, J., Simonyan, K. *et al.* [2017] Mastering the game of go without human knowledge, *Nature* **550**, 354−359.

Skora, L. I., Yeomans, M. R., Crombag, H. C. and Scott, R. B. [2021] Evidence that instrumental conditioning requires conscious awareness in humans, *Cognition* **208**, 104546, doi: 10.1016/j.2020.104546.

Slonim, N., Bilu, Y., Alzate, C. *et al.* [2021] An autonomous debating system, *Nature* **591**, 379−384, doi: 10.1038/s41586-021-03215-w.

Solms, M. [2021] *The Hidden Spring: A Journey to the Source of Consciousness.* (Profile Books, London).

Solvi, C., Gutierrez Al-Khudhairy, S. and Chittka, L. [2020] Bumble bees display cross-modal object recognition between visual and tactile senses, *Science* **367**(6480), 910−912, doi: 10.1126/science.aay8064.

Thompson, E. [2007] *Mind in Life: Biology, Phenomenology, and the Science of Life.* (Harvard University Press, Cambridge, MA).

West-Eberhard, M. J. [2003] *Developmental Plasticity and Evolution* (Oxford University Press, Oxford).