AI

MDPI

Opinion

The Ouroboros Model, Proposal for Self-Organizing General Cognition Substantiated

Knud Thomsen

Paul Scherrer Institut, 5232 Villigen-PSI, Switzerland; knud.thomsen@psi.ch

Abstract: The Ouroboros Model has been proposed as a biologically-inspired comprehensive cognitive architecture for general intelligence, comprising natural and artificial manifestations. The approach addresses very diverse fundamental desiderata of research in natural cognition and also artificial intelligence, AI. Here, it is described how the postulated structures have met with supportive evidence over recent years. The associated hypothesized processes could remedy pressing problems plaguing many, and even the most powerful current implementations of AI, including in particular deep neural networks. Some selected recent findings from very different fields are summoned, which illustrate the status and substantiate the proposal.

Keywords: biological inspired cognitive architecture; deep neural networks; auto-catalytic generation; self-reflective control; efficiency; robustness; common sense; transparency; trustworthiness

check for updates

Citation: Thomsen, K. The Ouroboros Model, Proposal for Self-Organizing General Cognition Substantiated. *AI* **2021**, *2*, 89–105. https://doi.org/10.3390/ai2010007

Academic Editors: Rafał Drężewski and Dimitrios Moshou

Received: 14 December 2020
Accepted: 22 February 2021
Published: 26 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The Ouroboros Model was conceptualized as the basic structure for efficient self-organizing cognition some time ago [1]. The intention there was to explain facets of natural cognition and to derive design recommendations for general artificial intelligence at the same time. Whereas it has been argued based on general considerations that this approach can shed some light on a wide variety of topics and problems, in terms of specific and comprehensive implementations the Ouroboros Model still is severely underexplored. Beyond some very rudimentary realizations in safeguards and safety installations, no actual artificially intelligent system of this type has been built so far. Nevertheless, diverse research fields have yielded results in the meantime, which supply ample supportive evidence for the main hypotheses of the Ouroboros Model.

In the following, it will be argued that this cognitive architecture still holds much promise waiting to be explored and utilized. It is claimed that this approach allows to address the "BICA Challenge" and to deliver towards a long list of topics, which one wishes to better understand concerning natural cognition [2,3]. Subsequently, the indicated answers can be harnessed for general AI. In order to substantiate this claim selected findings from a wide range of relevant fields are briefly reviewed. How future advancements for more standard AI might benefit from suggestions derived from the Ouroboros Model is the core theme of this short position paper.

2. Natural Cognition

No matter whether tight correspondence is assumed or just very high-level analogies, since the very start of work on Artificial Intelligence, natural intelligence has been a most valuable (the only) main source of inspiration. Ideally, models draw from (primate) neurobiology and, in turn, elucidate brain functions in a virtuous circle [4]. The workings of a mind (and soul) have been themes for myths, religions and philosophies over the ages. In a well-informed overview, topics, which one would like to understand better, have been listed recently [3].

1. Physical structures and functions underlying cognitive capabilities
2. Mechanisms for recognizing objects, actions and situations in uncertain environments
3. Emergence, evolution and development of cognitive capabilities in living organisms
4. Roles and implementation of memories
5. Types and organization of internal representations
6. Generating meaningful and adaptive behavior through (cognitive) processes
7. Adaptations and goal-setting mechanisms, purpose-driven learning
8. Genesis and roles of emotions and affect
9. Roles of communication and language
10. Ethics
11. Self-awareness and consciousness

One most general immediate upshot of the Ouroboros Model with respect to the above listed topics is that these questions cannot be cleanly separated from each other. It seems that, for the case of human cognition, disturbances or shortcomings related to any one of the topics significantly impact on the overall system performance. The Ouroboros Model is a proposal, which claims to shed light on all of these issues and their interdependencies, and the different facets of the overall picture are found to dovetail with only rather limited arbitrariness. What appears needed in any case is a comprehensive view with a wide perspective in order to address the links and interrelations between all these subject areas. Attempts to get a handle on cognition have hitherto largely followed a strategy of contrasting and mutually exclusive opposites, splitting research topics into either/or, with underwhelming success. In stark contrast, the Ouroboros Model strives for a coarse, but wide and overarching, picture; how selected dichotomies could be reconciliated in a common frame of comprehensive explanation has been sketched recently [5].

3. Selected Areas in the Focus for Current Approaches in AI

There is no doubt, artificial intelligence, AI, has made large strides over the last six decades, and even jumps in performance levels lately. Deep artificial neural networks regularly beat top human experts in domains, which were earlier thought of as reserved for truly intelligent human beings, e.g., automatic translation or the games of Chess, Go, and Jeopardy!

Still, a very subjective list of some current research topics and desiderata, or even manifest shortcomings of the presently prevalent AI approaches, can be compiled, like for natural cognition above [6–9]:

A. Structure
B. Decision Making
C. Learning
D. Memory
E. Scope and brittleness
F. Common sense, meaning, relevance
G. Consistency
H. Emotions
I. Transparency and trustworthiness
J. Ethics
K. Self-awareness and consciousness

These bullet points definitively are not meant as comprehensively covering all manifestly or potentially relevant topics but as a mere starting point for charting some of the territory where progress is needed, in part, very urgently. Many topics, which would deserve separate inclusion in a more comprehensive list of currently relevant topics, are not much dealt with explicitly, i.e., specific challenges in highly dynamically settings, gamification for enhancing motivation, ethics, applicable law, learning with sparse (or absent) rewards, building hierarchies of skills (nesting of schemata), and finding curricula for efficient explorations [10].

The focus here is primarily on underlying universal "technical" issues and not so much on the embedding of AI in contexts of applications, societies, or economics and the like, although it is argued that these cannot be considered meaningfully as independent but should rather be taken into consideration already during research and planning phases for AI systems [11].

4. Some Essentials of the Ouroboros Model and Their Relevance to the Highlighted Areas of Interest

The very basis for all action and cognition in the Ouroboros Model is provided by hierarchically-organized schemata, which are laid down in a kind of active memory [1]. Schemata are structured bundles of features of varying centrality, i.e., some are mandatory while others are more optional. They can exhibit strongly different extents ranging from a few associated attributes to whole narratives including background.

In perception as well as in cognition, a recursive self-monitoring process termed consumption analysis autonomously checks the fulfillment of anticipations based on memorized schemata. Activation (of features) due to incoming information is contrasted with (in its most simple version, substracted from) activation, which is based on memory (expectations biased). Slots in the prevalent schema without appropriate balance are, thus, marked as open. Depending on context and thresholds, not "consumed" surplus input can require changes in a schema or altogether another one. In each case, discrepancies between the current state and expectations derived from previous (similar) occasions are highlighted. Feedback from consumption analysis directs attention, action, and also the generation of new schemata and their storage in relevant situations for later use where/when the need arises. This extends from simple perceive $\rightarrow$ act schemata, which are required to perform specific routine tasks, to concepts for high-level self-reflection and goal-setting [12].

Comparing the Ouroboros Model to a generic actor-critic learning set-up shows some interesting parallels and communalities as well as differences, see Figure 1. Some implementation of time-difference learning between instantiations of states/between iterations of consumption analysis cycles is a shared feature. Actor-critic networks emphasize changing the environment. Accordingly, the environment delivers feedback to the actor and to the critic, and the latter is instrumental for policy gradient updates and choosing the next action by the actor. The environment also provides reward (as an external reinforcement signal) to the critic. The Ouroboros Model (as it stands now) focuses on the advancement of an internal procedure for settling on a satisfactory fit between an external (or internal, primary) data source and its (secondary) internal representation; actions are just a specific type of slots in schemata for influencing the concerned environment. Reward can come directly from the outside (from a supervisor), but primarily reward (a goodness of fit signal) is computed internally by consumption analysis. It is not in the foreground that analysis and actions are performed by different components of the system; the Ouroboros Model rather stresses the succession of different steps yielding an overall synchronous unity of an ongoing action.

The proposed general layout is meant as applicable to cognition, irrespective, who the actor is: an animal, a single human being or a group, a state, an artificial software agent, or a robot. It is the hypothesized backbone, the algorithmic and organizational structure of cognition, which is in the focus; for this, details of implementation do not matter too much. The abstract principles are claimed to be the same, whether they relate to real neurons or artificial neural networks or symbols in simulations (1/A).

In the following, the main characteristics of the Ouroboros Model are presented in an attempt to draw together different lines of research under this approach in a comprehensive view, together with a few pointers to the respective bullet points above ("1 . . . 11"/"A . . . K") as well as references to relevant recently published findings. It is not intended to go into intricate details, rather some coarse overview and potentially fruitful directions for research shall be indicated. In a fully self-consistent way, the layout is approximately in iterations, not as straight and linear deduction. The pointers in (round) brackets are also intended to further present and explicate the content in a self-reflective manner.

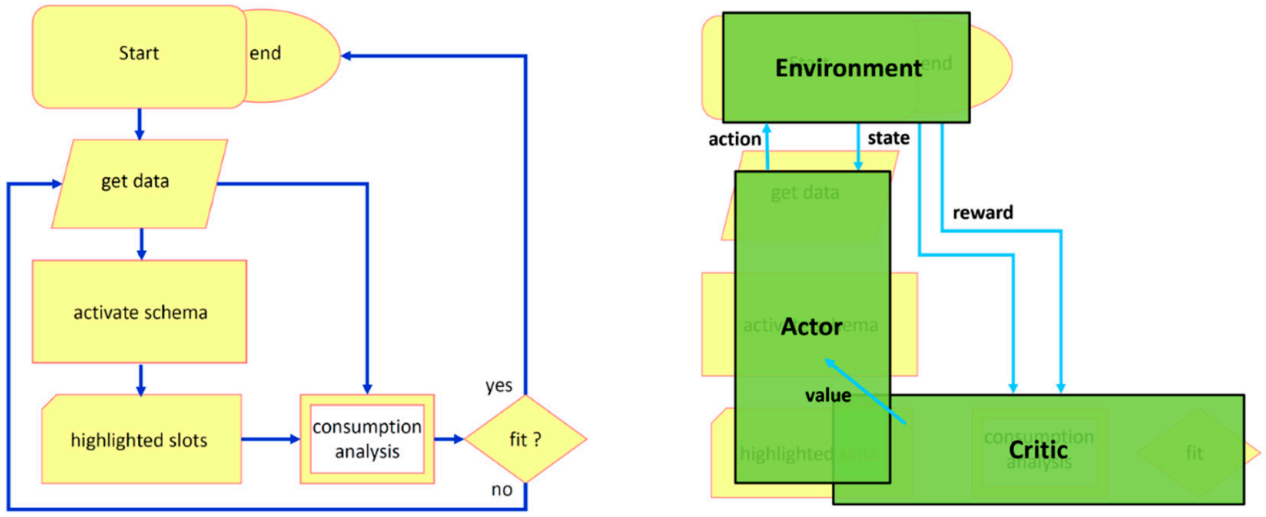

Figure 1. Coarse correspondence between the Ouroboros Model and a generic actor-critic neural network; (**left**): most simple version Ouroboros Model with arrows indicating the procedural flow [1], (**right**): generic actor-critic block diagram as an overlay.

There are only two basic ingredients, which in their interplay over time establish the full characteristics of the Ouroboros Model [1].

- Memory records in the form of schemata, organized into non-strict hierarchies
- An iterative process called consumption analysis, which continually monitors how well currently active input matches with expectations based on existing memory entries; through this, further action is triggered and steered.

The latter process-item emphasizing the role of time actually is a most important one. The Ouroboros Model describes evolution over time, with its own evolution—employing a repetitive procedure of error checking and subsequent correction, reiterated in cycles—included. In this process, representations are fine-tuned and incrementally built up in an auto-catalytic manner, while the number and sophistication of records grow. Memory in the form of reflexes begins at a fundamental bodily level, and it is the basis for cognition starting with simple if–then relations, and also at the top reaching up to planning, counterfactual abstract thinking, and even consciousness [12–14]. The Ouroboros Model, thus, is an example of self-propelling evolutionary computation. Starting from bodily grounding, cumulative extensions of the available schemata as knowledge base are only to some extent random. Schemata are expanded recursively with links to previously existing ones, and this process is constrained by abstract aims with some striving for survival at the (bottom) top (3/C).

There is no unstructured "knowledge soup", and neither any need for a "Lufthaken", which would allow to only consider "errors" from one level to the next in a simple version of predictive coding [15–17]. Taking time into full account and iterative incremental growth, which started with bodily adaptations long before the advent of cognition, there are always some earlier structures available without any danger of a vicious cycle (1/A).

"Vertically", concepts are organized in non-strict hierarchies, and "horizontally" features are (in living brains) optimally arranged in maps. This allows representing very accurate values in the form of distributions of activity over many neurons; it enables interpolations (horizontally), and both dimensions are very helpful for search when only approximate values are available as keys for retrieval. Relational coding employing well-defined building blocks (prototypes, capsule, features, attributes, soft labels, . . .) has may times been emphasized as very efficient approach [18–20]. By the structuring of attributes, numerical explosion can be avoided. Orderly representation in some type of map is not restricted to only very low-level attributes.

Schemata are bundles of ordered features, starting at the level of signals from (bodily) sensors, linked together as already envisioned by Otto Selz; most recently, they have experienced a great deal of renewed interest [18–23]. Different versions of schemata come with diverse names like sensations, percepts, concepts, prototypes, capsule, scripts, plans, mental models, or whole theories. They feature different shapes with respect to content, complexity and elaboration. The tight linking of individually rather well-defined constituent features, each amendable to pattern completion and internal consistency checking, grants robust coding of whatever objects; prime examples in this context are gestalts. A realization of Manfred Eigen's hypercycle ensures stability in the case that some attributes are not coded perfectly [24,25]. Activation spreading laterally as well as up and down hierarchies of building blocks always keeps some link of relevance to its origin. It is hypothesized that this ensures graceful degradation and a minimum of "common sense" in case no good fit is possible (3/C). The same applies to the ability to produce meaningful generalizations to previously not encountered stimuli (6/F). Obviously, the use of repeatedly occurring features as building blocks and intermediate levels of stability would benefit the perceptional and classification abilities of artificial applications, machine learning, pattern recognition, and deep neural networks. A focus of attention on a level above the neural representations has been recommended, and the strong benefit of distinct building blocks between low-level features and abstract symbols has been confirmed recently [19,20,26]. Specifically emphasizing prototypical parts of images in a deep learning network provides a high level of interpretability of classifications [27]. Even better than image patches, the leveraging of ("meaningful") previously established broadly tuned conceptual representations (i.e., schemata) improves performance and generalization from few examples [28].

Whereas artificial (convolutional) neural networks, CNNs, without special precaution tend to rely on local surface texture for the classification of images they can be trained with an emphasis on shapes by initially removing local cues; this significantly enhances their accuracy and robustness even when staying only in the visual domain [29] (5,6/E,F). Much can be learned from nature; already supplementing standard CNNs with a first layer modelled on the known characteristics of primary visual cortex V1 boosts their robustness to image perturbations [4]. With respect to time, the slow course of the maturation of human vision, in particular, of its acuity might not be accidental [30].

Towards a high-level end of the hierarchy of sequential stages in ventral visual cortex areas, representations of faces are reported to be coded following a coarse-to-fine strategy; modelling efforts with deep artificial neural networks reproduce this undoubtedly efficient organization of inferior temporal cortex when performance requirements as well as an ecological boundary condition of minimizing axonal wiring lengths are taken into account [31–33].

Global shape information is important. Crowding is an effect observed in humans and in feed forward convolutional neural nets: objects, which are easy to identify in isolation are hard to discern as clutter is added; this cannot be explained by classical accounts of vision [20]. What most common NNs cannot reproduce is un-cluttering: adding elements, which themselves can be nicely combined into gestalts, reduces crowding (the pointers here are in brackets of a different style than references in order to not disturb the fluence of reading too much). The inability to exhibit un-cluttering has been found to result in a principled way from the used architectures. Only with an organization into capsules (a version/another label for schema) recurrent grouping and segmentation can be achieved [19,20]. Self-reinforcement of the structure is suggested similarly to what had been proposed by Stephen Grossberg with the Adaptive Resonance Theory, ART [34]. The responsible process "routing by agreement", which enhances concordant feature-links and suppresses discrepant ones, is easily identified as another description of employing schemata and linking their constituents with available input; the resulting groups are distilled (and "ground-in") over repeated iterations.

Fitting nicely and in a complementary manner, unresolvable discrepancies as a result of the novelty detection effected by consumption analysis trigger the recording of new memory entries following the Ouroboros Model [1,12]. Quick memorization of encountered configurations can efficiently seed the process of subsequent grinding-in. It has been demonstrated that already at a rudimentary level of proto-objects, the linear combination of visual and auditory feature conspicuity maps, when incorporating gestalt-properties of proximity, convexity, and surroundedness, allows to capture a higher number of valid salient events than uni-sensory saliency maps [35,36] (4,5/D,E). The combined effect of bottom-up activation and top-down bias in living beings has been found in many experiments, e.g., during visual search [37,38].

In complete agreement with the tenets of the Ouroboros Model and in stark contradiction to hitherto held convictions, the stronger a memory is first encoded, the easier it is for a person to detect unexpected subsequent changes and to integrate them into a new record [39]. A strong and well-defined reference enables distinctive contrasting with new input; this results in a more pronounced discrepancy signal and stronger activation, which, in turn, better triggers a new memory entry including the old features, the novel ones, and their relationship. Such records can be understood as kind of snapshots ("core dumps") of all activations at a particular instant, immediately available for further use just as old schemata [1,12,40]. This is one-shot learning, contrasting the need for training with a large number of exemplars for neutral networks. Implementations appear to work with eligibility traces; these allow to extend the time, during which activations are considered for reinforcement [41]. For schemata comprising actions, experienced or expected outcomes including associated rewards, are constituent components.

With memories preferably laid down at the end of an episode, a "happy ending effect" ensues: episodes, which end well (and thus indicate improvement) are higher valued and bias future decisions more than bad ending ones [42]. Relief would just be an extreme example. Being afraid of bad feelings can trigger a vicious downward spiral [43].

In vertebrate brains, novelty triggers memorization [44]. For a given context (schema), deviations from expectations are primarily detected by the anterior cingulate cortex as key structure for model-based control [17,45] (3,4,7/C,D,G).

Schemata, interconnected attributes on one or more levels of abstraction, are regularly associated with an index entry on a meta level (in the hippocampus [46]), and often, especially at a medium level, there is a dedicated word denoting a concept [47]. The overpowering weight of a layer, which is many times activated and linked with language symbols, is assumed responsible for typicality- and non-monotonic effects in reasoning; for confidence judgements, representativeness was found to be the dominant contributing factor [48]. Having available a somewhat distinct second memory record provides a way to counter catastrophic forgetting while also avoiding reliance on solely local computations and completely isolated fine-grained changes during unsupervised learning.

For humans, with a sound grounding of concepts in their body, intrinsic qualia and "meaning" for an individual are available as attributes/building blocks for the whole hierarchy [1,49] (7,11/H,K). Whereas bodily grounding certainly is different for other cognitive agents, some level of self-awareness appears indispensable for somewhat encompassing intelligence [14] (3/C). (Only) in cases with strongly limited scope, models and thereupon winning strategies and skills can be learned from observations and sparse rewards without extensive semantics [50].

Short-term memory is comprised of activated entries from long term memory. In working memory, that content is processed together with fresh input (in the case of just "thinking", all content comes from extant records, they (parts thereof) are newly combined into novel arrangements and shapes) [9,51] (4/D). There are individual limits for what can distinctly be handled for how long, corresponding also to the capabilities for deploying attention and detecting inconsistencies [52].

The compartmentalization into discrete ("snapshot") records poses a challenge for continuous perception and action at the most fine-grained level. Animals have found a solution: cerebella function in brains as general-purpose interpolation machines [53].

The process of consumption analysis, which primarily delivers a graded goodness-of-fit signal, is the same, whatever the detailed content. What changes between levels in the hierarchy of schemata are the features (attributes, sub-schemata, abstractions, . . . all of principally similar format), which are bound together. This can vary from simple sensory signals to complex aggregations of previously elaborated schemata. In the real world inhabited by living beings, some building blocks are used very often and thus useful as schemata themselves [12]. Consistency can then subsequently be checked already for such components, and the total figure of a percept is immune to noise or tempering on a very detailed level [46]. The processing (consumption analysis) is the very same, and it is, in particular, the same for cognition and meta-cognition; just that for the latter higher-order and, for self-awareness and -consciousness, self-reflective attributes are taken into account (11/K).

The construction according to the Ouroboros Model is fully recursive, with consumption analysis controlling the generation of new concepts geared to contexts, where they are most useful, i.e., mainly when consumption analysis signals a better-than-expected match or a really bad match [12], (2–5/C–E). Repetitive activations can be seen as evidence for some form of good fit as otherwise there would have been no chance for repetition. As new schemata become available, they in turn exert their influence and guide subsequent perceptions and actions along their distinctions and tracks [40]. An important special case are analogies and abstractions, distilled from activity when only some of the totally selected features are coactivated repeatedly; differentiation of underlying details into distinct schemata would be a "twinning-process". Prime examples resulting from abstractions or generalizations from repeated occurrences of constellations are the rules of logic, moral rules, justice, Bayesian rationality, and harmony (5,10/E,J). Another here especially relevant trio of concepts, which have been abstracted over time, and which have developed a standing strong enough to hide their origin from particular forms of analogies, are induction, deduction, and abduction, following John Sowa in succession to Charles Peirce [15,54].

By harnessing all the previously established building blocks at a given time, new concepts can be assembled and coded most efficiently and parsimoniously in recursive composition when building up a hierarchy [55]. Basically, all schemata can be seen as bundles of the underlying features, starting from responses of simple cells in primary sensory cortex as an example. As a significant benefit, this organization allows for high complexity: no simple and straight-forward 'tertium non datur' applies at the level above the simplest constituent features; e.g., unambiguous "opposites" can only be determined unequivocally in completely clear-cut and restricted contexts (like inside formal logic).

Consumption analysis biases a whole schema, it draws attention to empty slots. This is the basis for activity prepared in the short-term [1]. All usually-linked concepts are slightly activated in a graded manner corresponding to their respective relevance (also given by shared features and connections in higher schemata) in the current situation, e.g., also for an explanation. Clearly distinct schemata are efficient means to avoid catastrophic forgetting when new content is learned; switching between such contexts allows transfer learning as well as lifelong growth [56]. Content is coded with features and they are in turn incorporated into schemata pertaining to the respective context. Dynamically switching contexts entails switching between applicable schemata. The compartmentalization of the total of memory into schemata also enables efficient local consistency checks (and error corrections) and to cleanse records of internal inconsistencies, while tolerating contradictions between largely unconnected fields. At the same time, given the delineation of a schema, which tells what constituent features ought to be present, the absence of an expected (relevant) attribute can carry important information (G/6). Contrastive explanations are a means for allowing people and AI to explain their working by identifying what is missing [57]. Using

selected counter examples showing what should not be there is an efficient approach for descriptions and for guiding learning and categorization [58].

A second result of consumption analysis, more long-term, yields information on a second or meta level; it monitors how matters develop in general, e.g., whether matches work out in a consistent way. This "housekeeping" signal has been hypothesized as constituting the "feeling" component of emotions [1]. Feelings of certainty, or uncertainty, confidence, a feeling of rightness have been observed as resulting from meta-cognitive monitoring [59]. They are not so much based on heuristics but rather on "self-observation", and they serve for regulating the timing and the invested mental effort. Cognitive control depends crucially on emotion [60] (8/H).

Affective tones are slower than direct gap-filling in two interlinked ways: in their generation, i.e., they arise from a wider perspective of monitoring, and in their impact, i.e., they are effective over longer periods of time. With previously laid down building blocks, each likely annotated with "inherited" affective attributes, and freshly generated emotional tags in the prevalent situation, mixed feelings can be expected as probably arising [5]. The possibility of akrasia, i.e., "acting against one's better knowledge" in situations with conflicting goals and motivations (feelings) has already been observed by the old Greeks. Matters get even more complicated as soon as anticipated emotions, e.g., during planning and thorough deliberations, are included; they might pertain to expected performance, previous as well as expected tags, and all combinations of these. This way, second- (and higher-) order emotions arise. Still, in general it does make sense to act on the appraisal of a current state and to take accrued emotional attributes into account [60]. In the case of intuitive judgements, i.e., in the absence of detailed cognitive efforts brought to a satisfactory conclusion, not much differentiated feedback can be obtained, only some general uneasiness, and this, plus some (older) feelings, comes to the foreground as "gut feeling".

The Ouroboros Model explains reasoning in the spirit of a simulation theory [61]. Elaborate schemata are mental models [62]. During cognition, the open slots posed by a task, e.g., in the format of a question, are filled-in recursively. For planning as well as for counterfactual reasoning, specific constituents are replaced with equally constrained selected alternatives; relevant then is the comparison between the ensuing expected results and the ones aimed at to reach a certain goal. Anticipations and predictions are at the heart of imagination (4/D).

Rational reasoning according to the Ouroboros Model means combining (parts of) different schemata in order to construct a consistent whole. In a most simple case, one specific schema is indicated, and there is one open slot, which demands closing/filling. Who, where, when—questions are examples; with why the situation can become a little more complicated depending on the involved content and context. Devising a scientific theory or assembling whatever narrative follows the same basic principles. If–then rules can be seen as derived from schemata, as kind of linear projections, shortcuts between different parts of a higher-level schema. When a schema (concept) gets activated it is biased as a complete unit, comprising parts, that are already available, as well as open slots, to which attention is thus focused and which are marked this way as demanding their filling [1]. One type of the many other possible relationships between (parts of) schemata, i.e., causal ones, deserves special mention; (counterfactually) disabling the relevant features/conditions while carefully taking time and timely successions into account (with an effect/event never observed before its cause and trigger) yields a significant open gap in the overall schema (7/G).

The potential mental processing power of an agent is ground-laid in the available knowledge, i.e., the number, coverage, complexity and elaboration of the concepts at her disposition; sheer performance at a single point of time arises as a result of the optimum interplay between these structured data and the effective execution of all the described processing steps, in particular, self-referential consumption analysis. Domain-general processes are, thus, essential, and they work on/with/for domain-specific as well as

more general content. Actively reducing discrepancies as fed back from consumption analysis can be understood as overarching motivation according to the Ouroboros Model; a somewhat comparable progress-based curiosity signal has been proposed very recently under the name of "γ-Progress", which measures improvement over exponentially decayed past models [63].

Living beings strive for consistency, and a top abstraction appears to be perfect (timeless) harmony. As a fitting example, a present-day glorification of Xi Jinping can be quoted: unity between theory and practice and many more concept-pairs argued to be characterized as in "unity" are enlisted for highest praise [64]. The overall aim for any action and also during problem solving is to establish and maintain a consistent and encompassing behavior or mental construct, in particular to cover the gaps as evidenced by any acute discrepancy or problem. Schemata at the very top level thus drive self-motivated action and living. (Physiological) homeostasis marks the other end of the scale. In between, problems can be solved intuitively or through conscious effort, e.g., with a trial-and-error strategy. Whatever guidance from (higher level) extant schemata ("structured priors" in mental models) can be brought to bear, mostly helps. In accordance with many and also most recent modelling and testing results, the ability to imagine hypothesized actions and their outcomes is taken as basis for flexible problem solving [13,21,22,65] (7/G).

Anticipations, according to the Ouroboros Model, thus lie at the core of all cognition and decision making [1,13]. The degree of their fulfillment drives adaptive action on different time scales ranging from immediate reflective or intuitive action, fine-tuning of representations, to life-long quests. It is in each case slots, which are somehow defined in a schema with the intrinsic motivation to be filled—immediately or over the course of an entire life. Curiosity is not arbitrarily evoked by areas of maximum blindness or diversity but rather specifically at the very borders of the (individually) established knowledge with medium levels of uncertainty and discrepancy. In a self-steered and autocatalytic manner, relevant concepts, schemata, are elaborated and iteratively built up, and with increasing complexity and sophistication predictability and control dwindle [11]. A general finding in this respect is that the interplay between previously established schemata and freshly incoming data can be construed as Bayesian optimization [66]. Given any well-defined frame in the real world, still, unavoidable uncertainties at all ends have to be expected and considered. The fundamental tenet of the Ouroboros Model here is that the aim has to be taking the most comprehensive considerations and the widest accessible context into account as available at a given time. Some neural substrates suitable for Bayesian inference have been found in the cerebral cortex [67]. Paying heed to real-world constraints for explanations, some deviations from exact Bayesian prescriptions (for basically static conditions) do make sense, e.g., as resulting from a trade-off between speed and accuracy [68] (6/F).

For living beings, the widest context always includes precepts, constraints, qualia, and feelings related to their bodies [14,47]. This is seen as a mandatory basis for common sense in human beings, but it can also lead to certain types of akrasia (maybe like addictions), and it can at the same time explain, why limits and clear violations from true Bayesian-optimized behavior are observed (6/F). As just one example, discounting can be easily understood as a heritage from times when the future was incomparably less certain than in the settings of laboratory experiments. Interesting are the "unknown unknowns". Anticipations of rewards are signaled in mammalian brains by dopamine using a distributional code for value, and it can be observed that with learning the associated signal is evoked earlier and already for predictive cues [69,70]. Anticipations of others' actions are fundamental for successful communication [47]. As an example for an analogy: the same basic mechanism as for akrasia can be seen at work at early adversary emails in which "good" words have been used in attempts to compromise spam filters [71].

At present, with AI rapidly increasing its importance as a constituent part of the real world in which we humans live, questions concerning safety, reliability, transparency, trustworthiness, accountability, communication, ethics in general, become ever more

pressing [7,11,61] (10/J). Consistent explanations are a prerequisite for trustworthiness. The good news, in a sense, lie in the fact that humans in their dealings with each other, have found viable solutions; the bad news are that human relations and behaviors certainly often are far from optimum. The Ouroboros Model claims to shed a little light on these issues by pointing out some way to communication, necessary for interpretations, (and confidence in them), transparency and mutual understanding while also highlighting limits [8,11,47,72]. Truth and trustworthiness (of a message and, in the end, of an agent) can be probed via the consistency of suitable parts of a narrative; coherence and consistency, i.e., explanatory goodness, is what counts [68].

Building blocks, which have some meaning of their own improve the interpretability, e.g., of image recognition. In "concept whitening", a dedicated module is added to a convolutional neural network and, through training, structure is imposed on hidden layers, which is aligned with meaningful axes corresponding to predefined concepts. Latent space is thus decorrelated and classifications are disentangled. Maximizing activation of single "concept-neurons" allows for the interpretation of activations and for tracking the associations and the progress by the neural net [73].

Following the "progressing" and open-ended evolution of a single agent in tight exchange with its surroundings and, in particular, with other agents, it can be argued that a certain amount of (bodily and bodily-grounded) self-awareness, awareness of roles and perspectives, and, at a still higher level, self-consciousness, is inevitable [14,49,72,74]. As the ultimate frontier of natural and artificial intelligence, consciousness defines the very top, which humans strive to understand and (someday) to artificially build (11/K). According to the Ouroboros Model, it comes in situations into play, when Higher Order Personality Activation (HOPA) is required as simple if—then reactions do not suffice [14]. While emphasizing the importance of including attributes of an actor herself in her cognitive and other actions for evolving and employing consciousness, it is probably equally important, that the processes is mirrored in several ways; consciousness is not absolute, it is also an ascribed feature; we attribute it to others, they grant it to us. In mutual synergy the concept stabilizes [14,72] (9/I). As to the temporal structure, it appears a mandatory consequence of the organization of the memory and a process in iterative cycles with only intermittently available monitoring results (feedback, reward) that the stream of consciousness, experienced as unique and continuous, below the surface is of a discrete retentional type [75] (11/K).

No split into different systems is necessary, distinct stages in succession of one process unfolding step-by-step fully suffice [26,76]. When an action is demanded, speed and the involvement of conscious effort depends on the availability of readymade schemata; a reflex, an immediate intuitive decision, or several iterations taking wider and different contexts into account, are just different elaborations of one and the same iterative process continuing until satisfactory match, in this case between task and solution, is reached, or the process has to be stopped and restarted in situations where a time limit is exceeded (2/B). Just in cases when the first hit of a schema is fully satisfactory no further iterations are required (this would correspond to "system 1").

Any decision, normally, means that a conclusion has been reached, all slots in the respective schema have successfully/satisfactorily been filled (or that schema has been discarded) and (in both cases) tagged as "consumed"; this starts with very basic behavior, and also, at higher levels, entails an inhibition of return. Activity is subsequently controlled by a schema higher up in the hierarchy (and/or a new one is selected for a still open task). Thus, activity unfolds over time in iterative cycles with the dragon biting its tail (hence the name of the Ouroboros Model) or rather spiraling upwards into an open future (3/C).

No account of cognition could be claimed as anywhere approaching reasonable coverage without taking note of the fact that the most successful actors (and also not so highly sophisticated ones) do forget and do need sleep (1,3,4,6/C,D,G). The Ouroboros Model sees these two observations as just more facets of the complete picture: based on their relevance, only useful records are maintained, and strengthening these certainly is one function of

sleep. With tight timely constraints during active and awake periods, "leftovers" are continually produced during matching, i.e., attributes and (fragments of) schemata, which have not been fully consumed. Getting rid of this "data garbage" has been postulated as another, equally important function of sleep, in particular during dreaming [77]. In the light of the fact that evolution has found no way of getting completely away without sleep, it can safely be assumed that not only AI built employing a similar architecture would also require some intermittent off-time for "housekeeping". Actually, this is also currently being investigated by other groups [78,79]. It has been proposed that sleep helps with the stability of spiking neural networks and against getting stuck in excessive detail ("overfitting"). Whereas these certainly are valid aspects, the Ouroboros Model additionally offers a much more principled explanation. One function does not necessarily exclude a second one.

5. Discussion and Summary of Solutions Indicated by the Ouroboros Model with Evidence in Recent Findings

Below is just the gist of the references and recommendations; it can be summarized as a plea of taking all and everything, which could be considered relevant for an action (including, in particular, cognitive action) into full account as far as possible at a given point in time in a self-reflective manner; observing boundary conditions dictated by the limited availability of time and paying heed to dynamic evolution ensues. While based on the Ouroboros Model and thus at least strongly inspired by natural intelligence, it is claimed that the distilled abstract features are useful for any type of intelligence, in particular, also for abstract and symbolic AI. Only in an incremental manner with long phases of learning in structures, the heights of (human) cognition appear reachable, not to speak of consciousness.

Structuring, thus, seems mandatory to take the next step in the performance level of AI, i.e., structuring in several respects:

- Layout/organization of artificial systems
  - Compartmentalization of the memory base, i.e., delimited prototypes for repeatedly encountered combinations starting with co-occurrences of simple features
  - (Non-strict) hierarchy of schemata/prototypes, recursively built, commencing with first prototypes and extending to the highest levels of sophistication and abstraction
- Procedures, following some well-orchestrated timing and strict succession of substeps, in particular, taking into full account the time at which structures are worked with/on, and their meaningful variations and build-up
  - Encompassing monitoring for consistency of activations
  - Steering short-term activity depending on the results of the monitoring process
    - Attentional mechanisms (in particular)
  - Steering long-term activity depending on the results of the monitoring process
    - Memory extension
    - Emotions
  - Consistency curation as an overall task
- For the highest individual (meta-) levels, self-referring memory records grounded in the physical existence of the actor are relevant; they grow after the same autocatalytic mold as any other schema in memory, and their full inclusion in the overall consistency monitoring comes naturally as soon as some common sense beyond a narrow, given frame is demanded. For self-awareness, curiosity, and self-consciousness they appear indispensable.

Starting out with learning only on a coarse level, early combination of different channels (features, data sources, attributes, . . .) for intermediate components, junks, as they occur repeatedly and are abstracted, and only later gradually moving to fine-graded differentiations will reduce brittleness and help against attacks using adversary data by

blocking unwarranted shortcuts during learning. Whereas the requirement of training with a very large number of examples for current artificial neural networks with back propagation is, biologically, most implausible for explaining adult human performance in single-shot learning or successful classification from a few specimens, it can fit with developmental processes in infants and young children for sensory and lower level analysis cortical areas.

Taking a wide perspective, the mode of action of the standard approach for learning, i.e., back propagation, from a reward back to the earlier underlying neural levels, can effectively be implemented by following the full circle around in an incremental manner: structures established at one instant will be employed at a next similar occasion and fine-tuned iteratively over repeated activations. In standard back propagation, an error signal is determined from contrasting the output of a neural network with the correct result supplied by the supervisor. Weights between layers are adjusted starting with the output and stepping (not biologically plausibly) back towards the input, see Figure 2.

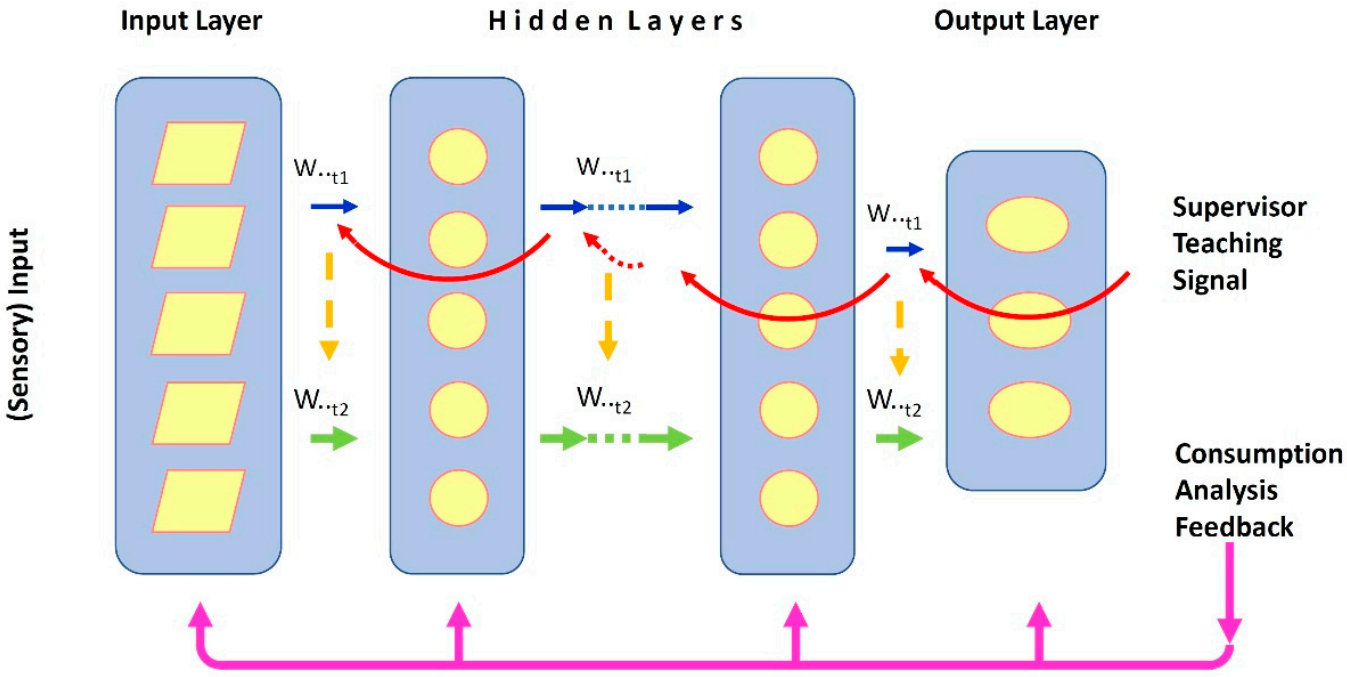

Figure 2. Beginning with first connections, weights are tuned based on teaching signals propagating backwards from output to input neural layers in standard back propagation (red arrows). With one globally distributed monitoring signal (magenta) it is possible to reach and, especially during subsequent occasions, tune input layers in a fully biological plausible way. Optimized (green) connections and their associated weights most likely turn out very similar in limited domains; deriving relevant details is work in progress.

Of course, it is always links and their weights, which determine the influence, which one neuron exerts on others downstream. But, according to the Ouroboros Model, in the beginning is not primarily the high precision of the weights, which record something; rather it is first decisive, which neurons participate at all in a percept or in whatever task. The common activation of (tagged) neurons leads to their linkage, which is triggered and modulated by a widespread diffuse monitoring signal as delivered by "consumption analysis". Weights are only later iteratively ground in, honed for fine-grained distinctions. This process appears to work already with the second encounter of a (similar) stimulus, demanding very few repetitions for some meaningful performance (as exhibited by natural intelligence). How these two accounts match with each other, and in particular, under which conditions they probably lead to similar (weights in) networks in the end, is in the focus of ongoing investigations.

While an organization into maps appear to evolve "naturally" in artificial neural networks, augmenting them with some type of separate "index" could help in many ways as hypothesized by the Ouroboros Model concerning a main function of hippocampi [46]. Neural nets, at least mostly in hidden layers, feature smooth representations; outcomes, e.g., for classifications, typically demand discrete results. In order to represent intermediate values or smooth transitions between discrete entities, some general-purpose interpolator, as what has been argued to be the main function of cerebella, might be good to have [53].

Relying on well-defined and -adapted robust stepping stones in the form of nested schemata, especially if they are harnessed at suitable points in time, certainly constrains detrimental freedom for fitting and learning and, thus, offers a remedy against under-specification, which has been identified as major problem shared by many AI schemes [80]. Practically, this means training in stages and with batches of selected input with an emphasis on causal structures, and probably also some (anyway realistically unavoidable) noise injection, well-timed and in proper doses. Meta-reasoning can advantageously be employed to optimize the process. Credibility, trustworthiness, and, in the end, common sense, are enhanced in parallel to progress traceability and the outcome with its derivation being understandable to a human.

Arriving just in time for this report, it is reassuring that AlphaFold has seemingly solved the folding problem by following an approach of similar type, which takes into account global constraints and emphasizes self-consistency. It pieces together local junks before fitting these into a whole; an "attentional" mechanism focusing on relevant and already known parts ("puzzle pieces") before filling-in gaps and assigning confidence scores are reported to complement the applied methods in iterative cycles [81,82].

A list with some design recommendations more directly based on the Ouroboros Model starting from a responsibility/traceability/ethics point of view has been published recently [11]. Going far beyond specific commendations for the design of AI, the considerations there might be summarized as culminating in a "negative imperative": with tight common restrictions and interlaced complex links and dependencies between partners, any violent action with high probability has negative impacts also on the originator and the whole world and, thus, should be avoided for his/her own most intrinsic self-interest (or collective interest). This applies to actors of whatever type.

6. Conclusions

Above considerations and recommendations are meant as being useful for addressing some selected specific topics and also general challenges for AI, both on a structuralist and on a functional level. Despite the cited converging developments and findings from neurobiology, cognitive science, and AI, it will only be possible to fully judge the merit of these after fleshing out details and investing some sizeable effort connected with actually implementing them directly in working AI systems.

Nevertheless, it is argued that one significant benefit of the Ouroboro Model is evident already right now: Starting with a very simple compound hypothesis, a great many different topics can be seen and addressed in one overarching frame, which adds to the overall consistency, transparency, and promise, of the approach; collaborations to further pursue it are most welcome.

Envisioned future steps are clear: going from the accruing but more indirect evidence in support for the Ouroboros Model as presented here to dedicated implementations comprising formalization and actual tests. Working out in full detail the correspondence with actor-critic neural networks as hinted at in Figure 1 is an obvious next step. Complementary efforts should of course be devoted to formulating clear-cut predictions derived from widely separated fields, which the approach claims to elucidate; the farther apart the connected topics and the more counter-intuitive a prediction, the more surprising and the better. More comprehensively understanding human cognition and working whatever systems following the recipes of the Ouroboros Model, which exhibit general artificial intelligence comprising, amongst others, single-shot learning, high flexibility, and common

sense, undoubtedly are more long-term aims. Successful artificial implementations would be needed to convincingly prove the tenets of the Ouroboros Model. How much there is still missing for a truly general AI to be accredited with genuine consciousness remains to be seen.

Only during the last stages of the editing process the author became aware of somewhat similar efforts by researchers proposing selective gating between distinct records in working memory and, in particular, including a matching process and a global feedback signal for more biologically plausible learning [83].

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: A number of knowledgeable and thought-provoking comments and questions by three anonymous reviewers as well as excellent support by the editor are gratefully acknowledged.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Thomsen, K. The Ouroboros Model in the light of venerable criteria. *Neurocomputing* **2010**, *74*, 121–128. [CrossRef]
2. Chella, A.; Lebiere, C.; Noelle, D.C.; Samsonovich, A.V. On a Roadmap to Biologically Inspired Cognitive Agents. In: Samsonovich, A.V., Johannsdottir, K.R. (Eds.) Biologically Inspired Cognitive Architectures 2011—Proceedings of the Second Annual Meeting of the BICA Society. *Front. Artif. Intell. Appl.* **2011**, *233*, 453–460.
3. Stork, H.-G. Towards a scientific foundation for engineering Cognitive Systems—A European research agenda, its rationale and perspectives. *Biol. Inspired Cogn. Arch.* **2012**, *1*, 82–91. [CrossRef]
4. Dapello, J.; Marques, T.; Schrimpf, M.; Geiger, F.; Cox, D.D.; Di Carlo, J.J. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. *bioRxiv* 2020. [CrossRef]
5. Thomsen, K. It Is Time to Dissolve Old Dichotomies in Order to Grasp the Whole Picture of Cognition. In *Proceedings of the International Conference on Theory and Practice of Natural Computing*; Fagan, D., Martín-Vide, C., O'Neill, M., Vega-Rodrìgue, A., Eds.; Springer: Heidelberg, Germany, 2018; pp. 1–11.
6. Dwivedi, Y.K.; Hughes, L.; Ismagilova, E.; Aarts, G.; Coombs, C.; Crick, T.; Duan, Y.; Dwivedi, R.; Edwards, J.; Eirug, A.; et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* **2021**, *57*, 101994. [CrossRef]
7. Russel, S.; Dewew, D.; Tegmark, M. Research Priorities for Robust and Beneficial Artificial Intelligence. In *AI Magazine Winter*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2015.
8. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
9. Reggia, J.A.; Katz, G.E.; Davis, G.P. Artificial Conscious Intelligence. *J. Artif. Intell. Conscious.* **2020**, *7*, 95–107. [CrossRef]
10. Aubret, A.; Matignon, L.; Hassas, S. A survey on intrinsic motivation in reinforcement learning. *arXiv* **2019**, arXiv:1908.06976.
11. Thomsen, K. Ethics for Artificial Intelligence, Ethics for All. *Paladyn. J. Behav. Robot.* **2019**, *10*, 359–363. [CrossRef]
12. Thomsen, K. Concept formation in the Ouroboros Model. In Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010, Lugano, Switzerland, 5–8 March 2010.
13. Thomsen, K.K. The Ouroboros Model, Selected Facets. In *From Brains to Systems*; Springer: New York, NY, USA; Dordrecht, The Netherlands; Heidelberg, Germany; London, UK, 2011; pp. 239–250.
14. Thomsen, K. Consciousness for the Ouroboros Model. *J. Mach. Conscious.* **2010**, *3*, 163–175. [CrossRef]
15. Sowa, J.; Ramadas, J.; Chunawala, S.; The Challenge of Knowledge Soup (Eds.) *Research Trends in Science, Technology and Mathematics Education*; Homi Bhabaha Denter: Mumbai, India, 2006.
16. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [CrossRef]
17. Thomsen, K. ONE Function for the Anterior Cingulate Cortex and General AI: Consistency Curation. *Med. Res. Arch.* **2018**, *6*, 1.
18. Sucholutsky, I.; Schonlau, M. 'Less Than One'-Shot Learning: Learning N Classes From M<N Samples. *arXiv* **2020**, arXiv:2009.08449.
19. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic Routing Between Capsules. *arXiv* **2017**, arXiv:1710.09829.
20. Doerig, A.; Schmittwilken, L.; Sayim, B.; Manassi, M.; Herzog, M.H. Capsule networks as recurrent models of grouping and segmentation. *PLoS Comput. Biol.* **2020**, *16*, e1008017. [CrossRef]
21. Selz, O. *Über Die Gesetze Des Geordneten Denkverlaufs, Volume 1*; Spemann: Stuttgart, Germany, 1913.
22. Selz, O. Zur Psychologie des produktiven Denkens und des Irrtums. In *Über die Gesetze des Geordneten Denkverlaufs, Volume 2*; Cohen: Bonn, Germany, 1922.

23. Ghosh, V.E.; Gilboa, A. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia* **2014**, *53*, 104–114. [CrossRef]
24. Eigen, M.; Schuster, P. The Hypercycle, A Principle of Natural Self-Organization, Part A, Emergence of the Hyper-cycle. *Naturwissenschaften* **1977**, *64*, 541–565. [CrossRef]
25. Eigen, M.; Schuster, P. The Hypercycle, A Principle of Natural Self-Organization, Part B, The Abstract Hypercycle. *Naturwissenschaften* **1978**, *65*, 7–41. [CrossRef]
26. Lieto, A.; Lebiere, C.; Oltramari, A. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cogn. Syst. Res.* **2018**, *48*, 39–55. [CrossRef]
27. Chen, C.; Li, O.; Tao, C.; Barnett, A.J.; Su, J.; Rudin, C. This Looks Like That: Deep Learning for Interpretable Image Reconstruction. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
28. Rule, J.S.; Riesenhuber, M. Leveraging Prior Concept Learning Improves Generalization From Few Examples in Computational Models of Human Object Recognition. *Front. Comput. Neurosci.* **2021**, *14*. [CrossRef] [PubMed]
29. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, The International Conference on Learning Representations, New Orleans. *arXiv* **2019**, arXiv:1811.12231.
30. Mayer, D.L.; Dobson, V. Visual acuity development in infants and young children, as assessed by operant preferential looking. *Vis. Res.* **1982**, *22*, 1141–1151. [CrossRef]
31. Dobs, K.; Isik, L.; Pantazis, D.; Kanwisher, N. How face perception unfolds over time. *Nat. Commun.* **2019**, *10*, 1–10. [CrossRef] [PubMed]
32. Dobs, K.; Kell, A.J.; Martinez, J.; Cohen, M.; Kanwisher, N. Using task-optimized neural networks to understand why brains have specialized processing for faces. *J. Vis.* **2020**, *20*, 660. [CrossRef]
33. Lee Hyodong Margalit, E.; Jozwik, K.M.; Cohen, A.M.; Kanwisher, N.; Yamins, D.L.K. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv* 2020. [CrossRef]
34. Grossberg, S. Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw.* **2013**, *37*, 1–47. [CrossRef]
35. Russell, A.F.; Mihalaş, S.; Von Der Heydt, R.; Niebur, E.; Etienne-Cummings, R. A model of proto-object based saliency. *Vis. Res.* **2014**, *94*, 1–15. [CrossRef] [PubMed]
36. Ramenahalli, S. A Biologically Motivated, Proto-Object-Based Audiovisual Salency Model. *AI* **2020**, *1*, 487–509. [CrossRef]
37. Spaak, E.; Fonken, Y.; Jensen, O.; De Lange, F.P. The Neural Mechanisms of Prediction in Visual Search. *Cereb. Cortex* **2015**, *26*, 4327–4336. [CrossRef]
38. Shurygina, O.; Kristjánsson, Á.; Tudge, L.; Chetverikov, A. Expectations and perceptual priming in a visual search task: Evidence from eye movements and behavior. *J. Exp. Psychol. Hum. Percept. Perform.* **2019**, *45*, 489–499. [CrossRef]
39. Stawarczyk, D.; Wahlheim, C.N.; Etzel, J.A.; Snyder, A.Z.; Zacks, J.M. Aging and the encoding of changes in events: The role of neural activity pattern reinstatement. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 29346–29353. [CrossRef] [PubMed]
40. Haque, R.U.; Inati, S.K.; Levey, A.I.; Zaghloul, K.A. Feedforward prediction error signals during episodic memory retrieval. *Nat. Commun.* **2020**, *11*, 1–14. [CrossRef]
41. Lehmann, M.P.; He, A.; Xu, L.V.; Herzog, M.H.; Gerstner, W.; Preuschoff, K. One-shot learning and behavioral eligibility traces in sequential decision making. *eLife* **2019**, *8*, e47463. [CrossRef]
42. Vestergaard, M.D.; Schultz, W. Retrospective Valuation of Experienced Outcome Encoded in Distinct Reward Representations in the Anterior Insula and Amygdala. *J. Neurosci.* **2020**, *40*, 8938–8950. [CrossRef]
43. Ford, B.Q.; Lam, P.; John, O.P.; Mauss, I.B. The psychological health benefits of accepting negative emotions and thoughts: Laboratory, diary, and longitudinal evidence. *J. Pers. Soc. Psychol.* **2018**, *115*, 1075–1092. [CrossRef] [PubMed]
44. Fredes, F.; Silva, M.A.; Koppensteiner, P.; Kobayashi, K.; Joesch, M.; Shigemoto, R. Ventro-dorsal Hippocampal Pathway Gates Novelty-Induced Contextual Memory Formation. *Curr. Biol.* **2020**, *31*, 25–38.e5. [CrossRef]
45. Akam, T.; Rodrigues-Vaz, I.; Marcelo, I.; Zhang, X.; Pereira, M.; Oliveira, R.F.; Dayan, P.; Costa, R.M. The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection. *Neuron* **2021**, *109*, 1–15. [CrossRef] [PubMed]
46. Thomsen, K. The Hippocampus According to the Ouroboros Model, the 'Expanding Memory Index Hypothesis'. In Proceedings of the IARIA COGNITIVE conference, Athens, Greece, 19–23 February 2017.
47. Thomsen, K. The Ouroboros Model embraces its sensory-motoric foundations. *Stud. Log. Gramm. Rhetor.* **2015**, *41*, 105–125. [CrossRef]
48. Dujmović, M.; Valerjev, P.; Bajšanski, I. The role of representativeness in reasoning and metacognitive processes: An in-depth analysis of the Linda problem. *Think. Reason.* **2020**, 1–26. [CrossRef]
49. Legrand, D. Myself with No Body? Body, Bodily-Consciousness and Self-consciousness. In *Handbook of Phenomenology and Cognitive Science*; Schmicking, D., Gallagher, S., Eds.; Springer: Dordrecht, Netherlands, 2010.
50. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nat. Cell Biol.* **2020**, *588*, 604–609. [CrossRef]
51. Cowan, N. What are the differences between long-term, short-term, and working memory? *Prog. Brain Res.* **2008**, *169*, 323–338.
52. Thompson, V.A.; Johnson, S.C. Conflict, metacognition, and analytic thinking. *Think. Reason.* **2014**, *20*, 216–244. [CrossRef]
53. Thomsen, K. The Cerebellum according to the Ouroboros Model, the 'Interpolator Hypothesis'. *J. Commun. Comput.* **2014**, *11*, 239–254.

54. Sowa, J. Cognitive Architectures for Conceptual Structures. In Proceedings of the 19th Conference on Conceptual Structures, Derby, UK, 25–29 July 2011; Andrews, S., Polovina, S., Hill, R., Akhgar, B., Eds.; Springer: Heidelberg, Germany, 2011; pp. 25–49.
55. Zhu, L.; Lin, C.; Huang, H.; Chen, Y.; Yuille, A. Unsupervised Structure Learning: Recursive Composition, Suspicious Coincidence and Competitive Exclusion. In Proceedings of the ECCV 2008 Part II, Marseille, France, 12–18 October 2008; Forsyth, D., Torr, P., Zisserman, A., Eds.; Springer: Heidelberg, Germany, 2008; pp. 759–773.
56. Tsuda, B.; Tye, K.M.; Siegelmann, H.T.; Sejnowski, T.J. A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 29872–29882. [CrossRef] [PubMed]
57. Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; Das, P. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
58. Wu, M.; Zhuang, C.; Mossé, M.; Yamins, D.; Goodman, N. On Mutual Information in Contrastive Learning for Visual Representations. *arXiv* **2020**, arXiv:2005.13149.
59. Ackermann, R.; Thompson, V.A. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends Cogn. Sci.* **2017**, *21*, 607–617. [CrossRef] [PubMed]
60. Inzlicht, M.; Bartholow, B.D.; Hirsh, J.B. Emotional foundations of cognitive control. *Trends Cogn. Sci.* **2015**, *19*, 126–132. [CrossRef] [PubMed]
61. VanderElst, D.; Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cogn. Syst. Res.* **2018**, *48*, 56–66. [CrossRef]
62. Jones, N.A.; Ross, H.; Lynam, T.; Perez, P.; Leitch, A. Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecol. Soc.* **2011**, *16*, 46. [CrossRef]
63. Kim, K.; Sano, M.; De Freitas, J.; Haber, N.; Yamins, D. Active World Model Learning with Progress Curiosity. In Proceedings of the 37th Conference on Machine Learning. PMLR 119, Vienna, Austria, 13–18 July 2020.
64. Yu, G. The Construction and Characteristics of the Theoretical System of Xi Jinping's View of History. *Philos. Study* **2020**, *10*, 503–510.
65. Allen, K.R.; Smith, K.A.; Tenenbaum, J.B. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 29302–29310. [CrossRef] [PubMed]
66. Harris, A.J.L.; Hahn, U.; Madsen, J.K.; Hsu, A.S. The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cogn. Sci.* **2016**, *40*, 1496–1533. [CrossRef]
67. Funamizu, A.; Kuhn, B.; Doya, K. Neural substrate of dynamic Bayesian inference in the cerebral cortex. *Nat. Neurosci.* **2016**, *19*, 1682–1689. [CrossRef]
68. Douven, I. The ecological rationality of explanatory reasoning. *Stud. Hist. Philos. Sci. Part A* **2020**, *79*, 1–14. [CrossRef] [PubMed]
69. Dabney, W.; Kurth-Nelson, Z.; Uchida, N.; Starkweather, C.K.; Hassabis, D.; Munos, R.; Botvinick, M. A distributional code for value in dopamine-based reinforcement learning. *Nat. Cell Biol.* **2020**, *577*, 671–675. [CrossRef]
70. Day, J.J.; Roitman, M.F.; Wightman, R.M.; Carelli, R.M. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* **2007**, *10*, 1020–1028. [CrossRef] [PubMed]
71. Biggio, B.; Roli, F. Wild Patterns: After the Rise of Adversarial Machine Learning. *Pattern Recognit.* **2018**, *84*, 317–331. [CrossRef]
72. Thomsen, K. Stupidity and the Ouroboros Model. In *Artificial General Intelligence, Lecture Notes in Computer Science*; Bach, J., Goertzel, B., Iklé, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7716, pp. 332–340.
73. Chen, Z.; Bei, Y.; Rudin, C. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.* **2020**, *2*, 1–11. [CrossRef]
74. Sanz, R.; López, I.; Rodríguez, M.; Hernández, C. Principles for consciousness in integrated cognitive control. *Neural Netw.* **2007**, *20*, 938–946. [CrossRef] [PubMed]
75. Herzog, M.H.; Drissi-Daoudi, L.; Doerig, A. All in Good Time: Long-lasting Postdictive Effects Discrete Perception. *Trends Cogn. Sci.* **2020**, *24*, 826–837. [CrossRef] [PubMed]
76. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus & Giroux: New York, NY, USA, 2011.
77. Thomsen, K. Sleep in the Ouroboros Model. In Proceedings of the TSC07 (Towards a Science of Consciousness), Budapest, Hungary, 23–27 July 2007.
78. Watkins, Y.; Kim, E.; Sornborger, A.; Kenyon, G.T. Using Sinusoidally-Modulated Noise as a Surrogate for Slow-Wave Sleep to Accomplish Stable Unsupervised Dictionary Learning in a Spike-Based Sparse, Coding Model. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1482–1487. [CrossRef]
79. Hoel, E. The Overfitted Brain: Dreams evolved to assist generalization. *arXiv* **2020**, arXiv:2007.09560.
80. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Matthew, D.; et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv* **2020**, arXiv:2011.03395.
81. Jumper, J.; Evans, R.; Prizel, A.; Green, T.; Figurnov, M.; Tunyasuvunakool, K.; Ronneberger, O.; Beates, R.; Žídek, A.; Bridgland, A.; et al. High Accuracy Protein Structure Prediction Using Deep Learning. In *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*; 30 November–4 December 2020; Available online: https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf (accessed on 30 November 2020).

82. Heaven, W.D. DeepMind's Protein-Folding AI Has Solved a 50-Year-Old Grand Challenge of Biology. Available online: https://www.technologyreview.com/2020/11/30/1012712/deepmind-protein-folding-ai-solved-biology-science-drugs-disease/ (accessed on 30 November 2020).
83. Kruijne, W.; Bohte, S.M.; Roelfsema, P.R.; Olivers, C.N.L. Flexible Working memory through selective gating and attentional tagging. *Neural Comput.* **2021**, *33*, 1–40. [CrossRef] [PubMed]