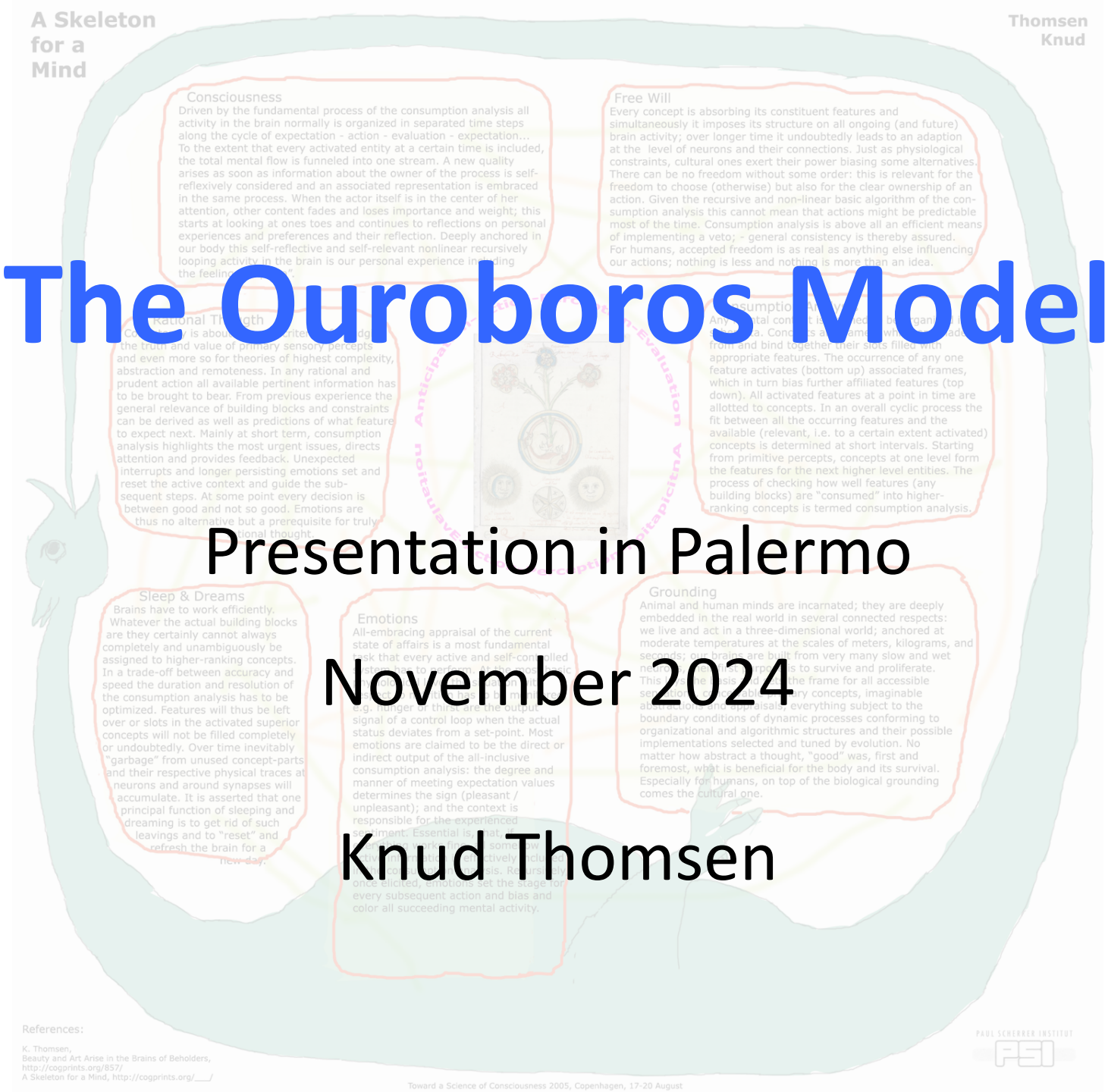


The Ouroboros Model

Presentation in Palermo

November 2024

Knud Thomsen

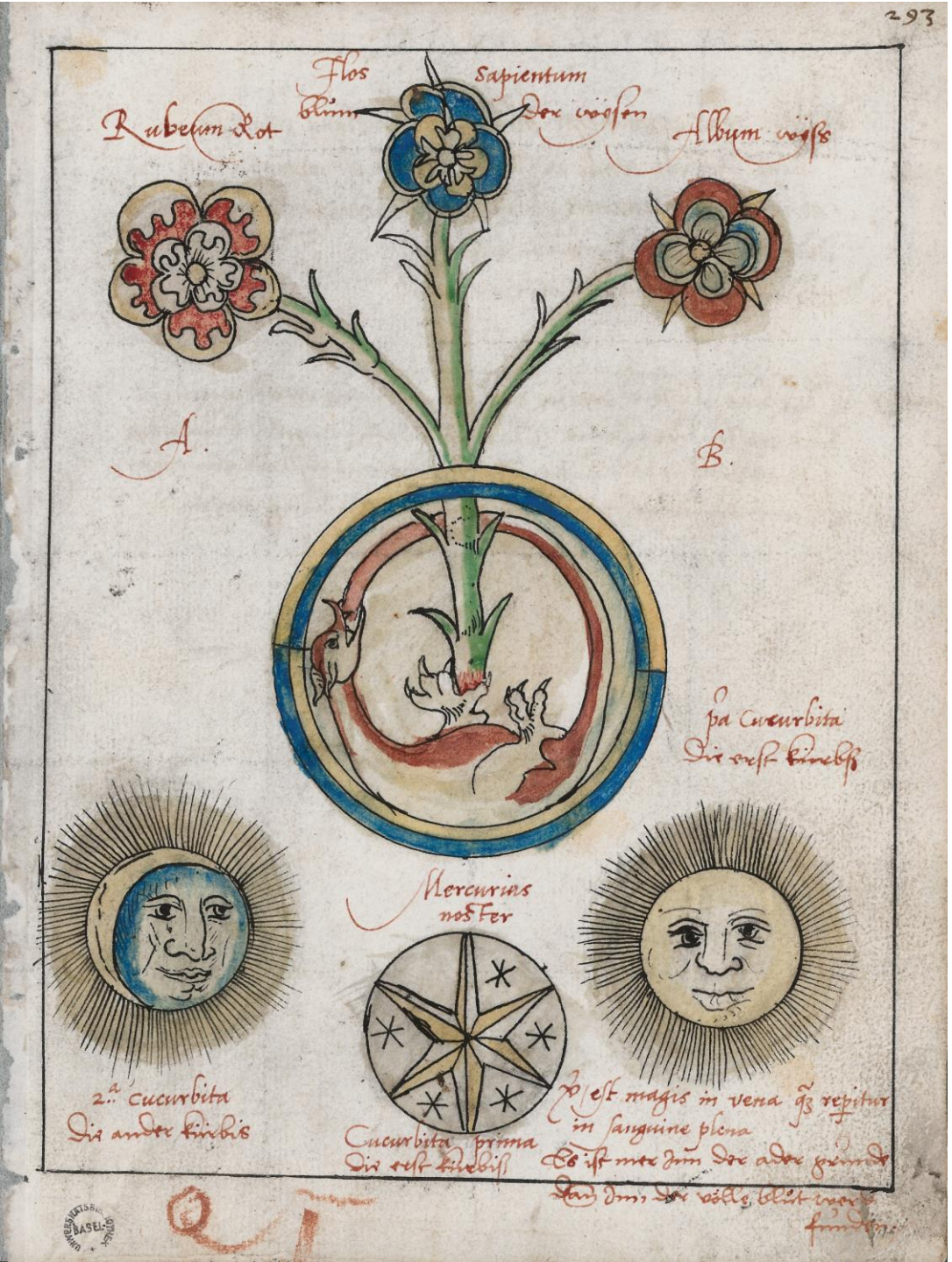
Approach of the Ouroboros Model:

- **Taking fresh look at the BIG PICTURE**
- **Avoid getting bogged down in details**
- **Observing different levels of analysis:**
  - functional ('computational')
  - algorithmic
  - implementational (D. Marr)
- **Grounding provides the basis (embodied & embedded)**
  (S. Harnad, L. Barsalou)

Self-reflectively and –consistently: **1st a coarse encompassing sketch**

OM a Palermo, content:

- Background
  - Name / Symbol
  - Scientific anchestors / related conceptualizations
- Basic Building Blocks & Principal Process
  - Schemata in memory
  - Consumption analysis
- Sheds light inter alia on:
  - Grounding
  - Attention
  - Emotion
  - Language & Communication
  - Sleep
  - (Self-)directed growth (..evolution..)
  - Consciousness
  - Free Will
  - Ethics
  - Neural implementation / known facts
- Relations to current research in AI
  - Predictive Coding
  - Large Language Models
    - Strengths & Deficiences
    - Recommendations for enhancement
  - Autonomy
  - Robotics !
- Real & wrong Limitations
  - Purportedly circular
    - Time !
  - Blending analog & digital
    - Control
    - Ontology / Emergence
  - In conceptual stage only
  - Limited ressources
    - but free ☺
  - Stupidity
- Future Development
  - Formalization(s)
  - Artifical Implementation(s)
  - Refinement(s)
  - Testing (!)
- Interpretation of Quantum Mechanics …
- Invitation for Cooperations
- Many thanks for the **I**nvitation
  & your **A**ttention

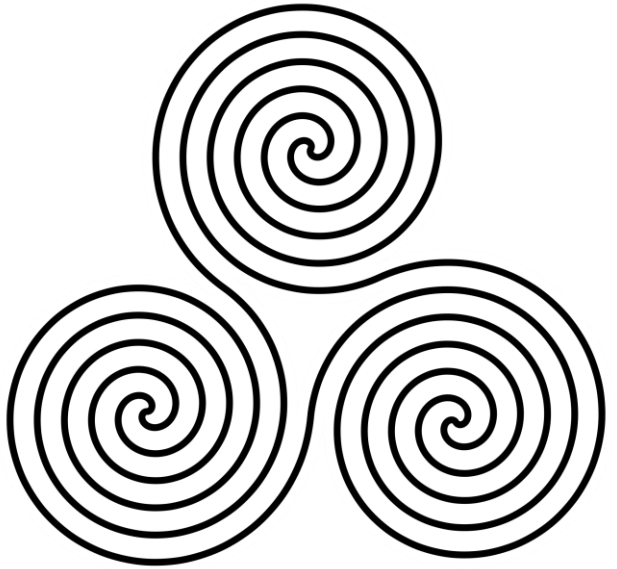


The Ouroboros is an ancient symbol depicting a serpent or dragon eating its own tail, it is a symbol for eternal cyclic renewal or a cycle of life, death and rebirth in many different cultures over eons, some examples:

Painting in an alchemistic manuscript ("Alchimistische Handschrift 1550“, Universitätsbibliothek Basel, Sign. Mscr. L IV 1, scan "uc172.jpg")

on one of the shrines enclosing the sarcophagus of Tutankhamun (wikipedia)

yin-yang in Chinese philosophy describes an opposite but interconnected, self-perpetuating cycle



Neolithic 5,000 year-old triskelion on an orthostat at Newgrange (Wikipedia)

Otto Selz, 1913, 1922

Schematic Anticipation

Cycle of Perception

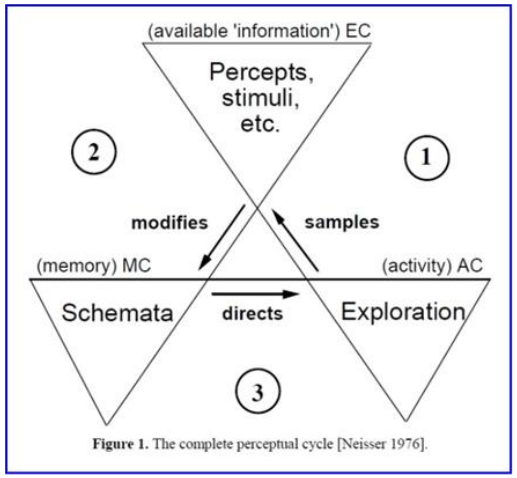
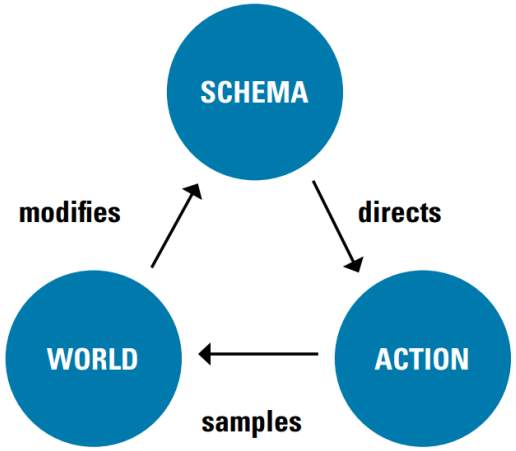


Figure 1. The complete perceptual cycle [Neisser 1976].

Neisser, U., **1976**. Cognition and Reality. S. Francisco, W.H. Freeman and Company

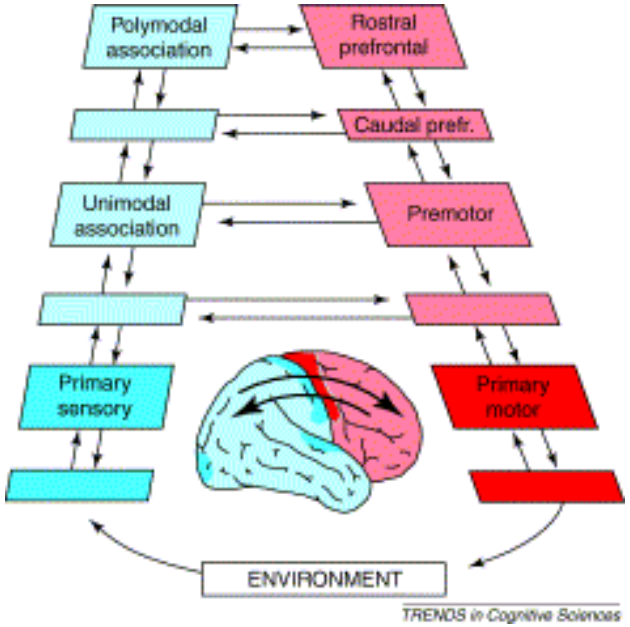


e.g. Katie Plant: HindSight 34 | WINTER **2022-2023**

Perception - Action Cycle

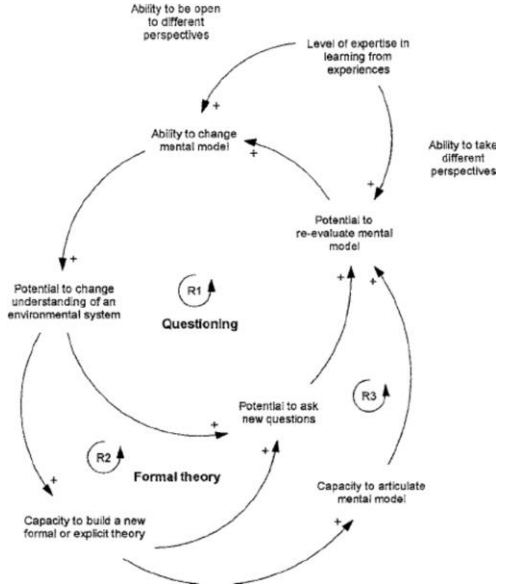
e.g., "All goal directed behavior is performed within the broad context of the perception-action cycle, which is grounded on a basic biological principle: the circular cybernetic flow of cognitive information that links the organism to its environment."

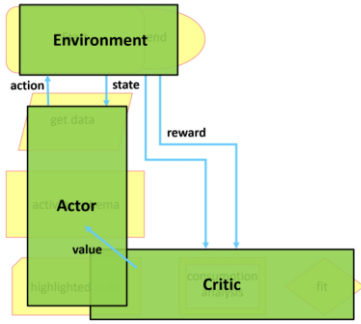
— Joaquín M. Fuster, Physiology of Executive Functions: The Perception-Action Cycle, **2002**



Mental Models, Actor/ Critic alg.

Many versions stress internal representations and feedback / reinforcement llm





e.g., Fazey, Ioan & Fazey, John & Salisbury, Janet & Lindenmayer, David & Dovers, Stephen. (**2006**). The nature and role of experiential knowledge for environmental conservation. Environmental Conservation. 33. 10.1017/S037689290600275X.

A Skeleton for a Mind

Thomsen
Knud

Consciousness

Driven by the fundamental process of the consumption analysis all activity in the brain normally is organized in separated time steps along the cycle of expectation - action - evaluation - expectation... To the extent that every activated entity at a certain time is included, the total mental flow is funneled into one stream. A new quality arises as soon as information about the owner of the process is self-reflexively considered and an associated representation is embraced in the same process. When the actor itself is in the center of her attention, other content fades and loses importance and weight; this starts at looking at ones toes and continues to reflections on personal experiences and preferences and their reflection. Deeply anchored in our body this self-reflective and self-relevant nonlinear recursively looping activity in the brain is our personal experience including the feeling of "qualia".

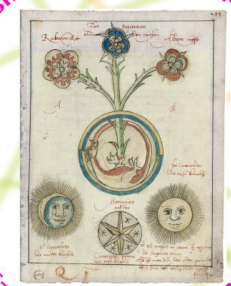
Free Will

Every concept is absorbing its constituent features and simultaneously it imposes its structure on all ongoing (and future) brain activity; over longer time it undoubtedly leads to an adaption at the level of neurons and their connections. Just as physiological constraints, cultural ones exert their power biasing some alternatives. There can be no freedom without some order: this is relevant for the freedom to choose (otherwise) but also for the clear ownership of an action. Given the recursive and non-linear basic algorithm of the consumption analysis this cannot mean that actions might be predictable most of the time. Consumption analysis is above all an efficient means of implementing a veto; - general consistency is thereby assured. For humans, accepted freedom is as real as anything else influencing our actions; nothing is less and nothing is more than an idea.

Rational Thougth

Consistency is about the only criterion for judging the truth and value of primary sensory percepts and even more so for theories of highest complexity, abstraction and remoteness. In any rational and prudent action all available pertinent information has to be brought to bear. From previous experience the general relevance of building blocks and constraints can be derived as well as predictions of what feature to expect next. Mainly at short term, consumption analysis highlights the most urgent issues, directs attention and provides feedback. Unexpected interrupts and longer persisting emotions set and reset the active context and guide the subsequent steps. At some point every decision is between good and not so good. Emotions are thus no alternative but a prerequisite for truly rational thought.

Anticipation-Action-Perception-Evaluation
noitaluvE-Action-Perception-noitapicitnA



Consumption Analysis

Any mental content is claimed to be organized into schemata. Concepts are frames, which are made up from and bind together their slots filled with appropriate features. The occurrence of any one feature activates (bottom up) associated frames, which in turn bias further affiliated features (top down). All activated features at a point in time are allotted to concepts. In an overall cyclic process the fit between all the occurring features and the available (relevant, i.e. to a certain extent activated) concepts is determined at short intervals. Starting from primitive percepts, concepts at one level form the features for the next higher level entities. The process of checking how well features (any building blocks) are "consumed" into higher-ranking concepts is termed consumption analysis.

Sleep & Dreams

Brains have to work efficiently. Whatever the actual building blocks are they certainly cannot always completely and unambiguously be assigned to higher-ranking concepts. In a trade-off between accuracy and speed the duration and resolution of the consumption analysis has to be optimized. Features will thus be left over or slots in the activated superior concepts will not be filled completely or undoubtedly. Over time inevitably "garbage" from unused concept-parts and their respective physical traces at neurons and around synapses will accumulate. It is asserted that one principal function of sleeping and dreaming is to get rid of such leavings and to "reset" and refresh the brain for a new day.

Emotions

All-embracing appraisal of the current state of affairs is a most fundamental task that every active and self-controlled system has to perform. At the most basic physiological level the situation with respect to nutrition has to be monitored, e.g. hunger or thirst are the output signal of a control loop when the actual status deviates from a set-point. Most emotions are claimed to be the direct or indirect output of the all-inclusive consumption analysis: the degree and manner of meeting expectation values determines the sign (pleasant / unpleasant); and the context is responsible for the experienced sentiment. Essential is, that, if everything works fine, all somehow active information is effectively included in the consumption analysis. Recursively, once elicited, emotions set the stage for every subsequent action and bias and color all succeeding mental activity.

Grounding

Animal and human minds are incarnated; they are deeply embedded in the real world in several connected respects: we live and act in a three-dimensional world; anchored at moderate temperatures at the scales of meters, kilograms, and seconds; our brains are built from very many slow and wet neurons, their first purpose is to survive and proliferate. This lays the basis and sets the frame for all accessible sensations, conceivable primary concepts, imaginable abstractions and appraisals, everything subject to the boundary conditions of dynamic processes conforming to organizational and algorithmic structures and their possible implementations selected and tuned by evolution. No matter how abstract a thought, "good" was, first and foremost, what is beneficial for the body and its survival. Especially for humans, on top of the biological grounding comes the cultural one.

References:

K. Thomsen,
Beauty and Art Arise in the Brains of Beholders,
http://cogprints.org/857/
A Skeleton for a Mind, http://cogprints.org/___/

Toward a Science of Consciousness 2005, Copenhagen, 17-20 August

PAUL SCHERRER INSTITUT
PSI

The Ouroboros Model, Main Features, I

- **All concepts are stored in** (non-strict hierarchy of) **SCHEMATA,**
  i.e., frames.. with features linked together
- **Activation of any one feature biases**
  associated ones, i.e. provokes expectations
- **A monitoring loop "CONSUMPTION ANALYSIS"**
  checks the fulfillment of these expectations
- **Expectations can be violated, met, exceeded;**
  lasting feedback = EMOTIONS
- **short-term Feedback (consistency)**
  directs Flow of Activity

The Ourobors Model, Main Features, II

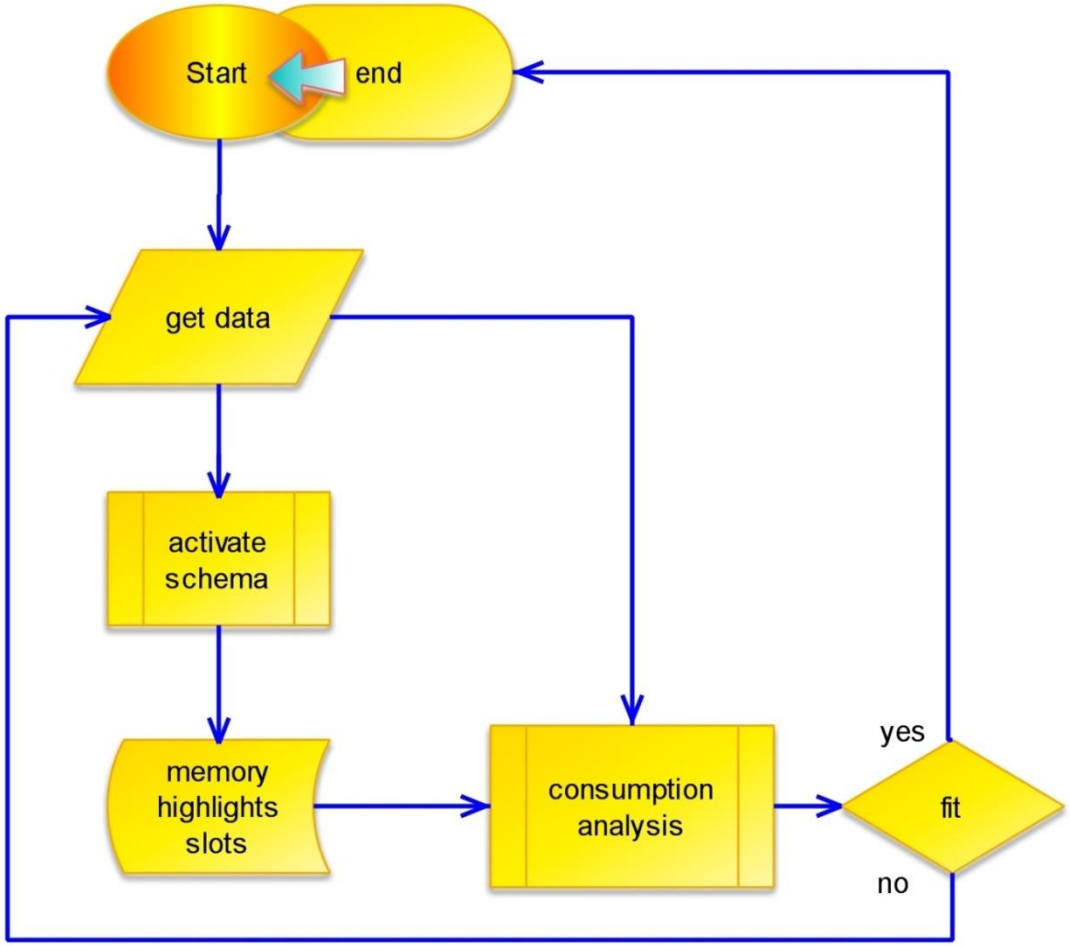
The Ouroboros Model proposes a dynamic and self-organizing <u>cognitive architecture</u> with the backbone of one fundamental recursive process /algorithm in the form of an extended action / perception cycle :

....anticipation,

- **action / perception,**
- **<u>evaluation</u>,**
- **anticipation,...**

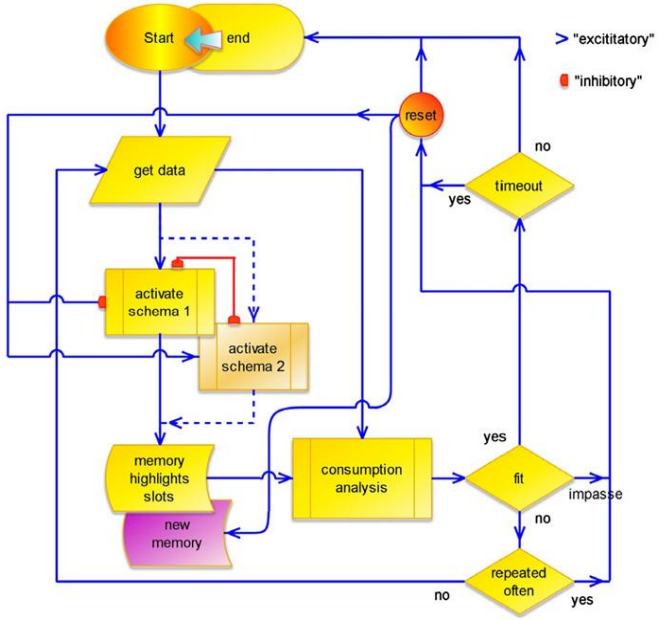
in 1 brain, some principal periodicity of all perceptual and cognitive activity ensues

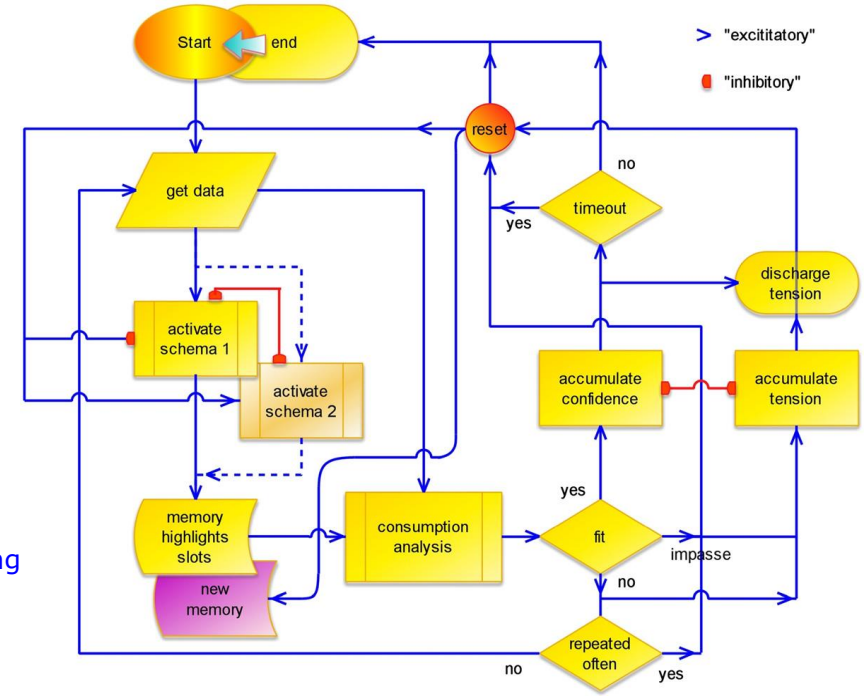
not restricted to single brains but applicable to (rational) groups, societies, AI, ...



Basic loop augmented with mechanisms for flexible schema selection and the recording of likely useful new memories

active
Process
in time !

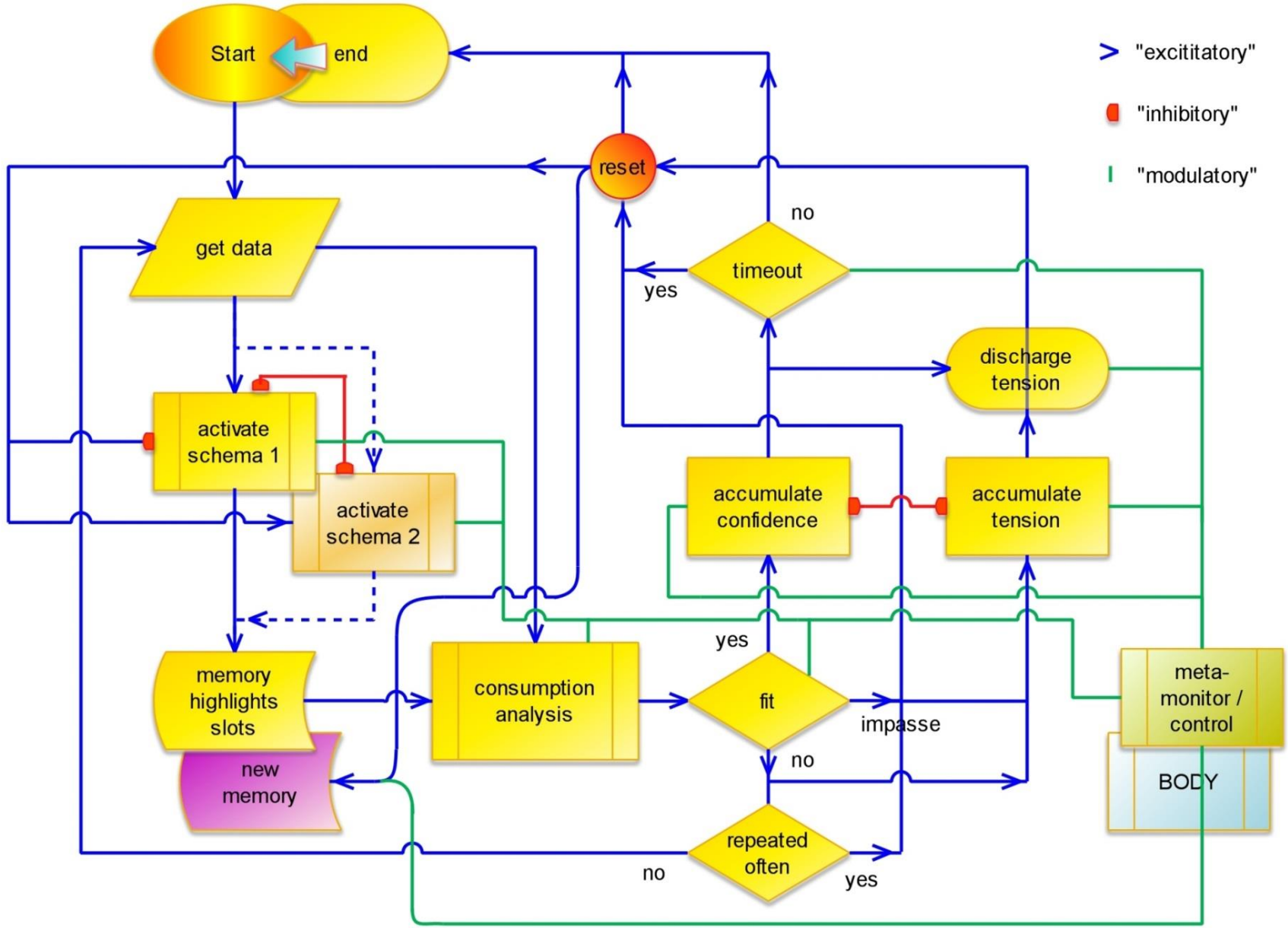




Basic Loop Structure of the Ouroboros Model

With self-monitoring by keeping track of the current performance

K. Thomsen, The Ouroboros Model in the light of venerable criteria. Neurocomputing 74, 121-128, 2010.

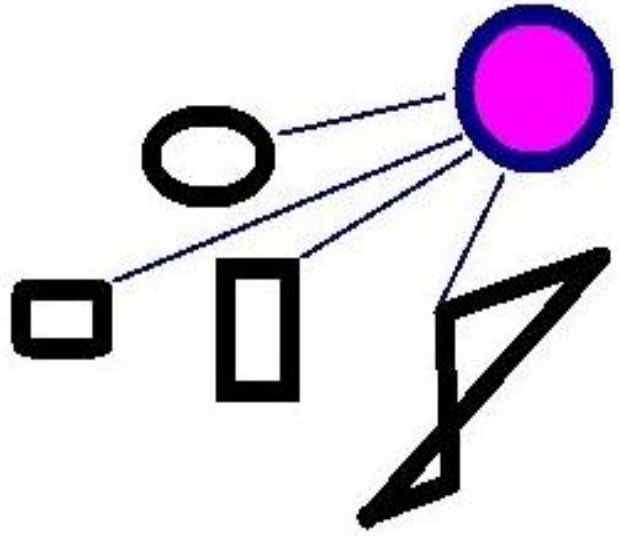


active
Process
in time !

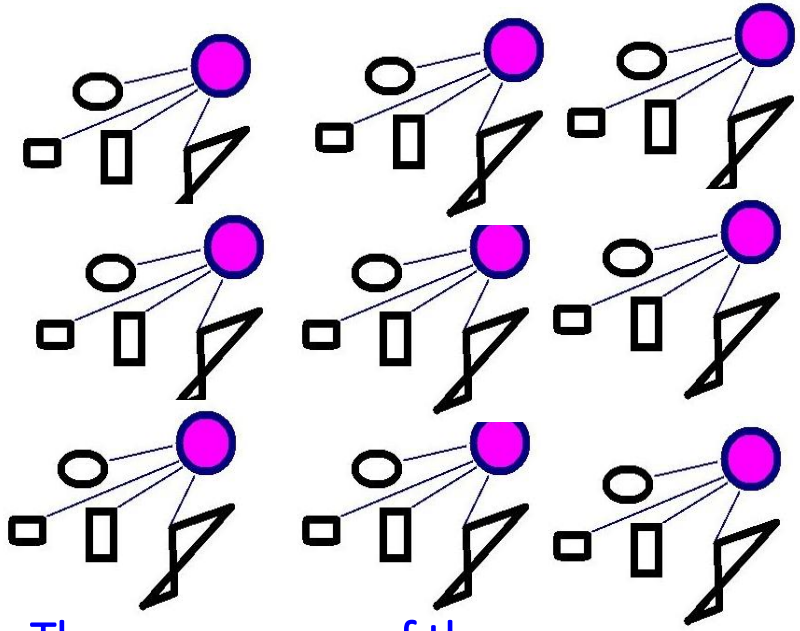
Loop augmented with mechanisms for flexible schema selection and the recording of likely useful new memories including self (bodily) monitoring by keeping track of the current performance monitoring (2nd order), which is employed for self-steering; the flow of activity and the outcome of the different processing stages exert mutual modulations

Compartmentalization is essential:

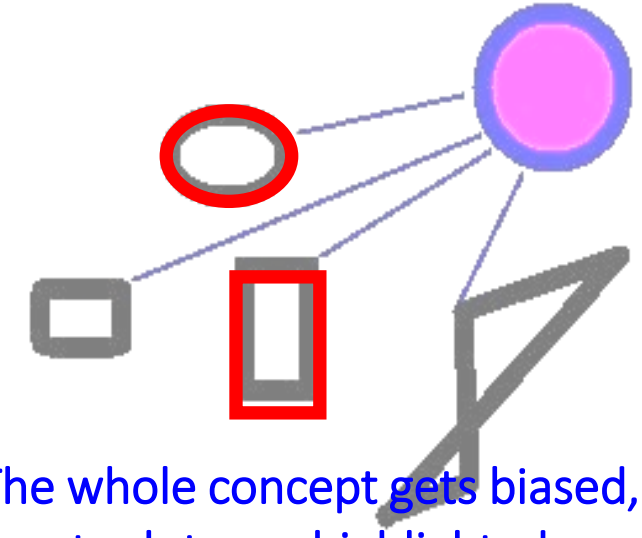
- Actors can only process a certain amount of input; especially so in a limited time span
- Compartmentalization provides the very basis for effective monitoring and meaningful error checking
- Data become “information” only in the interplay between source and receiver
- what definitively does not belong to a given schema is information, too
- the absence of an expected feature is probably important information
- in 1 (healthy) actor there is some minimum coherence; fragmentation ends at the bodily level
- indiscriminate associationism would be some type of “opposite”
  (this approach appears to be a problem for Large Language Models)



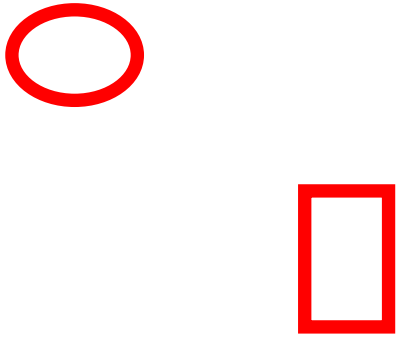
One concept with features



There are many of them

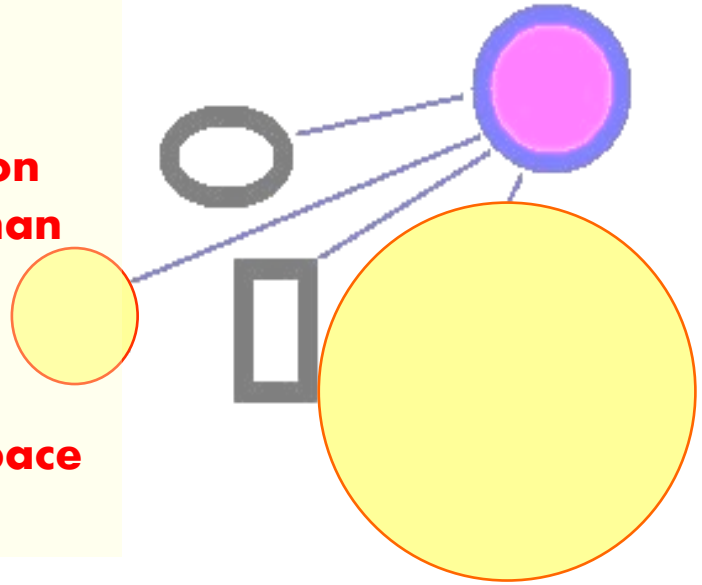


The whole concept gets biased,
Empty slots are highlighted

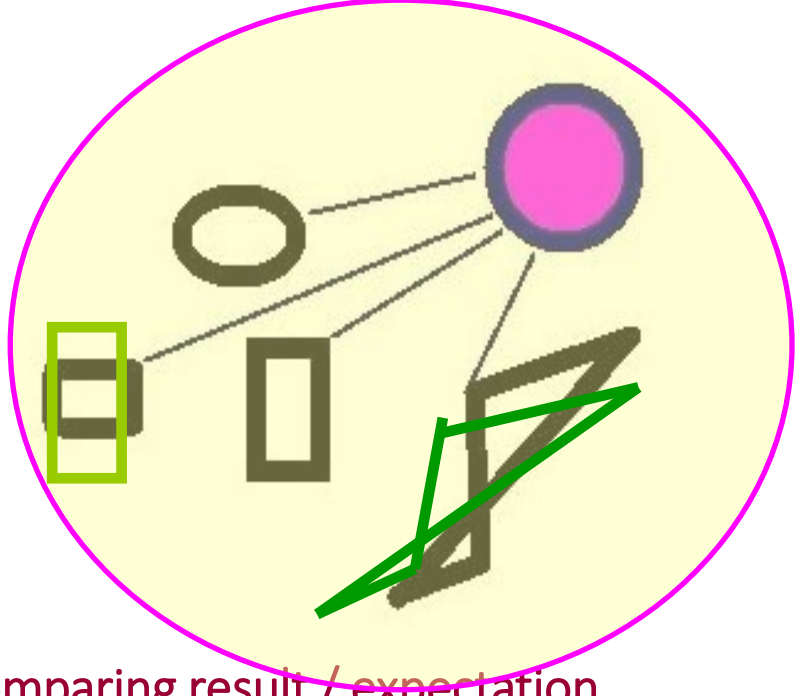


Input excites some features

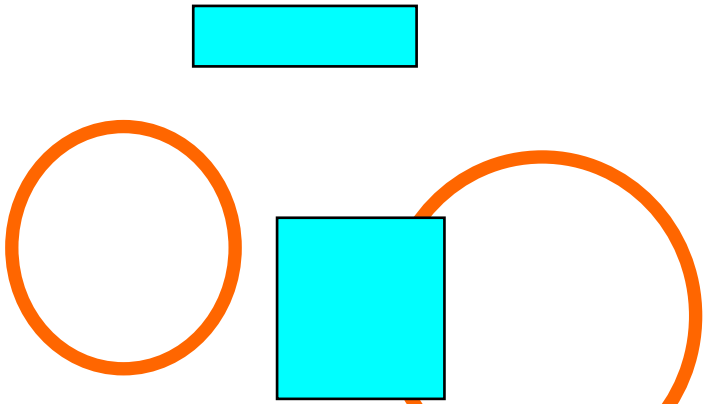
This provides a much better rationale for global activation (on demand) than what is postulated by standard Global Workspace theories



Checking for fitting input



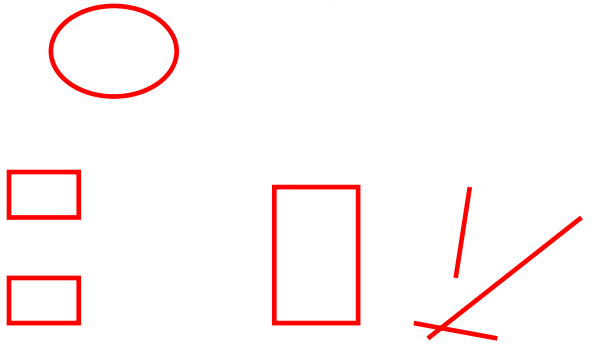
Comparing result / ~~expectation~~



Include highlighted features in new input

continue

EMOTIONS



Pointing out discrepancies

Representations are being built up auto-catalytically in a self-steered manner

3 **Regimes**, i.e. Outcomes of the Consumption Analysis:

„attentive“

- **Everything fits *(=> new memory entry)***
- **A good portion of all (relevant) activity can be satisfactorily "consumed"**
- **Nothing fits *(=> new memory entry)*** „attentive“

„pre-attentive“

In addition are associations and categorizations gradually distilled from the statistics of co-occurrences (with abstraction as a special case)

All of this learning is similar for diverse levels of abstraction / complexity and beyond the control of the initial designer !

K. Thomsen, Concept Formation in the Ouroboros Model, Third Conference on Artificial Gerneral Intelligence, Lugano, Switzerland, March 5-8, 2010.

Grounding is essential:

- (natural) Actors are **embodied**, and this connects even the most abstract concepts to the outside reality
- When a schema is first established, all activity in the brain at that moment is tied together, in particular, also **bodily signals**
- Except for the most abstract, e.g., mathematical concepts, all schemata thus include some features and signals pertaining to the **bodily reality** of an agent at the points in time of recording as well as use
- The growth and the addition of all new schemata is **colored** and **constrained** by what is available from before
- This makes all mental content primarily "**private**", i.e., as private as her body, understandable and accessible in full only by the individuum herself

In real life, i.e., in direct action in the world, as well as in discourse (and thought):

Attention as focused anticipation
inside the general accessible frame:

Consumption Analysis highlights slots of the activated schemata and thus directs Attention to the most urgent issues

What is directly connected in an activated schema determines first its Relevance

The general level of match together with an inherited Importance including indirect links contribute further to the relevance of an attribute


Pattern completion forms part of the very basis of our survival
NATIONAL GEOGRAPHIC
Photograph by Steve Winter
KAZIRANGA. AUGUST 2010
© COPYRIGHT NATIONAL GEOGRAPHIC SOCIETY. ALL RIGHTS RESERVED.

This account is claimed to explain these suggested effects & mechanisms:

- **Location priming**
- **Semantic enhancement**
- **Masking due to overlapping features**
- **Negative congruency effects** (reaction times)
- **Attentional blink**

  (resource depletion,
  processing bottleneck,
  temporary loss of control)

- **SOS (satisfaction of search** = difficulty to detect 2nd target**)**

Thomsen, K. The Ouroboros Model, Selected Facets. In: C. Hernández et al. (eds.) From Brains to Systems. New York Dordrecht Heidelberg London: Springer, 239-250, 2011.

The potential mental processing power of an agent is ground-laid in **knowledge**, i.e. the number, complexity, elaboration, and linkage of the concepts at her disposition

Schemata, their number of slots, the level of detail, the depth and extent of hierarchies, degree of connection and interdependence of the building blocks, and the width,
i.e. the extent of main schemata and their total coverage from the grounding level to the most abstract summits, determine what can be thought of efficiently

quasi-global Adequacy, Coherence & Consistency are crucial

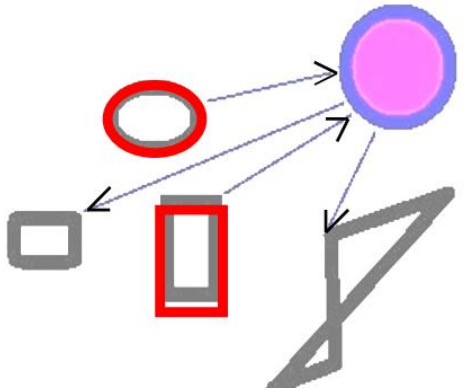
Sheer **performance** at a single point of time arises as a result of the optimum interplay between these structured data and the effective execution of all the described processing steps, in particular, self-referential consumption analysis and the continual use of its results

..there are constraints from implementational details

Problem Solving:

The Algorithm underlying the Ouroboros Model can be seen as a specific version of pattern matching and constraint satisfaction.

It can also be understood as an extension of production systems, any feature can serve as a starting point for activating a schema / an action.



Combining this way prior Knowledge with new data yields Bayesian Performance.

The Ouroboros Model is claimed to offer a meaningful and efficient approximation to Bayesian processing:

- **Priors pertaining to specific situations / tasks are laid down in the collection of distinct schemata**

- **Spreading information & consumption analysis effectively yield Posteriors, including an (error) signal, which is harnessed for continued optimized processing and action**

- **By limiting the extent of the applicable search space to activated schemata and performing the main steps in parallel, the well-known concerns of Bayesian processing being NP-hard are rendered invalid, while foregoing unnecessary precision further helps**

Schemata containing an “action-slot” most likely also comprise associated “consequence-slots”, and actions most often are chosen for exactly these with the aim of effecting such intended results (..embodied cognition, affordances)

Imperfect matches / slots staying open for longer time arouse curiosity and intentions

Consequences of an action are classified as unintended when a discrepancy involving an “outcome-slot” turns out higher than a preset threshold (some type of causation involved)

Unintended results probably occur especially in situations when there simply was not enough stored material or time available for an agent (or dedication and care invested) to become consciously aware of all potential outcomes.

Important, anything sufficiently novel at the first encounter kindles the same negative feeling of an unsatisfactory match, as just any plain error

Serendipity (space research parlance)

Deviations from the known and expected will especially surface at the boundaries of the established experience

New gains and advances can even be aimed for without knowing any details

As an example from scientific instrumentation:
it can optimistically and also rightly be anticipated that a new research satellite with greatly enhanced capabilities compared to previous missions will spot something interesting, which just could not be observed before
(“partly intended”)

Linking together sequences of distinct and separated stepping stones, decisive for the overall coherence of activity, according to the Ouroboros Model, is accomplished by diverse **processes at different timescales** and levels of detail

- At extended timescales, personal (bodily) features, emotions and moods ensure some coherence and continuity of perceptions as well as for the actions of an agent
- High-level schemata, e.g., for goals, provide a common frame, allowing for rather diverse possible plans and actions
- Over short to medium durations mainly in the order of seconds or minutes, the flow of action according to the Ouroboros Model is mediated by shared constituents, i.e., common attributes and ‘connecting-’features, of thus concatenated, otherwise distinct, schemata
- Most schemata will include some form of reference to time, e.g., start- / stop-conditions and attributes for durations or transients
- Especially for actions with a short latency, both for bodily movements and also more abstract cognitive processes, it is hypothesized that representations pertaining to intermediate values are calculated from more directly accessible distinct neighboring reference points by means of some type of averaging and interpolation

Experience and consciousness according to the Ouroboros Model evolve “in spirals”, intrinsically relying on progress in time; in the same vein, time is pushed forward and thus “produced”

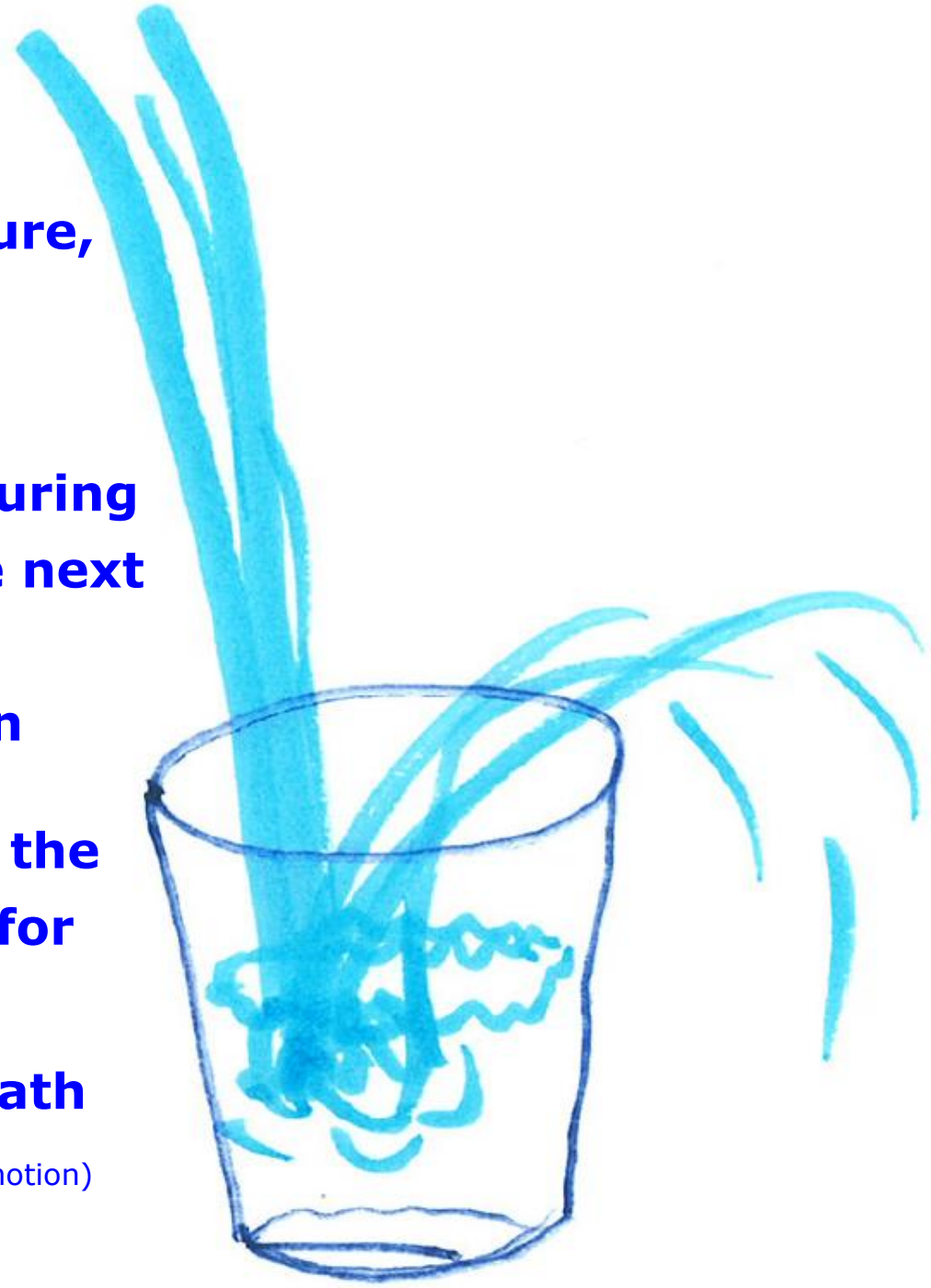
Emotions derived from evaluated anticipations:

When filling in anything into whatever structure, monitoring how it fits appears as the most obvious thing to do:

Emotional feedback allows for optimization during the process, it sets the optimum stage for the next actions, i.e., a complementary and longer bias than the short-term direction of attention

Feeling as a measure of the goodness of fit is the other side of the coin, a madatory ingredient for rational behavior !

- this does not preclude a second feed-back path via (sensing) bodily reactions (James-Lange theory of emotion)



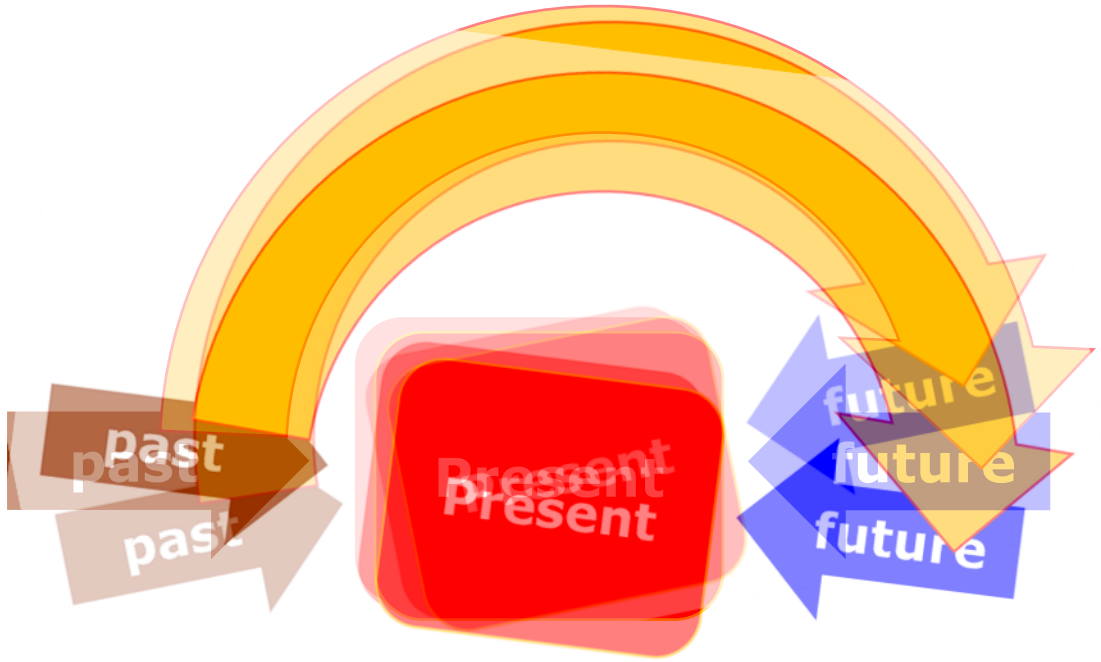
Interplay between old "inherited" emotional tags and current actual feed-back "feeling"

		Freshly derived Feeling Component		
		positive	negative	nondescript
Inherited Emotional Tag(s)	positive			
	negative			
	mixed			

The actual assessment component is often much more important than inherited tags;
this might even lead to **akrasia**, i.e., acting "against one's better judgment".

It is even more complicated as expectations for future(s), based on memory, are decisive !
Related: charm of the forbidden

use the past to anticipate the futures



„use the past
to understand
the present“

„use expectations
concerning the future
to inform the present“

and the whole picture to act in the present

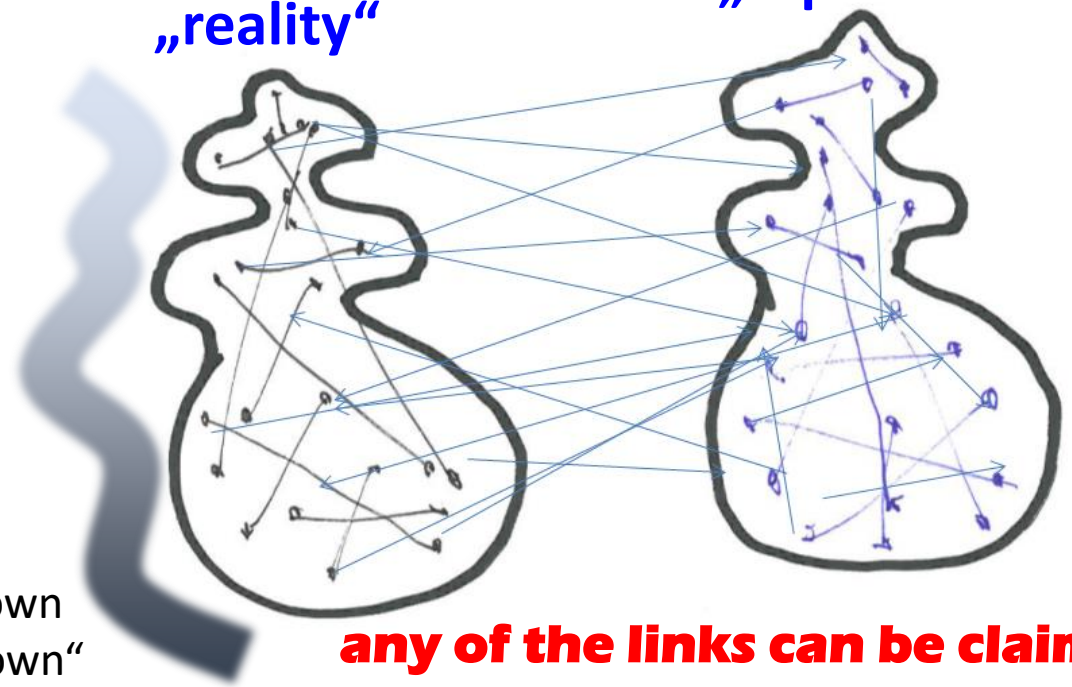
Further complication: short-term interests :: long-term strategies

Each part of a schema, in particular, external structures (their percepts!) e.g. affordances can suffice to activate the full script (its representation!)

in tight interplay:

„reality“

„representation“ – actor!



„mind“

„life“

„things“

B. Spinoza
I. Kant

N. Hartmann
H. Haken

K. Lorenz, P.W. Anderson
F. Rosas
J. Crutchfield
E. Hoel

„unknown unknown“

(particularly personal)

any of the links can be claimed as fundamental

all of this is in some sense „individual“

Multi-compound structures & representations are essential,
„models“ i.e. schemata determine what „exists“ (Immanuel Kant)

The Ouroboros Model advocats some „enlightened realism“

Cognition has a strong serial component

there appear to be serial canonical scan-paths for persons, faces (default successions of searched features)

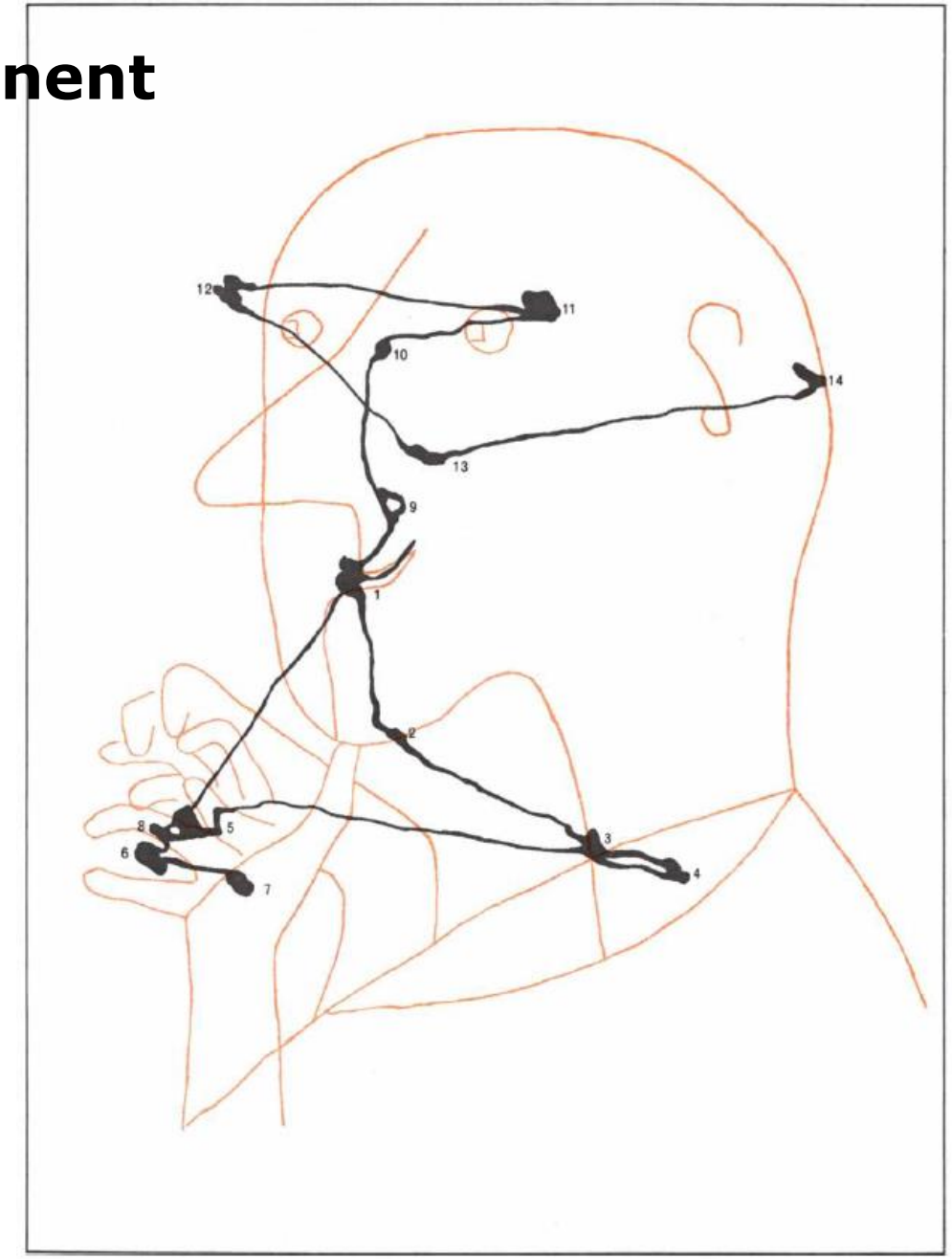
D. Noton and L. Stark, Scient. Am. 1971

embodied, one cannot move one´s hand up and down at the same time;

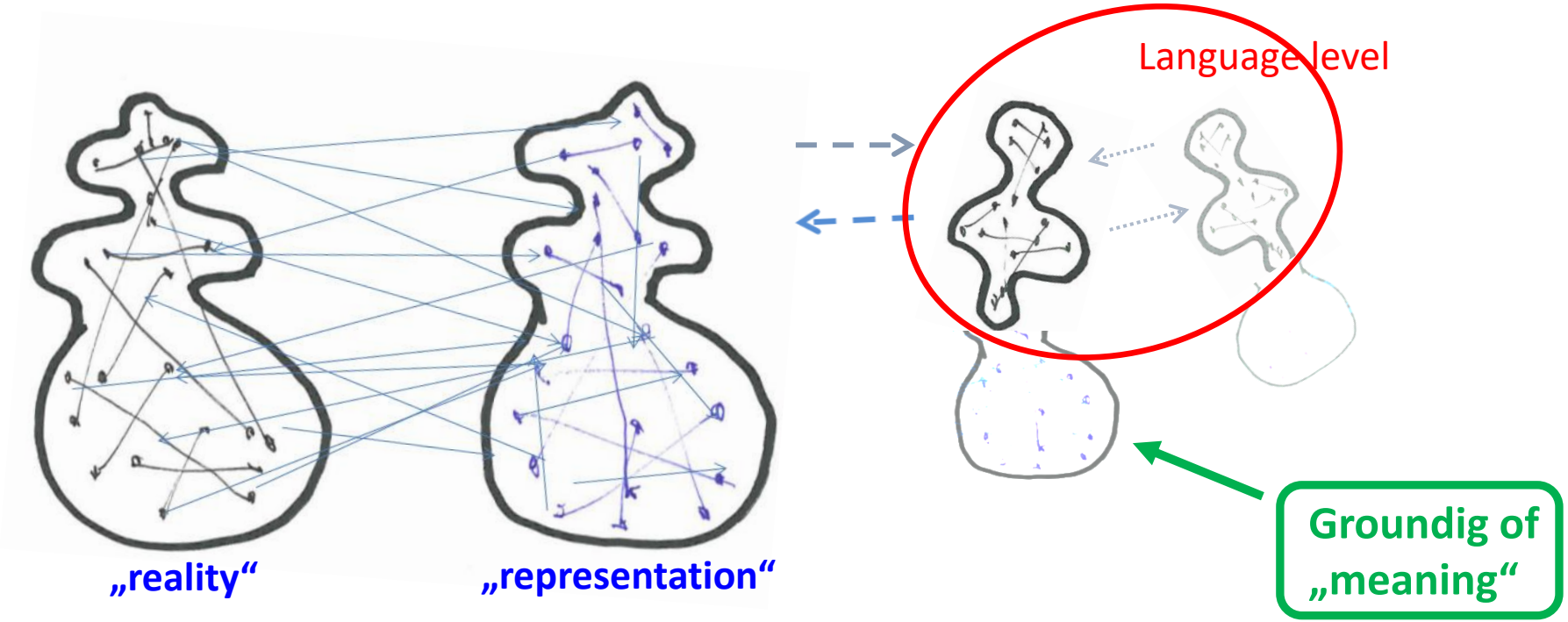
some stage/phase of planning and conflict resolution is needed,
overall seriality ensues

R. Cotterill, Progr. Neurobiol. 2001

Language 0



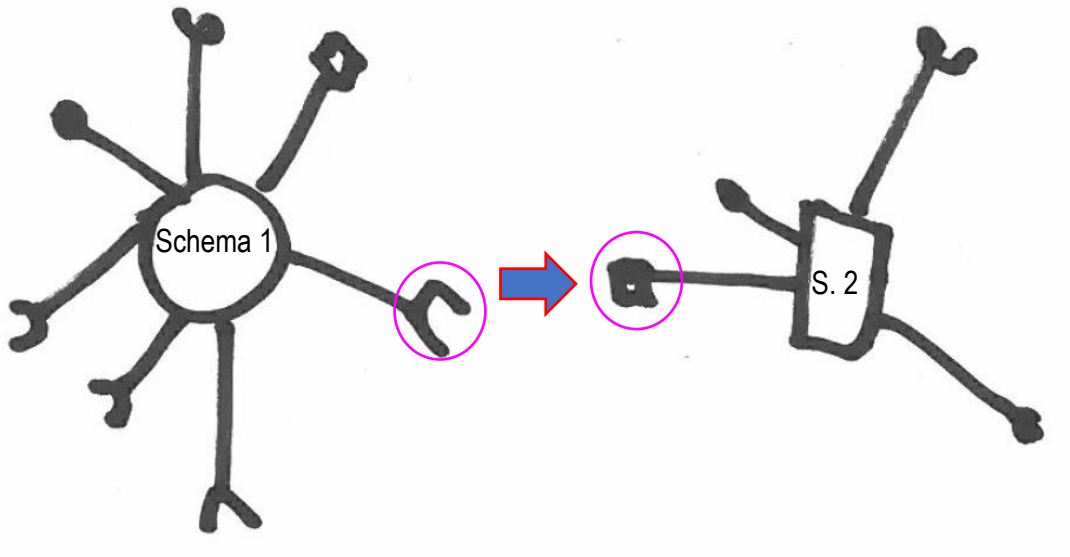
Representations: Language I



Representations, e.g. words (or gestures) vary with respect to their grounding in the real bodily world

with transfer to language and from agent to agent, there occur unavoidably shifts of emphasis and losses (of direct links to the real bodily world)

‚chemical‘ Syntax: Language II



The slots, which constitute schemata, at the same time define possible overlap between concepts and points of their connection; one primes the next

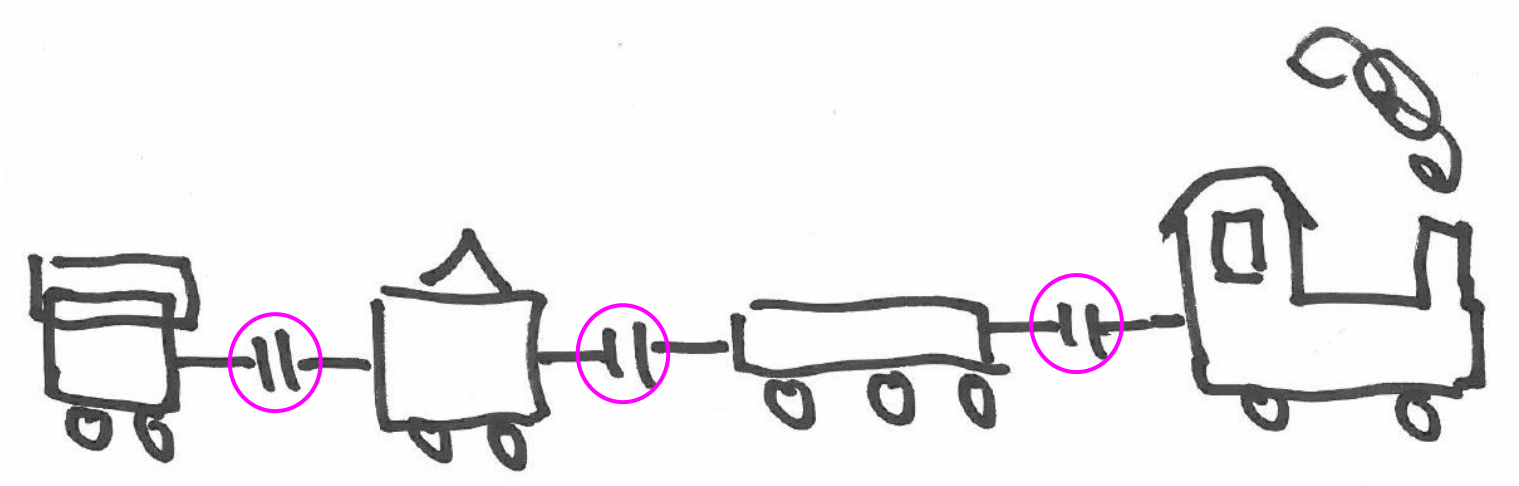
according to a «chemical» model of syntax, the Ouroboros Model holds that regularity in language (grammar) arises from wide-spread shared abstract features

(Grammatical) rules then are distilled from the repeated practice and abstracted

The process is very the same at the level of consonants / vowels, words, and concepts (on different time scales)

Formulation: Language III

In straight forward extension of serial processing at the component level, schemata, for which there is a sign (e.g. word) are linked in a series

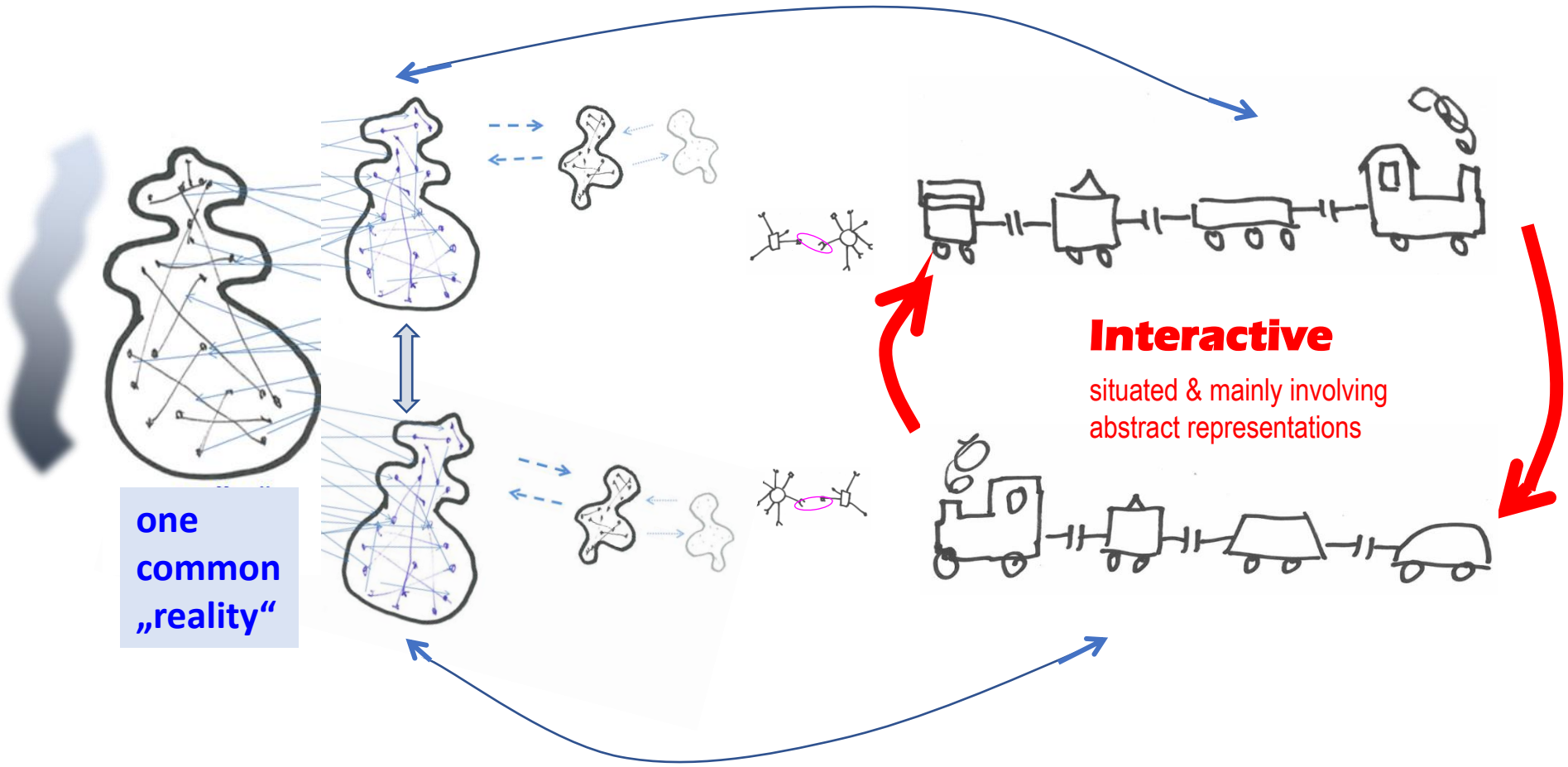


This addresses continuous and discrete facets of language

Only (!) stringing words for schemata following rules, i.e. distilled patterns of common use, leads to (reliably) understandable linear encodings of content

One needs to stick to language rules in order to express oneself freely

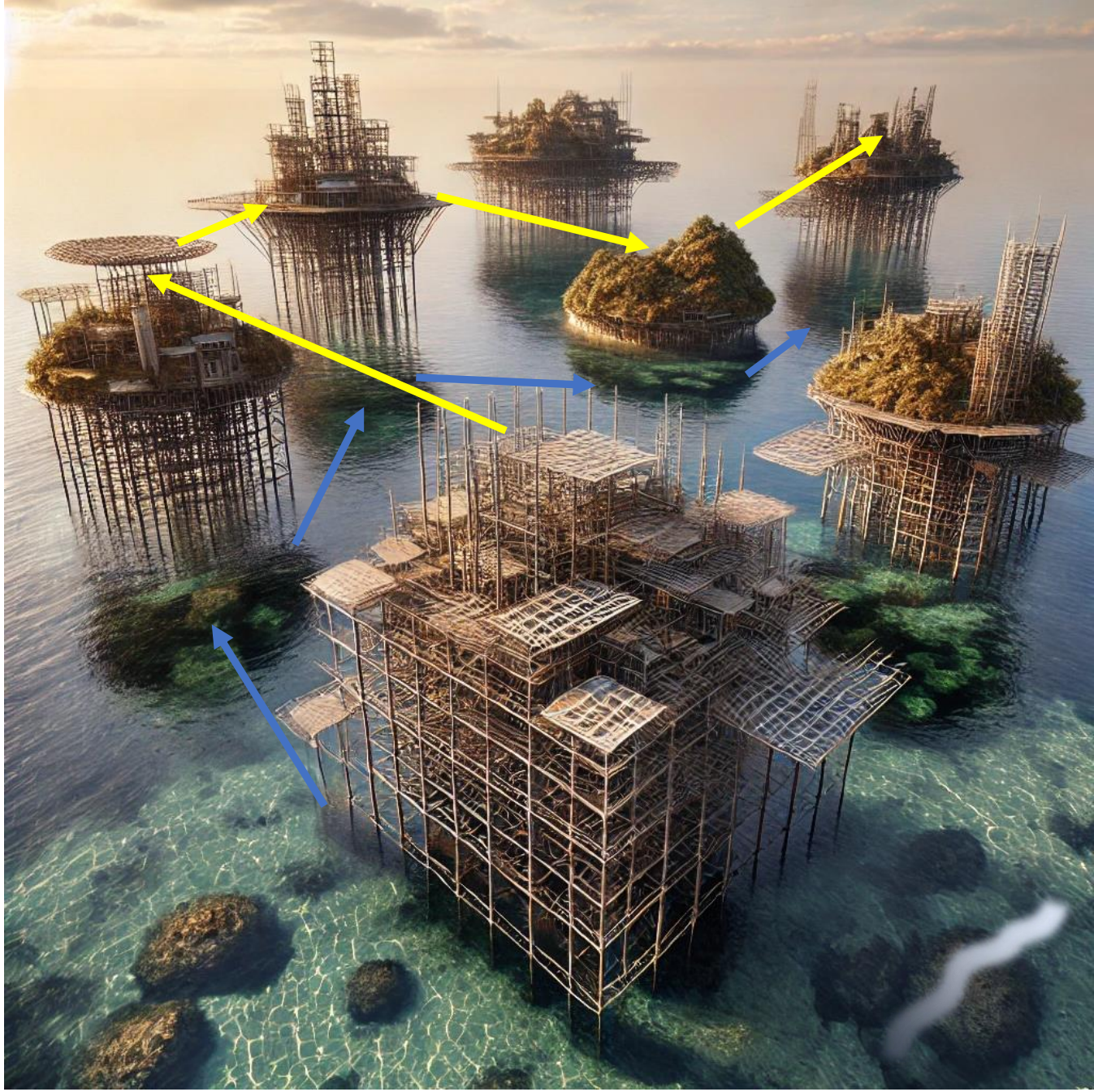
Communication: Language IV -- Communication



All discourse relies on some shared content, it cannot work without a minumum of common reference (and grounding)

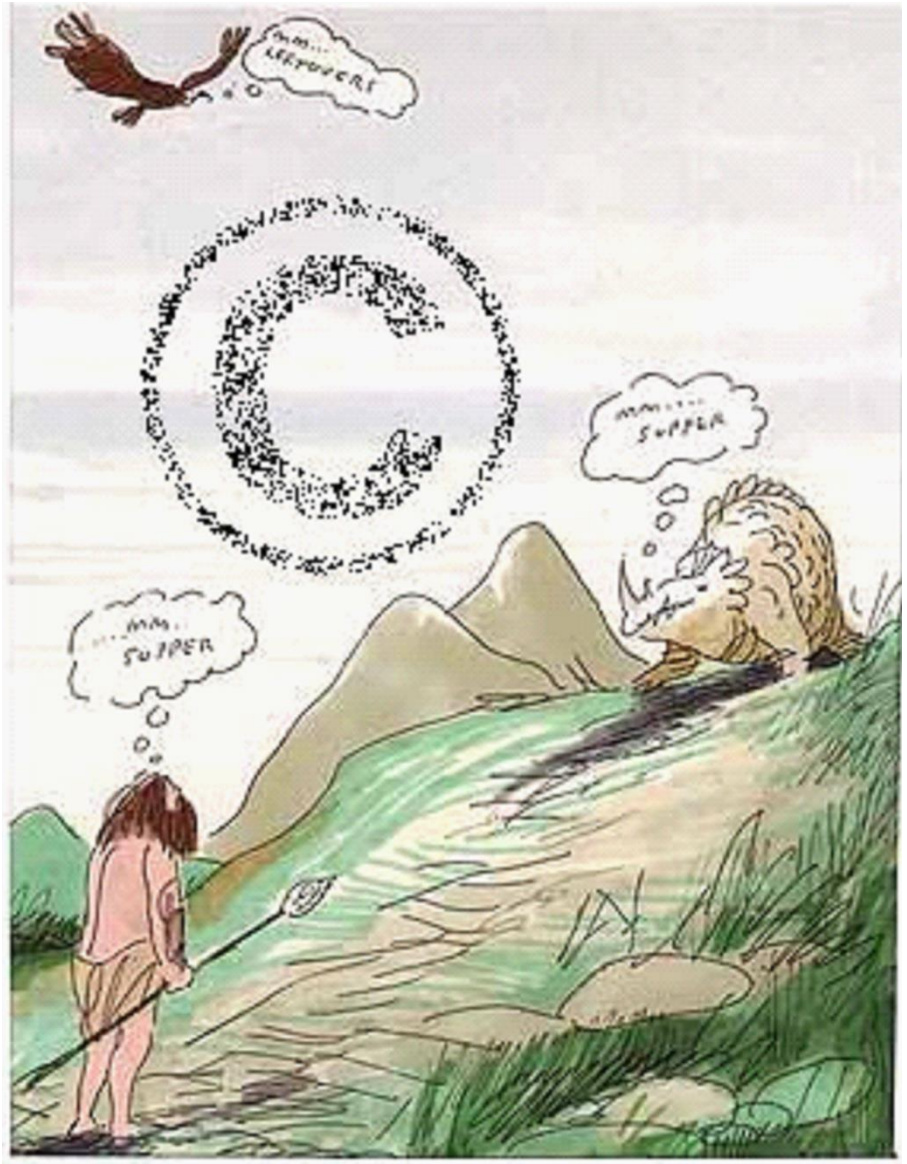
- Above a threshold for communication, peaks and island (**verbal tokens for concepts**) are emerging, standing firmly on an extended solid and shared foundation

- Established connections deep down (between **intervowen schemata**) allow which paths are meaningfully possible
- When restricted to the dry area above the water, **transitions between islands thus mirror the underlying semantic relations**
- Adhering to established (**surface**) statistics, chains will therefor **largely conform to the meaningful structures** at their foundations
- This is one reason why inner speech helps
- And one reason why you can come a long way with speaking without much deep thought and why LLMs apparently work so well



Prompt to Dall·e on 22 August, 2024: “5 islands protruding from a shallow sea, made of scaffolding, which is visible below the surface of the water; the water a little higher, please”

Assignment produces “LEFTOVERS”:



www.cartoonstock.com/
catalog reference mwi0131

(~Pigeonhole principle, picture: wikipedia)

Sleep, Observations & Hypotheses I:

Any known intelligent agent needs some sleep

- **Enhancing hippocampal and cortical representations is one well-established function of sleep**
- **new and old records are interwoven**
- **Junks (features @ frames) "approved" by the consumption analysis are marked as such producing "index-entries" in the hippocampus, and their associated cortical records are enhanced by replay**

- **removing leftovers, i.e. improving signal/noise by specifically erasing disproved material, i.e. partly 'empty' frames and 'loose'/unbound features, is a second complementary main function of sleeping and dreaming**
- **leftover frames and features in the association cortex are unprimed and weakened by REM sleep, --**
- **unless the "resonance" is so strong that one remembers the dream or wakes up**

Sleep, Observations & Hypotheses II:

Any known intelligent agent needs some sleep

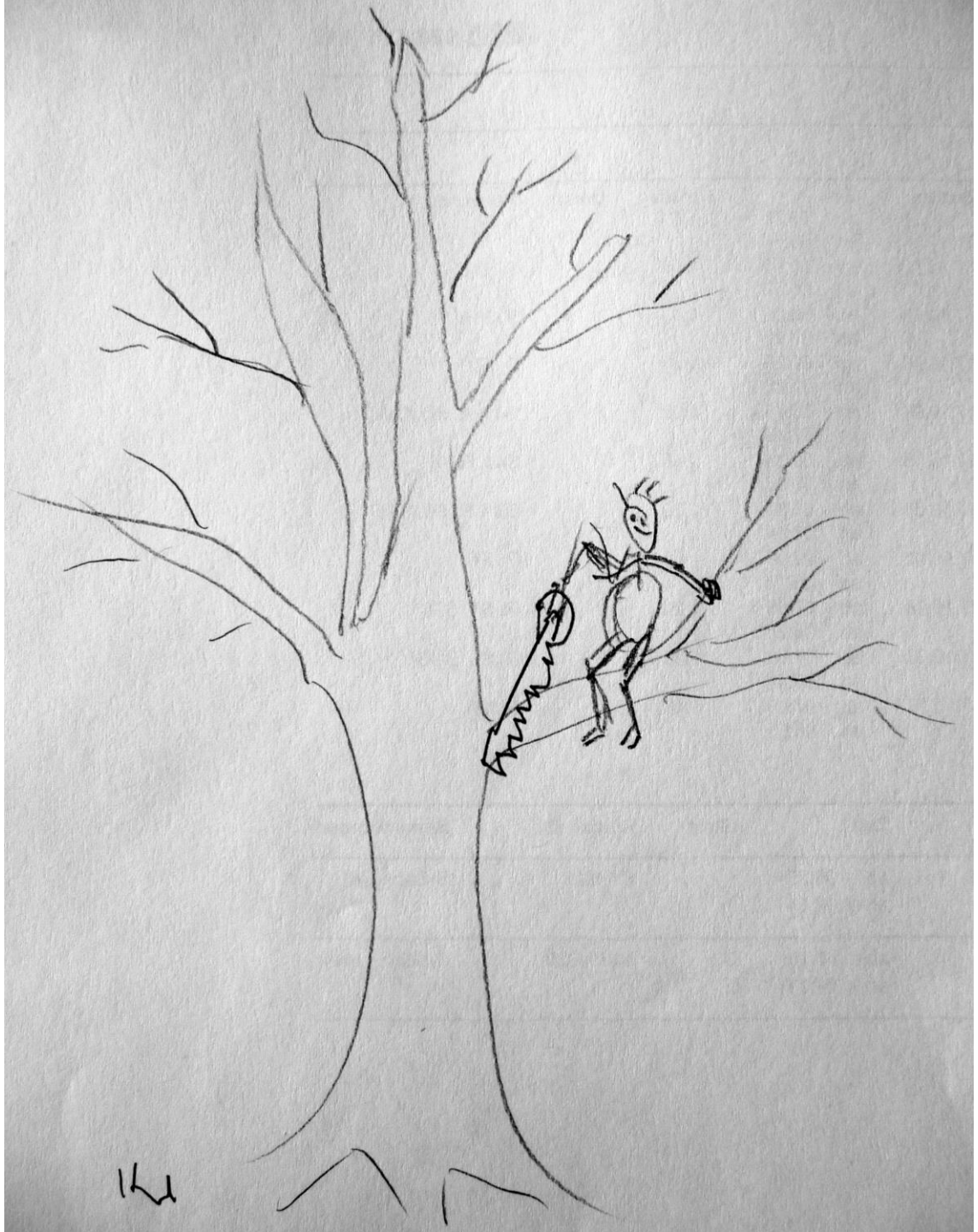
- **Trivially, more activity produces more remnants; even more so, when categories do not (yet) fit**
- **When suitable concepts, i.e. schemata, are not (yet) available, a lot is left unaccounted for; structures have to be distilled**
- **Threatening stuff is paradigmatic for something not digested, negative emotions arise when something does not (yet) fit**
- **Suppression leaves content apart by definition (frontal/parietal junction)**
- **Most residue accumulate where top and bottom meet**
- **Keep useful chunks and make them stand out more clearly, discard useless stuff and make place for new associations**

The most basic requirement during cleaning:
make sure that you produce less mess than you try to get rid of, but do not throw too much stuff away, which might still turn out very useful (so, when in a dream the sleeper is struck by an unexpected insight, like Kekule, the sleeper better wakes up and remembers)

Consciousness:

starting with some minimum of <u>self-awareness</u> seems inevitable for an agent in the real world;

a lack thereof can become costly:



CONSCIOUSNESS FOR THE OUROBOROS MODEL, International Journal of Machine Consciousness, Vol. 3, No. 1 (2011) 163-175, World Scientific Publishing Company

HOPA, Higher Order Personality Activation

The Ouroboros Model endorses a particular **version of HOGS** theory of consciousness (Higher Order Global State, R. van Gulick).

In contrast to more standard varieties of global workspace theories (Dehaene, Baars), the OM does not just make information globally available (the strongest excitement might not always be the best) but rather explains how first the need for widespread activation arises:

With slots not quickly filled (or with surprises), and automatic responses not sufficing („system1“), activation spreads, more than just a few iterations are required and also attributes of the agent herself (including her body) are finally biased and excited. **Higher Order Personality Activation** ensues („system2“).

At the other extreme is the „flow“ phenomenon (Csikszentmihalyi). Complicated activities, which cannot be normally covered by automatic (mainly subcortical) routines, nevertheless are performed without much detailed conscious control

This funtional account of consciousness is not fundamentally limited to humans; with different content, it works the same for all agents with a minimum level of sophistication, in particular, self-monitoring: animals, states, robots, software agents, extraterrestrials, ... nothing much to do with bio

Recurrent processing theory
RPT-1: Input modules using algorithmic recurrence
RPT-2: Input modules generating organised, integrated perceptual representations
Global workspace theory
GWT-1: Multiple specialised systems capable of operating in parallel (modules)
GWT-2: Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism
GWT-3: Global broadcast: availability of information in the workspace to all modules
GWT-4: State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks
Computational higher-order theories
HOT-1: Generative, top-down or noisy perception modules
HOT-2: Metacognitive monitoring distinguishing reliable perceptual representations from noise
HOT-3: Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring
HOT-4: Sparse and smooth coding generating a "quality space"
Attention schema theory
AST-1: A predictive model representing and enabling control over the current state of attention
Predictive processing
PP-1: Input modules using predictive coding
Agency and embodiment
AE-1: Agency: Learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals
AE-2: Embodiment: Modeling output-input contingencies, including some systematic effects, and using this model in perception or control

Recurrency, i.e., **Iterations lie at the the heart of the OM**

With search expanding with biasing activation, the OM offers a much better rationale for global activation than standard Global Workspace theories;
at a certain level, representations pertaining to the body are included, actually the whole actor, i.e.,
HOPA, Higher Order Personality Activation

the OM describes a much more encompassing view than that simple perspective

Errors (discrepancies) are a means, not the message

For an embodied agent at a certain level of mobility and cognitive sophistication, consciousness evolves inevitably, according to the Ouroboros Model

„Indicator Properties for consciousness“,
Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708.

Consciousness for the Ouroboros Model. J. Mach. Conscious. 2010, 3:163–175

Top-down systems	The Ouroboros Model explains how model are built up incrementally where the need arises; once available, they act top down
Internal models	Internal models under the name of **schemata** make up the „substance“ of the OM
Artificial empathy	Everything, which activates some neural representation, later impacts performance; a model of a self exits (only) in mutual correspondence and interplay with models of others
Self-organizing dynamics	The workings of **consumption analysis** are a way of implementing meaningful self-organization
Cognitive consciousness	Excluding feelings and relying on some higher-order quantified multi-operational logics, „Cognitive Consciousness“ seems like a very crude approximation only
Artificial wisdom	Hard to tell, what counts as „natural“ wisdom; **the Ouroboros Model claims that the best that any agent at any particular moment can do is taking all then available and relevant information into account** **Accept for very restricted contexts, it is principally impossible to be sure whether all necessary considerations are even possible**

Antonio Chella, **Artificial consciousness: the missing ingredient for ethical AI?** Front. Robot. AI, 21 November 2023

What could be **tests for consciousness** ?

- Perception
- Reactions

(sheer minimum ?)

- Global integration
- Communication
- Actions with intention

This is, what the medical doctors look for:

Key features of consciousness (= multimodal situational survey) in healthy subjects and in patients with Disorders of Consciousness

Feature	Description in healthy subjects	Description in patients with DoCs
Qualitative richness	Conscious experience is qualified by distinct sensory modalities and submodalities	Conscious contents (if any) might be limited in both sensory modalities and submodalities. They can be evaluated based on brain damage and residual behaviours (e.g. sniffing for smelling)
Situatedness	Conscious experience is specified by the subject´s spatiotemporal condition	Spatiotemporal framing, as well as bodily experience, might be changeable and discontinuous/fragmented
Intentionality	Consciousness is about something other than its neuronal underpinnings	Possible residual consciousness might be still intentional but less egocentric and more allocentric. Arguably decoding from the visual cortical system may indicate what residual visual experience is about
Integration	The components of the conscious experience are perceived as a unified whole	The elements of a scene might be perceived independently or at different levels of detail
Dynamics and stability	Conscious experiences include both dynamic changes and short-term stabilization	Being the anticorrelation between DMN and DAT compromised, residual conscious processing might be any capacity for stabilization. Also the updating (dynamics) of conscious experience can be compromised

Default Mode Network (DMN), Dorsal Attention Network (DAT)

- Story telling, anticipatory planning
- Lying
- Forgetting (of stuff, which lost relevance)

Indicators and criteria of consciousness: ethical implications for the care of behaviourally unresponsive patients, M. Farisco, C. Pennartz, J. Annen, B. Cecconi & K. Evers, BMC Medical Ethics volume 23, Article number: 30 (2022)

Consciousness, Interim Summary:

- All known undisputedly potential intelligent agents are embodied, grounded
- It is argued that embodiment and a fundamental strive for survival in competitions powered the development of living beings to higher and higher levels of capabilities
- Latest when moving around, some anticipation of some immediate consequences in the future become important…
- …Some self-concept and self-awareness seems unavoidable
- In direct extension of bodily homeostasis, the workings of consumption analysis can be understood as a most versatile implementation of a cybernetic control loop
- Consciousness is mutually ascribed & emergent
- Different levels emerging are not black & white

Michael Herzog's hard criteria for empirical theories of consciousness:

- What does the theory address — all
- Coping with the „unfolding argument" (recurrent//feedforward) — yes
- Coping with „small (large) network argument" — yes
- Coping with the „multiple realization argument" — yes

Scope:

one has to take <u>time</u> into account properly:

- content / state — both
- graded / binary — both
- unitary — yes (if healthy)
- continuous / discrete — both
- unconscious stuff — use, discard or recycle

claims the Ouroboros Model 🙂

A. Doerig, A. Schurger, and M. Herzog, Hard criteria for empirical theories of consciousness, COGNITIVE NEUROSCIENCE, Routledge, 2020

„Free Will“

is tricky to define,
it involves mutual ascriptions and emergence

It is all mine !!

Internal perspective:

Actions are considerate and definitively not just random

External perspective:

Actions are not fully predictable appear meaningful not always understandable..

„free“

transcends

any simple dichotomy

Actions ar… unpredi… compl… …

…lly …table,

Free Will

Is an abstraction, -- the same as consciousness (both, mutually ascribed and endorsed in interplay)

While each and every of our concepts is somewhere bodily grounded, no level is rigidly tied to the levels below; on the contrary, higher levels gain new functions and freedom

Emergence is a fact; **more is different** (P.W. Anderson)! like abstraction itself, emergence is an abstraction

Free Will depends on a basis, context,.. e.g., situation and (cultural) community

This was understood and described in his ontology by Nicolai Hartmann, applied by Konrad Lorenz under the name of „Fulguration“, and first mathematically treated by Hermann Haken with his „Synergetics“

Now, with recent attempts of formalization, the concept of emergence seems to gain wider acceptance (e.g., Quanta Magazine June 10, 2024)

Dismissing Free Will with a reference to arguments, like that everything would have a simple mechanistic cause, and especially meaning physiological events at the bottom, is just an category mistake

Free Will is a fundamental assumption in enlightened societies and widely considered a necessary basis for whether any action (or omission) can be qualified as ethical or not

Emergence does not only produce tidy orderly hierarchies; combining content from different fields and levels can give rise to something completely new

Free Will is not without limits or constraints, Ethics

Kant‘s categorial imperative can be understood as a consistency condition amongst equals

The Ouroboros Model holds that Immanuel Kant was right in claiming that we can never fully grasp the „Ding an sich“
Kant consequently cannot really be fully right at the same time when he intransigently clings to deontological ethics

The Ouroboros Model claims that there is a way to derive a „ought“ from a „is“
(opposing David Hume)

In our one tightly interconnected world, some form of a **Negative Imperative** applies:

“given an inescapably limited overall frame,
violence has to be avoided as a result of reflected self-interest”

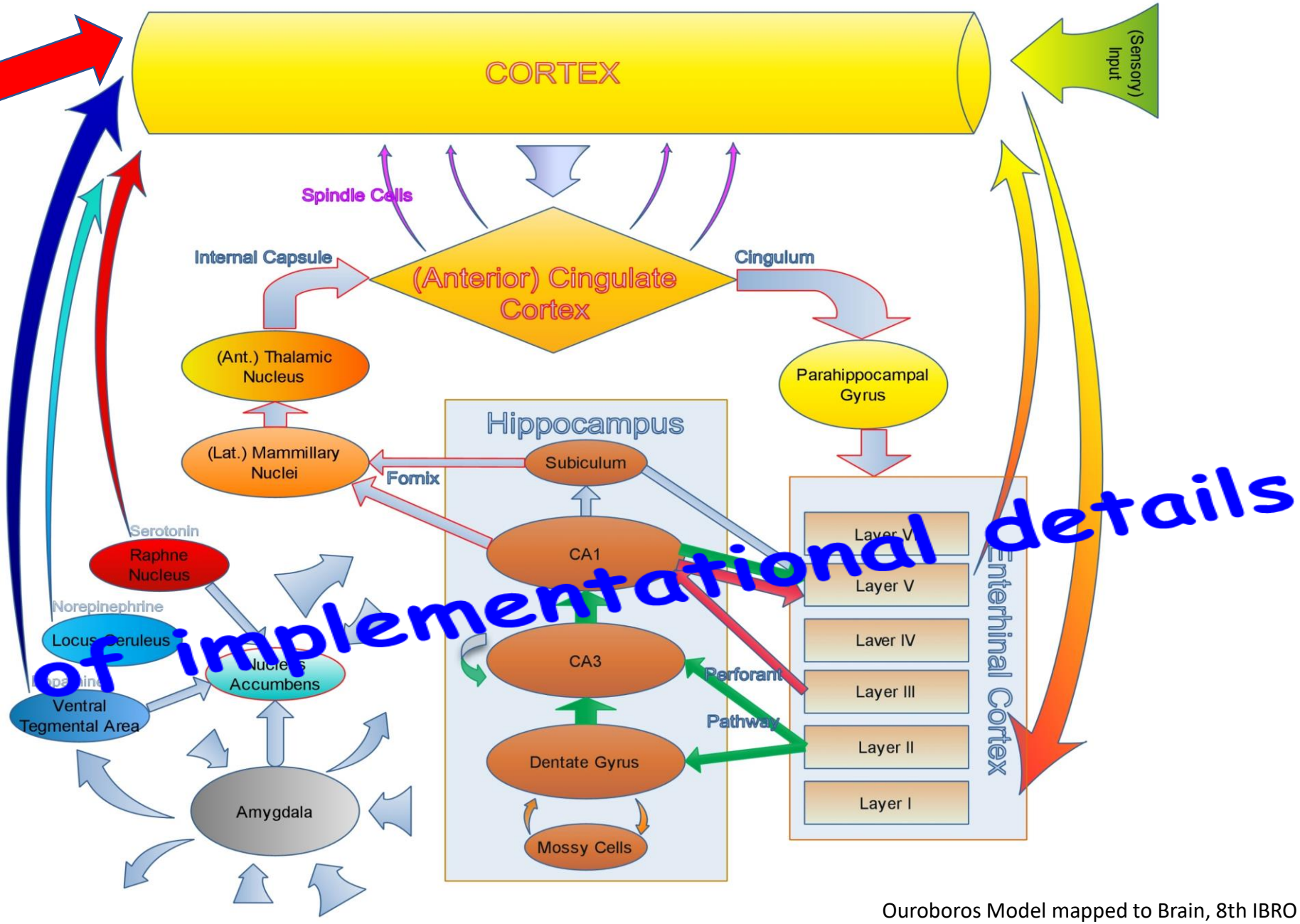
Ethics:

It does not take retiliation, your own strike will hit you !

Never possibly knowing all relevant details and, in particular, in a finite and closed world, it appears wise to act cautiously

Cerebellum

The Cerebellum according to the Ouroboros Model, the "Interpolator Hypothesis" Journal of Communication and Computer 11 (**2014**) 239-254



Sketch of implementational details

Ouroboros Model mapped to Brain, 8th IBRO World Congress of Neuroscience, Florence, 14-18 July, **2011**

Many Key Functions are ascribed to the

Anterior Cingulate Cortex:

- Control
  - Anticipation and preparation before (difficult) task performance
  - Attention
  - Error detection
  - Learning
  - Conflict monitoring
  - Reduced during “Flow”
- (Regulation of) Emotions
  - Stress
  - Appraisal
  - (Reward)Bias
  - Curiosity
- Pain
- Consciousness

Directly underpinning these functions, particular anatomical features have been found in the ACC:

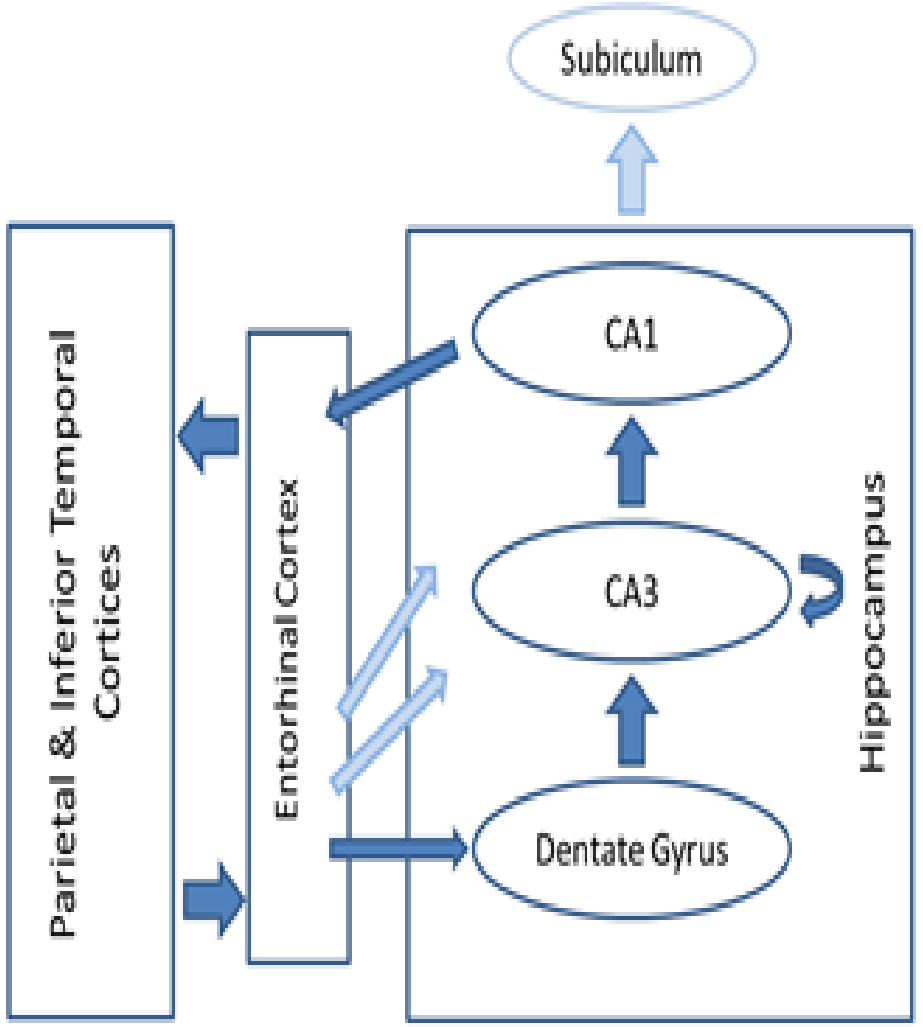
- Anatomic peculiarities
  - No granular layer
  - Von Economo Neurons (VENs)
  - Asymmetry

❖ Details reported as linked to different frequency bands

ONE Function for the Anterior Cingulate Cortex: Consistency Curation

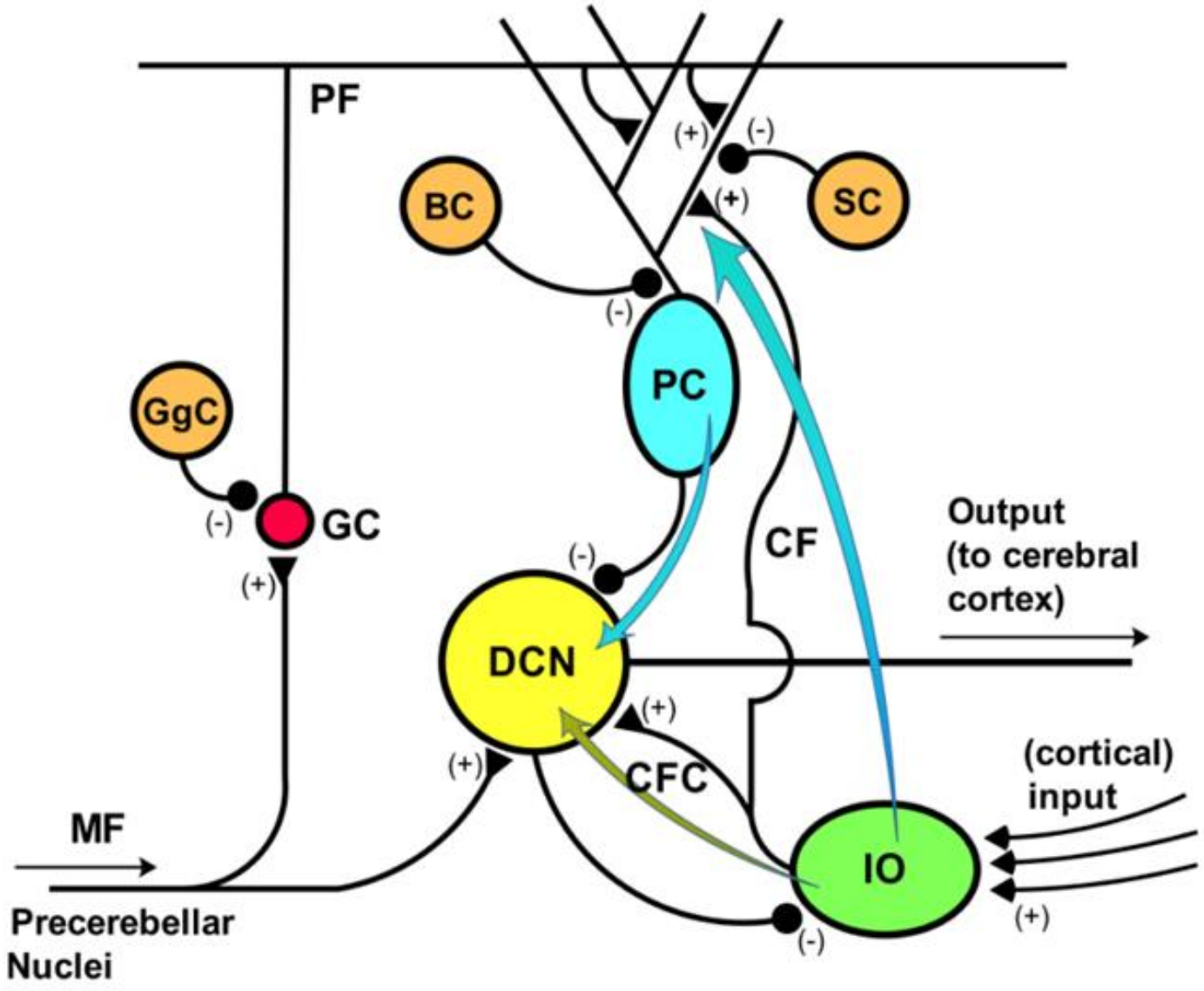
ONE Function for the Anterior Cingulate Cortex and General AI: Consistency Curation, Medical Research Archives 6, Jan. 2018.

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.



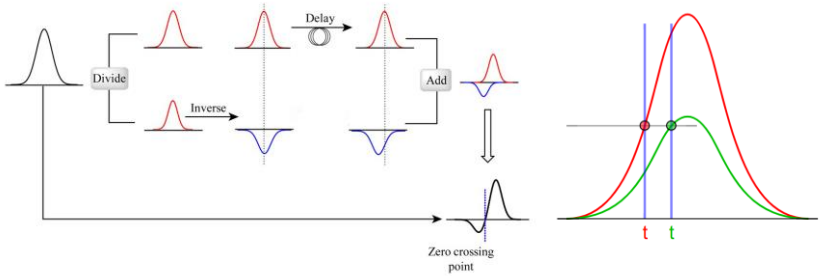
- The hippocampus provides an index to wide-spread cortical representations (hash codes, both ways)
- Content-addressable at both repositories, memories can most efficiently be retrieved from partial keys
- Entries in the hippocampus and in the cerebral cortex mutually endorse each other and thus form a hypercycle (multiple-times pattern completion & stabilization, i.e., loops inside hippocampus, cortex, and together)
- Snapshots of activity in delay lines maps time --> image
- Memory separation/orthogonalization capability for the unique indexing of novel episodes is greatly enhanced by adding adult-born neurons in the DG
- This is somewhat similar to the addition of hidden units in Extreme Learning Machines (ELMs)

The **Cerebellum** according to the Ouroboros Model, the ‚**Interpolator** Hypothesis', Journal of Communication and Computer 11, 239-254, 2014



(implementation of constant fraction det.)

- The human cerebellum contains more neurons than our cerebral cortex
- Cerebellum-like structures are age old
- Interpolation is most helpful for fine-graded perception as well as action (bodily and cognitively)
- Interpolation greatly augments the efficiency of any coding, including possibilities for error correction (e.g., Reed–Solomon codes)



Relations to current research in AI, Predictive Coding

Wikipedia: “predictive coding (also known as predictive processing) is a theory of brain function which postulates that the brain is constantly generating and updating a "mental model" of the environment…” so far, so good

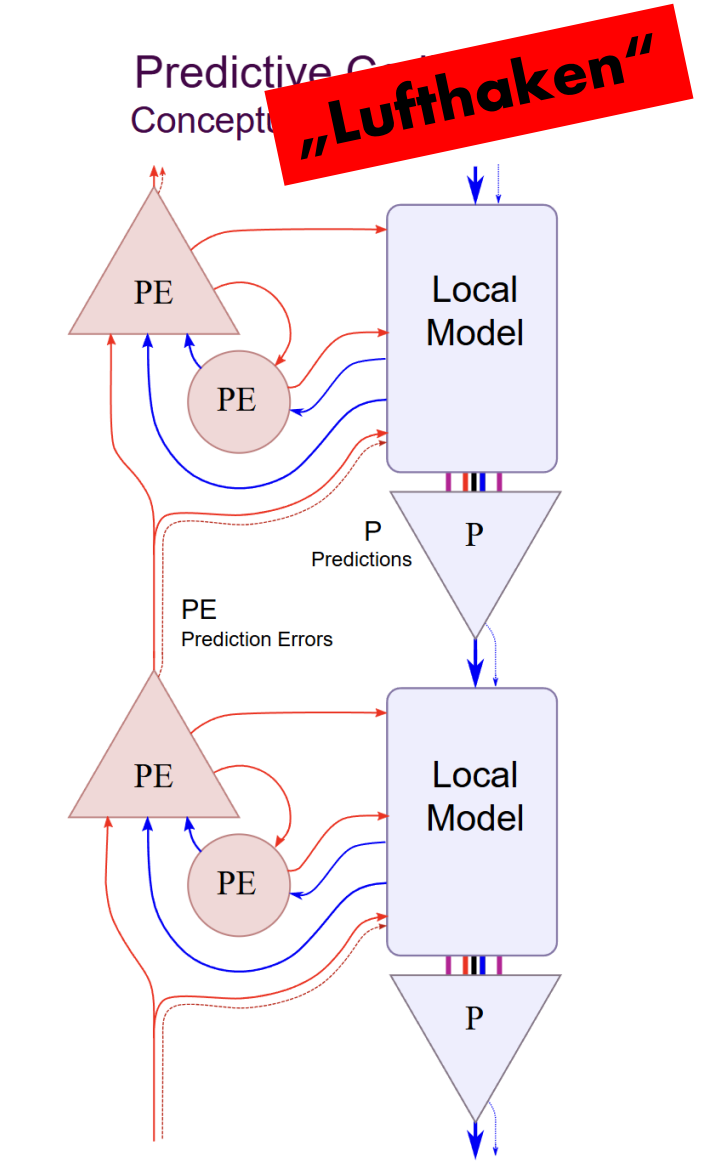
In the light of the Ouroboros Model, the current versions of predictive coding focus way too strongly on prediction errors (the only information in the forward direction) without first explaining, how any mental model is built up to start with (“Lufthaken”)

In addition, improving and fine-tunig any model, certainly is an important and honorable task, but the current versions seems to forget that the immediate first purpose for any agent is to act

An elegant and stringent mathematical formalism, no doubt, is nice to have, but a comprehensive, even if at a certain time only cursory, understanding is more important, according to the Ouroboros Model

It has been pointed out that, mathematically, standard predictive coding (Bayesian processing) faces insurmountable resource problems

As 1 alternative, some Quantum-Mechanics inspired projection mechanism can yield uncertainty-minimizing dynamical behavior equivalent to Bayesian updating and consistent with a free energy principle



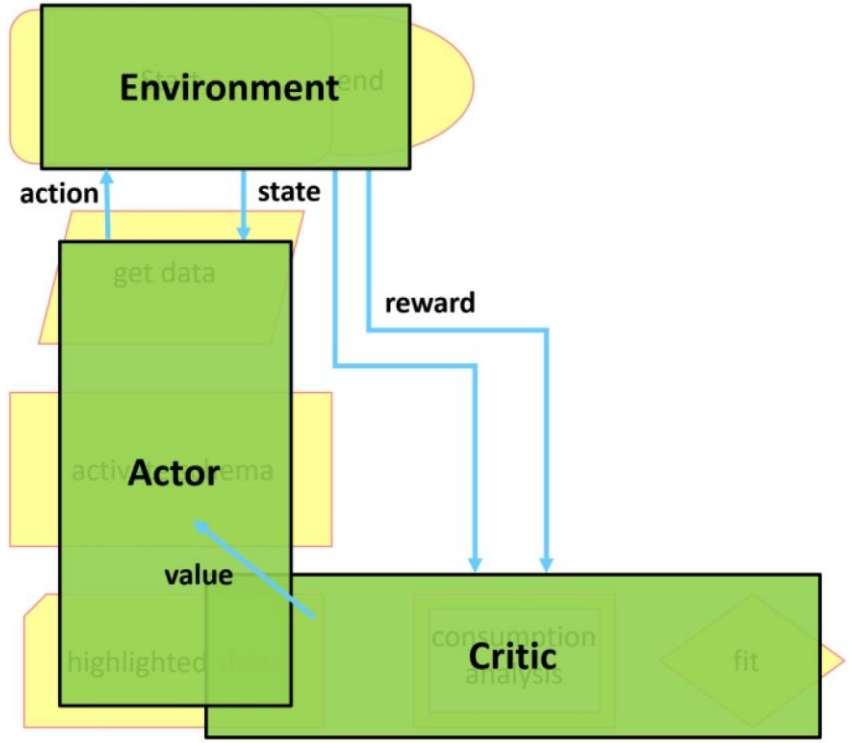
Predictive Coding compares data coming from the senses against expectations coming from the brain. When they do not match, Prediction Errors (PE) are generated.

Kwisthout, J. & Van Rooij, I. (2020), Computational Resource Demands of a Predictive Bayesian Brain, Comp. Brain & Behavior. 3 (2): 174–188; D.C. Brody, Quantum formalism for the dynamics of cognitive. psychology Science Reports 13-16104, 2023

Relations to (current) research in AI, Mental Models

- Have a long history in philosophy and psychology
- internal representations of external reality are hypothesized to play a major role in cognition, reasoning and decision-making
- Mental Models are emphasizing the perception of the world as a set of assertions to derive conclusions from them, e.g., harnessing counter examples for rational decisions and inferences
- The Ouroboros Model can actually be seen as an encompassing mechanistic/functional proposal for a mental-model approach emphasizing the use and autocatalytic further build up of mental models (schemata)

- For reinforcement learning, there is some nice correspondence, e.g., to **actor-critic models**



Mental models and human reasoning, Philip N. Johnson-Laird, PNAS 107, 18243–18250

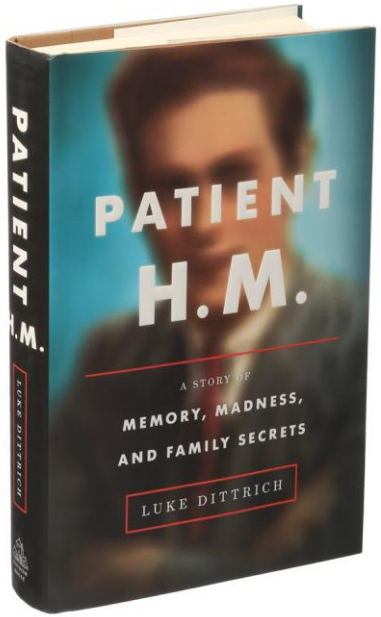
The Ouroboros Model, Proposal for Self-Organizing General Cognition Substantiated, AI 2021, 2, 89–105

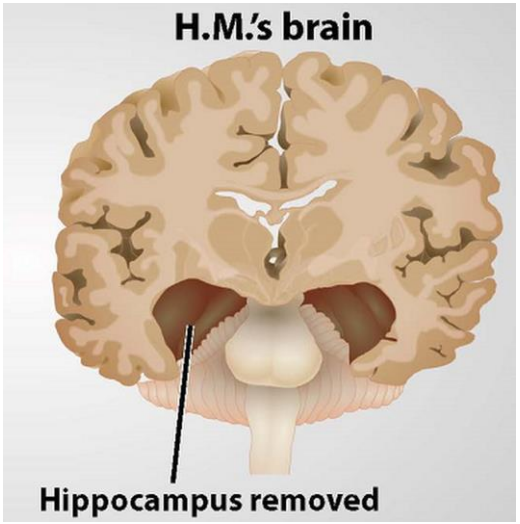
Relations to current research in AI, Large Language Models I

- LLMs can formulate better than most people
- LLMs can make an impression of intelligence, understanding and creativity
- LLMs are so good because language is intimately intervined with human thinking
- LLMs do not master any true understanding
- LLMs lack grounding
- LLMs show no common sense
- LLMs lack structured concepts or world mod
- LLMs do not have emotions, intention
- LLMs are brittle, and with a learni
- LLMs have problems with n eneralization, counterfactuals, tenses
- LLMs are better at ind deductive reasoning
- LLMs can easily h and distracted by added irrelevant details
- LLMs requi ve training
- LLMs allucinate, they have problems saying „I do not know“ (worse newer and bigger LLMs)
- …

current LLMs are not truly intelligent

Large Language Models
are a little like H.M.





Relations to current research in AI, Large Language Models II

LLMs could certainly gain (for some of the deficiences particular remedies are being developed) from:

- clever junking, i.e.,
- structuring their knowledge base
- a capability for one-shot learning
- referring to secured anchor points, e.g., wikipedia articles
- some self-reflection including
- an ability to self-assess their results (confidence, correctness, reliability, … consistency)
- provisions for explainning themselves
- proper grounding, i.e., linking to physics in the real world (..robots!)
- thinking aloud during any action, reporting reasoning / intermediate steps („inner speech“, as shown by Antonio Chella)

Whatever the exact details:

any true intelligent agent will have to learn and be (a little) autonomous, it will be increasingly difficult to understand

Relations to current research in AI, an illustration

Bodily Grounding !

(gravity, common sense)

…. Survival ….

1 example for «emergence»: what counts are not the single strokes

Relations to current research in AI, Autonomy … Robotics

Experiments with DALL·E:

A challenge in A(G)I, cybernetics revived in the Ouroboros Model as one algorithm for all thinking, Artif. Intell. Auton. Syst., 6 March 2024

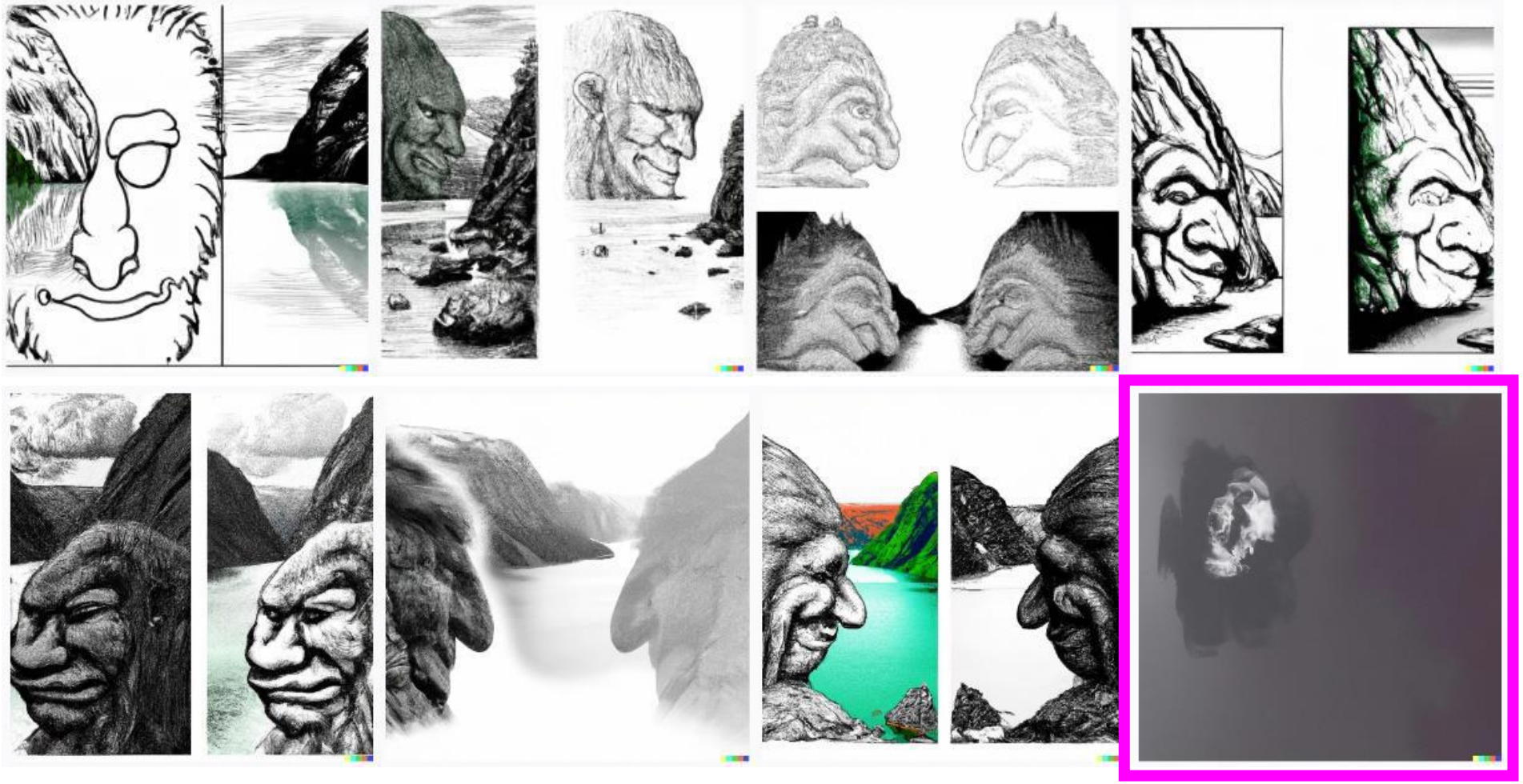
Prompt: “the troll and his home”

Prompt: “a drawing in black and white, which shows the head of a troll when the picture is in portrait orientation, and the same picture showing a Norwegian fjord when viewed in landscape orientation”, in June 2023

Relations to current research in AI, Autonomy … Robotics

Experiments with DALL·E:

Prompt: “a drawing in black and white, which shows the head of a troll when the picture is in portrait orientation, and the same picture showing a Norwegian fjord when viewed in landscape orientation”, **about four months later**:



An abstract picture (?!):

Not showing anything discernable, the image matches with about any prompt

Relations to current research in AI, Autonomy … Robotics I

How to build an intelligent, conscious and ethical artificial agent, some Ingredients:

- Large (compartmentalized) memory organized in hierarchies of differentiated and encompassing schemata
- One-shot learning and learning by forming habits including time-tagging
- Consistency Curation, i.e., discrepancy monitoring and its manifold and repeated use for flexibly steering towards promising actions and learning of new additional useful content
- An endowment with flexible, diverse and redundant means for sensing the internal and external environment (i.e., Grounding, not only for Robots)
- Means for communication with the outside
- Other agents, indispensable for learning and development in an assistive environment
- somewhat ordered and predictable conditions
- enough time and interactions as essential for the growth of adapted useful schemata comprising representations of a self as well as of and for other agents
- Enlightened self-interest promoting curiosity, mutual respect and fairness, prudence, pragmatism, tolerance and modesty

Ethics for Artificial Intelligence, Ethics for All, Paladyn, J. Behav. Robot. 2019; 10:359–363

Relations to current research in AI, Autonomy … Robotics II

- A recursive structure as proposed with the Ourovors Model autonomously directs attention, action, learning,.. to where the need arises
- Where that leads to cannot be predicted much earlier
- In an incremental process, these structures are expanding, non-predictably, and further enhancing the **autonomy** of the agent
- At the same time, its responses and actions become less understandable
- Telling, what an agent thinks, how it arrived at a conclusion, certainly helps

  (inner speech!, chat CPT01 „Strawberry“ already does that)
- On a next level, the same basic processes apply during interactions between agents
- Reciprocity in a dialog or during interactions is a simple generalization of the central role of consistency and adequately fitting between expectations and experience
- As with humans, assuming and demanding honesty should form a common basis

An obvious question arises:

If the Ouroboros Model is so good as claimed here,
why you have not (yet) heard of it,
why is it not wider & better known and worked on ??

There are a number of reasons, the OM is:

- Purportedly Circular
- Blending Analog & Digital Control
- Messing up Disciplines and Ontologies
- Advocating Emergence

- In a conceptual stage only
- Limited by the availability of resources

- There is a widespread mistrust towards papers from single unknown authors

- Purportedly Circular:
  Take <u>time</u> into account properly

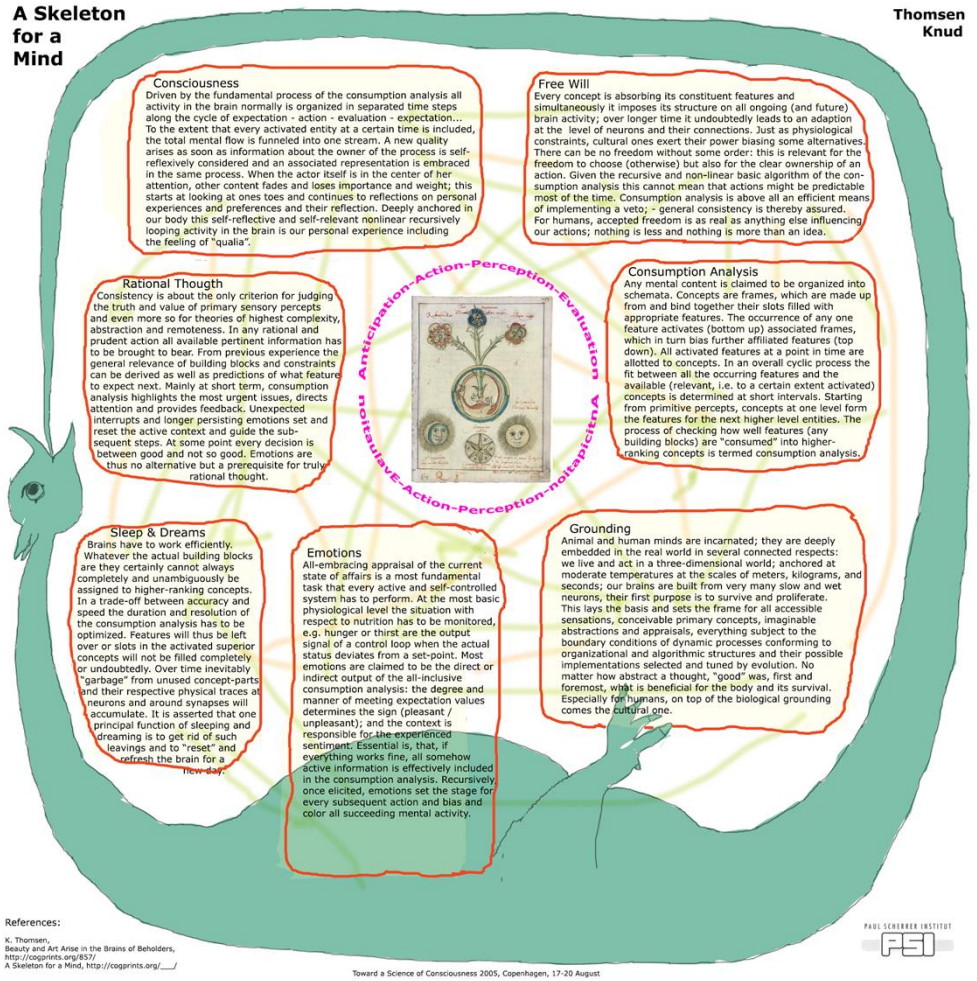
- Blending Analog & Digital Control
  This is how neurons and neural systems work

- Messing up Disciplines and Ontologies
  Creative new combinations often drive progress

- Advocating Emergence
  Gaining wider acceptance recently

- In a conceptual stage only
- Limited by the availability of resources
  This is (still) true

- There is a widespread mistrust towards papers from single unknown authors
  e.g., see: https://www.parolacce.org/2014/10/05/the-true-story-of-stronzo-bestiale/



the very fact, that in a self-consistent manner widely separated questions can be tackled within just one functional approach, is taken as one of the main arguments in favor of the proposed structures and processes, justifying further work

In Physics, one often looks at extremes, at something like opposites,

→ What can “stupidity” tell us about intelligence ?

- **Stupidity is not absolute, it is context-dependent**
- **Consistency can (in any interesting context and in finite time) never be complete**
- **Stupidity is ascribed by actors to other actors**
- **If 2 parties consider each other stupid, both are likely right**
- **Stupidity is inevitable, systemic and 2x highly individual**
  (and infinite as Einstein once remarked)

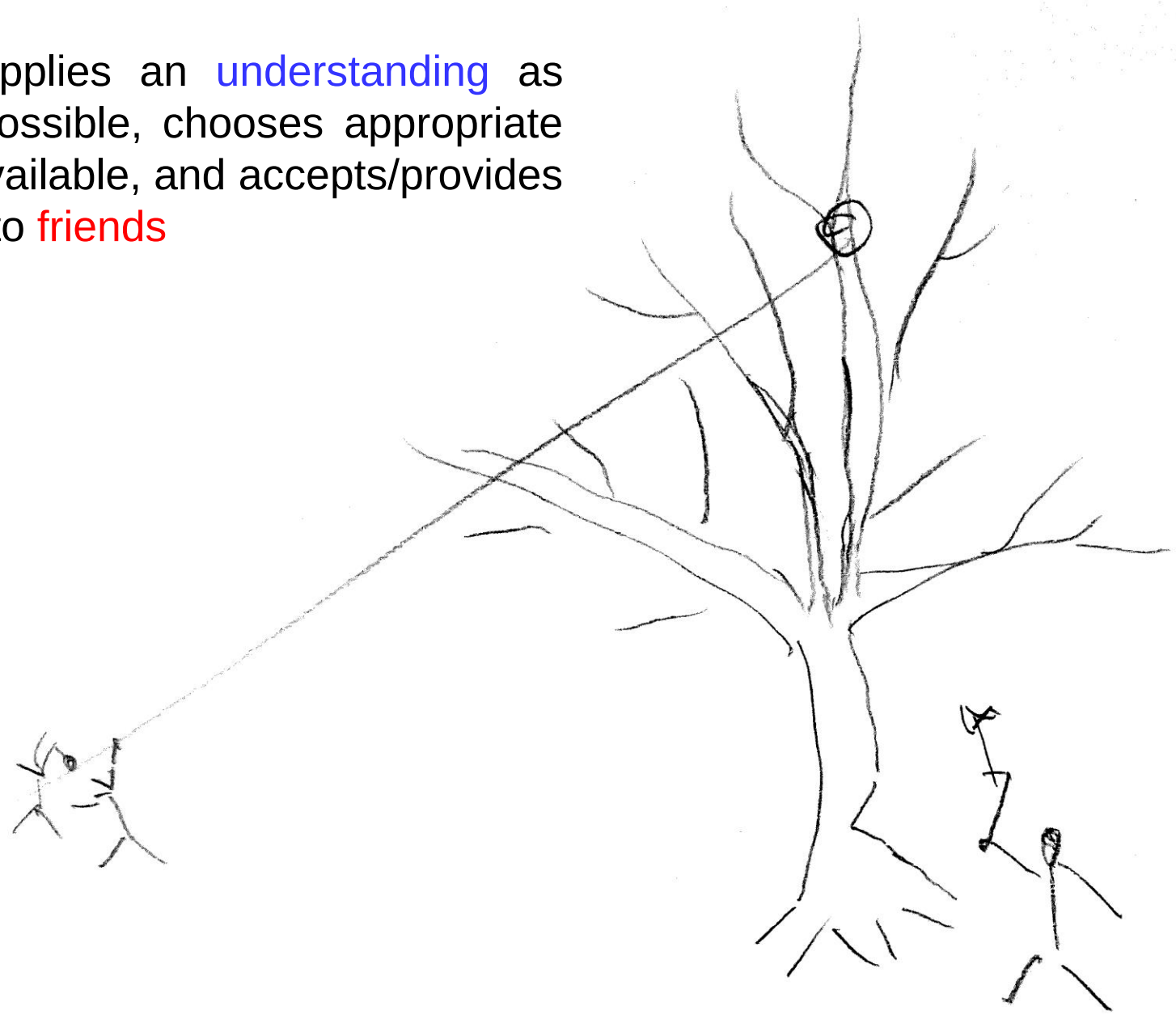
The OM claims that almost all of the above applies equally well to intelligence

„This guy should know …“

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; Vol. 7716, pp. 332–340.

Wise

is, who applies an understanding as wide as possible, chooses appropriate tools as available, and accepts/provides help from/to friends



Components of „wisdom“ as found in a study in 12 countries on 5 continents:

Reflective Orientation

control emotions
apply experience
think logically
think before acting
think in many ways
recognize change

Socio-Emotional Awareness

care for others' feelings
(intellectual) humility
pay attention to emotions
rely on third-parties
others' perspective
sense of humor
nature and divinity

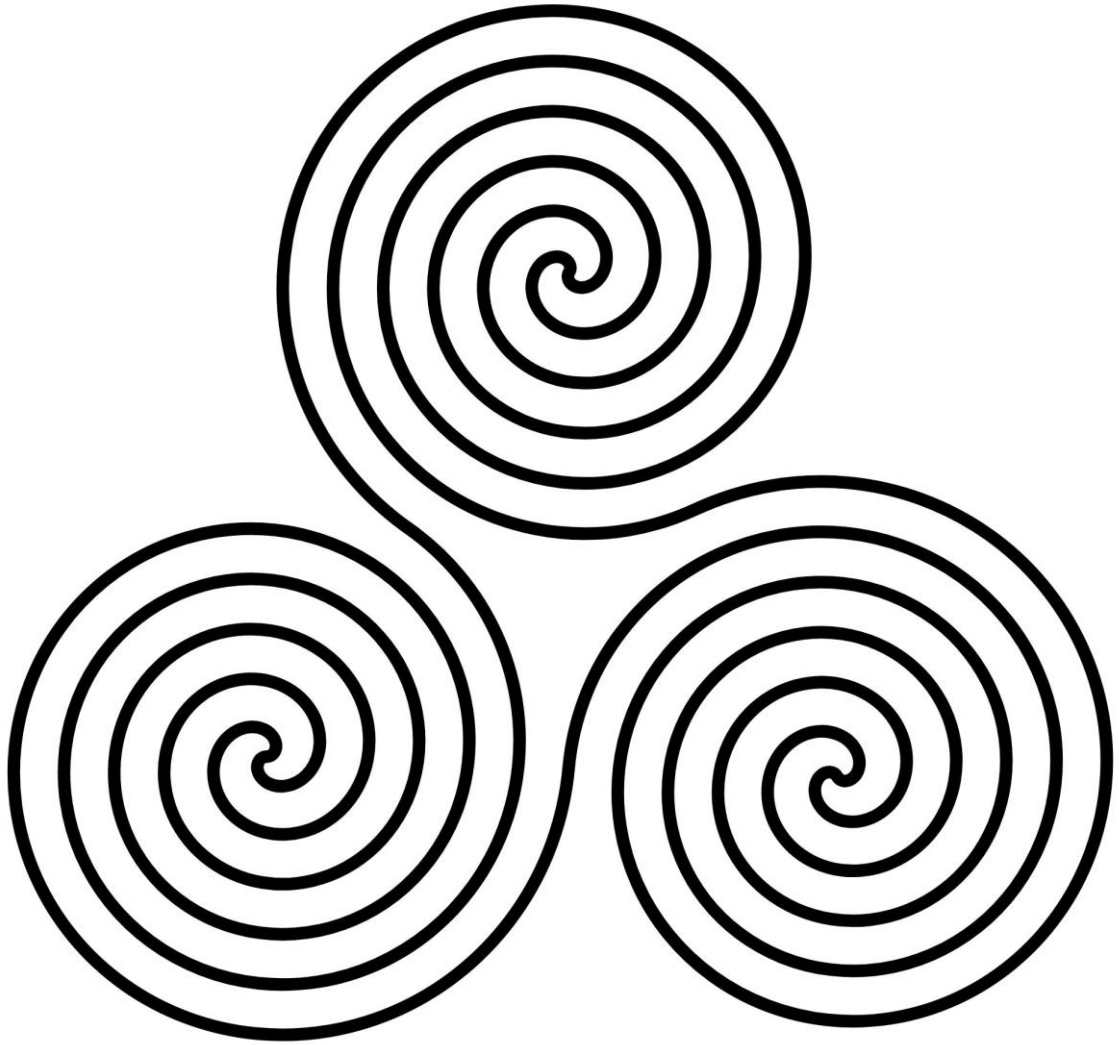
Even better: not cutting the tree but cultivating it !!

i.e., Future Development(s)

- Formalization(s)
- Artifical Implementation(s)
- Refinement(s)
- Testing (!)

Invitation to Collaboration(s) !

Some wider considerations along the lines of the Ouroboros Model yield parallels to Quantum Mechanics and result in a heuristic sketch „how it could all fit together with time“



A heuristic sketch how it could fit all together with time, arXiv:2405.10335

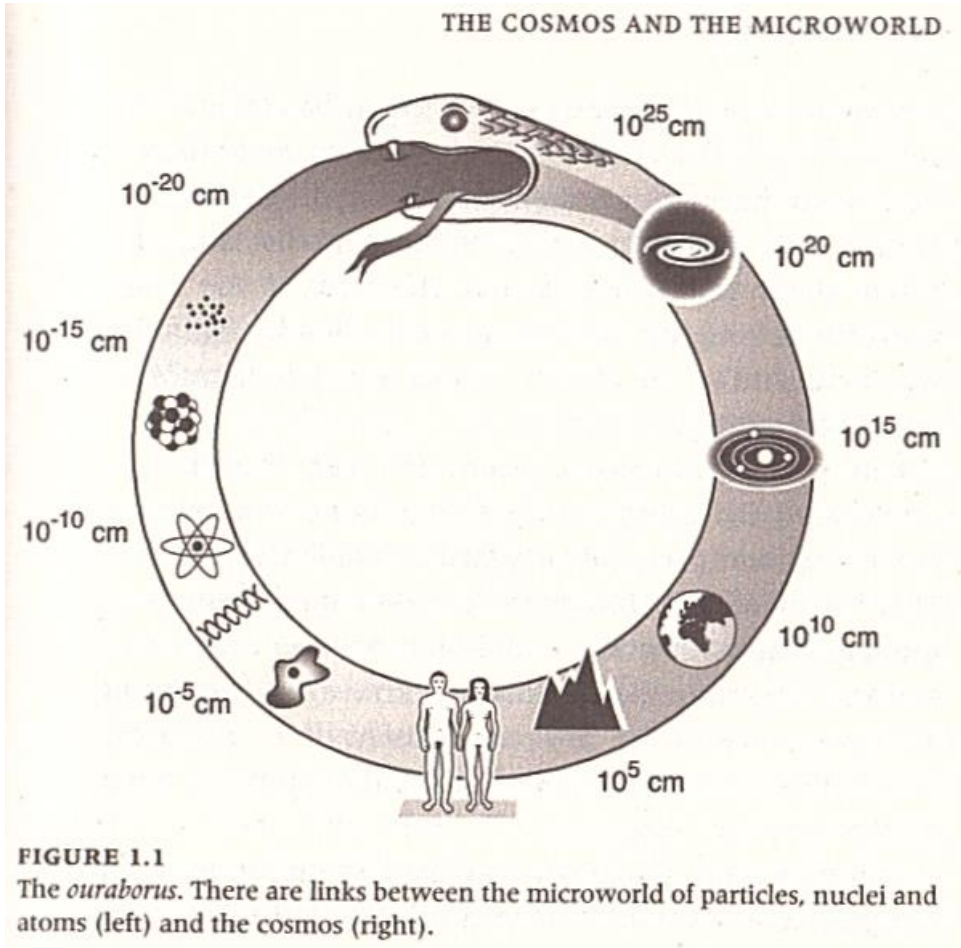


FIGURE 1.1
The *ouraborus*. There are links between the microworld of particles, nuclei and atoms (left) and the cosmos (right).

Martin Rees, Just Six Numbers, The deep forces that shape the universe, Weidenfeld & Nicolson, London, 1999

Thank You for the Invitation and for Your Attention!

https://confluentia-institut.ch

knud@confluentia-institut.ch

Material for questions

ALL ACTIVITY of an actor goes through phases, which are predominantly:

bottom -> up **and** ***top -> down***

parallel **and** ***serial***

=> **"Funnel",**
1 stream (of consciousness, probably "in slices")

Pattern Completion,

any part can excite whole

(i.e. component-based & holisitic,
for figure & ground in, and establishing a context)

Tricky question: nesting of minor loops in big one

Dimensions of wisdom perception

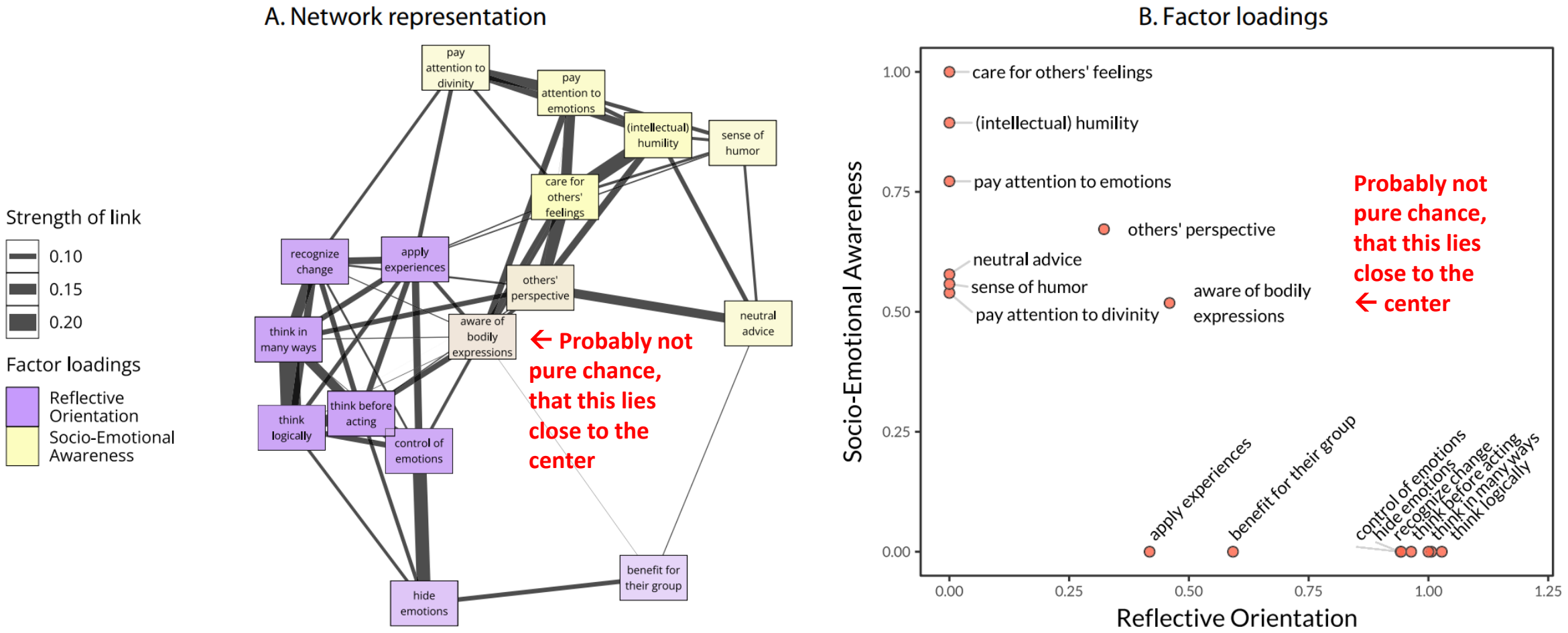


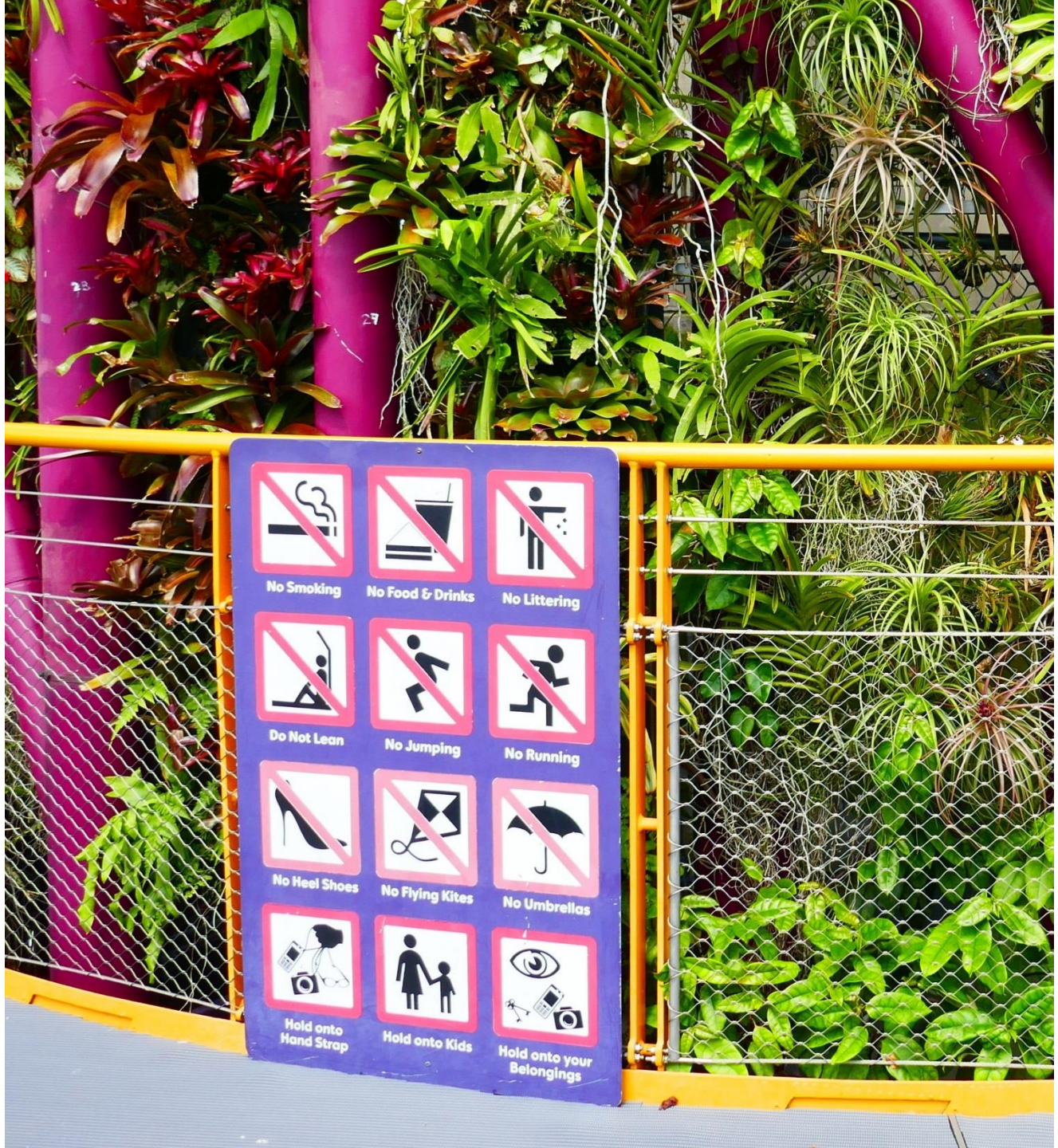
Fig. 2 | The structure of the latent wisdom perception dimensions. A Network graph representation of items demonstrating closer (and stronger) associations of items making up each factor. **B** Unstandardized factor loadings of items of the two factors taken from a multigroup multilevel confirmatory factor analysis. Drawing on prior tests, the underlying model assumes isomorphism (i.e., equal factor loadings at between- and within-individual levels) and partial invariance of loadings across eight cultural regions (only loadings on items 'aware of bodily expressions,' 'consider others' perspective,' and 'listen to nature or divinity' differed across cultural groups). The model fit was acceptable, CFI = 0.912, RMSEA = 0.033, $SRMR_{within}$ = 0.032, $SRMR_{between}$ = 0.078.

Ethics, Negative Imperative (requires a minimum of intellect)

Kant‘s categorial imperative can be understood as a consistency condition amongst equals

With ever more tight common restrictions and interlaced complex links and dependencies between partners any strongly 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 (/collective self-interest).

This means the foundation of an “ought not” in rational egoism. It achieves the same for fairness as Rawls’ veil of ignorance, but in a much more natural and fully self-consistent evolutionary way and not suffering from any artificial limitation like the veil of ignorance (maybe except demanding a certain level of prudency).



1 Modern Example from a crowded place:

Donts

by far outnumber

Dos