Metacognition the Mathematical Way:
Trying to Nest Constructs

Zippora Arzi-Gonczarowski
Typographics, Ltd.
46 Hehalutz Street
Jerusalem 96222, Israel
zippie@actcom.co.il
www.actcom.co.il/typographics/zippie

Abstract

Nesting of computational constructs is prevalent in computers. If one had a rigorous and general formal model of cognition, a high-level programmable and computable schema, then it would be possible to provide a cognitive AI system with that schema, let the system apply the schema to its own cognition as a substitution instance, thus turning the system into a metacognitive system. Concerns would still include infinite nesting and 'first person' grounding.

Introduction

Cognizing about cognition is what researchers in philosophy, psychology, and then cognitive science and AI, have been doing. Assume that we had a rigorous and general formal model of cognition, a high-level schema that can be implemented computationally. It would be possible to provide substitution instances of that formal model as phenomena to cognitive AI systems, to cognize about these substitution instances. If an AI system can self refer to its own cognition as a substitution instance, then it has the potential of becoming a metacognitive system. The proposal may perhaps sound like somebody trying to be clever, 'diamond cut diamond', but the essential general idea is prevalent, taking different shapes in different contexts: In humans introspection and self reflection; In computers bootstrapping and compilers being written in the source language that they compile; Various computational problems contain sub-problems of the same kind, and they are typically modeled by recursion (or iteration): fractals, parsing, sorting, calculating the factorial of an integer or the determinant of a matrix.

Mathematical theory analyzes and applies meta-languages and meta-constructs. They are used to formally model and discuss object-languages (object-constructs), which need not be different from the meta-language (meta-construct) itself. To model and to discuss metacognition that way, one needs to: ($\imath$) Propose a formalized schema of cognition. This is done below in the section entitled ISAAC. ($\imath\imath$) Show how to nest the model in itself, paying particular attention to concerns about infinite nesting, and about 'first person' grounding. That is done in the sections thereafter.

ISAAC: A Mathematical Schema

ISAAC ('Integrated Schema for Affective Artificial Cognition')[1] is a proposal for a mathematical model and a formal theory that could be implemented computationally (though it has no programmed implementation yet). Similar to the natural evolutionary context, the schema starts from a simple model of corresponding sensations and reactions. It then structures 'upgrades' (e.g. handle conflicting reactions, internal representation, and so on) on top of that, using generative reasoning to systematically obtain and study the properties of these upscaled structures. Among other things, this approach models a continuous bridge from low level reactive embodiments to high-level, abstractive and reflective intelligence. In the course of a few years of ongoing research, a variety of cognitive processes have been modeled on the basis of uniform, yet flexible, premises, capturing mental activities from streams of interpretations, through behavior development and integration, representation formation, to imaginative design, anticipation, and analogy making. Lastly, the modeling of social and self perception within ISAAC applies nesting and provides premises for metacognitive capabilities. In addition to modeling single aspects of cognition and affect, the collection features a further value of an integrated whole: because they share uniform modeling premises, the various processes can be neatly composed and alternated between, modeling multifaceted intelligences. A brief outline of the components of ISAAC follows. Theoretical results and constructs, that have not been anticipated at the outset of ISAAC, support the proposal as being on a promising track, towards a unifying theory and, hopefully one day, ($\imath$) Cognitive science becoming a science with a sound formalism, and ($\imath\imath$) Artificial intelligence that integrates embodiment with high level cognitive and affective processes, not losing the Big Picture by over fragmentation.

ISAAC in a Nutshell

'A prerequisite for something to be intelligent is that it has some way of sensing the environment and then selecting and performing actions.' says (Allen 1998). Biological systems started evolving from the earliest nerve cell that was probably a combined receptor (receiving environmental stimuli) and motor unit (producing muscle or gland response). While

[1] Papers in the author's web site (Arzi-Gonczarowski 2004b)

low-level organisms of this type do not have 'minds', that was the beginning. Intelligence in that context is about suitable reactions being conjured up by discriminating sensations that are, in turn, tailored for the forms of behaviour that are afforded by a system. If all extras and subsequent evolution are eliminated, the stuff that should still remain is about a basic marriage between behaviour and circumstances. An abstraction of that forms the basic building blocks of ISAAC:

Forms of Behaviour These are packed in one bag: a set $\mathcal{Z}$ of elements, where each element z abstracts a behaviour/action that a system is able to perform. The abstraction avails an open ended diversity of substitution instances, including, for example, non-overt mental conduct (ISAAC eventually gets to that), and perhaps even procedures and methods that no one has yet conceived of. (e.g. the second part below discusses a reaction that consists of an instruction on how to pass parameters to a nested construct.) For now, as a simple working example, consider $z = \texttt{drink}$.

Sensations These are slightly more complex, being intentional, *about* something environmental. An account is needed of both the sensation and the environmental entity that it is related to. (Recall (Gibson 1977)'s *affordances*: the resources that the environment offers an animal, and the animal needs to possess the capabilities to perceive them and to use them.) $(\imath)$ The set $\mathcal{E}$ models a collection of environmental chunks, typically objects or events, (*world elements*, *w-elements*) that could be perceived. Environments exist independent of their perceptions, but instantiations of the set $\mathcal{E}$ vary between points of view. For example, the carving up of environments into individuated w-elements vary, as one may perceive a single case where another perceives many bottles. This abstraction also avails an open ended diversity of environmental phenomena. As a simple working example, consider w in $\mathcal{E}$ that stands for a bottle (aiming to eventually get to a w-element that stands for someone's cognition). $(\imath\imath)$ The set $\mathcal{I}$ models a collection of sensed discriminations afforded by a perceiver. This avails an open ended diversity of substitution instances as well, including sensations that are not effable and that nothing external may even approximate 'what it is like' to sense them. As simple examples, consider such elements in $\mathcal{I}$ as $\alpha = wet$, or $\beta = empty$. (It is easiest to discuss discriminations using words, but anything recognizable could do, such as an icon or a diagram or an internal state, and so on.) $(\imath\imath\imath)$ A basic role of perception, or precognition, is to relate between w-elements and sensations about them, that potentially distinguish them from one another. For any w-element w in $\mathcal{E}$, and for any discrimination α in $\mathcal{I}$, it associates the pair (w, α) with a truth value. Formally, ϱ is the 3-valued *perception predicate* (*p-predicate*): $\varrho : \mathcal{E} \times \mathcal{I} \longrightarrow \{t, f, u\}$. $\varrho(w, \alpha) = t$ stands for 'α is perceived true about w'. $\varrho(w, \alpha) = f$ stands for 'α is perceived not-true about w'. $\varrho(w, \alpha) = u$ indicates that perception, for some reason, does not tell whether the discrimination α is sensed about w. For example: $\varrho(w, contains\_water) = t$, $\varrho(w, empty) = f$, and $\varrho(w, cold) = u$ (e.g. don't know/ don't care if it is cold).

Reactions: Marrying Perception with Behaviour To formalize the way in which behaviours are conjured, sensations are mapped to reactions: $\mathcal{R}$ is a mapping: $\mathcal{R} : \mathcal{I} \rightarrow \mathcal{Z}$. When we soon pack all the definitions together, it will be the case that if, for some w and α it so happens that $\varrho(w, \alpha) = t$, then $\mathcal{R}(\alpha)$ is conjured. In a thirsty substitution instance, for example, $\mathcal{R}(contains\_water) = \texttt{drink}$. (Reactions that are conjured when a sensation is *absent*, or reactions to combinations of sensations, constitute upscaled capabilities. They are modeled and studied in the cited papers.)

The Definition of Perception-Reaction Packing the building blocks together one gets:

A *Perception-Reaction* is a 5-tuple $\mathcal{M} = \langle \mathcal{E}, \mathcal{I}, \varrho, \mathcal{Z}, \mathcal{R} \rangle$
$(\imath)$ $\mathcal{E}$, $\mathcal{I}$, $\mathcal{Z}$ are finite, disjoint sets
$(\imath\imath)$ ϱ is a 3-valued predicate: $\varrho : \mathcal{E} \times \mathcal{I} \longrightarrow \{t, f, u\}$.
$(\imath\imath\imath)$ $\mathcal{R}$ is a function: $\mathcal{R} : \mathcal{I} \longrightarrow \mathcal{Z}$

Specific substitution instances of the five co-ordinates would model instances of embodiments in authentic environments. Out of the open ended diversity of possible instantiations of $\mathcal{M}$, some are likely to make more sense than others, in the same manner that not every well formed formula can be applied to compute something of interest. In the biological context, certain successful sensory motor neural apparatuses survived in their ecological niches, when lucky combinations of environments, sensations, and reactions supported endurance of a species.

This basic formalism models a fixed state of something deterministic. For now, it lacks dynamics, learning and adaptation, behavioural control, and so on. So far we have premises for modeling low-level systems, maybe an abstracted amoeba. This basic schema could be programmed simply, using a loop that checks the sensors and reacts deterministically, practically conflating ϱ with $\mathcal{R}$, and *using the world as its own model* in the spirit of (Brooks 1999).

Perseverant Vitality An embodied implementation of a perception-reaction, situated in a real environment, is unlikely to work without physical robustness of the construction and its performance. For example, an embodiment that cannot lift a wet slippery bottle is unlikely to manage drinking from that bottle. That embodiments should not be made of, say, paper foldings, is so obvious, that theoretical cognitive models seldom bother to say that, overlooking a consequential premise. First of all, it would be more suitable to rename the elements of $\mathcal{Z}$ 'action tendencies' rather than actions: Internal incitements to perform behaviours, which are either consummated or not, depending on an open ended diversity of reasons. The term is associated with (Frijda 1986), who defines his conceptualization of emotions as the readiness to act in a certain way.

To model robustness (rather than 'leave that for implementations' and ignore the consequences), ISAAC associates with every behaviour z of an embodied system a *Vitality Value* $V(z)$, to be evaluated by an open ended diversity of resources: time, energy, injury, tear-and-wear, and so on. For example, if the action $\texttt{drink}$ is hindered, a 'price' in resources $V(\texttt{drink})$ is paid, modeling the vitalily involved in really striving to drink. (This is neither the energy consumption for the act of drinking, nor the physiological implication of the lack of liquids, though $V(\texttt{drink})$ may be related to these issues). Further discussions of implications on the modeling of higher-level cognitive and affective phenomena, appear in (Arzi-Gonczarowski 2004a).

Intelligence Steers between Perceptions-Reactions When its habitat changes, an organism is unlikely to survive unless it adjusts its perceptive-reactive gear. That is one evolutionary pressure to allow change, a capability that seems to be at the basis of higher-level cognition. The idea is that minor changes in single co-ordinates can be composed and recomposed into more elaborate ones, (like a movement of a cartoon character that is made of a consecution of basic movements of every joint). We start with the formal definition, and after that discuss its meaning.

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two perceptions-reactions: $\mathcal{M}_1 = \langle \mathcal{E}_1, \mathcal{I}_1, \varrho_1, \mathcal{Z}_1, \mathcal{R}_1 \rangle, \mathcal{M}_2 = \langle \mathcal{E}_2, \mathcal{I}_2, \varrho_2, \mathcal{Z}_2, \mathcal{R}_2 \rangle$ Define a *Perception morphism* (also called *p-morphism*, or *arrow*): $h : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ by the set mappings: $h : \mathcal{E}_1 \rightarrow \mathcal{E}_2$, $h : \mathcal{I}_1 \rightarrow \mathcal{I}_2$, $h : \mathcal{Z}_1 \rightarrow \mathcal{Z}_2$, With two structure preservation conditions: $(\imath)$ *No–Blur:* For all *w* in $\mathcal{E}$ and all α in $\mathcal{I}$: If $\varrho_1(w, \alpha) \neq$ *u*, then $\varrho_2(h(w), h(\alpha)) = \varrho_1(w, \alpha)$. $(\imath\imath)$ *Disposition:* For all α in $\mathcal{I}$, If $\mathcal{R}_1(\alpha)) \neq \texttt{null}$, then $\mathcal{R}_2(h(\alpha)) = h(\mathcal{R}_1(\alpha))$.

Set maps are formal tools to parallel, respectively, between environments, discriminations, and behaviours. They afford the modeling of $(\imath)$ Replacements/translations of elements, $(\imath\imath)$ Extensions of sets ('onto' vs. not 'onto'), $(\imath\imath\imath)$ Generalizations by amalgamating elements, thus losing distinctions between them ('one-to-one' vs. 'many-to-one'), $(\imath\imath\imath\imath)$ Identity maps model non-transitions (quite often, some of the maps will be identities). Certain capabilities may not be supported in certain agents: If an agent is unable to learn new discriminations, then the $\mathcal{I}$ map will always be 'onto', and if it is unable to generalize, then its maps will never be 'many-to-one', and so on. When these options are applied to the maps in the definition, one gets: $(\imath)$ The 'literal-analogical' map $h : \mathcal{E}_1 \rightarrow \mathcal{E}_2$ models environmental changes. If not 'onto', then $\mathcal{E}_2$ features a new w-element, that is not part of $\mathcal{E}_1$ (for example, the bottle was not there, or $\mathcal{M}_1$ did not attend to it). If not 'one-to-one', then $\mathcal{E}_2$ features a more general chunking (for example, ignoring distinctions between individual bottles). Replaced w-elements model an analogical similarity (for example comparing/replacing the bottle with another source of water). (Arzi-Gonczarowski 1999b) studies ISAAC's formal modeling of analogies and metaphors. $(\imath\imath)$ The 'interpretive' map $h : \mathcal{I}_1 \rightarrow \mathcal{I}_2$ models changes in discriminations. If not 'onto', then $\mathcal{I}_2$ features a new discrimination, that is not part of $\mathcal{I}_1$, but it has been embedded into $\mathcal{I}_2$ (for example, learn to discriminate *potable*). If not 'one-to-one', then $\mathcal{I}_2$ makes more generalized discriminations, ignoring finer ones, such as merging degrees of fullness *half_full, quarter_full, full* into a single *nonempty*. Replaced discriminations model interpretive translations (for example *water* may map to H_2O). (Arzi-Gonczarowski & Lehmann 1998b) study ISAAC's formal modeling of interpretive transitions. $(\imath\imath\imath)$ The 'behavioural-adaptation' map $h : \mathcal{Z}_1 \rightarrow \mathcal{Z}_2$ models behavioural changes. If not 'onto', then $\mathcal{Z}_2$ features a new behaviour, that is not part of $\mathcal{Z}_1$, but it has been embedded into $\mathcal{Z}_2$, for example using a drinking straw. If not 'one-to-one', then a few behaviours (for example drink directly from the bottle, or use a drinking straw, or pour into a glass first) are replaced by a generalized one. Replaced behaviours could model adaptation and tuning of reactions, for example transiting to drinking more slowly or more quickly. These modeling tools are flexible, yet not infinitely so, designed as a spacious mould that would still not contain certain things. The three maps are bound together by the structure preservation conditions. That is related to the notoriously evasive core invariable aspect of meaning that one would like to preserve across transitions. $(\imath)$ *No-blur* binds change in the environment $\mathcal{E}$ with change in the interpretation $\mathcal{I}$: Transitions between w-elements need to be justified by commensurate sensations, and transitions between discriminations need to be grounded by commensurate experience. $(\imath\imath)$ Behaviours and reactions may be modified along arrows, but that is confined by *Disposition*, which binds change in interpretation with change in behaviour. $(\imath\imath\imath)$ Specific contexts may add additional structure preservations beyond these minimal constraints.

Typical examples of arrows involve time: the perception-reaction on the left of the arrow intuitively happens 'before', while the one on the right of the arrow intuitively happens 'after'. However, *there is nothing in the formal construction that obliges us to that interpretation*. When the chronological functionality is abstracted away, an arrow becomes a structural tool that can be applied to model other commensurations between perceptions-reactions, where time is either irrelevant, or perhaps even proceeds in an opposite direction to the structural arrow. A regularity of all functionalities and modules in ISAAC is that they are based on (structured compositions of) arrows, modeling the workings of minds.

An autonomous perception-reaction would have to activate the proper transition *by itself*. Benefits of the construction with building blocks can be deployed now: Since an engine for reactions has been formalized, and arrow transitions can be defined as a legitimate behaviour for a system, then all that needs to be done at the formal level is to define actions z that stand for $\texttt{Activate the arrow } h$.

A Word about the Formalism In order to model a myriad of cognitive processes with shared high-level theoretical units, a higher level of abstraction has been applied. The challenge is a balance between abstraction that is not detached, and grounding that is not over deterministic. Categorization is a tool that has been developed precisely for such purposes within mathematics itself. Technically, perceptions-reactions with arrows make a mathematical category, providing a formal infrastructure to capture the structural essence of cognitive processes, without being over deterministic. The proposed category has been presented in (Arzi-Gonczarowski & Lehmann 1998b). A salient property of categorization is that a chain of arrows can be composed into one arrow. An arrow can also be factorized into a consecution of composable arrows, each one of them making a part of the transition. That way, one can model transitions from small alterations, to transformations so profound that $\mathcal{M}_1$ and $\mathcal{M}_2$ may appear to have little in common.

ISAAC's formal perspective is about cognitive processes that perform uniformly by construction and navigation of edge paths in a (sub)graph, where states of mind are vertices, and arrows are edges that lead from one state to another. Autonomous intelligent systems do that construction and navigation pro-actively. The perspective of the following sections is that a metacognitive system $\mathcal{M}$ should be able to perceive such (sub)graphs in its own environment $\mathcal{E}$, using suitable discriminations, correctly ascribing them to itself and/or to others, and eventually also behaving accordingly by adjusting and improving its own (sub)graph.

With possibly multiple ways of getting from one vertex (i.e. perception-reaction) to another, one needs to make sure where arrow paths lead. That is where theorems about commutative diagrams come into the picture. (Barr & Wells 1995, p.83) entitle commutative diagrams *'the categorist's way of expressing equations'*. The sides of these equations are arrow paths, and an equation formally warrants that they get to the same vertex. Such theorems are ISAAC's theoretical results. They state the rules of paraphrasing (namely: meaning) in the proposed 'language' of cognition.

Further Upscaling The framework of this extended abstract does not permit a detailed discussion of constructss and specially trimmed arrow paths that lead heel-and-toe to states of mind with higher-level capabilities, and readers are referred to the cited works for details: Arrows to, and between, perception-reactions with various degrees of structure on the set of discriminations formalize upscaling to monitor behavioral conflicts (giving rise to emotions (Arzi-Gonczarowski 2002)), these structures can also support representations and analytic thought (Arzi-Gonczarowski & Lehmann 1998a); Arrows to, and between, perception-reactions with conceived environments formalize memory, anticipation, and design processes (e.g. to monitor the outcome of actions in the world) (Arzi-Gonczarowski 1999a); Arrows to, and between, perception-reactions with a structure on the environment formalize upscaling to further creativity and analogy making (Arzi-Gonczarowski 1999b).

Metacognition with ISAAC

We start with ISAAC's modeling of social and self perception. That will provide premises for metacognition.

The Evolutionary Pressure of Social Contexts A social environment features other agents as w-elements. An obstacle that stops one's way to drinking water may surrender to a displacing push, but if it has a mind of its own, then more than its mass inertia may have to be grappled with. A natural evolutionary pressure follows to upscale perceptions-reactions to apply their gear to other perceptions-reactions. Subject perceptions-reactions should, for example, predict that in such cases, objects typically show a bit of their own minds when contacted physically. If objects' perceptions-reactions do the same, then additional iterations may follow, with some being possibly ahead of others. Evolution naturally selected for humans a gift to relate to agents with minds (Humphrey 1984). An example is animism in young children, who intuitively assign mental states to nonhumans (Piaget 1926). A negative example: the cognitive explanation of autism has become known as the 'theory of mind deficit', because autistics seem to lack the intuitive understanding that people have mental states (Baron-Cohen 1995).

The Evolutionary Pressure of Self Contexts Similar to social contexts as above, why not let an object perception-reaction be eventually instantiated by the subject perception-reaction itself, letting it behave also on the basis of a perception of itself, yielding a more sophisticated kind of higher-order control. Such agents can monitor their own behavior with forethought, flexibility, and creativity. That is their competitive edge (Ballonoff 2000).

Structural Requirements Evolution theorists use the term *exaptations* (Gould & Vrba 1982) to refer to minor changes that make use of already existing capabilities to create new behaviours. The significance of the capability naturally grows together with the number of behaviours that it supports. The two evolutionary pressures above require that a perception-reaction should be able to apply its gear to perceptions-reactions as part of its environment. The difference between the two is that one is done from a 'third person', and the other from a 'first person', perspective. It requires that self-modeling should have a distinct 'feel' (i.e. a discrimination) that tweaks the behavior towards its own self, and that would constitute the 'exaptation'. Both variants share the structural composition of building blocks, which is to let perceptions-reactions perceive-react to perceptions-reactions (and eventually also to arrows and entire subgraphs), their own and others. Before going into some thorny issues, observe that, with no additional definitions, one could let, for example, (subparts of) the co-ordinates of a perception-reaction $\mathcal{M}_\imath = \langle \mathcal{E}_\imath, \mathcal{I}_\imath, \varrho_\imath, \mathcal{Z}_\imath, \mathcal{R}_\imath \rangle$ be w-elements in the environment $\mathcal{E}_\jmath$ of $\mathcal{M}_\jmath = \langle \mathcal{E}_\jmath, \mathcal{I}_\jmath, \varrho_\jmath, \mathcal{Z}_\jmath, \mathcal{R}_\jmath \rangle$. It may be that $\imath = \jmath$. (Extending the schema with such higher-order constructs has also been discussed in (Arzi-Gonczarowski 2001a; 2001b).)

Thorny Issue: Infinite Regress The proposed structure raises theoretically problematic issues that go back to the paradoxes that led to an overhaul of the foundations of modern math. These paradoxes typically originate in circular references. If $\imath = \jmath$, or if $\mathcal{M}_\jmath$ also perceives $\mathcal{M}_\imath$, and each one of the behaviours $\mathcal{R}_\imath, \mathcal{R}_\jmath$ depends on the perception of the other behaviour, one may get into a vicious circle. That would challenge the *iterative hierarchy* of the construction: Begin with some primitive elements (w-elements, discriminations, behaviours), then form all possible perceptions-reactions with them, then form all possible perceptions-reactions with constituents formed so far, and so on. In programs with recursion, it is essential to avoid an infinite sequence of procedure calls by having a 'start of recursion' for which no further call of the procedure is necessary; Iterative loops need to have a halting condition; In set theory, the *axiom of foundation* is normally added to the five original axioms of Zermelo, to warrant an iterative hierarchy. Non-classical solutions have been proposed in (Aczel 1987; Barwise & Etchemendy 1987; Barwise & Moss 1991). It is noted that not all self references produce theoretical paradoxes, and not all perceptions-reactions of perceptions-reactions involve vicious regress. Some are benign and bottom out neatly. The theoretical difficulties lie in forming conditions that exclude the pathological cases only.

The motivation to go ahead with that risky construction, is that these theoretical difficulties are precisely those that model difficulties of social and self modeling, and hence metacognition as well. Vicious circles do happen in real situations, and they need to be modeled. A 'no free lunch' price needs to be paid. Straight line computations that always converge would have provided a reason for worries concerning the validity of a model. (Sloman 2000) also remarks that: *'Self-monitoring, self-evaluation, and self-control are all fallible. No system can have full access to all its internal states and processes, on pain of infinite regress'*. Even recursions that eventually halt can consume a lot of resources and interfere with performance, if a procedure calls itself many times. In humans, the analog of that is sometimes avoided when one says *'Never mind'*, applying a natural reluctance to reflect and to wonder too much about things, to preserve resources and to divert attention. A proposed solution to the next thorny issue below will suggest another way out of reflective infinite regress.

Thorny Issue: 'First Person' or 'Third Person'? Let a perception-reaction $\mathcal{M}_\imath$ be perceived by $\mathcal{M}_\jmath$ as suggested above, and let $\imath = \jmath$. That can still be done from a 'third person' or a 'first person' perspective, namely $\mathcal{M}_\jmath$ does not necessarily 'know' that it is referring to itself. Also, what if $\imath \neq \jmath$, but they describe the same agent at different points in time, so that $\mathcal{M}_\imath$ should be treated as 'third person'? When the nesting is from a 'third person' perspective, then the nested $\mathcal{M}_\imath$ is accessed as a separate instantiation of a perception-reaction. When the nesting involves 'first person' self reference, then the co-ordinates of $\mathcal{M}_\imath$ should be accessed in the caller's perception-reaction, similar to the 'call by name' parameter passing (à la Algol60). It is suggested that strong-self-reference à la (Perlis 1997), or core consciousness à la (Damasio 1999), is captured when $\mathcal{M}_\jmath$ discriminates that it should access the co-ordinates of $\mathcal{M}_\imath$ in its own perception-reaction, rather than copy something into a new stack frame. In ISAAC's terminology $\mathcal{M}_\jmath$ has a discrimination $\mathit{my\_core\_self\_now} \in \mathcal{I}_\jmath$, and the reaction involved is defined by $\mathcal{R}_\jmath(\mathit{my\_core\_self\_now})=$`call_by_name`. The discrimination, perhaps Perlis's *ur-quale* sensation, may be grounded in Damasio's proto-self, or in spatio-temporal distinctions. Its goal is to trigger a mode of 'parameter passing'. That happens prior to any actual access and evaluation of the nested constituents in $\mathcal{M}_\imath$, thus perhaps capturing Perlis's suggestion that mere self modeling 'has a feel' which is prior to any specific experience.

Among other things, that form of 'parameter passing' *would not generate infinite data structures*. On the other hand, repeated reevalutation of the required co-ordinates, upon each access, could be inefficient and resource consuming, motivating a 'lazy' use of stored values where possible, relying on assumptions about things that rarely change. It has to be decided, ad hoc, whether to reevaluate the nested co-ordinates over and over again or just once, perhaps applying 'call-by-need'. What about 'mutually nesting' perception-reactions?: $\mathcal{M}_\jmath$ perceives $\mathcal{M}_\imath$ that perceives $\mathcal{M}_\jmath$ in return…In the natural context as well, this form of 'self consciousness' is not a comfortable one.

The ad-hoc distinction of access to co-ordinates in one's own perception-reaction is not an all-or-none capability. Intelligent, 'conscious', beings could occasionally make mistakes, too: If one 'remembers' that horses are typically brown, does one bother to 'reevaluate' that that one looks purple in the sunlight? This example is inspired, of course, by the impressionist painters who noticed that we often use a stored value where they would suggest reevaluation. In yet other examples reevaluation in one's own perception-reaction is applied where a stored value would have been more appropriate: (Perlis 1997) discusses (following Gopnik) an example mistake made by 3-year-olds, but adult humans can also get disoriented: When we empathize with another person, for instance, or try to figure out what is going on in other minds, we often substitute our own co-ordinates for the other, sometimes knowingly ('consciously'), but sometimes by mistake ('unconsciously'). Hence confusion between 'first person' and 'third person' could occasionally occur in otherwise conscious beings, too. (In (Arzi-Gonczarowski & Lehmann 1998b) it is shown how ISAAC methodically joins two perceptions/reactions using categorical coproducts and pushouts. That can be applied to model empathy that does not smudge boundaries.)

Ontological Distinctions The ontology that emerges avails distinction between various types of self modeling and the extent of metacognition that they may support: ($\imath$) Whether a perception-reaction $\mathcal{M}$ discriminates all of its co-ordinates $\langle \mathcal{E}, \mathcal{I}, \varrho, \mathcal{Z}, \mathcal{R} \rangle$, or just (portions of) some of them, could provide a distinction between a sense of the 'perceiving self', 'the reacting self', and so on. ($\imath\imath$) Reflective perception of the vitality values of own reactions, also the level of own vitality resources, may capture aspects of emotional awareness. If vitality is drained below a safety margin ('one is getting out of one's mind'), then one might react by giving up, or diverting to other activities that use more available resources, or taking a maintenance rest, and so on. ($\imath\imath\imath$) Metacognition at its best seems to require the capability to perceive-react to arrows and subgraphs as well, beyond one's own current vertex on the graph. Does a percepion-reaction feature (some) memory or anticipation of past or future paths? How far? That seems to coincide with what Damasio calls *extended consciousness* of an *autobiographical self*. Pro-active metacognition is about reacting to such perceptions by prediction and modification of future paths. At the theoretical limit, a metacognitive $\mathcal{M}$ would have the same bird's eye view as the proposed theory, and would be able to relate to the entire categorical graph of perceptions-reactions, describe and analyze itself and others, with their possible pasts and contingent futures, as sub-graphs.

An Example Chapter 12 in the book 'The Little Prince'[2]: *The next planet was inhabited by a tippler. This was a*

[2] By Antoine de Saint Exupéry, translated by Katherine Woods `www.angelfire.com/hi/littleprince/frames.html`

very short visit, but it plunged the little prince into deep dejections. 'What are you doing here?' he said to the tippler, whom he found settled down in silence before a collection of empty bottles and also a collection of full bottles. 'I am drinking,' replied the tippler, with a lugubrious air. 'Why are you drinking?' demanded the little prince. 'So that I may forget,' replied the tippler. 'Forget what?' inquired the little prince, who already was sorry for him. 'Forget that I am ashamed,' the tippler confessed, hanging his head. 'Ashamed of what?' insisted the little prince, who wanted to help him. 'Ashamed of drinking!' The tippler brought his speech to an end, and shut himself up in an impregnable silence. And the little prince went away, puzzled. The grown-ups are certainly very, very odd, he said to himself, as he continued on his journey. The tippler perceives himself drinking, reacts by being ashamed, to which he reacts by further drinking. He has a perception-reaction about his current state, but he is trapped in a vicious circle. One metacognitive way out of that would be a perception-reaction that relates to arrows as well: It would then be possible to perhaps introduce and follow an appropriate arrow where `tipple-to-forget`↦`stop-shameful-behavior`. (If nothing else changes, then the disposition condition is preserved.)

summary

ISAAC is a theoretical mathematical model of cognition, a high-level schema that could be implemented computationally. The rigorous and general formalism can be nested in itself, letting cognitions cognize about cognitions, others as well as themselves. A 'first person' perspective may be modeled using a 'call-by-name' of the nested construct, which also avoids the generation of infinite data structures. However, similar to natural metacognition, confusion between 'first person' and 'third person', also infinite regress and vicious circles, could occasionally occur.

References

Aczel, P. 1987. *Lectures in Nonwellfounded Sets*. Number 9 in CSLI Lecture notes. CSLI.

Allen, J. F. 1998. AI growing up – the changes and opportunities. AI *Magazine* 19(4):13–23.

Arzi-Gonczarowski, Z., and Lehmann, D. 1998a. From environments to representations–a mathematical theory of artificial perceptions. *Artificial Intelligence* 102(2):187–247.

Arzi-Gonczarowski, Z., and Lehmann, D. 1998b. Introducing the mathematical category of artificial perceptions. *Annals of Mathematics and Artificial Intelligence* 23(3,4):267–298.

Arzi-Gonczarowski, Z. 1999a. Categorical tools for perceptive design: Formalizing the artificial inner eye. In Gero, J. S., and Maher, M. L., eds., *Computational Models of Creative Design IV*. University of Sydney, Australia: Key Centre of Design Computing and Cognition. 321–354.

Arzi-Gonczarowski, Z. 1999b. Perceive this as that - analogies, artificial perception, and category theory. *Annals of Mathematics and Artificial Intelligence* 26(1-4):215–252.

Arzi-Gonczarowski, Z. 2001a. Perceptions that perceive themselves – a mathematical schema. *IJCAS: International Journal of Computing Anticipatory Systems* 8:33–51.

Arzi-Gonczarowski, Z. 2001b. Self, empathy, manipulativity: Mathematical connections between higher order perception, emotions, and social cognition. In Cañamero, D., ed., *Emotional and Intelligent II – The Tangled Knot of Social Cognition*. AAAI Press Technical Report FS-01-02. 9–14.

Arzi-Gonczarowski, Z. 2002. AI emotions: Will one know them when one sees them? In *Cybernetics and Systems 2002*, volume 2, 739–744. Austrian Society for Cybernetic Studies, Vienna.

Arzi-Gonczarowski, Z. 2004a. From embodiments back to their models: An affective abstraction. In Schultz, A. C.; Breazeal, C.; Anderson, J. R.; and Trafton, G., eds., *The Intersection of Cognitive Science and Robotics: From Interfaces to Intelligence*. AAAI Press Technical Report FS-04-05. 76–81.

Arzi-Gonczarowski, Z. 2004b. Home page: Papers. http://www.actcom.co.il/typographics/zippie.

Ballonoff, P. 2000. On the evolution of self-awareness. In *Cybernetics and Systems 2000*, volume 1, 347–352. Austrian Society for Cybernetic Studies, Vienna.

Baron-Cohen, S. 1995. *Mindblindness*. MIT Bradford.

Barr, M., and Wells, C. 1995. *Category Theory for Computing Science*. Prentice Hall.

Barwise, J., and Etchemendy, J. 1987. *The Liar*. Oxford University Press.

Barwise, J., and Moss, L. 1991. Hypersets. *The Mathematical Intelligencer* 13(4):31–41.

Brooks, R. A. 1999. *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.

Damasio, A. R. 1999. *The Feeling of What Happens*. Harcourt Brace & Company.

Frijda, N. H. 1986. *The Emotions*. Cambridge: Cambridge University Press.

Gibson, J. J. 1977. The theory of affordances. In Shaw, R., and Bransford, J., eds., *Perceiving, Acting, and Knowing*. New-York: Wiley. 67–82.

Gould, S. J., and Vrba, E. 1982. Exaptation: A missing term in evolutionary theory. *Paleobiology* 8:4–15.

Humphrey, N. 1984. *Consciousness Regained*. Oxford University Press.

Perlis, D. 1997. Consciousness as self-function. *Journal of Consciousness Studies* 4(5/6):509–525.

Piaget, J. 1926. *La Representation du Monde chez l'Enfant*. Paris: Presses Universitaires de France.

Sloman, A. 2000. Architectural requirements for human-like agents, both natural and artificial (what sorts of machines can love?). In Dautenhahn, K., ed., *Human Cognition and Social Agent Technology*. John Benjamins Publishing. 163–195.