A Formal Model of Cognitive Synergy

Ben Goertzel(✉)

OpenCog Foundation, Sha Tin, Hong Kong
ben@goertzel.org

Abstract. "Cognitive synergy"– a dynamic in which multiple cognitive processes, cooperating to control the same cognitive system, assist each other in overcoming bottlenecks encountered during their internal processing. – has been posited as a key feature of real-world general intelligence, and has been used explicitly in the design of the OpenCog cognitive architecture. Here category theory and related concepts are used to give a formalization of the cognitive synergy concept. Cognitive synergy is proposed to correspond to a certain inequality regarding the relative costs of different paths through certain commutation diagrams. Applications of this notion of cognitive synergy to particular cognitive phenomena, and specific cognitive processes in the PrimeAGI design, are discussed.

1 Introduction

In [4] one possible general principle of computationally feasible general intelligence was proposed – the principle of "cognitive synergy." The basic concept of cognitive synergy, as presented there, is that general intelligences must contain different knowledge creation mechanisms corresponding to different sorts of memory (declarative, procedural, sensory/episodic, attentional, intentional); and that these different mechanisms must be interconnected in such a way as to aid each other in overcoming memory-type-specific combinatorial explosions.

In this paper, cognitive synergy is revisited and given a more formal description in the language of category theory. This formalization is a presented both for the conceptual clarification it offers, and as a hopeful step toward proving interesting theorems about the relationship between cognitive synergy and general intelligence, and evaluating the degree of cognitive synergy enabled by existing or future concrete AGI designs. The relation of the formal notion of cognitive synergy presented to the OpenCog/PrimeAGI design developed by the author and colleagues [4,5] is discussed in moderate detail, but this is only one among many possible examples; the general ideas proposed here should be applicable to a broad variety of AGI designs.

This paper relies on concepts and terms introduced in the prequel paper [2], which outlines a series of formal models of generally intelligent agents.[1]

[1] The preprint [3] contains the present paper and the sequel, plus a bit of additional material.

T. Everitt et al. (Eds.): AGI 2017, LNAI 10414, pp. 13–22, 2017.
DOI: 10.1007/978-3-319-63703-7_2

2 Theory of Stuckness

In the **PGMC Agent** model introduced in [2], one has a collection of Cognitive Control Processes (CCPs) working together to update a representational hypergraph, and guiding their cognitive activities via probabilistic pattern mining of prior cognitive activities on the hypergraph. Within this framework, we now introduce a series of concepts that will allow us to formalize what it means for a group of CCPs to interact synergetically.

In a real-world cognitive system, each CCP will have a certain limited amount of resources, which it can either use for its own activity, or transfer to another cognitive process. In OpenCog, for instance, space and time resources tend to be managed somewhat separately, which would mean that a pair of floats would be a reasonable representation of an amount of resources. For our current theoretical purposes, however, the details of the resource representation don't matter much.

Let us say that a CCP, at a certain point in time, is "stuck" if it does not see any high-confidence, high-probability transitions associated with its own corresponding cognitive process, from current state h-patterns to future state h-patterns that have significantly higher goal-achievement values. If a CCP is stuck, then it may not be worthwhile for the CCP to spend its limited resources taking any action at that point. Or, in some cases, it may be the best move for that CCP to transfer some of its allocated resources so some other cognitive process. This leads us straight on to cognitive synergy. But before we go there, let us pause to get more precise about how "getting stuck" should be interpreted in this context.

A Formal Definition of Stuckness. Let G_A denote the CPT graph corresponding to cognitive process A. This is a subgraph of the overall cognitive process transition graph of the system, and it may be considered as a category unto itself, with object being the subgraphs, and a Heyting algebra structure.

Given a particular situation S ("possible world") involving the system's cognition, and a time interval I, let e.g. $G_A^{S,I}$ denote the CPT graph of A during time interval I, insofar as it exists explicitly in the system (not just in the meta-system).

Where P is a h-pattern in the system, and (S, I) is a situation/time-interval pair, let $P(S, I)$ denote the degree to which the system displays h-pattern P in situation S during time-interval I. Let $g(S, I)$ denote the average degree of goal-achievement of the system in situation S at time during time interval I. Then if we identify a set $\mathcal{I}$ of time-intervals of interest, we can calculate

$$g(P) = \frac{\sum_{(S,I),I\in\mathcal{I}} g(S,I)P(S,I)}{\sum_{(S,I),I\in\mathcal{I}} P(S,I)}$$

to be the degree to which P implies goal-achievement, in general (relative to $\mathcal{I}$; but if this set of intervals is chosen reasonably, this dependency should not be sensitive).

On the other hand, it is more interesting to look at the degree to which P implies goal-achievement across the possible futures of the system as relevant in a particular situation at a particular point in time. Suppose the system is currently in situation S, during time interval I_S. Then $\mathcal{I}$ may be defined, for instance, as a set of time intervals in the near future after I_S. One can then look at

$$g_{S,I_S,\mathcal{I}}(P) = \frac{\sum_{(S',I),I\in\mathcal{I}} g(S',I)P(S',I)Prob((S',I)|(S,t))}{\sum_{(S',I),I\in\mathcal{I}} P(S',I)Prob((S',I)|(S,t))}$$

which measures the degree to which P implies goal-achievement in situations that may occur in the near future after being in situation S. The confidence of this value may be assessed as

$$c_{S,I_S,\mathcal{I}}(P) = f(\sum_{(S',I),I\in\mathcal{I}} P(S',I)Prob((S',I)|(S,t)))$$

where f is a monotone increasing function with range $[0,1]$. This confidence value is a measure of the amount of evidence on which the estimate $g_{S',I_S}(P)$ is based, scaled into $[0,1]$.

Finally, we may define $e_{C,I_R,S,I_S}(P,I,I_P)$ as the probability estimate that the CCP corresponding to cognitive process C holds for the proposition that: In situation S during time interval I_S, if allocated a resource amount in interval I_R for making the choice, C will make a choice leading to a situation in which $P(S,I) \in I_P$ during interval I (assuming I is after I_S). A confidence value $c_{C,I_R,S,I_S}(P,I,I_P)$ may be defined similarly to $c_{S',t}(P)$ above.

Given a set $\mathcal{I}$ of time intervals, one can define $e_{C,I_R,S,\mathcal{I}}(P,I,I_P)$ and $c_{C,I_R,S,\mathcal{I}}(P,I,I_P)$ via averaging over the intervals in $\mathcal{I}$.

The confidence with which C knows how to move forward toward the system's goals in situation S at time t may then be summarized as

$$\text{conf}_{C,S,I_S,\mathcal{I}} = max_P\,(g_{S',I_S,\mathcal{I}}(P)c_{S',I_S,\mathcal{I}}(P)e_{C,I_R,S,\mathcal{I}}(P,I,I_P)c_{C,I_R,S,\mathcal{I}}(P,I,I_P))$$

with

$$\text{stuck}_{C,S,I_S,\mathcal{I}} = 1 - \text{conf}_{C,S,I_S,\mathcal{I}}$$

3 Cognitive Synergy: A Formal Exploration

What we need for "cognitive synergy" between A and B to exist, is for it to be the case that: For many situations S and times t, exactly one of A and B is stuck.

In the metasystem, records of cases where one or both of A or B were stuck, will be recorded as hypergraph patterns. The set of (S,t) pairs in the metasystem where exactly one of A and B was stuck to a degree of stuckness in interval $I_d = (L,U)$, has a certain probability in the set of all (S,t) pairs in the metasystem. Let us call this set stuck_{A,B,I_d}.

The set $G^{\text{stuck}}_{A,B,I_d}$ of CPT graphs $G_A^{S,t}$, $G_B^{S,t}$ corresponding to the (S,t) pairs in stuck_{A,B,I_d} can also be isolated in the metasystem, and has a certain probability considered as a subgraph of the metasystem (which can be calculated according to the intuitionistic graph probability distribution). An overall index of cognitive synergy between A and B can then be calculated as follows.

Let $\mathcal{P}$ be a partition of $[0, 1]$ (most naturally taken equispaced). Then,

$$\text{cog-syn}_{A,B,\mathcal{P}} = \frac{\sum_{I_d \in \mathcal{P}} w_{I_d} Prob(G^{\text{stuck}}_{A,B,I_d})}{\sum_{I \in \mathcal{P}} w_{I_d}}$$

is a quantitative measure of the amount of cognitive synergy between A and B.

Extension of the above definition to more than two cognitive processes is straightforward. Given N cognitive processes, we can look at pairwise synergies between them, and also at triple-wise synergies, etc. To define triplewise synergies, we can look at stuck_{A,B,C,I_d}, defined as the set of (S, I) where all but one of the three cognitive processes A, B and C is stuck to a degree in I_d. Triplewise synergies correspond to cases where the system would be stuck if it had only two of the three cognitive processes, much more often than it's stuck given that it has all three of them.

This may seem a somehow anticlimactic formalization of such an exciting-sounding quality as "cognitive synergy." However, exciting higher-level emergent phenomena often occur as a result of more prosaic-looking lower-level interactions. Mutual exclusion regarding where two cognitive processes get stuck, at the micro-level of very small cognitive steps, is what enables the two cognitive processes to work together creatively (including helping each other become unstuck) at the meso-level of slightly bigger cognitive steps.

3.1 Cognitive Synergy and Homomorphisms

The existence of cognitive synergy between two cognitive processes will depend sensitively on how these cognitive processes actually work. However, there are likely some general principles at play here. For instance we suggest

Conjecture 1. *In a PGMC agent operating within feasible resource constraints: If two cognitive processes A and B have a high degree of cognitive synergy between them, then there will tend to be a lot of low-cost homomorphisms between subgraphs of $G_A^{S,t}$ and $G_B^{S,t}$, but* **not** *nearly so many low-cost isomorphisms.*

The intuition here is that, if the two CPT graphs are too close to isomorphic, then they are unlikely to offer many advantages compared to each other. They will probably succeed and fail in the same situations. On the other hand, if the two CPT graphs don't have some resemblance to each other, then often when one cognitive process (say, A) gets stuck, the other one (say, B) won't be able to use the information produced by A during its work so far, and thus won't be able to proceed efficiently. Productive synergy happens when one has two processes, each of which can transform the other one's intermediate results, at somewhat low cost, into its own internal language – but where the internal languages of the two processes are not identical.

Our intuition is that a variety of interesting rigorous theorems likely exist in the vicinity of this informal conjecture. However, much more investigation is required.

Along these lines, recall Conjecture 1 above that most cognitive processes useful for human-like cognition, are implemented in terms of rules that are mostly homomorphisms or inverse homomorphisms. To the extent this is the case, it fits together very naturally with Conjecture 1.

Suppose $G_A^{S,t}$ and $G_B^{S,t}$ each consist largely of records of enacting a series of hypergraph homomorphisms (followed by weight updates), as Conjecture 1 posits. Then one way Conjecture 2 would happen would be if the homomorphisms in $G_A^{S,t}$ mapped homomorphically into the homomorphisms in $G_B^{S,t}$. That is, if we viewed $G_A^{S,t}$ and $G_B^{S,t}$ as their own categories, the homomorphisms posited in Conjecture 2 would take the form of functors between these two categories.

3.2 Cognitive Synergy and Natural Transformations

Further interesting twists emerge if one views the cognitive process A as associated with a functor F_A that maps G^S into $G_A^S \subseteq G^S$, which has the property that it maps $G^{S,t}$ into $G_A^{S,t} \subseteq G^{S,t}$ as well. The functor F_A maps a state transition subgraph of S, into a state transition subgraph involving only transitions effected by cognitive process A. So for instance, if X represents a sequence of cognitive operations and conclusions that have transformed the state of the system, then $F_A(X)$ represents the closest match to X in which all the cognitive operations involved are done by cognitive process A. The cost of $F_A(X)$ may be much higher than the cost of X, e.g. if X involves vision processing and A is logical inference, then in $F(X)$ all the transitions involved in vision processing need to be effected by logical operations, which is going to be much more expensive than doing them in other ways.

A natural transformation $\eta^{A,B}$ from F_A to F_B associates to every object X in G^S (i.e., to every subgraph of the transition graph G^S of the system S) a morphism $\eta_X^{A,B} : F_A(X) \to F_B(X)$ in G^S so that: for every morphism $f : X \to Y$ in G^S (i.e. every homomorphic transformation from state transition subgraph X to state transition subgraph Y) we have $\eta_Y^{A,B} \circ F_A(f) = F_B(f) \circ \eta_X^{A,B}$.

This leads us on to our final theoretical conjecture:

Conjecture 2. *In a PGMC agent operating within feasible resource constraints, suppose one has two cognitive processes A and B, which display significant cognitive synergy, as defined above. Then,*

1. *there is likely to be a natural transformation $\eta^{A,B}$ between the functor F_A and the functor F_B – and also a natural transformation $\eta^{B,A}$ going in the opposite direction*

2. *the two different routes from the upper left to the bottom right of the commutation diagram corresponding to* $\eta^{A,B}$*,*

$$\begin{tikzcd} F_A(X) \arrow[r, "F_A(f)"] \arrow[d, "\eta_X^{A,B}"'] & F_A(Y) \arrow[d, "\eta_Y^{A,B}"] \\ F_B(X) \arrow[r, "F_B(f)"'] & F_B(Y) \end{tikzcd} \tag{1}$$

will often have quite different total costs

3. *Referring to the above commutation diagram and the corresponding diagram for* $\eta^{B,A}$*,*

$$\begin{tikzcd} F_B(X) \arrow[r, "F_B(f)"] \arrow[d, "\eta_X^{B,A}"'] & F_B(Y) \arrow[d, "\eta_Y^{B,A}"] \\ F_A(X) \arrow[r, "F_A(f)"'] & F_A(Y) \end{tikzcd} \tag{2}$$

   – *often it will involve significantly less total cost to*
     – *travel from* $F_A(X)$ *to* $F_B(Y)$ *via the left-bottom path in Eq. 2, and then from* $F_B(Y)$ *to* $F_A(Y)$ *via the right side of Eq. 2; than to*
     – *travel from* $F_A(X)$ *to* $F_A(Y)$ *directly via the top of Eq. 2*

   That is, often it will be the case that

$$\begin{aligned} &\mathit{cost}(F_A(X) \xrightarrow{\eta_X^{A,B}} F_B(X)) + \mathit{cost}(F_B(X) \xrightarrow{F_B(f)} F_B(Y)) \\ &+\mathit{cost}(F_B(Y) \xrightarrow{\eta_Y^{B,A}} F_A(Y) < \mathit{cost}(F_A(X) \xrightarrow{F_A(f)} F_A(Y)) \end{aligned} \tag{3}$$

Inequality (3) basically says that, given the cost weightings of the arrows, it may sometimes be significantly more efficient to get from $F_A(X)$ to $F_A(Y)$ via an indirect route involving cognitive process B, than to go directly from $F_A(X)$ to $F_A(Y)$ using only cognitive process A. This is a fairly direct expression of the cognitive synergy between A and B in terms of commutation diagrams.

To make this a little more concrete, suppose X is a transition graph including the new conclusion that Bob is nice, and Y is a transition graph including additionally the even newer conclusion that Bob is helpful. Then f represents a homomorphism mapping X into Y, via – in one way or another – adding to the system's memory the conclusion that Bob is helpful. Suppose A is a cognitive process called "inference" and B is one called "evolutionary learning." Then e.g. $F_A(X)$ refers to a version of X in which all conclusions are drawn by inference, and $F_B(Y)$ refers to a version of Y in which all conclusions are drawn by evolutionary learning. The commutation diagram for $\eta^{A,B} = \eta^{\text{inference,evolution}}$, then looks like

$$\begin{tikzcd}
F_{\text{inference}}(\text{BobNice}) \arrow[r, "F_{\text{inference}}(f_{\text{nice}\to\text{helpful}})"] \arrow[d, "\eta^{\text{inference,evolution}}_{\text{BobNice}}"'] & F_{\text{inference}}(\text{BobHelpful}) \arrow[d, "\eta^{\text{inference,evolution}}_{\text{BobHelpful}}"] \\
F_{\text{evolution}}(\text{BobNice}) \arrow[r, "F_{\text{evolution}}(f_{\text{nice}\to\text{helpful}})"'] & F_{\text{evolution}}(\text{BobHelpful})
\end{tikzcd} \tag{4}$$

and the commutation diagram for $\eta^{\text{evolution,inference}}$ looks like

$$\begin{tikzcd}
F_{\text{evolution}}(\text{BobNice}) \arrow[r, "F_{\text{evolution}}(f_{\text{nice}\to\text{helpful}})"] \arrow[d, "\eta^{\text{evolution,inference}}_{\text{BobNice}}"'] & F_{\text{evolution}}(\text{BobHelpful}) \arrow[d, "\eta^{\text{evolution,inference}}_{\text{BobHelpful}}"] \\
F_{\text{inference}}(\text{BobNice}) \arrow[r, "F_{\text{inference}}(f_{\text{nice}\to\text{helpful}})"'] & F_{\text{inference}}(\text{BobHelpful})
\end{tikzcd} \tag{5}$$

The conjecture states that, for cognitive synergy to occur, the cost of getting from $F_{\text{inference}}(\text{BobNice})$ to $F_{\text{inference}}(\text{BobHelpful})$ directly via the top arrow of Eq. 4 would be larger than the cost of getting there via the left and then bottom of Eq. 4 followed by the right of Eq. 5. That is to get from "Bob is nice" to "Bob is helpful", where both are represented in inferential terms, it may still be lower-cost to map "Bob is nice" into evolutionary-programming terms, then use evolutionary programming to get to the evolutionary-programming version of "Bob is helpful", and then map the answer back into inferential terms.

4 Some Core Synergies of Cognitive Systems: Consciousness, Selves and Others

The paradigm case of cognitive synergy is where the cognitive processes A and B involved are learning, reasoning or pattern recognition algorithms. However, it is also interesting and important to consider cases where the cognitive processes involved correspond to different scales of processing, or different types of subsystem of the same cognitive system. For instance, one can think about:

- A = long-term memory (LTM), B = working memory (WM)
- A = whole-system structures and dynamics, B = the system's self-model
- A and B are different"sub-selves" of the same cognitive system
- A is the system's self-model, and B is the system's model of another cognitive system (another person, another robot, etc.)

Conjecturally and intuitively, it is natural to hypothesize that

- Homomorphisms between LTM and WM are what ensure that ideas can be moved back and forth from one sort of memory to another, with a loss of detail but not a total loss of essential structure
- Homomorphisms between the whole system's structures and dynamics (as represented in its overall state transition graph) and the structures and dynamics in its self-model, are what make the self-model structurally reflective of the whole system, enabling cognitive dynamics on the self-model to be mapped meaningfully (i.e. morphically) into cognitive dynamics in the whole system, and vice versa
- Homomorphisms between the whole system in the view of one subself, and the whole system in the view of an other subself, are what enable two different subselves to operate somewhat harmoniously together, controlling the same overall system and utilizing the knowledge gained by one another
- Homomorphisms between the system's self-model and its model of another cognitive system, enable both theory-of-mind type modeling of others, and learning about oneself by analogy to others (critical for early childhood learning)

Cognitive synergy in the form of natural transformations between LTM and WM means that when unconscious LTM cognitive processing gets stuck, it can push relevant knowledge to WM and sometimes the solution will pop up there. Correspondingly, when WM gets stuck, it can throw the problem to the unconscious LTM processing, and hope the answer is found there, later to bubble up into WM again (the throwing down being according to a homomorphic mapping, and the bubbling up being according to another homomorphic mapping). As WM is closely allied with what is colloquially referred to as "consciousness" [1] – meaning the reflective, deliberative consciousness that we experience when we reason or reflect on something in our "mind's eye" – this particular synergy appears key to human conscious experience. As we move thoughts, ideas and feelings back and forth between our focus of attention and the remainder of our mind and memory, we are experiencing this synergy intensively on an everyday basis – or so the present hypothesis suggests; i.e. that

- When we pull a memory into attention, or push something out of attention into the "unconscious", we are enacting homomorphisms on our mind's state transition graph.
- When the unconscious solves a problem that the focus of attention pushed into it, and then the answer comes back into the attentional focus and gets

deliberatively reasoned on more, this is the action of the natural transformation between unconscious and conscious cognitive processes – it's a case where the cost of going the long way around the commutation diagram from conscious to unconscious and back, was lower than the cost of going directly from conscious premise to conscious conclusion.

Cognitive synergy in the form of natural transformations between system and self mean that when the system as a whole cannot figure out how to do something, it will map this thing into the self-model (via a many-to-one homomorphism, generally, as the capacity of the self-model is much smaller), and see if cognitive processes acting therein can solve the problem. Similarly, if thinking in terms of the self-model doesn't resolve a solution to the problem, then sometimes "just doing it" is the right approach – which means mapping the problem the self-model's associated cognitive processes are trying to solve back to the whole system, and letting the whole system try its mapped version of the problem by any means it can find.

Cognitive synergy in the form of natural transformations between subselves means that when one subself gets stuck, it may map the problem into the cognitive vernacular of another subself and see what the latter can do. For instance if one subself, which is very aggressive and pushy, gets stuck in a personal relationship issue, it may map this issue into the world-view of another more agreeable and empathic and submissive subself, and see if the latter can find a solution to the problem. Many people navigate complex social situations via this sort of ongoing switching back and forth between subselves that are well adapted to different sorts of situations [6].

Cognitive synergy in the form of natural transformations between self-model and other-model means that when one get stuck in a self-decision, one can implicitly ask "what would I do if I were this other mind?"..."what would this other mind do in this situation?" It also means that, when one can't figure out what another mind is going to do via other routes, one can map the other mind's situation back into one's self-model, and ask "what would I do in their situation?"..."what would it be like to be that other mind in this situation?"

In all these cases, we can see the possibility of much the same sort of process as we conjecture to exist between two cognitive processes like evolutionary learning and logical inference. We have different structures (memory subsystems, models of various internal or external systems, systematic complexes of knowledge and behavior, etc.) associated with different habitual sets of cognitive processes. Each of these habitual sets of processes may get stuck sometimes, and may need to call out to others for help in getting unstuck. This sort of request for help is going to be most feasible if the problem can be mapped into the cognitive world of the helper in a way that preserves its essential structure, even if not all its details; and if the answer the helper finds is then mapped back in a similarly structure-preserving way.

Real-world cognitive systems appear to consist of multiple subsystems that are each more effective at solving certain classes of problems – subsystems like particular learning and reasoning processes, models of self and other, memory systems of differing capacity, etc. A key aspect of effective cognition is the ability for these various subsystems to ask each other for help in very granular ways, so that the helper can understand something of the intermediate state of partial-solution that the requestor has found itself in. This sort of "cognitive synergy" seems to be reflected, in an abstract sense, in certain "algebraic" or category-theoretic symmetries such as we have highlighted here.

References

1. Goertzel, B.: Characterizing human-like consciousness: an integrative approach. Procedia Comput. Sci. **41**, 152–157 (2014)
2. Goertzel, B.: From abstract agents models to real-world AGI architectures: bridging the gap. In: Proceedings of AGI 2017. Springer, Heidelberg (2017, submitted)
3. Goertzel, B.: Toward a formal model of cognitive synergy (2017). https://arxiv.org/abs/1703.04361
4. Goertzel, B., Pennachin, C., Geisweiller, N.: Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy. Atlantis Thinking Machines, New York (2013). Springer
5. Goertzel, B., Pennachin, C., Geisweiller, N.: Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI. Atlantis Thinking Machines, New York (2013). Springer
6. Rowan, J.: Subpersonalities: The People Inside Us. Routledge Press, New York (1990)

Springer
http://www.springer.com/978-3-319-63702-0

Artificial General Intelligence
10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings
Everitt, T.; Goertzel, B.; Potapov, A. (Eds.)
2017, XI, 275 p. 38 illus., Softcover
ISBN: 978-3-319-63702-0