Towards a mathematical theory of consciousness

Comparing IIT axioms and categorical (universal) constructions

Steven Phillips[1], Naotsugu Tsuchiya[2, 3]

National Institute of Advanced Industrial Science and Technology AIST

[1]Human Informatics and Interaction Research Institute, AIST, Tsukuba, JAPAN

[2]School of Psychological Sciences, Turner Institute, Monash University, Melbourne, Australia

[3]Laboratory of Qualia Structure, ATR, Kyoto, Japan

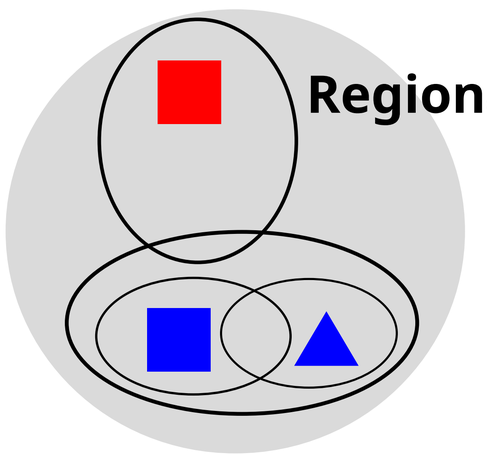
Slogan: Consciousness is a *universal construction*

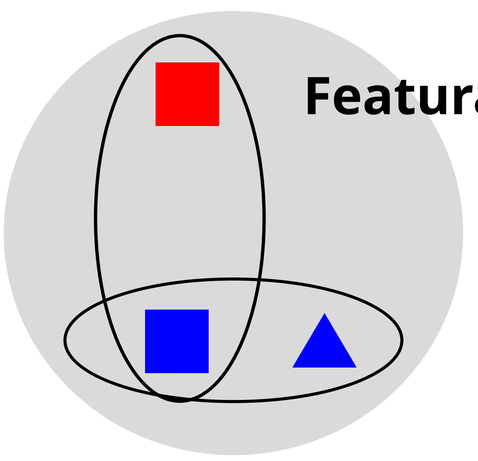
Synopsis: A category theory view of IIT axioms

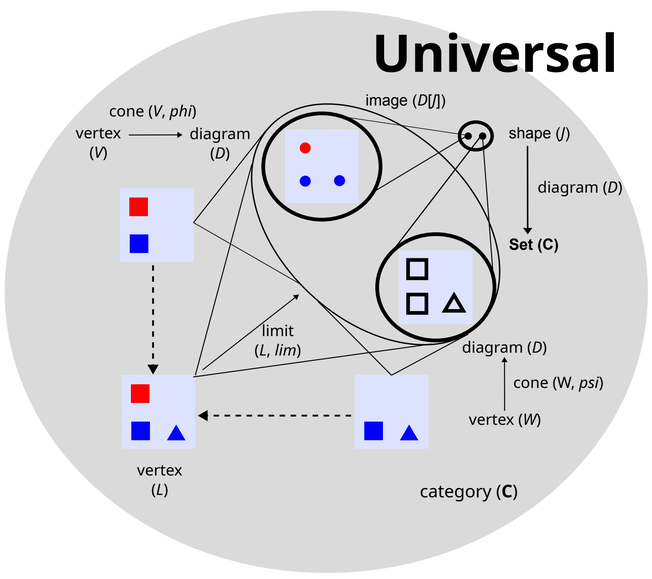
- **Background** Axioms are necessary/sufficient for conscious experience (Albantakis, et al, 2023)
  - (0) existence, (1) intrinsicality, (2) information, (3) integration, (4) exclusion, (5) composition
- **Problem** Informal; questionable status as axioms or constraints on theory (Bayne, 2018)
- **Method** Comparison with *category theory* (see, e.g., Leinster, 2014)
- **Result** Axioms comparable to category theory (universal) construction, called *(co)limit*
- **Discussion** Aspects of consciousness from *universal mapping properties* (UMPs)

Composition (binding)

"... composed of *distinctions* and *relations* that bind them..."

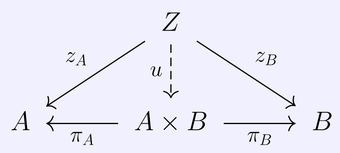






Intersection (of regions) is a product

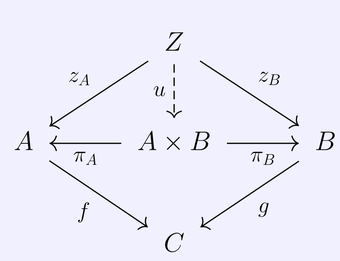
A *product* of A and B is an object P, denoted $A \times B$, and maps π_A and π_B such that for every object Z and maps z_A and z_B there **exists a unique** map u making the diagram commute:



The product of sets U and V in $\mathbf{Set}^{\subseteq}$ is their intersection, $U \cap V$.

Feature (colour-shape) binding is a pullback

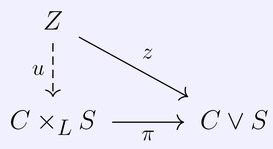
A *pullback* of $f : A \to C$ and $g : B \to C$ is an object P, denoted $A \times_C B$, and maps π_A and π_B such that for every object Z and maps z_A and z_B there **exists a unique** map u making diagram



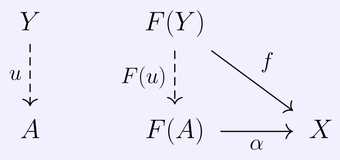
commute. The pullback of colour and shape maps, $cl : C \to L$ and $sh : S \to L$ is set $C \times_L S = \{(c, s) | cl(c) = sh(s)\}$ and (π_C, π_S).

Products/Pullbacks (limits) are universal constructions

A pullback (limit) is a *universal cone* to a $\vee$ (J)-shaped *diagram* (functor) $D : J \to \mathbf{C}$, e.g., colour-shape is limit to $C \vee S$:



A *universal morphism* from functor F to object X is an object A and a map α such that for every object Y and map f there **exists a unique** map u making the diagram commute:



A limit is a universal morphism from diagonal functor Δ to D.

Glossary of some comparable terms

IIT	**CT** (category theory)
system/substrate	category (cf. directed graph)
unit(s)	object (set)
cause-effect purview	arrow ($f : A \to B$)
purview	subobject ($A \subseteq B$)
self-relation	identity ($1_A : A \to A$)
composition	product ($A \times B, \pi$)
composition (dual)	coproduct ($A + B, \iota$)
relation	pullback ($A \times_C B, \pi$)
relation face	pullback object ($A \times_C B$)
min-max principles	unique-existence property (UMP)
unfolding	coalgebra (corecursion)

UMP: unique-existence condition; (co)products and pullbacks are (co)limits, i.e. universal constructions satisfying UMP.

0. Existence: limit

IIT "Experience *exists*: there is *something*."

CT Conscious experience corresponds to existence of limit.

1. Intrinsicality: subject category (C)

IIT "Experience is *intrinsic*: it exists for *itself*.": first-person and independent of external observers.

CT Consciousness as a category of subjective elements (objects) and their relations (arrows).

2. Information: objects and arrows in C

IIT "Experience is *specific*: it is this *one*." Consciousness picks out one experience from many other possible experiences.

CT A diagram (functor) picks out a specific collection of objects and arrows in that category.

3. Integration: vertex (*L*)

IIT "Experience is *unitary*: it is a *whole*, irreducible to separate experiences.": cannot split as independent left and right experiences.

CT Limit cannot be reduced to other objects/arrows without failing to be a universal construction.

4. Exclusion: not in image of D

IIT "Experience is *definite*: it is this *whole*.": excludes all other phenomena that we could have experienced at that time.

CT The limit is universal with respect to objects and arrows in the image of the diagram.

5. Composition: vertex (*L*) and legs (*lim*)

IIT "Experience is *structured*: it is composed of *distinctions* and the *relations* that bind them together, yielding a *phenomenal structure* that feels *the way it feels*."

CT A limit is composed of an object and arrows satisfying a certain universal mapping property.

Discussion (consciousness and (co)limits)

- existence: start (**IIT**) vs "end" (**CT**) point
  - **CT**: *terminal object* in comma category
- integrate/unify: synonym (**IIT**) vs duality (**CT**)
  - **CT**: integrate (colimit: $A \to C \leftarrow B$) vs unify (limit: $A \leftarrow C \to B$)
- predictions pertain to those of (co)limits: e.g., constituent independence
  - cf. compositionality and *Language of Thought* (Fodor, 1975; Huan & Tononi, 2024; Phillips, 2024; Phillips & Wilson, 2010; Quilty-Dunn, et al, 2023)
  - conceptual $\leftrightarrows$ phenomenal structure?
- computation (corecursion): search for limit (terminal) in comma category

References

Albantakis, et al (2023). Integrated information theory (IIT) 4.0. *PLoS Computational Biology, 19(10)*, e1011465.

Bayne, T. (2018). On the axiomatic foundations of IIT. *Neuroscience of Consciousness, 4(1)*.

Fodor, J. A. (1975). The language of thought. New York, NY: Crowell.

Huan & Tononi (2024). The unfathomable richness of seeing.

Leinster (2014). *Basic category theory*. Cambridge University.

Phillips (2024). A category theory perspective on Language of Thought: LoT is universal. *Frontiers in Psychology, 15*.

Phillips & Wilson (2010). Categorial compositionality: A category theory explanation for the systematicity of human cognition. *PLoS Computational Biology, 6(7)*, e1000858.

Quilty-Dunn, et al (2023). The best game in town: The re-emergence of the Language of Thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences, 46*, e261.

Funding

Japanese Society for the Promotion of Science Grant-in-aid (23H04829, 23H04830).