

Categorical AI phenomenology: A first-person approach

Robert Prentner*

*Institute of Humanities, ShanghaiTech University,
No. 393 Middle Huaxia Road, Shanghai, 201120, China*
✉

*Association for Mathematical Consciousness Science,
Geschwister-Scholl-Platz 1, Munich, 80539, Germany*
robert.prentner@amcs.science

This paper develops a phenomenology-first approach to artificial consciousness by reframing consciousness as the subjective experience enacted through an agent’s interface with the world. We shift the methodological focus to first-person structures, modeled mathematically by categories derived from Q -networks to capture actions and phenomenological invariants. In this framework, Q -networks are conceptualized as relational interfaces encoding agent-world interaction, analogous to how the dynamical states of a computer depend on its sensory inputs, previous states, and actions.

Our work provides a rigorous framework for *interface consciousness* to describe computational systems that embed information-processing into phenomenological structure. The approach aligns with 4E approaches to cognition by emphasizing enactive, embedded, and extended dimensions of experience. The paper thus offers a principled, relational, and phenomenological account of artificial phenomenology grounded in categorical mathematics.

Keywords: AI consciousness; (applied) category theory; phenomenological interfaces; Q -networks; structural unification

1. First-person computation

What does it mean to say that computers could be conscious? Can a machine subjectively experience its surroundings and inner workings? And what is the best way to study this phenomenon? [Association for Mathematical Consciousness Science, 2023; Gamez, 2008; Dehaene *et al.*, 2017; Bach, 2019; Kanai *et al.*, 2019; Ruffini & Lopez-Sola, 2022; VanRullen & Kanai, 2021; Butlin *et al.*, 2023; Blum & Blum, 2024] We propose that these questions translate to the question of whether computers can have their *own* subjective perspective on the world. And to study this means doing *phenomenology from the machine’s point of view*.

Consciousness science differs from conventional scientific fields such as the study of chemical reaction networks or genetic regulation. The core difference is that

*ORCID-ID: 0000-0003-1890-0827; please cite the published article: R. Prentner (2026). Categorical AI Phenomenology: A First-Person Approach, *Journal of Artificial Intelligence and Consciousness*, <https://doi.org/10.1142/S2705078526500013>

studying consciousness, unlike the latter subjects, requires one to integrate a first-person perspective with conventional third-person science: how is it that it feels like anything for a physical system that processes information, such as a complex molecular system, a human brain, or a computer? This question gave rise to plenty of debates about the underlying metaphysical picture [Chalmers, 2002], which still continue today [Frankish, 2016; Goff, 2019; Friston *et al.*, 2020; Ellia *et al.*, 2021; Graziano, 2022; Prentner & Hoffman, 2024; Kuhn, 2024]. Some of these views seem to make artificial consciousness likely. Others would rule it out.

In this contribution, by contrast, we want to primarily focus on the *methodological issues* related to the question whether machines could be conscious. Our answer entails a re-conceptualization of consciousness, not as “inner theater”, but as the subjective experience enacted through an agent’s interface with the world. Conceivably, such interfaces could be instantiated by various kinds of computational processes, natural and artificial ones. This statement is made more precise in the course of the paper. Talking about agency, at least as it is related to “enacting a subjective experience,”¹ requires one to consider the first-person perspective. We thus specifically focus on those interfaces that enable an agent to organize computational processes according to phenomenological principles.

We can first take some inspiration from the scientific study of consciousness in neuroscience [Signorelli *et al.*, 2021a; Seth & Bayne, 2022; Albantakis, 2025]. It has been proven challenging to exclusively deal with third-person data (such as brain scans) and construct a meaningful science of consciousness. An alternative is to invert the methodology and instead start from the first-person perspective, that is, to begin with an understanding of how appearances are given to us, thus following a “phenomenology-first approach” to consciousness. One model in the neuroscience of consciousness, integrated information theory [Tononi *et al.*, 2016; Albantakis *et al.*, 2023], purports to do precisely this:² start with basic, almost self-evident and indubitable, introspective features of conscious experience, and infer how these are in turn represented in certain physical (neural) systems.

We argue that we have to go even deeper, beyond the confines of viewing introspection as if it were just a perception of “inner objects.” We need to let the “subject speak” and understand how consciousness shapes the experienced world around us. Arguably, the most systematic philosophical tools for this are provided by phenomenology [Gallagher & Zahavi, 2008; Yoshimi, 2016; Smith, 2018], the study of subjective experience as it was conceived by Edmund Husserl and others in the early 20th century.

¹Whether the concept of “agency” in fact obviates the possibility of machine consciousness is an important discussion that we do not wish to enter at this point, but see discussions in [Franklin & Graesser, 1997; Barandiaran *et al.*, 2009; Roli *et al.*, 2022; Dung, 2024; Floridi, 2025].

²Integrated information theory presents a notable methodological advance, but in our opinion the methodology needs to be pursued with even more candor.

At the same time, we also want to adopt a mathematical method [Prentner, 2025b]. Eventually, our goal is to integrate first-person studies with contemporary models of computation. By doing this, we inevitably reach beyond classical Husserlian phenomenology. Husserl, for example, emphasized that consciousness lacks the exactness of mathematical objects (Ideas I, §73-75), and his analyses were often descriptive rather than formal. Our proposal should therefore be understood as a philosophically-informed extension of these insights, not a reconstruction of Husserl’s own position together with a faithful translation into the language of computer science.

The specific mathematical framework chosen for this work is *category theory* [Lawvere & Schanuel, 2009; Mac Lane, 1998]. While highly abstract, category theory has already been applied to theoretical computer science and programming [Mileski, 2019], but also implications for the study of consciousness have been hinted at (some examples are listed in [Signorelli *et al.*, 2021c; Tull & Kleiner, 2021; Tsuchiya & Saigo, 2021; Taguchi & Saigo, 2023; Prentner, 2024]).

Any mathematical framework for studying consciousness should start from the fact that we undergo experiences and can relate some of those experiences. Our hypothesis is that subjectivity lives in the ever-changing relations between different experiences. It is the task of (mathematical) phenomenologists to elucidate how this process unfolds in a systematic way. Our specific choice of category theory is motivated by the idea that phenomenology is relational, which is illustrated by at least the following: (i) many traditional phenomenological concepts such as “intentionality” or “constitution” are inherently relational in spirit [Taguchi, 2019; Voltolini, 2024; Prentner, 2025b], often correlating a subjective act and their corresponding objects, and (ii) phenomenological analyses often proceed in terms of relating potential and actual experiences that underlie our manifest (i.e., introspectively available) experience [Yoshimi, 2016; Prentner, 2025b]. Such an approach to phenomenology can be distinguished from a purely descriptive one that equates the study of consciousness with introspection.

The key contribution of this paper is to introduce *Q*-networks as a minimal formalism for modeling first-person structures of experience,³ thereby providing a rigorous framework for artificial phenomenology. Our project can be seen as an elaboration on previous work on “phenomenal spaces” [Prentner, 2019] to mathematically treat spaces of experience, enriched by insights from above-mentioned work on category theory in consciousness studies and phenomenology.

Q-networks and their relation to category theory are discussed in more detail in in section 2, where it is proposed to derive categorical structures from them for phenomenological analyses. We briefly discuss the relation of this approach to 4E cognition and, more generally, to the project of “interface consciousness” in section

³It is important to clarify that “*Q*-networks” as used here are not specifically related to Q-learning or Q-function approximators as commonly understood in the reinforcement learning community (see, e.g., Sutton & Barto [2018]).

3 (including a computational toy model), before we revisit phenomenology in section 4 to discuss further organizational principles of Q -network-derived structures. We then conclude in section 5.

The “phenomenological hierarchy” laid out in this paper can be inferred from Fig. 1.

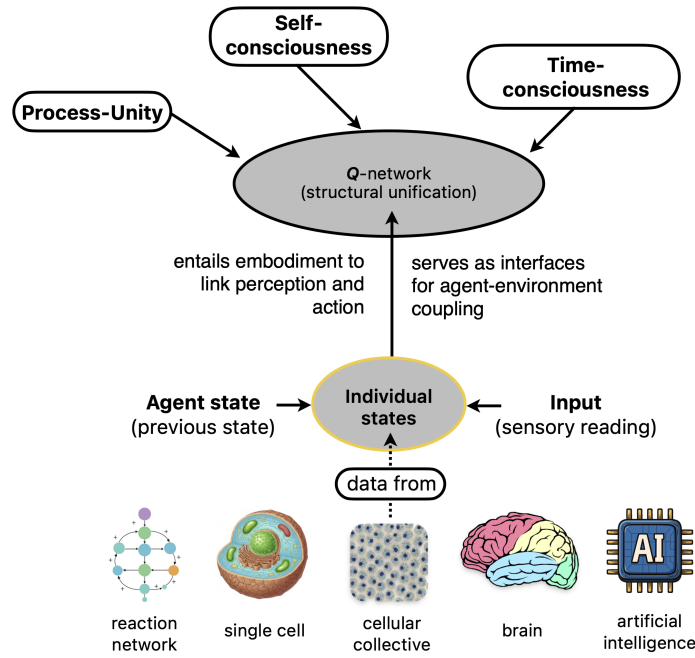


Fig. 1: The “phenomenological hierarchy” of this paper: data about various natural and artificial information-processing systems (e.g. reaction networks, single cells, cellular collectives, brains or AIs) can be embedded into Q -networks that encode relations between actual and potential states of an agent. Q -networks can be seen as exemplars of the larger class of *interfaces* encoding agent-environment interaction, inherently involving embodiment. A phenomenological approach investigates how such interfaces could mediate the structural conditions for subjectivity, potentially giving rise to more complex phenomenological properties mirrored in, for example, process-unity, time-, or self-consciousness. Collectively this defines how *subjective experiences are enacted through an agent’s interface with the world*. (Diagram created with the help of Microsoft Copilot.)

2. Categorical phenomenology

2.1. A doubly relational structure: $\Phi = R^2$

Before we appeal to the technical specification of our model, we first want to highlight two senses in which phenomenology should be understood as being relational. This further motivates our use of a particular mathematical formalism (i.e., category theory).

The first sense, in which phenomenology is relational, is hardly controversial to researchers familiar with the historical material since it involves many of the traditional concepts employed in almost any phenomenological investigation, albeit sometimes differing in terminology:

- Intentionality,
- Constitution,
- Self,
- Time-consciousness,
- Intersubjectivity,
-

Another, perhaps less obvious, sense in which phenomenology is primarily relational involves the question of potentiality. Accordingly, phenomenology is an analysis not of actual conscious experiences (though this is undoubtedly their starting point) but of the relation between actual and possible experiences. In the *Crisis*, Husserl wrote that “[each object that I experience] indicates an ideal general set of actual and possible experiential manners of givenness.” [Husserl, 1970]

Yoshimi [2011] provided a physics-metaphor for this. Experiences correspond to points in a state space, and traditional phenomenological analyses are (mostly qualitative) statements about possible trajectories through these spaces. Each trajectory thus connects countless potential experiences and defines a common relational property (technically, a trajectory can be defined as a one-parameter family of points).

We now further generalize this idea to what we call a minimal mathematical model of phenomenology. Whereas this requirement of minimality follows a general advice for scientific modeling known as Occam’s razor, it strikes us as particularly relevant when scientifically studying consciousness, which has a public-facing image of being “unmathematizable in principle.” We disagree. Let us list some basic desiderata of such a mathematization:

- (1) It should start with some (but not too many) explicit assumptions and stay general enough to allow for very different formal statements about experience [Signorelli *et al.*, 2025].
- (2) It is universal [Kanai & Fujisawa, 2024] yet metaphysically neutral [Chen & Yoshimi, 2023]. Hence, it could be used to understand the minds of a wide variety of entities, not just humans or animals, but also AIs or aliens.

- (3) It is amenable to the idea that a phenomenological system can reflect on itself, which is further related to the notion of “self-consciousness” [Gallagher & Zahavi, 2021, cf. also Section 4 of the present paper].

We, thereby, should base it on as few essential properties of consciousness as possible. These features are typically modeled on how *our* (limited form of) consciousness appears to *us* in introspection as (limited) observers. While it may be tempting to assume that we have transparent access to our own conscious mind, a substantial body of philosophical and psychological research shows otherwise: humans are notoriously unreliable at reporting what they just perceived, how they achieved their goals, or why they acted or desired in a particular way [Nisbett & Wilson, 1977; Usher *et al.*, 2023]. This tendency toward confabulation and misreporting has a striking parallel in large language models, which often generate plausible but inaccurate self-explanations of their reasoning [Stechly *et al.*, 2025; Shojaee *et al.*, 2025]. We thus stick to a very limited amount of (highly general) fundamental properties, and in turn reconstruct “typical” features of consciousness, such as unity or selfhood.

A historical antecedent for our structure-first approach is Husserl’s early (“pre-transcendental”) work on the formal character of mathematical objects. (See Husserl [2003] for the original discussion and Hartimo [2010] for recent scholarship on the historical relation between Husserl and mathematics.) For example, Husserl’s notion of a “definite manifold” emphasizes the primacy of relations and structures over particular contents: what matters is the formal pattern that can be instantiated across different ontological domains. This resonates with our methodological stance of deriving categorical structures from relational data rather than privileging any particular substrate (whether brain or AI). This further motivates looking for sources of structure that can be abstracted and compared via tools from category theory.

A category is defined by (Def. 2.1):

- (1) A collection of objects.
- (2) For any two objects, a collection of morphisms between them.
- (3) For each pair of morphisms, there exists a composite morphism if the target object of the first equals the source object of the second, i.e., $f : x \rightarrow y$, $g : y \rightarrow z \Rightarrow g \circ f : x \rightarrow z$. Composition is also associative.
- (4) For each object, there exist an identity morphism from the object to itself, 1_x , which satisfies unit laws, i.e., $f \circ 1_x = 1_y \circ f = f$.

In category theory, since there must be a unique identity morphism to any object, we could replace talk about objects by talk about morphisms. Indeed, because of compositionality, we could go even one step further and conceive of the objects as mere interfaces for how morphisms should be composed.

In our model, we interpret phenomenological objects as structures derived from relational data (e.g., cliques, trajectories, or other higher-order patterns), irrespec-

tive of whether those data come from biological or artificial systems. For example, such an object could be derived from a network of possible sensory states of a cell, a subset of brain states, or some of the memory states in a computer. Morphisms correspond to relations between these phenomenological objects. Our (“intentional”) form of consciousness is then but a special way of relating them.

2.2. Categories and graphs

A simple construction proceeds from a relational set of states:⁴

- (1) At any moment, there are potential experiential states, represented as elements $x \in X$.
- (2) States stand in relations to one another, encoded by a function

$$Q : X \times X \rightarrow R. \quad (2.1)$$

- (3) The choice of R determines whether the relation is probabilistic, weighted, or binary. We represent this by a *kernel* Q , which in the probabilistic⁵ case assigns transition probabilities, but may also encode weighted or binary relations depending on the application.

We call the resultant object a “ Q -network.” The definitional flexibility is intentional. Phenomenology requires only that experiential states stand in relations; it does not commit us to a specific numerical type. Temporal scaffolding, similarity structure, or entailment are all admissible relational forms. What makes a structure a “ Q -network” is that the kernel Q encodes how *potential experiential states are related*, from which phenomenological objects can later be derived.

We thereby specifically focus on (i) *relations of states*, and (ii) the influence of *action* on those, for example, by discussing action-indexed families of such kernels.

We illustrate this with the simple case where relations between states can be described deterministically and where the total number of states is finite. In this case, any kernel can be written as a finite (binary) matrix with entries zero or one. This need not be the case more generally.

⁴The mathematical model described here is similar to that of Hoffman *et al.* [2023], who developed a model of consciousness as simplification of a previous theory by Hoffman & Prakash [2014]. The original “conscious agent theory” models consciousness as network of conscious agents which are each specified via a 6-tuple, $\langle X, G, \mathbf{P}, \mathbf{D}, \mathbf{A}, n \rangle$: two measurable spaces for the experience and actions (X and G) of an agent, three kernels that relate them to themselves and the world, and an integer number that counts kernel executions. Note that we can recover our formalism after choice of R and sufficient integration. The counter n is thereby being neglected.

A historical precursor is the *Monadology* of Leibniz [2005].

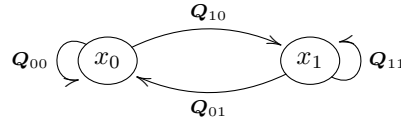
⁵So one can meaningfully talk about *this* state, as opposed to *that* one, $x_i \neq x_j$, at least with a certain probability. Formally, this requires us to postulate that X is measurable. To keep things simple, we assume that X is finite and discrete, and that each x_j just picks out exactly one element of the set. In other words, the relevant algebra of events is the power set $\mathcal{P}(X)$ and the kernel is Markovian (normalized either row- or column-wise, depending on convention).

We interpret relations as the possibility of one potential experiential state being followed by another: x_j relates to x_k if and only if x_j is followed by x_k according to the kernel $\mathbf{Q}$, for example, modeling an attentional process or sensory dynamics.⁶

It is instructive to first look at a system that has only two possible states, x_0 and x_1 . A kernel $\mathbf{Q}$ on this set is a 2×2 matrix that has the form

$$\mathbf{Q} = \begin{bmatrix} Q_{00} & Q_{01} \\ Q_{10} & Q_{11} \end{bmatrix}, \quad (2.2)$$

which corresponds to the following graphical representation:



A possible choice of $\mathbf{Q}$ would be realized by the “*not*-kernel”, $\mathbf{Q} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, where x_0 is followed by x_1 , and vice versa. Another choice would be the kernel that finds that any state will lead to the same state, say x_0 , regardless of where it started, $\mathbf{Q} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$. Under the constraints that (i) each row contains exactly one ‘1’ (deterministic transition) and (ii) every state has at least one outgoing transition, there are four possible kernels, corresponding to the matrices $\{\mathbf{id}, \mathbf{0}, \mathbf{not}, \mathbf{1}\} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \right\}$.

As an illustration, one could think of a robot with a single memory state. The transition between an “off” and an “on” state could be represented by the matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Whenever the state is off now, it will be on in the future; and vice versa (discounting for any external effects.)

Is it fair to say that such a system would be “minimally conscious” as in the example of the photodiode in [Oizumi *et al.*, 2014]? The system lacks *practically any* interesting phenomenological structure, so this question is hard to answer. More generally, we would maintain that it is about uncovering *relations* between networks, not the intrinsic properties of networks that let us connect this to phenomenology. The question of AI consciousness is more akin to the question *to what degree* does the computer exhibit certain relational structures, rather than the question whether it is endowed with a “magic sauce” (that we anyway cannot hope to define).

⁶Note that this suggests an interpretation on the level of “earlier/later,” which presupposes an ordering scheme that might not yet be given *from the perspective of the system* itself but only in the external observer’s spatiotemporal framework. However, we will adopt this interpretation for what follows for simplicity and concreteness. The reader is advised to keep in mind that the kernel more generally expresses a (potentially asymmetric) relation between different (actual or potential) experiences. Other notable interpretations are similarity-structures (symmetric) or entailment relations (asymmetric). For a general discussion on the varieties of the notion of “structure” as used in consciousness studies, see Kleiner [2024].

Concretely, one may study kernels on a set of n elements, for which there are n^n (deterministic) kernels. Among these, exactly $n!$ are invertible, corresponding to the permutations of the set $\{1, 2, \dots, n\}$. Within this class, one can further classify kernels according to cycle components, i.e. the disjoint cycles in the permutation decomposition.

For example, a $\mathbf{Q}$ -network with $n = 3$ elements may admit an invertible kernel that decomposes into two components (one loop and one singleton). A representative case is shown below:

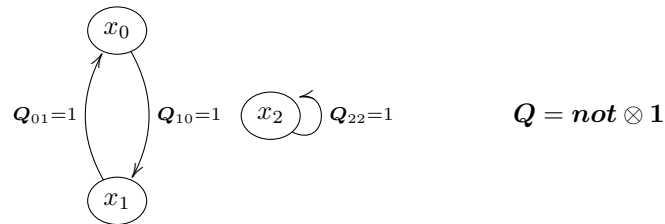
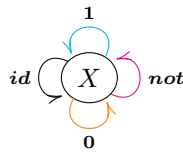
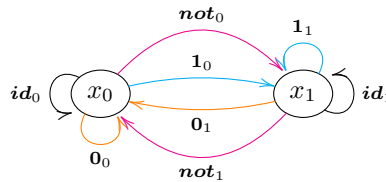


Table 1 summarizes the distribution of invertible kernels by cycle components for small n . Results for larger n can be obtained combinatorially, and efficient algorithms exist for computing the cycle decomposition of arbitrary kernels.

Another feature of this representation, which also aligns with the potentiality-interpretation of phenomenology, is that actions (such as movements) can be taken to *parametrize* the space of possible relations between potential experiential states. We assume that the specific subjective perspective encoded by the $\mathbf{Q}$ -network depends not only on what states are currently available to the system but also, crucially, on the way it could transition between them upon acting a certain way. One could, for example, represent our simple system as a “monoid”, illustrating a collection of four different kernels from a set to itself:



or, written a little less concisely:



Colors indicate one of the four possible kernels. Categorical phenomenology is typically based on the co-presence of many such kernels. Indeed, as the number of states grows, the number of possible kernels quickly becomes very large.

In order to move the discussion from the analyses of state transitions to phenomenology, we appeal to the framework of category theory [Lawvere & Schanuel,

Table 1: Number of invertible kernels on n elements with at least one cycle of size k . The notation $(c_j, c_k, \dots)$ indicates the sizes of cycles.

nk	$(c_j, c_k, \dots)$	number of kernels
$n = 3$		
31	(1,1,1)	1
32	(2,1)	3
33	(3)	2
total:		$3! = 6$
$n = 4$		
41	(1,1,1,1)	1
42	(2,2), (2,1,1)	$\binom{4}{2}/2 + \binom{4}{2} = 9$
43	(3,1)	$2 \cdot \binom{4}{3} = 8$
44	(4)	$3! = 6$
total:		$4! = 24$
$n = 5$		
51	(1,1,1,1,1)	1
52	(2,2,1), (2,1,1,1)	$3 \cdot \binom{5}{1} + \binom{5}{3} = 25$
53	(3,2), (3,1,1)	$2 \cdot \binom{5}{2} + 2 \cdot \binom{5}{2} = 40$
54	(4,1)	$6 \cdot \binom{5}{1} = 30$
55	(5)	$4! = 24$
total:		$5! = 120$
$\vdots$		

2009; Mac Lane, 1998; Ehresmann & Vanbreemersch, 2007]. Note that we do not treat the $\mathbf{Q}$ -network itself as a category (e.g., via the free-category construction). Instead, we investigate categories derived from $\mathbf{Q}$ -networks, depending on the structural features under consideration.⁷ In the toy model of Section 3.2, for example, objects are cliques and morphisms are inclusions, yielding a poset-like category. More generally, one should distinguish between the transition graph of states and the categorical structures derived from it.

⁷A brief phenomenological-methodological remark: the transition from $\mathbf{Q}$ -networks to derived categorical structures can be read in both a static and a genetic sense. In the static reading, a $\mathbf{Q}$ -network specifies structural relations among states that let us derive categorical objects, interpreted to be related to intentional experience. In a more genetic reading, such relations can be interpreted as giving rise to higher-order experiential structures, echoing the distinction in phenomenology between descriptive (static) analyses and those that trace how higher-order acts emerge from lower-order ones. An explicitly genetic interpretation is an interesting direction for future work.

Category theory provides tools that go beyond standard graph theory, allowing us to capture, for example, invariants, actions, and various kinds of universal phenomenological structures:

- (1) *Invariants.* Phenomenologists often look for structures that remain stable across different modes of experience. One way, already emphasized by [Husserl \[2012\]](#), proceeds by systematically uncovering such invariants (equivalence classes).

For example, invariants can then be identified by functorial correspondences that preserve structure across thresholds. In a practical example (compare Section 3.2), this can amount to computing persistent homology of $\mathbf{Q}$ -network-derived structures. As detailed in this section, such invariants provide summaries of the global structure of phenomenological objects, for instance, via Betti numbers that track connectedness and cyclic organization.

- (2) *Actions.* Beyond perceptual invariants, phenomenology also studies how consciousness changes under action. One example from phenomenology is Husserl’s method of variation in imagination, where experience changes upon mentally changing some of its properties, e.g., inverting the colors of a picture we see before our mind’s eye. Some variations might also happen if we physically engage in certain activities. Related, in models of “active inference” [[Clark, 2017](#); [Parr et al., 2022](#)], it is assumed that what an organism perceives depends on the way it acts in the world. In applications to computational systems, one might look for the invariants that the computer “sees” as a result of a computational process.

One could model such action with functors that act on categories. A functor can be described as a mapping from objects in one category to objects in a target category (including the category itself):

$$F : \mathcal{C} \rightarrow \mathcal{D}, \quad (2.3)$$

Thereby, functorial mappings preserve categorical structure, such as composition and identity laws. In particular, it is the case that if a morphism exists in $\mathcal{C}$, it also exists in $\mathcal{D}$:

$$A \rightarrow B \Rightarrow F(A) \rightarrow F(B). \quad (2.4)$$

Moreover, actions can be composed: sequential functors correspond to sequential transformations of experiential structure. Fig. 2 illustrates this with abstract nodes A, B, C , emphasizing that we are now working at the categorical level (functors between categories), rather than at the level of the individual states x_i of a $\mathbf{Q}$ -network. A particularly simple type of action is represented by functors corresponding to “loosening” and “sharpening” attention (cf. the toy model in Section 3).

- (3) *Structural conditions* for phenomenological unification. Category theory often appeals to universal constructions, for example (co)limits [[nLab authors, 2024b](#)]. Concretely, a colimit unifies a pattern of relations into a single minimal object

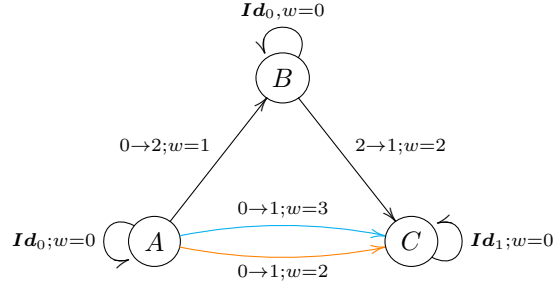


Fig. 2: From A to B via C , one can compose the mapping in two distinct ways. In one case, the weight of the composite is the sum of the parts (cyan); in another, it is the maximum (orange). This flexibility illustrates how categories can accommodate different rules for composing actions. (This example has been inspired by the example of a transport-network discussed in [Ehresmann & Vanbreemersch, 2007]).

through which all such relations factor. Fig. 3 illustrates this idea: the colimit C serves as the minimal cone through which all morphisms from the pattern must pass. In the colimit diagram, the objects X, Y, W, Z represent local sub-structures of the category, together with the mappings between them; the labels w denote arbitrary weights assigned to arrows to demonstrate different composition rules (e.g., additive vs. max-composition). The colimit C then serves as the minimal unifying object through which all such relations factor. In the toy model presented in Section 3.2, this abstract idea acquires a concrete computational marker when certain Betti numbers collapse or stabilize (e.g., β_0 reducing to 1), indicating that disparate relational components have merged into a unified structure.

In philosophical terms, we can think of the colimit as the unified locus of perception-action loops. Each object in the diagram represents a partial trajectory of experience, parametrized by action, but the colimit gathers them into a single point of convergence. It is the minimal object through which all such loops must pass, and thus encodes the conditions under which perception and action are integrated. In this sense, the colimit functions as the “meeting point” of all experiential dynamics, rather than merely summarizing relations. Furthermore, this object could itself be accessible to reflection, even if normally not part of immediate awareness, suggesting representation in terms of higher category theory [nLab authors, 2022b].

Together, these examples suggest that certain mental properties can be represented categorically and related functorially to physical data, once they have been embedded into $\mathcal{Q}$ -networks. This conception aligns with earlier proposals in the literature

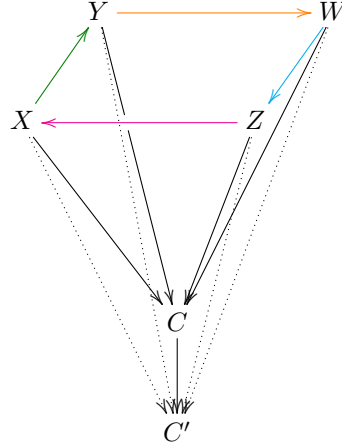


Fig. 3: Illustration of the concept of colimit in a category. All mappings between objects in the pattern (colors) commute with mappings to C . Specifically, the colimit, if it exists, picks out the minimal cone C . Any mapping from the pattern to another cone C' factors uniquely through C via a mapping from C to C' .

[Ehresmann & Gomez-Ramirez, 2015; Tsuchiya *et al.*, 2016; Atmanspacher & Rickles, 2022; Prentner, 2024].

3. Interface consciousness and a concrete toy model

3.1. Interface consciousness

The outlined approach dovetails with a broader research program on *interface consciousness*, which conceives of consciousness as the enactment of subjective experience through interfaces. $\mathbf{Q}$ -networks provide the formal scaffolding for such interfaces, which are defined by relations between the states of an agent.

Recall the “phenomenological hierarchy” presented in Fig. 1. Here we distinguished (i) between states, (ii) relations between states that define $\mathbf{Q}$ -networks, and (iii) networks that are organized according to further phenomenological constraints, which all correspond to different aspects of interfaces. This models how third-person data are embedded into a “phenomenal space” [Prentner, 2019], how actual and possible states are related to define experience, and how such relations are organized into forms of intentional consciousness.

What matters for consciousness is relational, and relations *prima facie* do not dependent on the nature of the substrate they are relating.⁸ By modeling *interfaces*

⁸Of course, this leaves open that certain types of relations could only exist between certain types of objects.

rather than *substrates*, we align our study with received approaches in the vicinity of 4E cognition [Menary, 2010; Newen *et al.*, 2018], applied to consciousness in a non-reductive way. Below are just three examples from the contemporary literature:

- (1) *Active Inference* approaches to consciousness [Seth & Bayne, 2022; Albarracín *et al.*, 2022; Ramstead *et al.*, 2023] find that the content of consciousness is shaped more by an agent’s internal beliefs and possible actions than by the “raw” sensory signals it receives. From this perspective, one could say that consciousness research is about uncovering the *phenomenological* principles of (Bayesian) active inference [Parr *et al.*, 2022]. *Subjective experience is about harboring a generative model (related to phenomenology).*
- (2) The related *Interface Theory of Perception* [Hoffman *et al.*, 2015; Prakash *et al.*, 2020, 2021; Prentner, 2021; Prentner & Hoffman, 2024] argues that what an agent perceives is shaped by utility, not by true environmental structure. The structures we perceive are embedded in an agent-specific way of representing its environment. Applied to consciousness science, this means to understand how those agent-specific ways of representing align with phenomenological ideas. *Consciousness is fundamentally relational.*
- (3) The *Projective Consciousness Model* [Rudrauf *et al.*, 2017; Williford *et al.*, 2022] is premised on the idea that the contents of consciousness (all contents, not only the visual ones) amount to representations that are organized in a projective space. Moreover, the way how those contents are represented accounts for certain phenomenological insights that enables a representational system to carry out specific functions aimed at behavioral control and optimization. *Phenomenology is tied to function.*

This relational and projective character of experience has long been emphasized in phenomenology, for example in Aron Gurwitsch’s field theory of consciousness [Gurwitsch, 1964] and Alfred Schutz’s analyses of anticipatory, action-oriented structures of experience [Schutz, 1967]. Therefore, *interface consciousness* is not merely a metaphor but a philosophically-informed research program that can be computationally instantiated and experimentally probed. The following subsection illustrates this with a concrete toy model.

3.2. A toy interface

To illustrate how the abstract framework can be implemented, we sketch a simple computational toy model. The aim is not to capture the full richness of phenomenology, but to demonstrate how third-person data can be embedded into a phenomenal space, structured into *Q*-networks, and then analyzed with categorical and topological tools. Unlike in the examples of Section 2.2, we do not adhere to the requirement that transitions between states are (families of) deterministic kernels (although they could be, such as in the case of temporal adjacency discussed below).

3.2.1. Third-person data and phenomenal embedding

We begin with a synthetically generated time series of $n = 50$ points, each with 5 features (Fig. 4). In a realistic scenario, such features can refer to a wide range of empirical or simulated systems, for example molecular concentrations, sensor readings, neuronal activity patterns, internal memory states of an artificial agent, or higher-level structures such as IIT’s cause–effect repertoires [Tononi *et al.*, 2016; Albantakis *et al.*, 2023]. The specific choice of features is not essential for the present illustration; what matters is that they encode third-person data from which relational structure can be derived.

These data are then embedded into a three-dimensional space using principal component analysis (PCA), yielding a geometric representation in which each point corresponds to a potential or actual state of the resulting $\mathcal{Q}$ -network (Fig. 5). More sophisticated versions of the model could employ nonlinear or learned embeddings, such as autoencoders or contrastive neural embeddings.



Fig. 4: Evolution of features over time (“third-person data”).

3.2.2. Defining $\mathcal{Q}$ -networks

From the time series embedding, we construct a $\mathcal{Q}$ -network (Fig. 6) as follows:

- (1) Compute pairwise similarity among embedded states $\varphi(s)$, based on the cosine similarity matrix with entries

$$M_{ij}^{\text{sim}} = \frac{\varphi(s_i)\varphi(s_j)}{||\varphi(s_i)||||\varphi(s_j)||}. \quad (3.1)$$

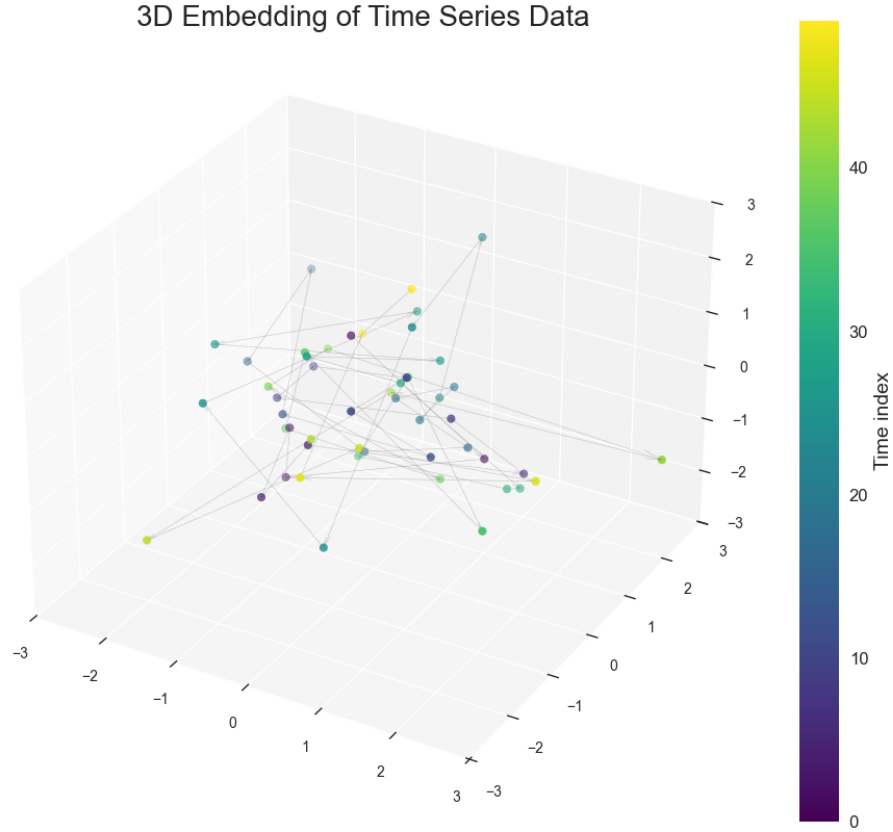


Fig. 5: 3D embedding of the time series into a phenomenal space. Each point corresponds to a state of a $\mathcal{Q}$ -network; arrows indicate temporal succession.

A similarity graph $G_{\text{sim}}(\tau)$ is then defined as undirected graph with vertex set S and an (undirected) edge $\{i, j\}$ iff $M_{ij}^{\text{sim}} \geq \tau$, where $\tau \in [-1, 1]$ is an arbitrarily chosen similarity threshold. Similarity graphs are later used to compute persistent homology.

- (2) Record temporal adjacency between successive states in the time series. A directed temporal adjacency graph is defined if there is an observed transition from state i to state j in the recorded trajectory.

This separation preserves interpretability and allows comparison between hypothesized “topological phenomenal properties” (similarity-based) and observed dynamics (temporal adjacency-based) in future, more realistic settings.

While recorded separately, both constructions are combined into a single graph for illustration (Fig. 6), where the similarity-based relation is treated as a bi-directional arrow (weighted by the threshold, indicated by colors).

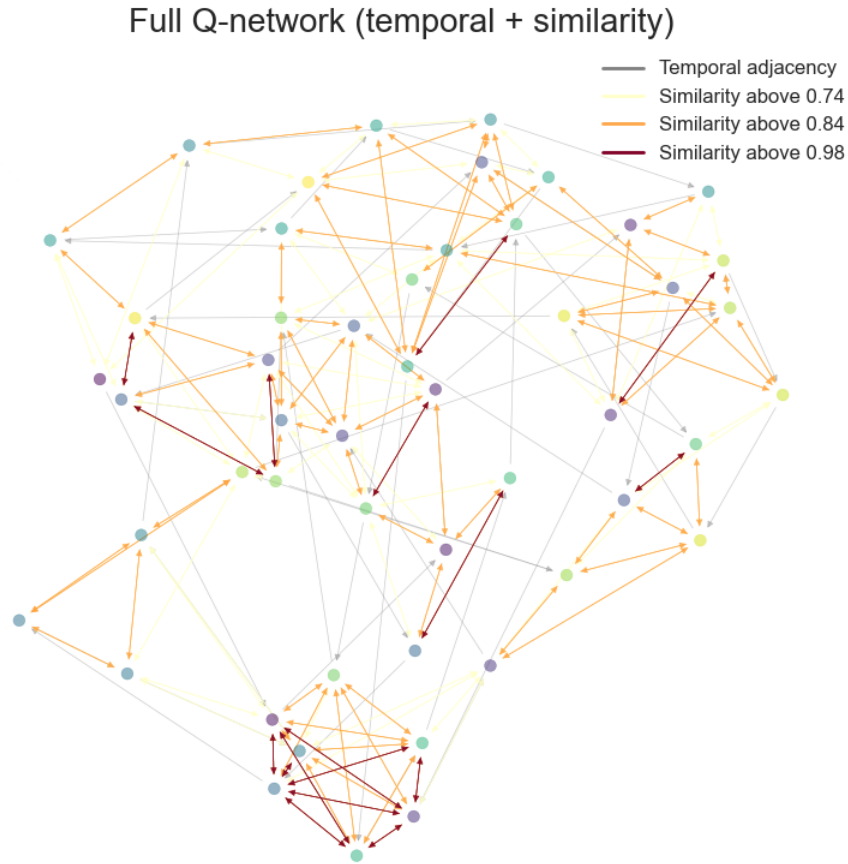


Fig. 6: Full Q -network combining temporal adjacency (gray edges) and similarity-based edges (colored). Node color encodes time index.

Subgraphs can be extracted for selected nodes [1, 6, 17, 20, 31, 32, 35] and similarity thresholds (Fig. 7), revealing different “graph motifs” [Newmans, 2010]. A exemplary discussion of some motifs is provided in Table 2.

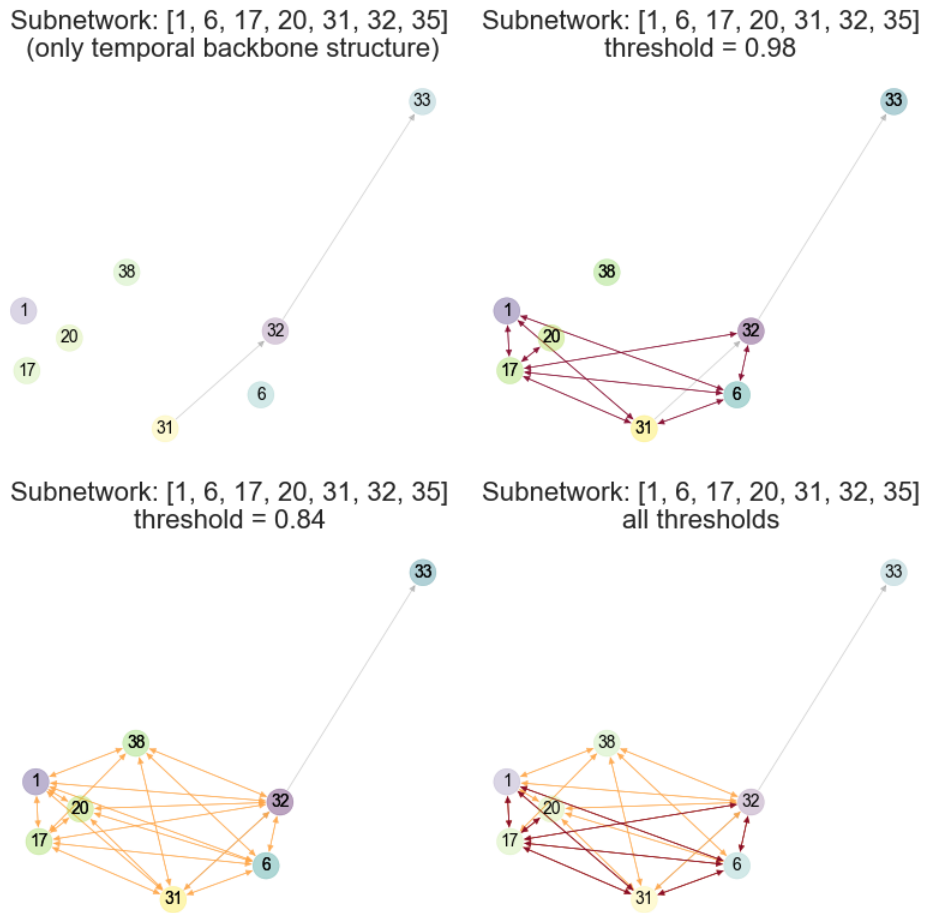


Fig. 7: Example subnetworks for selected nodes at different similarity thresholds.

Table 2: Motif analysis of similarity-based $\mathcal{Q}$ -network subgraphs across thresholds.

Threshold	# Max cliques	Avg clique size	# Triangles
0.98	11	2.27	1
0.84	32	3.00	16
0.74	32	3.97	9

Examples (representative subsets):

- **0.98:** Max cliques: [3, 37], [5, 44], [9, 40], [12, 35], [14, 40], [14, 48], [16, 33], [17, 20], [17, 6, 32], [17, 6, 1, 31], [30, 34]. Triangles: [17, 6, 32].
- **0.84:** Max cliques (subset): [1, 32, 6, 17, 31, 20], [1, 32, 6, 17, 31, 38], [2, 16, 33], [2, 3, 37], [3, 8, 13], [3, 8, 37], [4, 24], [4, 44, 5], [5, 43, 44], [7, 41, 46], [8, 15, 9], [8, 15, 13, 39], [9, 40, 48, 14], [9, 40, 15], [10, 34, 30, 23], [10, 11], [11, 45, 47], [11, 29], [12, 46, 41], [12, 46, 35], [18, 21, 42], [18, 21, 19], [22, 26], [22, 28], [23, 49, 34, 30], [25, 43], [25, 27], [26, 48], [27, 39], [28, 34, 30], [29, 36], [43, 45, 47]. Triangles (subset): [2, 16, 33], [2, 3, 37], [3, 8, 13], [3, 8, 37], [4, 44, 5], [5, 43, 44], [7, 41, 46], [8, 15, 9], [9, 40, 15], [11, 45, 47], [12, 46, 41], [12, 46, 35], [18, 21, 42], [18, 21, 19], [28, 34, 30], [43, 45, 47].
- **0.74:** Max cliques (subset): [0, 22, 49], [0, 22, 26], [1, 20, 32, 38, 17, 6, 31], [1, 20, 19, 21], [2, 37, 8, 3], [2, 37, 33, 16], [4, 24, 44, 5], [4, 13], [5, 44, 43, 25], [5, 44, 43, 47], [6, 32, 7, 17, 20], [7, 41, 12, 46], [10, 11, 29], [10, 23, 49, 34, 30], [10, 23, 29], [11, 43, 45, 47], [12, 35, 41, 46], [12, 35, 45], [15, 8, 40, 9], [15, 8, 3, 9, 37], [15, 8, 3, 13, 37], [15, 8, 3, 13, 39], [15, 27, 13, 39], [15, 14, 40, 9], [18, 42, 19, 21], [22, 28, 49], [23, 28, 49, 34, 30], [25, 27, 36], [25, 27, 13, 39], [25, 29, 36], [26, 48, 14], [42, 40, 48, 9, 14]. Triangles: [0, 22, 49], [0, 22, 26], [10, 11, 29], [10, 23, 29], [12, 35, 45], [22, 28, 49], [25, 27, 36], [25, 29, 36], [26, 48, 14].

3.2.3. Categorification and invariants

In a next step, an “inclusion-category” should be derived from the $\mathbf{Q}$ -network. From now on, we focus only on similarity relations and ignore temporal scaffolding by only using the undirected similarity graph $G_{\text{sim}}(\tau)$. We thereby introduce the notion of a “clique” from algebraic topology: a clique is a set of vertices that are pairwise connected in the underlying similarity graph. We use the clique (flag) complex construction, in which cliques of the similarity graph are treated as simplices. This construction is standard in algebraic topology and topological data analysis (see, e.g., [Edelsbrunner & Harer \[2010\]](#)). To relate this to category theory (chosen to establish the link to phenomenology),⁹ we treat cliques as objects and inclusions between cliques as morphisms. Parameterized by the similarity threshold, this yields a filtered category [[nLab authors, 2025](#)]. Minimal inclusions can be visualized as Hasse diagrams.

At this point, it should be emphasized that one cannot extract much quantitative information from visual inspection of these Hasse diagrams, especially at lower threshold values where the number of cliques and inclusions grows rapidly. The diagrams quickly become visually cluttered, obscuring structural features. For this reason, one can resort to computing topological invariants, such as Betti numbers β_i derived from persistent homology, which provide concise and interpretable summaries of the underlying invariants across thresholds. Persistent homology is computed on filtrations of simplicial complexes. In this work we form the simplicial complexes via the clique (flag) construction: cliques in the underlying graph become simplices, and inclusions among cliques are recorded in an inclusion-category whose nerve is the resulting simplicial complex. We do not compute persistent homology directly on arbitrary categorical constructions; this would, for example, run the risk of introducing spurious homological features because compositional morphisms create formal cycles. Appendix A summarizes the construction and its limitations.

Formally, β_0 counts connected components, while β_1 captures cycles. In our implementation, Betti numbers were computed for different thresholds (Fig. 9) using the GUDHI library [[Maria et al., 2014](#)], which provides efficient algorithms for building simplex trees and extracting persistent homology.

Phenomenologically, we interpret β_0 as “islands of experience,” i.e. fragmented (micro-)contexts that remain disconnected at high thresholds. A high initial β_0 reflects a subjectivity scattered into many disjoint experiential units. As the threshold

⁹The observations in this subsection are heuristics based on the toy model rather than general mathematical theorems. In particular, the apparent coincidence of Betti-number crossings with qualitative regime shifts in the derived categories is an empirical pattern in this example, not a categorical statement about colimits, universal constructions, or functorial properties. Category theory is introduced here not to establish such correspondences, but because it provides the structural language needed to model how experiential structures relate, transform, and compose. Persistent homology captures the shape of relational data, whereas category theory captures the transformations between such shapes (e.g., inclusions, functors, and action-induced mappings).

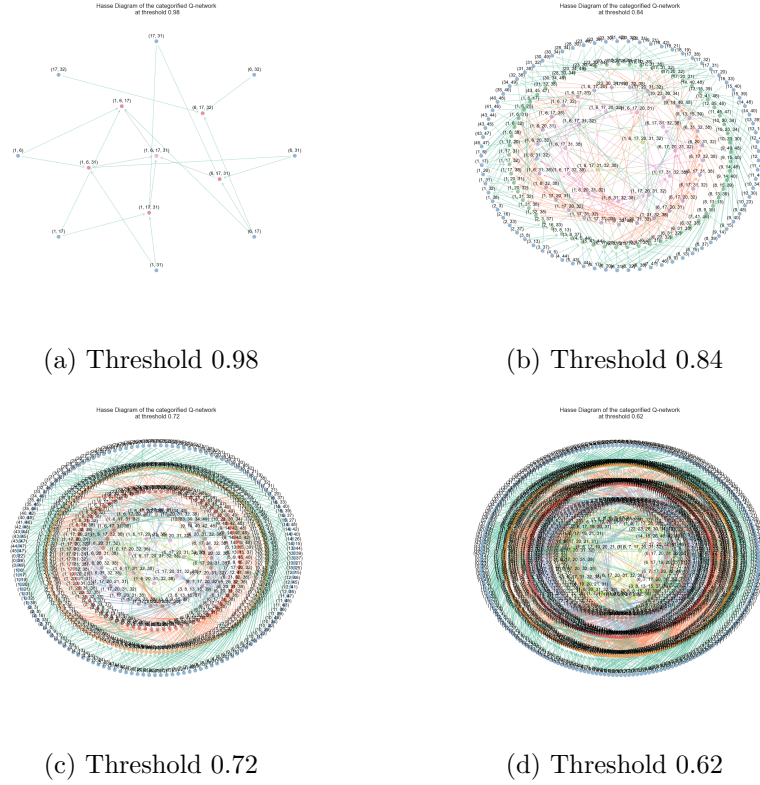


Fig. 8: Hasse diagrams of the inclusion-category defined over the Q -network at different thresholds. Each diagram shows minimal inclusions between cliques, treated as objects of the category. As the threshold decreases, the diagrams become increasingly dense and visually complex.

decreases, β_0 declines, indicating that these fragments begin to merge into larger experiential unities.

By contrast, β_1 is interpreted as sign of “experiential loops,” motifs where recurrent paths connect states. The rise of β_1 (peaking near threshold ≈ 0.65 in our example) marks a phase of maximal relational tension: the experiential field is unified enough to sustain recurrent motifs, yet still diverse enough to generate ambiguity and multi-stability. When $\beta_0 = 1$ but $\beta_1 > 0$, the system exhibits a single experiential field enriched by internal cycles, suggesting a unified but internally complex subjectivity. As the threshold decreases further, β_1 falls again, reflecting the collapse of these tensions into higher-order integration, where the experiential field becomes so interconnected that loops are absorbed into more global structures.

As a minimal robustness check, one could verify that the qualitative behavior of β_0 and β_1 is stable under small perturbations of the similarity threshold, invari-

ant under relabeling, and consistent under alternative embeddings. Specifically, we checked that (i) the Betti curves are invariant under random permutations of the similarity matrix, (ii) their qualitative behavior is robust when changing to a random linear projection, and (iii) the observed regime shifts in are not tied to a single fine-tuned threshold but persist across short threshold intervals in the filtration.

Higher Betti numbers (not observed in the toy model) admit further phenomenological speculation. A non-zero β_2 would correspond to “experiential surfaces,” enclosing relational flows in two-dimensional shells. A non-zero β_3 would indicate “phenomenal bubbles,” higher-order enclosures where tetrahedral relations wrap into sealed volumetric horizons. These higher-dimensional invariants would phenomenologically correspond to increasingly complex forms of experiential integration not reducible to simple connectedness or loops. It is an interesting question to what extent this points to a fundamental divergence in the phenomenal structure of, say, humans and future AI systems.

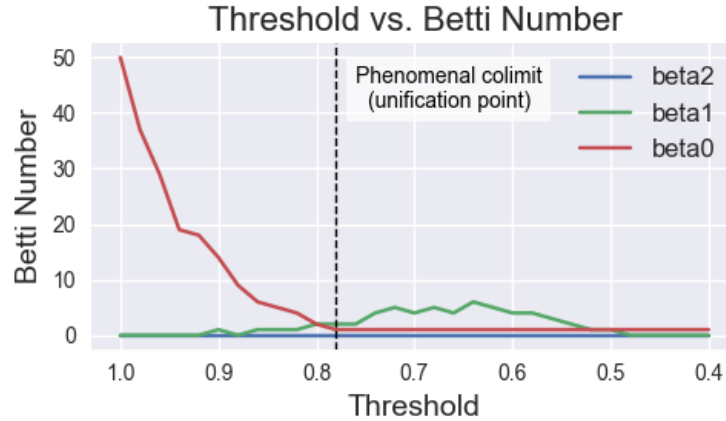


Fig. 9: Betti numbers as a function of threshold. These topological invariants provide phenomenological interpretations of connectedness and cyclic structure. In the example, only β_0 (red) and β_1 (blue) take non-zero values, standing for the presence of disconnected components and loops, respectively. The dashed vertical line marks the colimit of the filtered category, i.e. the threshold at which all components unify into a single experiential field.

3.2.4. Functorial actions, maximality, and colimits

Finally, we propose to represent actions by functors on the category (implemented here as endomap on the cliques). Particularly simple examples are inclusion functor and the functor obtained by restricting to the subcategory, corresponding to

“loosening” or “sharpening” attention:

$$F_{\text{inc}} : \mathcal{C}_{\text{high}} \hookrightarrow \mathcal{C}_{\text{low}}, \quad F_{\text{res}} : \mathcal{C}_{\text{low}} \rightarrow \mathcal{C}_{\text{high}}. \quad (3.2)$$

Maximal cliques under these functors can be visualized in heatmaps (Fig. 10, only shown only for the inclusion functor).

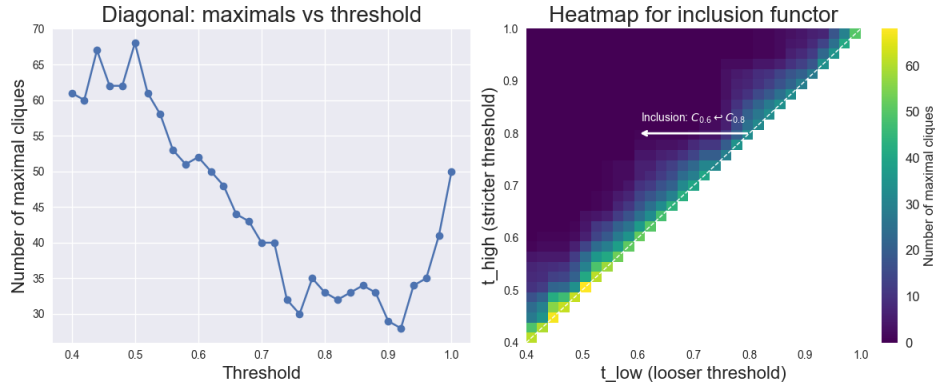


Fig. 10: Diagonal curve (left) and heatmap (right) showing the number of maximal cliques under inclusion functors across thresholds.

While these functors describe how structures persist or disappear as attention is varied, we can also ask about the *global* outcome of this process. Category theory captures this with the notion of a colimit: the minimal unifying object through which all clique-inclusions (morphisms) factor. Phenomenologically, the colimit corresponds to the point where fragmented “islands of experience” (high-threshold cliques) converge into a unified experiential field.

Since we already computed Betti numbers, we can easily determine the threshold value where β_0 (connected components) collapses to 1, indicated by the dashed line in Fig. 9. This provides a concrete marker for the colimit, i.e. the moment when all inclusions factor through a single connected object. Additional phenomenological structure may be revealed by considering “local” colimits of subcategories, which appear as crossings of different Betti numbers β_j and β_k ($j \neq k$) cross.

It is important to note that the examples discussed above instantiate “action” only from the perspective of the analyst, via changes in the similarity threshold that modulate the derived categorical structure. This should not be confused with action in the enactive, agent–environment sense.¹⁰ A genuinely agentic implementation would require that the data itself be modified as a consequence of the agent’s actions, and that the categorical construction track these changes.

¹⁰I thank one reviewer for highlighting this distinction.

While developing such a setup lies beyond the scope of the present paper, we still like to speculate on a possible direction. In a biological or embodied artificial system, one could record an “unperturbed” time series prior to action and a “perturbed” time series after the agent intervenes in its environment. Provided that the counterfactual case of “no action” does not itself induce comparable changes, the difference between the corresponding categorical structures could serve as a phenomenological marker of action. Another possibility is to train a predictive model on the unperturbed time series (not unlike next-state prediction in current language models) and compare its predictions with the actual changes induced by action. In both cases, the key idea is that actions would alter the relational structure encoded by the $\mathbf{Q}$ -network, and the categorical machinery would then capture how these alterations reorganize the agent’s phenomenal interface.

3.2.5. Summary and Limitations

The toy model demonstrates how third-person data can be systematically related to a first-personal structure, illustrating the phenomenological hierarchy (Fig. 1): from raw features to phenomenal embeddings that define a $\mathbf{Q}$ -network, and from networks to categories and invariants. While highly simplified, the model demonstrates how phenomenological properties can be mapped onto computational constructions, providing a bridge between the mental and the physical. Still, the toy model is limited, especially insofar as it connects to the phenomenological themes of self, time, and unity discussed in more details in the next Section 4. This is not very surprising, after all the data used for this example was drawn from a synthetic time-series.

- The collapse of β_0 was interpreted as structural marker of phenomenological unification in the toy model. By contrast, the notion of a *self* [Gallagher & Zahavi, 2021, whether “pre-reflective” or “reflective”]) requires more than mere connectedness: it presupposes a distinction between self and world, and potentially higher-order reflective structures. While the present toy model does not implement these, its categorical framework suggests how such distinctions might be formalized in future extensions, for example, by introducing a partition of the state space into “self-directed” and “world-directed” components and studying how functorial actions preserve or transform this partition. This would allow one to model the stability of self-related structures under changes of perspective or action.
- A further limitation concerns *temporality*. In the toy model, temporal adjacency was removed before computing the categorical invariants. Retaining temporal edges would have likely introduced artifacts. Still, one could imagine reintroducing adjacency relations in a more realistic model as a way to investigate a *locus classicus* of phenomenology, namely the temporal modes of consciousness [Dainton, 2023].

- The *unity* of consciousness can be related to constraints on how processes compose. In the toy model, we already speculated on the connection between attention and the inclusion functor. If consciousness is treated as a processes, then unity corresponds to the requirement that the resulting diagrams cannot be rewritten into disconnected parts. This resonates with the idea of “process-unity” that subsumes both synchronic and diachronic forms of unity under a single categorical condition.

4. More complex phenomenological properties

4.1. *Self-consciousness*

In the previous sections, we argued that data about information-processing systems, both natural and artificial, can be integrated with the study of a first-person perspective via the notion of a *Q*-network. Through deriving categorical structures, such *Q*-networks can be studied with respect to (i) invariants, (ii) action-dependency, and (iii) structural unification.

Beyond these relational features, there are further properties characteristic of the familiar, human form of (“intentional”) consciousness. The minimal relational organization captured by a *Q*-network is therefore not yet intentional consciousness, but a structural precondition for it. Intentional consciousness arguably results from additional constraints that are imposed on a *Q*-network (see Fig. 1, top).

One important constraint pertains to the notion of “self.” On a conventional reading, the self is sometimes believed to be yet another type of representational content or a relation between those (e.g. the “phenomenal model of the intentionality relation” of Metzinger [2003]). However, research into “minimal phenomenal experience” [Metzinger, 2020] or a literal reading of various “no-self” experiences, as they have been reported in multiple religious traditions, suggest that the self might not be necessary for consciousness *as such*. In our model we understand self-consciousness, not as a necessary (substantial) entity, but as a structural feature of *Q*-networks that allows an agent to distinguish its own states from environmental inputs. Phenomenologists often qualify this further by assuming the self to be a content of *reflective* consciousness only. By contrast, the notion of a *pre-reflective* self is not itself an object (or content) of consciousness but arises as structuring principle of intentional consciousness more generally.

Here, category theory plays nicely alongside phenomenology and provides tools to translate qualitative statements into a mathematical language. For example, phenomenologists would typically assume that the pre-reflective self could itself be reflected upon [Gallagher & Zahavi, 2021], i.e., made into the intentional object of a “higher” act of consciousness. Reflection, in category theory, could be conceptualized in the form of a “2-category” [nLab authors, 2024a] – for example a “category of categories” – and the (reflective) self would be a (universal) object in this higher category.

To speculate even further, the more than 30 empirical notions of “self” [Strawson, 1997] could be given a more systematic reading as correlates of an “ n -self”, which refers to a universal construction on an n -category. Moreover, there is a natural sense in which n -selves can be nested, though this need not be the case, cf. the intricacies of “ramification” [Ehresmann & Vanbremeersch, 2007].

4.2. Time-consciousness

The relations giving rise to self-consciousness are orchestrated throughout the dynamical evolution of a system. Metaphorically speaking, time is the glue between the self and the world. Perhaps, time-consciousness has been the most studied concept from phenomenology to date, which has attracted various empirical and computational studies, spanning neuroscience, predictive processing, and even psychopathology research [Wiese, 2017; Kent & Wittmann, 2021; Singhal & Srinivasan, 2021; Tewes & Stanghellini, 2020].

Phenomenologists have long speculated on the way how time-consciousness could be described at least since [Husserl, 1966]. Importantly, describing time-consciousness is not tantamount to describing the subjective experience of time in terms of, for example, “past”, “present” and “future” (i.e., tensed time). These are secondary aspects, “experienced labels,” that we attach to particular instances of consciousness (and their contents). In contrast, time-consciousness refers to an organizing principles of the broader dynamics of intentional consciousness itself. It is natural to interpret Husserl’s tri-partite model of “retention,” “primal impression,” and “protention” [Dainton, 2023] in terms of a Q -network that supports three systematically interrelated families of indexed functors. For example, in section 3.2, we separated temporal adjacency from similarity-based structures. One could now compare invariants computed on (i) a similarity-only graph and (ii) a temporally constrained graph (for example, restricting edges to a fixed lag window), thereby distinguishing structural features that are topological from those that are temporal. Such a comparison would be essential in applications to biological or embodied systems beyond simple toy models.

More generally, an investigation into time-consciousness would likely stay non-reductive. While we seem to know how time works in (classical¹¹) physics, e.g., how chemical reactions occur over a measurable period of time or how neurons fire relative to an external clock, one is hard-pressed to account for a basic puzzle in the philosophy of time: how are genuine change and its experience at all possible? Relatedly, phenomenologists typically try to stop presupposing an external, physical world. Even in this (perhaps only imaginary) situation, phenomenologists argue that one’s experience exhibits a sense of “time,” which is remarkably distinct from the observable physical dynamics. This issue has been probed repeatedly in

¹¹There exist serious puzzles in understanding time in fundamental physics [Price, 1996; Brukner, 2014].

neurophenomenology, starting with the seminal works featured in [Petitot *et al.*, 1999]. The state of the art approach often involves a dual description of physical (neuronal) and psychological systems, informed by phenomenological principles, see also [Horvat & Northoff, 2022; Singhal & Srinivasan, 2024] for a recent take on this. Distinguishing between the temporal profile of experience as it is given by physical transitions of the systems (mirrored by temporal adjacencies in the Q -network) and the phenomenological aspects that are given by other structural features (e.g. mirrored in the similarity-structure of the graph) would be consistent with this dual approach.

4.3. Process Unity

What further distinguishes the minimal relational organization of Q -networks from intentional forms of (e.g., human) consciousness is encoded by constraints on the latter such as the unity of consciousness [Bayne, 2010]. Spelling out this unity-constraint amounts to a major formal research question in the science of consciousness [Bayne & Chalmers, 2003; Tononi *et al.*, 2016; Prentner, 2019; Kleiner & Ludwig, 2024].

To translate this into the conceptual language of the current work we appeal to applied category theory [Abramsky & Coecke, 2008; Bradley, 2018]. In applied category theory, objects are often called “types”, which are taken to represent, in our case, categorical structures derived from Q -networks. In addition, applied category theory knows of “processes” that operate upon these types. Each process would then refer to a transformation of these structures.

Applied category theory studies the *composition of processes* [Coecke, 2021], which can be naturally interpreted as a single process. Graphically, this is represented with a string diagram, where categorical structures derived from Q -networks (types) are transformed by processes (boxes), and sequential or parallel compositions yield new processes (Fig. 11).

In applied category theory, emphasis is often placed on the equivalence of string diagrams according to so-called “re-writing rules.” If one diagram can be re-written in terms of a second, i.e., if there exist inverse (up to natural transformation) functors between them, they are equivalent. This lets us represent the idea of constraints as categorical equivalences on compositions of processes [nLab authors, 2022a]. It is here where we see the main relation to the unity of consciousness. In consciousness studies, one often distinguishes a “diachronic” unity of consciousness across time and a “synchronic” unity at the same point in time [Dainton, 2000]. But believing that either a diachronic or synchronic sense of unity is in some way more fundamental might stem from a conflation of single states of experiences (which are trivially unified) with intentional forms of consciousness (which can potentially be

¹²Here “o” denotes sequential composition of morphisms, while “ $\otimes$ ” denotes parallel composition. T denotes the transpose; m and p are horizontal composites of the identity and other generic objects (“cups” and “caps” [Signorelli *et al.*, 2021c]).

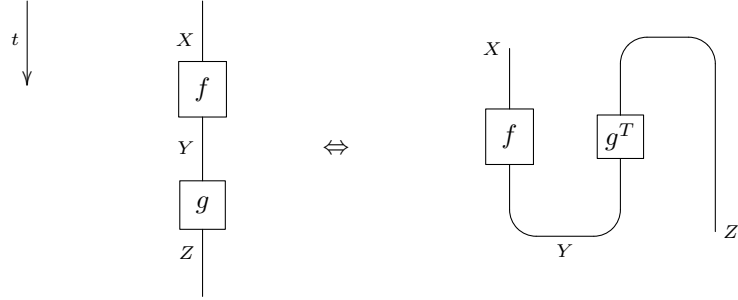


Fig. 11: Generic string diagram for a strict monoidal category illustrating the idea of “process-unity.” Types (wires; experiences described by categorical structures derived from $\mathbf{Q}$ -networks) are transformed by processes (boxes; interpreted as transformations of these structures). On the left side, one sees a sequential composition of processes ($g \circ f : X \rightarrow Z$). On the right, is an equivalent diagram with the respective processes not composing sequentially but horizontally ($m \circ (f \otimes g^T \otimes 1) \circ p : X \otimes 1 \rightarrow 1 \otimes Z$)¹²; time flows from top to bottom.

dis-unified). Following [Signorelli *et al.*, 2021b], it might be better to subsume both under the broader notion of a “process unity”: intentional consciousness is unified iff its representation as a string diagram cannot be written in an equivalent but disconnected way; compare Fig. 11.

5. Conclusions and outlook

In this article, we proposed a phenomenological framework for computational systems. We first argued on conceptual grounds that category theory is suitable for understanding phenomenology’s “doubly relational” nature ($\Phi = R^2$). We then outlined an exemplary way to base such a treatment on $\mathbf{Q}$ -networks, specifically claiming that some central notions of phenomenology can be represented by categorical objects derived from $\mathbf{Q}$ -networks.

Category theory appears promising not only for its relational character but also for the precision of its methods. We hypothesize that the information-processing in various systems (including AIs) can be embedded in $\mathbf{Q}$ -networks that define a first-person perspective insofar as they are organized around concepts that articulate phenomenological themes such as intentionality, subjectivity, etc. Category theory serves as a high-level link that facilitates interpretation and design of those networks.

What could falsify this? For example, one could examine whether features such as “structural unification” (Section 2.2) are present or absent in systems taken to be conscious. But this immediately raises a problem. Either we rely on intuitive judgments about which systems are conscious, which is highly problematic, since

consciousness is not a third-personally observable phenomenon.¹³ Alternatively, we compare observable behavior with scientific determinations of consciousness [Bayne *et al.*, 2024]. Yet, especially when it comes to conscious AI and the question of moral status [Shevlin, 2024], this is problematic too! For some systems, such as vertebrates, this can perhaps be done using a “theory-light” approach [Birch, 2022]. However, machines could conceivably produce sophisticated behavior without consciousness being involved.

One might thus think that without a good theory that explains the (metaphysical) relation between physical states and consciousness, we will ultimately not know whether conscious AI has a “realistic possibility of conscious experience” [Andrews *et al.*, 2024]. However, this presents the ultimate problem for empiricist approaches to consciousness: nothing in our experience seems to show that something else is or is not a locus of experience. At the same time, this suggests the need for a novel reasoning strategy beyond received empiricist approaches, no matter whether they are “theory-light” or “theory-heavy” [Butlin *et al.*, 2023] or whether they define “theory-derived” indicators of consciousness [Butlin *et al.*, 2025].

Instead, we recommend a different strategy. In his *Prolegomena*, Immanuel Kant [2004] asked how mathematical knowledge is possible, for example, how we can know that the angles of a triangle sum to 180° . He argued that we can know such things because our minds are set up to experience the world in terms of space. For Kant, space is the basic format built into how we sense anything. By analogy, we may ask: how is subjective experience possible? Our proposal is that computational processes, when subsumed under phenomenological concepts, are thereby constrained in ways that make subjective experience possible (Fig. 12). These constraints jointly determine the *format of the interface* through which a system enacts its first-person perspective, and thus relates to the world.

We therefore suggest focusing on evaluating the relational richness of the interfaces that could be instantiated by various systems, such as biological ones or AIs. The structural markers proposed here, such as the degree of structural unification (in the toy model revealed by topological invariants) or the way categorical structures reorganize under action-indexed transformations (self- and time-consciousness or process-unity), offer concrete means of assessing how richly an interface relates its possible states to one another. They do not operationalize consciousness itself, but they reveal whether a system’s relational architecture exhibits the kinds of multi-level coherence that phenomenology requires.¹⁴

Future work should extend this analysis of $\mathbf{Q}$ -networks to different kinds of third-person data — molecular, cellular, neuronal, or digital — in order to test how

¹³This is reminiscent of the unresolved debate about “conscious grids” in IIT [Merker *et al.*, 2022; Grasso *et al.*, 2021]. Some might find this so counter-intuitive so as to count as as falsification of the theory. Others find this to be a major predictive strength of the formalism.

¹⁴In a different context, this has recently been operationalized as “SLP-tests” that test for the subjective-linguistic, latent-emergent, and phenomenological-structural dimensions of AI-interfaces [Prentner, 2025a].

30 REFERENCES

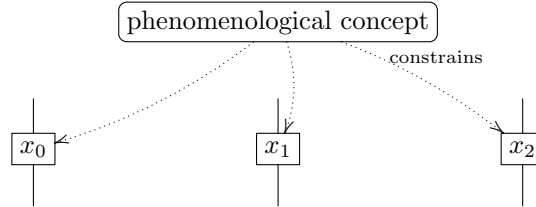


Fig. 12: Schematic illustration of how individual computational processes are organized under a phenomenological concept. The concept functions as a constraint on processing, not as another layer of computation, as indicated by the dotted arrows. This jointly defines the phenomenological structure of the interface instantiated by a computational system.

far the categorical framework generalizes across levels of organization. Such comparative analyses would help clarify whether the same phenomenological concepts can constrain processes at multiple scales, and whether artificial systems can be meaningfully embedded into this scheme.

Consciousness is not an event in the world; it is the event in which a world appears.

References

- Abramsky, S. and Coecke, B. [2008] Categorical quantum mechanics, in K. Engesser, D. M. Gabbay & D. Lehmann (eds.), *Handbook of Quantum Logic and Quantum Structures* (Elsevier), pp. 261–323.
- Albantakis, L. [2025] On the utility of toy models for theories of consciousness, <https://doi.org/10.48550/arXiv.2508.00190>.
- Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P. and Tononi, G. [2023] Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms, *PLoS Comp Bio* **19**(10), e1011465, [doi:10.1371/journal.pcbi.1011465](https://doi.org/10.1371/journal.pcbi.1011465).
- Albarracin, M., Pitliya, R. J., Ramstead, M. J. D. and Yoshimi, J. [2022] Mapping husserlian phenomenology onto active inference, <https://doi.org/10.48550/arXiv.2208.09058>.
- Andrews, K., Birch, J., Sebo, J. and Sims, T. [2024] Background to the New York Declaration on Animal Consciousness, <https://sites.google.com/nyu.edu/nydeclaration/declaration>.
- Association for Mathematical Consciousness Science [2023] The responsible development of ai agenda needs to include consciousness research, <https://amcs-community.org/open-letters/>.

- Atmanspacher, H. and Rickles, D. [2022] *Dual-Aspect Monism and the Deep Structure of Meaning* (Routledge, London).
- Bach, J. [2019] The cortical conductor theory: Towards addressing consciousness in AI models, in A. V. Samsonovich (ed.), *Biologically Inspired Cognitive Architectures 2018* (Springer, Cham), pp. 16–26.
- Barandiaran, X., Di Paolo, E. and Rohde, M. [2009] Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action, *Adaptive Behavior* **17**(5), 367 – 386, doi:10.1177/1059712309343819.
- Bayne, T. [2010] *The Unity of Consciousness* (Oxford University Press, Oxford).
- Bayne, T. and Chalmers, D. J. [2003] What is the Unity of Consciousness? in A. Cleeremans (ed.), *The Unity of Consciousness: Binding, Integration, Dissociation* (Oxford University Press, Oxford), pp. 23–58.
- Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A., Razi, A. and Mudrik, L. [2024] Tests for consciousness in humans and beyond, *Trends in Cognitive Sciences* **28**(5), 454–466, doi:10.1016/j.tics.2024.01.010.
- Birch, J. [2022] The search for invertebrate consciousness, *Noûs* **56**(1), 133–153, doi:10.1111/nous.12351.
- Blum, L. and Blum, M. [2024] AI consciousness is inevitable: A theoretical computer science perspective, <https://doi.org/10.48550/arXiv.2403.17101>.
- Bradley, T.-D. [2018] What is applied category theory? <https://doi.org/10.48550/arXiv.1809.05923>.
- Brukner, Č. [2014] Quantum causality, *Nature Physics* **10**, 259 – 263, doi:10.1038/nphys2930.
- Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A., Schwitzgebel, E., Simon, J. and VanRullen, R. [2025] Identifying indicators of consciousness in AI systems, *Trends in Cognitive Sciences*, S1364661325002864 doi:10.1016/j.tics.2025.10.011.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G. and et al. [2023] Consciousness in artificial intelligence: Insights from the science of consciousness, <http://arxiv.org/abs/2308.08708>.
- Chalmers, D. J. [2002] Consciousness and Its Place in Nature, in D. J. Chalmers (ed.), *Philosophy of Mind. Classical and Contemporary Readings* (Oxford University Press, Oxford), pp. 247–272.
- Chen, K.-C. and Yoshimi, J. [2023] The metaphysical neutrality of cognitive science, *Synthese* **201**(2), 63, doi:10.1007/s11229-023-04046-0.
- Clark, A. [2017] How to knit your own markov blanket, in T. Metzinger & W. Wiese (eds.), *Philosophy and Predictive Processing* (MIND Group, Frankfurt am Main), doi:10.15502/9783958573031.
- Coecke, B. [2021] Compositionality as we see it, everywhere around us, <https://doi.org/10.48550/arXiv.2110.05327>.

32 REFERENCES

- Dainton, B. [2000] *The Stream of Consciousness: Unity and Continuity in Conscious Experience* (Routledge, London).
- Dainton, B. [2023] Temporal Consciousness, in E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, Spring 2023 ed. (Metaphysics Research Lab, Stanford University).
- Dehaene, S., Lau, H. and Naccache, L. [2017] What is consciousness, and could machines have it? *Science* **358**(6362), 486–492, doi:10.1126/science.aan8871.
- Dung, L. [2024] Understanding artificial agency, *The Philosophical Quarterly* **75**(2), 450–472, doi:10.1093/pq/pqad056.
- Edelsbrunner, H. and Harer, J. [2010] *Computational Topology: An Introduction*, Graduate Studies in Mathematics, Vol. 69 (American Mathematical Society), ISBN 978-0-8218-4925-5.
- Ehresmann, A. C. and Gomez-Ramirez, J. [2015] Conciliating neuroscience and phenomenology via category theory, *Progress in Biophysics and Molecular Biology* **119**(3), 347 – 359.
- Ehresmann, A. C. and Vanbremeersch, J.-P. [2007] *Memory Evolutive Systems: Hierarchy, Emergence, Cognition* (Elsevier, Amsterdam).
- Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., P. Lang, J., M. Haun, A., Albantakis, L., Boly, M. and Tononi, G. [2021] Consciousness and the fallacy of misplaced objectivity, *Neuroscience of Consciousness* **2021**(2), niab032, doi:10.1093/nc/niab032.
- Floridi, L. [2025] AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis, *Philosophy & Technology* **38**(30), 1–18, doi:10.1007/s13347-025-00858-9.
- Frankish, K. [2016] Illusionism as a Theory of Consciousness, *Journal of Consciousness Studies* **23**(11-12), 11–39.
- Franklin, S. and Graesser, A. [1997] “Is it an agent, or just a program? a taxonomy for autonomous agents,” in J. Müller, M. Wooldridge & N. Jennings (eds.), *Intelligent Agents III: Agent Theories, Architectures, and Languages*, Lecture Notes in Computer Science, Vol. 1193 (Springer), pp. 21–35, doi:10.1007/BFb0013570.
- Friston, K. J., Wiese, W. and Hobson, J. A. [2020] Sentience and the origins of consciousness: From cartesian duality to markovian monism, *Entropy* **22**(5), 516, doi:10.3390/e22050516.
- Gallagher, S. and Zahavi, D. [2008] *The phenomenological mind: An introduction to the philosophy of mind and cognitive science* (Routledge, London).
- Gallagher, S. and Zahavi, D. [2021] Phenomenological Approaches to Self-Consciousness, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2021 ed. (Metaphysics Research Lab, Stanford University).
- Gamez, D. [2008] Progress in machine consciousness, *Consciousness and Cognition* **17**(3), 887–910.
- Goff, P. [2019] *Galileo’s Error. Foundations for a New Science of Consciousness* (Vintage Books, New York).

- Grasso, M., Haun, A. M. and Tononi, G. [2021] Of maps and grids, *Neuroscience of Consciousness* **2021**(2), niab022, [doi:10.1093/nc/niab022](https://doi.org/10.1093/nc/niab022).
- Graziano, M. S. A. [2022] A conceptual framework for consciousness, *Proc. Natl. Acad. Sci. USA* **119**(18), e2116933119.
- Gurwitsch, A. [1964] *The Field of Consciousness* (Duquesne University Press, Pittsburgh).
- Hartimo, M. [2010] *Husserl and Mathematics* (Springer, Dordrecht).
- Hoffman, D. D. and Prakash, C. [2014] Objects of consciousness, *Frontiers in Psychology* **5**, 577, [doi:10.3389/fpsyg.2014.00577](https://doi.org/10.3389/fpsyg.2014.00577).
- Hoffman, D. D., Prakash, C. and Prentner, R. [2023] Fusions of consciousness, *Entropy* **25**(1), 129, [doi:10.3390/e25010129](https://doi.org/10.3390/e25010129).
- Hoffman, D. D., Singh, M. and Prakash, C. [2015] The interface theory of perception, *Psychonomic Bulletin & Review* **22**(6), 1480–1506, [doi:10.3758/s13423-015-0890-8](https://doi.org/10.3758/s13423-015-0890-8).
- Horvat, S. and Northoff, G. [2022] The self and its time - a non-reductive neurophenomenological perspective on the brain's spontaneous activity, *Mind and Matter* **20**(2), 195–216.
- Husserl, E. [1966] *The Phenomenology of Internal Time-Consciousness* (Indiana University Press, Bloomington, IN).
- Husserl, E. [1970] *The Crisis of European Science and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy* (Northwestern University Press, Evanston, IL).
- Husserl, E. [2003] *Philosophy of Arithmetic: Psychological and Logical Investigations* (Kluwer Academic Publishers, Dordrecht), originally published 1891; English translation of *Philosophie der Arithmetik*.
- Husserl, E. [2012] *Ideas. General Introduction to Pure Phenomenology* (Routledge, London).
- Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M. and Guttenberg, N. [2019] Information generation as a functional basis of consciousness, *Neuroscience of Consciousness* **5**(1), niz016.
- Kanai, R. and Fujisawa, I. [2024] Toward a universal theory of consciousness, *Neuroscience of Consciousness* **1**, niae022, [doi:10.1093/nc/niae022](https://doi.org/10.1093/nc/niae022).
- Kant, I. [2004] *Prolegomena to Any Future Metaphysics* (Cambridge University Press, Cambridge).
- Kent, L. and Wittmann, M. [2021] Time consciousness: the missing link in theories of consciousness, *Neuroscience of Consciousness* **2021**(2), niab011, [doi:10.1093/nc/niab011](https://doi.org/10.1093/nc/niab011).
- Kleiner, J. [2024] Towards a structural turn in consciousness science, *Consciousness & Cognition* **119**, 103653, [doi:10.1016/j.concog.2024.103653](https://doi.org/10.1016/j.concog.2024.103653).
- Kleiner, J. and Ludwig, T. [2024] What is a mathematical structure of conscious experience? *Synthese* **203**, 89, [doi:10.1007/s11229-024-04503-4](https://doi.org/10.1007/s11229-024-04503-4).

34 REFERENCES

- Kuhn, R. L. [2024] A landscape of consciousness: Toward a taxonomy of explanations and implications, *Progress in Biophysics and Molecular Biology* **190**, 28–169, [doi:10.1016/j.pbiomolbio.2023.12.003](https://doi.org/10.1016/j.pbiomolbio.2023.12.003).
- Lawvere, F. W. and Schanuel, S. H. [2009] *Conceptual Mathematics: A First Introduction to Categories*, 2nd ed. (Cambridge University Press, Cambridge).
- Leibniz, G. W. [2005] *Discourse on Metaphysics and The Monadology* (Dover, Mineola, NY), originally published as *Discours de métaphysique* (1686) and *La Monadologie* (1714). Translated by G. Montgomery.
- Mac Lane, S. [1998] *Category Theory for the Working Mathematician* (Springer, Berlin/Heidelberg).
- Maria, C., Boissonnat, J.-D., Glisse, M. and Yvinec, M. [2014] Gudhi: Geometry understanding in higher dimensions, *International Congress on Mathematical Software*, 167–174.
- Menary, R. [2010] Introduction to the special issue on 4e cognition, *Phenomenology and the Cognitive Sciences* **9**(4), 459–463, [doi:10.1007/s11097-010-9187-6](https://doi.org/10.1007/s11097-010-9187-6).
- Merker, B., Williford, K. and Rudrauf, D. [2022] The integrated information theory of consciousness: A case of mistaken identity, *Behavioral and Brain Sciences* **45**, e41, [doi:10.1017/S0140525X21000881](https://doi.org/10.1017/S0140525X21000881).
- Metzinger, T. [2003] *Being No One. The Self-Model Theory of Subjectivity* (MIT Press, Cambridge, MA).
- Metzinger, T. [2020] Minimal phenomenal experience, *Philosophy and the Mind Sciences* **1**, 7 – 44.
- Mileski, B. [2019] Category Theory for Programmers, <https://github.com/hmemcpy/milewski-ctfp-pdf/>.
- Newen, A., De Bruin, L. and Gallagher, S. (eds.) [2018] *The Oxford Handbook of 4E Cognition* (Oxford University Press, Oxford).
- Newmans, M. [2010] *Networks. An Introduction* (Oxford University Press, Oxford).
- Nisbett, R. E. and Wilson, T. D. [1977] Telling more than we can know: Verbal reports on mental processes, *Psychological Review* **84**(3), 231–259, [doi:10.1037/0033-295X.84.3.231](https://doi.org/10.1037/0033-295X.84.3.231).
- nLab authors [2022a] equivalence of categories, <https://ncatlab.org/nlab/show/equivalence+of+categories>, Revision 43.
- nLab authors [2022b] higher category theory, <http://ncatlab.org/nlab/show/higher+category+theory>, Revision 78.
- nLab authors [2024a] 2-category, <https://ncatlab.org/nlab/show/2-category>, Revision 64.
- nLab authors [2024b] colimit, <https://ncatlab.org/nlab/show/colimit>, Revision 27.
- nLab authors [2025] filtered category, <https://ncatlab.org/nlab/show/filtered+category>, Revision 34.

- Oizumi, M., Albantakis, L. and Tononi, G. [2014] From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *PLoS Computational Biology* **10**(5), e1003588, [doi:10.1371/journal.pcbi.1003588](https://doi.org/10.1371/journal.pcbi.1003588).
- Parr, T., Pezzulo, G. and Friston, K. [2022] *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior* (MIT Press, Cambridge, MA), [doi:10.7551/mitpress/12441.001.0001](https://doi.org/10.7551/mitpress/12441.001.0001).
- Petitot, J., Varela, F. J., Pachoud, B. and Roy, J.-M. (eds.) [1999] *Naturalizing Phenomenology. Issues in Contemporary Phenomenology and Cognitive Science* (Stanford University Press, Stanford, CA).
- Prakash, C., Fields, C., Hoffman, D. D., Prentner, R. and Singh, M. [2020] Fact, Fiction, and Fitness, *Entropy* **22**(5), 514, [doi:10.3390/e22050514](https://doi.org/10.3390/e22050514).
- Prakash, C., Stephens, K. D., Hoffman, D. D., Singh, M. and Fields, C. [2021] Fitness Beats Truth in the Evolution of Perception, *Acta Biotheoretica* **69**, 319–341, [doi:10.1007/s10441-020-09400-0](https://doi.org/10.1007/s10441-020-09400-0).
- Prentner, R. [2019] Consciousness and Topologically Structured Phenomenal Spaces, *Consciousness and Cognition* **12**(1), 93–118, [doi:10.1016/j.concog.2019.02.002](https://doi.org/10.1016/j.concog.2019.02.002).
- Prentner, R. [2021] Dr Goff, Tear Down This Wall! The Interface Theory of Perception and the Science of Consciousness, *Journal of Consciousness Studies* **28**(9-10), 91–103, [doi:10.53765/20512201.28.9.091](https://doi.org/10.53765/20512201.28.9.091).
- Prentner, R. [2024] Category theory in consciousness science: going beyond the correlational project, *Synthese* **204**, 69, [doi:10.1007/s11229-024-04718-5](https://doi.org/10.1007/s11229-024-04718-5).
- Prentner, R. [2025a] “Artificial consciousness as interface representation,” in M. Iklé, A. Kolonin & M. Bennett (eds.), *Artificial General Intelligence (Lecture Notes in AI 16058)* (Springer Nature Switzerland, Cham), pp. 135–146, [doi:https://doi.org/10.1007/978-3-032-00800-8_12](https://doi.org/10.1007/978-3-032-00800-8_12).
- Prentner, R. [2025b] Mathematized phenomenology and the science of consciousness, *Phenomenology and the Cognitive Sciences* [doi:10.1007/s11097-025-10060-z](https://doi.org/10.1007/s11097-025-10060-z).
- Prentner, R. and Hoffman, D. D. [2024] Interfacing Consciousness, *Frontiers in Psychology* **15**, [doi:10.3389/fpsyg.2024.1429376](https://doi.org/10.3389/fpsyg.2024.1429376).
- Price, H. [1996] *Times’ Arrow and Archimedes Point: New Directions for the Physics of Time* (Oxford University Press, Oxford).
- Ramstead, M. J. D., Albarracin, M., Kiefer, A., Klein, B., Fields, C., Friston, K. and Safron, A. [2023] The inner screen model of consciousness: Applying the free energy principle directly to the study of conscious experience, <https://doi.org/10.48550/arXiv.2305.02205>.
- Roli, A., Jaeger, J. and Kauffman, S. A. [2022] How organisms come to know the world: Fundamental limits on artificial general intelligence, *Frontiers in Ecology and Evolution* **9**, 806283, [doi:10.3389/fevo.2021.806283](https://doi.org/10.3389/fevo.2021.806283).

36 REFERENCES

- Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K. and Williford, K. [2017] A mathematical model of embodied consciousness, *Journal of Theoretical Biology* **428**, 106 – 131, [doi:10.1016/j.jtbi.2017.05.032](https://doi.org/10.1016/j.jtbi.2017.05.032).
- Ruffini, G. and Lopez-Sola, E. [2022] Ait foundations of structured experience, *Journal of Artificial Intelligence and Consciousness* **9**(2), 153–191.
- Schutz, A. [1967] *The Phenomenology of the Social World* (Northwestern University Press, Evanston).
- Seth, A. K. and Bayne, T. [2022] Theories of consciousness, *Nature Reviews Neuroscience* **23**, 439–452, [doi:10.1038/s41583-022-00587-4](https://doi.org/10.1038/s41583-022-00587-4).
- Shevlin, H. [2024] Consciousness, machines, and moral status, <https://philarchive.org/rec/SHECMA-6>.
- Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S. and Farajtabar, M. [2025] “The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity,” in *Advances in Neural Information Processing Systems (NeurIPS)*, <https://arxiv.org/abs/2506.06941>.
- Signorelli, C. M., Cea, I. and Prentner, R. [2025] We need to explain subjective experience, but its explanation may not be mechanistic, *Neuroscience of Consciousness* In print.
- Signorelli, C. M., Szczotka, J. and Prentner, R. [2021a] Explanatory profiles of models of consciousness - towards a systematic classification, *Neuroscience of Consciousness* **2021**(2), niab021, [doi:10.1093/nc/niab021](https://doi.org/10.1093/nc/niab021).
- Signorelli, C. M., Wang, Q. and Coecke, B. [2021b] Reasoning about conscious experience with axiomatic and graphical models, *Consciousness and Cognition* **95**, 103168, [doi:10.1016/j.concog.2021.103168](https://doi.org/10.1016/j.concog.2021.103168).
- Signorelli, C. M., Wang, Q. and Khan, I. [2021c] A compositional model of consciousness based on consciousness-only, *Entropy* **23**, 308.
- Singhal, I. and Srinivasan, N. [2021] Time and time again: a multi-scale hierarchical framework for time-consciousness and timing of cognition, *Neuroscience of Consciousness* **2021**(2), niab020.
- Singhal, I. and Srinivasan, N. [2024] Just one moment: Unifying theories of consciousness based on a phenomenological “now” and temporal hierarchy, *Psychology of Consciousness: Theory, Research, and Practice* [doi:10.1037/cns0000393](https://doi.org/10.1037/cns0000393).
- Smith, D. W. [2018] Phenomenology, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2018 ed. (Metaphysics Research Lab, Stanford University).
- Stechly, K., Valmeekam, K. and Kambhampati, S. [2025] “On the self-verification limitations of large language models on reasoning and planning tasks,” in *International Conference on Learning Representations (ICLR)*, <https://arxiv.org/abs/2402.08115>.
- Strawson, G. [1997] The self, *Journal of Consciousness Studies* **4**(5-6), 405–428.
- Sutton, R. S. and Barto, A. [2018] *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press).

- Taguchi, S. [2019] Mediation-Based Phenomenology: Neither Subjective nor Objective, *Metodo* **7**(2), 17–44, [doi:10.19079/metodo.7.2.17](https://doi.org/10.19079/metodo.7.2.17).
- Taguchi, S. and Saigo, H. [2023] The Monoid-Now: A Category Theoretic Approach to the Structure of Phenomenological Time-Consciousness, *Frontiers in Psychology* **14**, 1237984, [doi:10.3389/fpsyg.2023.1237984](https://doi.org/10.3389/fpsyg.2023.1237984).
- Tewes, C. and Stanghellini, G. (eds.) [2020] *Time and Body: Phenomenological and Psychopathological Approaches* (Cambridge University Press, Cambridge).
- Tononi, G., Boly, M., Massimini, M. and Koch, C. [2016] Integrated information theory: from consciousness to its physical substrate, *Nature Reviews Neuroscience* **17**(7), 450–461, [doi:10.1038/nrn.2016.44](https://doi.org/10.1038/nrn.2016.44).
- Tsuchiya, N. and Saigo, H. [2021] A relational approach to consciousness: categories of level and contents of consciousness, *Neuroscience of Consciousness* **2021**(2), niab034, [doi:10.1093/nc/niab034](https://doi.org/10.1093/nc/niab034).
- Tsuchiya, N., Taguchi, S. and Saigo, H. [2016] Using category theory to assess the relationship between consciousness and integrated information theory, *Neuroscience Research* **107**, 1 – 7, [doi:10.1016/j.neures.2015.12.007](https://doi.org/10.1016/j.neures.2015.12.007).
- Tull, S. and Kleiner, J. [2021] Integrated information in process theory, *Journal of Cognitive Science* **22**(2), 92–123, the arXiv version can be accessed at <https://arxiv.org/abs/2002.07654>.
- Usher, M., Negro, N., Jacobson, H. and Tsuchiya, N. [2023] When philosophical nuance matters: safeguarding consciousness research from restrictive assumptions, *Frontiers in Psychology* **14**, 1306023, [doi:10.3389/fpsyg.2023.1306023](https://doi.org/10.3389/fpsyg.2023.1306023).
- VanRullen, R. and Kanai, R. [2021] Deep learning and the global workspace theory, *Trends in Neuroscience* **44**(9), 692–704.
- Voltolini, A. [2024] *Intentionality as Constitution* (Routledge, New York), [doi:10.4324/9781003299714](https://doi.org/10.4324/9781003299714).
- Wiese, W. [2017] Predictive processing and the phenomenology of time consciousness - a hierarchical extension of Rick Grush's trajectory estimation model. in T. Metzinger & W. Wiese (eds.), *Philosophy and Predictive Processing* (MIND Group, Frankfurt am Main), [doi:10.15502/9783958573277](https://doi.org/10.15502/9783958573277).
- Williford, K., Bennequin, D. and Rudrauf, D. [2022] Pre-reflective self-consciousness & projective geometry, *Review of Philosophy and Psychology* **13**(2), 365–396, [doi:10.1007/s13164-022-00638-w](https://doi.org/10.1007/s13164-022-00638-w).
- Yoshimi, J. [2011] Phenomenology and connectionism, *Frontiers in Psychology* **2**, 288, [doi:10.3389/fpsyg.2011.00288](https://doi.org/10.3389/fpsyg.2011.00288).
- Yoshimi, J. [2016] *Husserlian Phenomenology: A Unifying Interpretation* (Springer, Cham), [doi:10.1007/978-3-319-26698-5](https://doi.org/10.1007/978-3-319-26698-5).

Appendix A: Persistent Homology, Clique Complexes, and the Inclusion-Category

The aim of this appendix is to make explicit how to compute persistent homology, which is standardly defined for filtrations of simplicial complexes, and show how this is related to the inclusion-category used to connect this to phenomenology.

The standard way to derive a simplicial filtration from a $\mathcal{Q}$ -network goes as follows:

- (1) The **underlying graph** G is given by the $\mathcal{Q}$ -network, which resulted from embedding physiological data (in the toy model: a synthetic timeseries) into a phenomenal space; vertices correspond to embedding points, edges correspond to cosine-similarities between embeddings in the phenomenal space.
- (2) **Enumerate all cliques** C (up to a chosen maximum size if needed). A clique is defined as a set of vertices that are pairwise connected by edges. The clique complex (also called the flag complex) of a graph replaces every k -clique by a $(k - 1)$ -simplex [Edelsbrunner & Harer, 2010].
- (3) A nested sequence of simplicial complexes $K_0 \subset K_1 \subset \dots \subset K_T$ is obtained by varying a scale or threshold parameter. In our case, the threshold parameter is defined with respect to the cosine-similarities. The **filtration** $\{K_t\}$ is obtained by lowering the similarity threshold, thereby adding edges and hence new cliques.
- (4) **Persistent homology** is computed by evaluating homology groups $H_k(K_t)$ across the filtration and records the birth and death of homological features (connected components, loops, voids) as intervals (persistence). The output is typically barcodes or persistence diagrams. For our toy-model, we recorded the i -th Betti-number to map persistent features back to domain semantics (e.g., long-lived 1-cycles indicate robust loop structures among cliques.)

The inclusion-category is a categorical rephrasing of the simplicial complex structure.¹⁵ It encodes the same combinatorial data as the clique complex: every object corresponds to a simplex, and inclusions correspond to face relations. Because morphisms are inclusions, there is a canonical nerve/simplicial realization: the nerve of the inclusion-category is (isomorphic to) the clique complex. This gives a direct, canonical route to persistent homology.

While long persistence suggests robustness, not every persistent feature implies a direct phenomenological or causal structure without further validation. Moreover, if future work introduces richer categorical morphisms (beyond mere inclusions), then the nerve construction may no longer capture the intended semantics and a careful theory of persistent homology on such categories would be required. That is a substantial mathematical project that goes outside the present scope.

¹⁵Rather than, for example, a free categorical construction that would possibly introduce formal compositional cycles unrelated to geometric simplices.