Title: Applying Yoneda's lemma to consciousness research: categories of level and contents of consciousness

Naotsugu Tsuchiya

ORCID: 0000-0003-4216-8701

[1] School of Psychological Sciences and Turner Institute for Brain and Mental Health, Monash University, Melbourne, Victoria, Australia

[2] Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Suita, Osaka 565-0871, Japan

[3] Advanced Telecommunications Research Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan.

Hayato Saigo

ORCID: 0000-0002-4209-352X

Nagahama Institute of Bio-Science and Technology, 1266 Tamura-cho, Nagahama, Shiga 526-0829, Japan

Abstract

Characterizing consciousness in and of itself is notoriously difficult. Any effort to define consciousness seems to evade what it tries to achieve. In particular, definitions often involve comparisons of different kinds of “consciousness” in a self-referential manner. The tautological nature of characterizing consciousness has led some scholars to propose that establishing a science of consciousness is infeasible. Here, we propose an alternative approach to characterize, and to eventually define, consciousness through exhaustive descriptions of consciousness’ relationships to all other consciousness. This approach is mathematically founded in category theory. Indeed, category theory can prove two objects A and B in a category can be equivalent if and only if all the relationships that A holds with others in the category are the same as those of B; this proof is called the Yoneda lemma. To introduce the Yoneda lemma, we gradually introduce key concepts of category theory to consciousness researchers in this paper. Along the way, we propose several possible definitions of categories of consciousness, both in terms of level and contents, through the usage of simple examples. We also propose empirical research programs that can test the validity of our proposed categories of consciousness and to improve them. We propose to use the categorical structure of consciousness as a gold standard if one tries to empirically test some structural theories of consciousness, such as Integrated Information Theory of consciousness.

1. Problems in characterizing consciousness on its own

Over the last few decades, the question of the nature of consciousness is gaining a respected position as the target of scientific inquiry. Psychologists and cognitive neuroscientists have accumulated a massive amount of evidence that can be used to investigate the nature of the link between the physical world of materials with its electrical and chemical interactions and the mental world with subjective conscious experience. Much of the empirical research has originated from this research program, which has tried to identify the neural correlates of consciousness ([Koch, Massimini, Boly, & Tononi, 2016](#)). Building on the massive amount of empirical evidence, some models or theories have been proposed to explain the link between various features and aspects of the neural activity and the associated functions and phenomena of consciousness ([Brown, Lau, & LeDoux, 2019](#); [Engel & Singer, 2001](#); [Graziano & Kastner, 2011](#); [Lamme, 2015](#); [Mashour, Roelfsema, Changeux, & Dehaene, 2020](#); [G. Northoff & Huang, 2017](#)).

Another distinct approach to consciousness is to study it using a mathematical framework ([Fekete & Edelman, 2011](#); [Hoffman, 1966, 1980](#); [Kleiner, 2019](#); [Prentner, 2019](#); [Stanley, 1999](#); [Tononi, 2004](#); [Yoshimi, 2011](#)). Among these, the Integrated Information Theory (IIT) of consciousness by Tononi and colleagues ([Oizumi, Albantakis, & Tononi, 2014](#); [Tononi, Boly, Massimini, & Koch, 2016](#)) is arguably the most developed and discussed in the literature. IIT takes a unique approach, where it tries to first identify the essential properties of consciousness, which are always true to any experience. These essential properties are called “phenomenological axioms” and IIT tries to derive mathematical postulates that any physical system should satisfy to support these properties. IIT then assumes an explanatory “identity” between phenomenal consciousness and information structure ([Andrew Haun & Tononi, 2019](#)).

In the past ([Tsuchiya, Taguchi, & Saigo, 2016](#)) we, the authors, have suggested that, rather than assuming the “identity” between the two in the construction of the theory, it would be empirically more tractable to break up the IIT project into the following three sub-projects: 1) to characterize the structure of conscious phenomenology as a category; 2) to characterize the structure of information as a category; and 3) to assess the degree of similarity between the structures. For the second issue, we and others have made some initial efforts on the mathematical side ([Kleiner & Tull, 2020](#); [Georg Northoff, Tsuchiya, & Saigo, 2019](#); [Tull & Kleiner, 2020](#)). For the third issue, we also made an initial attempt ([A. Haun, Oizumi, et al., 2017](#)). Here, we deal with the first issue: how we can characterize the structure of consciousness as a category.

Characterizing or defining consciousness has been notoriously difficult or very controversial at least ([Dennett, 1988](#); [Sloman, 1991](#); [Stanovich, 1991](#); [Velmans, 2009](#)). Indeed, the essential properties of consciousness proposed by IIT have met various criticisms ([Tim Bayne, 2018](#); [Pautz, 2019](#)). While a precise definition of consciousness per se is indeed difficult, certain characterizations of consciousness are understood and accepted by many for use in empirical consciousness research ([Koch et al., 2016](#); [Mashour et al., 2020](#)). One

of the widespread consensus is to distinguish “level” (or quantity) and “contents” (or quality) of consciousness ([Boly et al., 2013](#); [Laureys, 2005](#)) (also see ([Van Gulick, 2018](#))). For the debates about the relationships between these concepts, see ([Hohwy, 2009](#))). For example, Searle ([Searle, 2000](#)) gives the following definition: “Consciousness consists of inner, qualitative, subjective states and processes of sentience or awareness. Consciousness, so defined, begins when we wake in the morning from a dreamless sleep and continues until we fall asleep again, die, go into a coma, or otherwise become “unconscious””. In the case of the contents of consciousness or qualia ([Balduzzi & Tononi, 2009](#); [Kanai & Tsuchiya, 2012](#); [Tye, 2018](#)), a particular conscious experience of red color is typically characterized as “redness of red” or “red like this wine”. Regardless of whether it is about level or contents, consciousness is usually defined or characterized in reference to or in analogy to another conscious experience. This self-referential nature of consciousness appears to be one of the fundamental characteristics of consciousness. The study of consciousness is fundamentally ill-suited to those that require a consensus definition of the phenomenon before commencing their research ([Dennett, 1988](#); [Sloman, 1991](#); [Stanovich, 1991](#)).

In this paper, by introducing several notions of category theory, we offer mathematical justification for the self-referential nature of defining and characterizing consciousness. In fact, situations where a particular object is not possible to define or even characterize on its own arise relatively often outside of consciousness research. In these situations, what has been effective is to rely on the relationships between the object to be defined and its surroundings. For example, some linguists consider that meanings of a word can be understood only through how the word is related to other words and how they are put into the context in the sentence (e.g., ([Frege, 1980](#))). In ecology, it is essential to characterize any life form within an ecosystem; defining a tree without mentioning the geological area in which it lives, and which animals, insects and other plants interact with it and in what way would miss the very essence of what that tree is. In quantum theory, the essence of quantum features is that they are only possible to explain through the interactions between the objects ([Coecke & Kissinger, 2017](#)). In cosmology, blackholes are, in themselves, impossible to measure and characterize, but their interaction with their neighbors can be measured and used to characterize their properties. In mathematics, various types of infinity can be distinguished through what types of relationships they have with other mathematical objects.

There is a deep mathematical foundation in why these theoretical fields can dispense with a direct definition of an object in favor of characterizing its interaction with its surroundings. That is what we introduce to consciousness research in this paper: the Yoneda lemma. In short, the Yoneda lemma allows us to equate A with B (up to isomorphism defined in Section 2.2) if the relationships between A and the rest of objects are the same (up to natural equivalence defined in Section 4.2.1) as those between B and the rest. Importantly, this is true, even if A and B themselves are difficult to define or characterize in and of themselves, as in the case of blackholes in cosmology or infinity in mathematics. While defining consciousness directly in a way where everyone agrees is not easy, characterizing consciousness through a rich set of relationships is much more feasible.

Note that we will not offer a definition(s) of consciousness in this article. Rather, our plan is to introduce a novel perspective on how we can do so through the application of the Yoneda

lemma. To apply the Yoneda lemma, we need to propose several possible categories of consciousness. After introducing key concepts in category theory, we will come back to the issue of how we can apply the Yoneda lemma in consciousness research and discuss what it means for consciousness research. While the Yoneda perspective has been implicitly taken by some consciousness researchers, we believe that making it explicit and giving mathematical footings will be critical for further progress in consciousness research.

2. What's the category in category theory?

2.1 Proposing categories of consciousness

The Yoneda lemma, which deals with the characterization problem of consciousness, is a very important result in category theory. In this section, we will explain the basics of category theory, which is necessary to explain the Yoneda lemma with a view to its application in consciousness research. In particular, we will propose two categories: category of level of consciousness, L , and category of contents of consciousness or qualia, Q . Note that we are not claiming that these examples are “the” only possible formulation of categories of consciousness. They are invoked here as a starting point. For an accessible introduction to category theory, see ([Bradley, 2017](#); [Lawvere & Schanuel, 2009](#); [Georg Northoff et al., 2019](#); [Spivak, 2014](#)).

The basis of category theory is three concepts: 1) category, 2) functor and 3) natural transformation. Let's explain category first. A category is a system consisting of “objects” and “arrows”. For a category of level of consciousness, we consider global states of consciousness, or the degree of wakefulness as an object (Note that one object can contain many elements). The level of consciousness is usually assumed to go from 0 in dead humans, which is lower than deeply anesthetized or dreamless sleep, and highest in fully wakeful states, which is largely accepted by the community and can be potentially verifiable subjectively ([Mormann & Koch, 2007](#)). We will discuss the subtlety later ([T. Bayne, Hohwy, & Owen, 2016](#); [Pautz, 2019](#)). For a category of qualia, we can consider any possible content of consciousness, such as color, sound and pain. As is usually the case in mathematics, however, starting from the most general situation, encompassing all levels of consciousness in all animals or all types of qualia, is not a wise strategy to make progress. Accordingly, we will consider a very concrete case of consciousness categories below and leave generalization of these categories to future work.

An arrow, $\rightarrow$, relates an object with another object. For example, $f: A \rightarrow B$ denotes a relationship f between object A and object B . In the case of level of consciousness category L , we can define the meaning of f to be the level of consciousness in A “is higher or equal to” B . Soon, it will become clear why we need “equal to”. Read on! In the case of qualia category Q , we can define the meaning of f to be that A “is experienced as nearly indistinguishable with” B in a certain aspect.

If the system consisting of objects and arrows satisfy the further three conditions (composition, associativity, and unit) as depicted in Figure 1, then it qualifies as a category. We visualize these three requirements concisely and intuitively using diagrams. This is one

of the most powerful aspects of category theory: simply visualizing complex relationships to facilitate understanding of the subject matter.

Figure 1. Requirements of a category. a) Composition: if A , B , and C are objects in category X , and $f: A \rightarrow B$ and $g: B \rightarrow C$ are arrows in category X , then we can compose (or combine) f and g to obtain an arrow, $f;g: A \rightarrow C$. Note $f;g$ reads as “ f then g ” and it is often denoted as $g \circ f$ ([Fong & Spivak, 2019](#)). b) Associativity: if f , g , and h are arrows in category X , then the order to compose the arrows do not matter: $(f;g);h = f;(g;h)$. A , B , C , and D are objects in the category X . c) Unit: For any object A in the category X , there is a self-referential arrow $A \rightarrow A$, which is called identity arrow: $1A$. For any arrow $f: A \rightarrow B$, the following is always satisfied: $1A;f = f = f;1B$.

2.1.1. A category of level of consciousness L.

Let's verify whether our proposed objects and arrows for level of consciousness, L , can qualify as a category. Objects are certain levels of consciousness supported by a human brain, such as A , B , C , etc. An arrow f exists between two conscious levels A and B , if level of consciousness A “is higher than or equal to” B .

Consider the requirement of composition. Suppose that $f: A \rightarrow B$ and $g: B \rightarrow C$ exist as arrows for objects A , B , and C in consciousness level category L . That is, the level of consciousness A is higher or the same as B , and B is higher or the same as C . Then it follows that A is higher than C or the same. In other words, there is an arrow from A to C , which is the combination of f and g , which we denote $f;g$. Therefore, the condition for composition is satisfied.

Next, let's consider associativity. We assume that for levels of consciousness A , B , C , and D , three arrows exist $f: A \rightarrow B$, $g: B \rightarrow C$, and $h: C \rightarrow D$. Then, $(f;g);h$ means that 1) we first confirm that A is higher than or the same as C (i.e., $f;g$) and 2) due to h (i.e., C is higher than or equal to D), we conclude that A is higher than or the same as D . On the other hand, $f;(g;h)$ means that 1) we first confirm $g;h$, that is, B is higher than or equal to D , and 2) we conclude A is higher than or equal to D . In other words, associativity requires that these two conclusions have to be exactly the same. In the diagram (Figure 1B), this translates to the fact that two conclusions via the two paths are exactly the same if the starting and the ending points are the same.

Finally, let's check the unit. This is the reason why we have added “or equal to” as a part of the definition of the arrow. For any level of consciousness, A , there exists an arrow $1_A: A \rightarrow A$ because A is always higher than or equal to A . If $f: A \rightarrow B$ exists, then $1_A \circ f$ means that 1) after confirming A is the same as A , then 2) conclude A is higher than or equal to B . Similarly, $f \circ 1_B$ means that 1) after confirming A is higher than or equal to B , then 2) confirm that B is equal to B . If we define $f: A \rightarrow B$ as A is strictly higher than B , our proposed system of objects and arrows for level of consciousness L does not qualify as a category.

Our category of level of consciousness L is an example of categories called “preordered sets”. Preordered sets are categories such that between any two objects there is at most one arrow. The arrows in preordered sets are called “preorder”. We believe that introducing the concept of preordered sets as a category of level of consciousness L has an important implication in consciousness research. Some authors have raised issues with the concept of level of consciousness ([T. Bayne et al., 2016](#); [Pautz, 2019](#)). One of the issues is that this concept seems to imply a gradual change from unconscious to full consciousness, which is not widely agreed on. Another issue is the existence of certain pairs of conscious states, where we can never subjectively compare which state has “higher” in conscious level. For example, which is higher in level of consciousness between vivid dreaming and drowsy awakening states? Or, what about between under deep general anesthesia and coma, where consciousness completely disappears? We believe that these issues arise because the scholars who criticize the concept of level of consciousness implicitly assume that “level” is something that is isomorphic to natural numbers (or positive real numbers), where it is always possible to rank the order between two objects. Such a strongly ordered structure is called “total order” in mathematics. We believe that the total order assumption is unnecessary at all for the concept of level of consciousness to be useful. We propose that a looser concept, preorder, is a useful and appropriate concept for level of consciousness. As for preorder, any two objects (levels) may or may not have a relationship “ $\leq$ ”.

Our proposed category L can be applied to many cases. Compared to the fully wakeful states, deeply anesthetized, dreamless sleep, or coma would be lower in the level of consciousness. In L , we will have no arrows between cases of lower level of consciousness, where subjects have no ability to compare their levels and do not show any behavioral outputs to be meaningfully compared. Likewise, L can be proposed across animals without proposing any arrows between animal species but only considering arrows within animals (between fully wakeful and deeply anesthetized, deep sleep or coma).

2.1.2. A category of qualia Q .

Next, let's examine if our proposed definition can make a system of contents of consciousness or qualia Q into a category. Here, objects are the contents of consciousness. Let's consider a system Q that consists of only three objects, A = red sunset, B = red crayon, and C = red wine. For simplicity, we consider a case of a preordered set; that is, between two objects, there is either one arrow or none at all. Suppose there is an arrow from A to B when the contents of the two consciousness are “nearly indistinguishable” in terms of their color. In other words, if there is an arrow f between the two contents of consciousness A

and B, then $f: A \rightarrow B$, and A is subjectively felt as nearly the same as B in terms of their color. It is clear that the composition holds. In other words, if the redness of sunset and crayon are similar, and the redness of crayon and wine are similar, then the redness of sunset and wine are also similar. Mathematically speaking, if $f: A \rightarrow B$ and $g: B \rightarrow C$, then $f \circ g: A \rightarrow C$ also holds. Associativity holds as well. The unit is also valid (these two proofs are simple and left to the reader). Therefore, our proposed system of objects and arrows Q can constitute a category. Here, since arrows mean “nearly indistinguishable”, the direction of an arrow does not matter. Now that we are considering a preorder, there is only one arrow in either direction. This implies that any arrow in this category is invertible. That is, any arrow has an arrow of a reverse direction, such that the composition of the two is the identity arrow. Invertible arrows are also called “isomorphism”, not only in preordered sets but also in categories in general. When all arrows are isomorphism, the category is called “groupoid”. In sum, we proposed an exemplar category of qualia Q, which is preorder and groupoid, where all arrows are isomorphism (If there are multiple isomorphism between objects, then such a category is groupoid but not preorder).

The two categories, L and Q, proposed here are just toy examples and we are not claiming that they are “the” categories of consciousness. Such a proposal requires further investigations that require interdisciplinary collaborations between mathematicians and consciousness researchers. In both level and contents of consciousness, it is possible to focus on other aspects of consciousness and consider the corresponding categories. For example, by focusing on an aspect of consciousness that changes over time, one can propose a “category of mobility” ([Saigo, Naruse, Okamura, Hori, & Ojima, 2019](#)). Subjective experience of “meaning” of a word is often considered as one kind of conscious contents (but see ([Kemmerer, 2015](#); [McClelland & Bayne, 2016](#)) for recent discussion on this issue). Another kind of category, called “co-slice category”, is effective in capturing association possibilities and metaphorical structures of meaning ([Fuyama & Saigo, 2018](#)). In Appendix, we explain a further framework to loosen the condition of arrows in Q, by making it in the variable degrees of similarity, using a concept of enriched categories ([Fong & Spivak, 2019](#); [Lawvere, 1973](#); [Leinster & Meckes, 2017](#)).

The point of this section was to provide the readers an understanding that it is not difficult to propose a category of consciousness. Some readers may think that as a mathematical theory, category theory cannot be applied to the problem of consciousness, especially because of the problem in composition ([Pitt, 2018](#)). As we explained, however, category theory can be applied if “objects” and “arrows” are appropriately defined and if they satisfy a few conditions.

2.2 What does it mean for objects to be the same in a category? Isomorphic objects in a category.

So far, we have introduced category L and Q for level and contents of consciousness. Next, to deepen our understanding of what it means to define an object, we will consider what it means for objects A and B to be the “same”.

When we say that A and B are the same in everyday life, it means the same with respect to some definitions or assumptions. For example, when there are two apples on the right and the left, they may be the same in terms of its color or a category within fruit, but they are different in their surface textures. In fact, they are not the same thing, being located in different places. Category theory provides a wealth of mathematical tools to handle these subtle differences in “sameness” ([Tsuchiya et al., 2016](#)) in that it characterizes different kinds of sameness using various types of relationships.

Let us explain the sameness between objects in a category, called “isomorphism”. The objects A and B in a certain category C are isomorphic if there is an “invertible” arrow (called isomorphism) between them. When the arrow $f: A \rightarrow B$ is “invertible”, there exists an arrow $g: B \rightarrow A$ with $f \circ g = 1_A$ and $g \circ f = 1_B$. In the case of category L, if coma level A is lower than minimally conscious level B, the reverse of the arrow $f: A \rightarrow B$ does not exist. In our proposed example of the category of qualia with three objects, all arrows are isomorphisms (see section 2.1.2), that is, any pair of objects is isomorphic if there exists an arrow between them.

This concept of isomorphism is much looser and more flexible than the usual concept of “equalities”. For example, under the framework of set theory, a set of alphabets {a, b, c, d, e} is not equal to a set of numbers {1, 2, 3, 4, 5}. The difference in terms of “elements” disqualifies them to be considered as the same sets. Now, consider a category Set, whose objects are sets and arrows are functions or mappings between sets. Within the category Set, two objects (=sets) {a, b, c, d, e} and {1, 2, 3, 4, 5} are considered as isomorphic objects, in the sense that each element can be mapped from one set to the other set in a one-to-one manner. In fact, a concept of a number, such as 5, is “defined” when we regard these “different” sets as the “same” in a certain respect. In this sense, isomorphism captures the concept of “essential sameness” in the category. As another example of isomorphism between objects in a category, consider a category of topological space, Top, which is the foundation of the mathematical field called topology. In Top, objects are topological spaces and arrows are continuous maps. While coffee cups and donuts are usually considered as “completely different” objects, they are isomorphic in Top.

In order to make progress in consciousness research, “sameness” in the sense of isomorphism is not powerful enough. As we discuss later, we believe that different types of sameness in category theory will be more useful and find wider application in consciousness research. In particular, isomorphism between categories, or even looser sameness, called categorical equivalence will be critical. To introduce these concepts, we need to introduce two concepts: functors and natural transformations, which are the topics of the next two sections.

3. What is a functor?

3.1 A functor is an arrow between categories (which are considered as objects).

In simple terms, a functor is an arrow between two categories. Importantly, a functor needs

to map one category to another category while keeping the structure of the category. From the viewpoint of “sameness” discussed in the last chapter, “the existence of a functor” is one important condition to consider as the sameness between categories. The sameness as existence of a functor is much weaker than other types of sameness discussed in this paper. Mathematically, we define a mapping F , which maps an object and an arrow in category C to an object and an arrow in category D , as a functor if it satisfies the three conditions summarized in the diagram in Figure 2.

Figure 2. What is a functor? A structural mapping F between category C to D is called a functor if the following three conditions are satisfied: 1) an arrow $f: X \rightarrow Y$ in category C is mapped onto a corresponding arrow $F(f): F(X) \rightarrow F(Y)$ in category D ; and 2) a composite arrow $f;g$ in category C is preserved as $F(f;g) = F(f);F(g)$; and 3) an identity arrow 1_X for an object X in category C is preserved as $F(1_X) = 1_{F(X)}$.

To see the relevance of functors in consciousness research, let's examine a functor from categories of level of consciousness L . To remind you, L is a preordered set; objects and arrows in L are level of consciousness and " $\leq$ " as introduced in Section 2.1.1. One of the aims of consciousness research is to establish a consciousness meter, which translates to a construction of a functor from category L to a convenient preorder category D . As an example, consider a category D whose objects and arrows are natural numbers and $\leq$. Readers can check easily that D is a preorder category; any pair of objects have at most one arrow (e.g., $1 \leq 2$), any arrows can be composed in an associative manner (e.g., $1 \leq 2 \leq 3$) and that each object has a unit arrow (e.g., $1 \leq 1$).

Let's consider a mapping F that maps from category L to category D and satisfies the conditions to be a functor. In other words, F maps any object of L , that is, any level of consciousness, into a natural number, while all $\leq$ relations in L are preserved in D . In other words, searching for a functor F is the same as the construction of a consciousness meter as strived for by some consciousness researchers.

Next, let's consider a functor F for category Q for contents of consciousness as considered in Section 2.1.2. Q is a groupoid, and its three objects are A = red sunset, B = red crayons, and C = red wine. Arrows are “nearly indistinguishable” in terms of color. Here, we consider

two persons' category QX and QY. For Ms X, all objects' colors are nearly indistinguishable, thus there are arrows for any pair of objects AX, BX and CX. However, for Mr Y, all objects are distinguishable, thus there are no arrows among AY, BY, and CY. (Figure 3a)

Figure 3. Exemplar functors between categories QX and QY. A) In category QX, there exist arrows for all possible object pairs (only three arrows are depicted here). In category QY, there are no arrows except for identity arrows for each object (not drawn). B) One of 27 ($=3^3$) functors from QY to QX that preserve structures of QY. C) One of three functors from QX to QY is shown. The functor has to map all objects in QX into one object in QY and all arrows in QX to the identity arrow in QY.

In this situation, there are 27 functors that map from category QY to category QX, since the destination of each object in category QY can be any of AX, BX, and CX (Figure 3b). For example, one functor F1 maps AY, BY, CY into AX. The only arrows in QY are the identity arrows and they all map onto the identity arrow for AX; $F(1AY) = F(1BY) = F(1CY) = 1AX$. The requirement of composition is also satisfied (because the only way to compose arrows in QX is to compose the identity with the identity, which is the identity).

What about functors from category QX to category QY? Like the above example, if F collapses all objects in QX into one object in QY, F can satisfy all conditions to be a functor. If F does not collapse them, what happens? For example, if AX is mapped to AY while BX is mapped to BY. Because Mr Y can distinguish AY and BY, there is no arrow between AY and BY. However, Ms X sees AX as indistinguishable from BX. Thus, the arrow that

connects AX and BX is lost in translation, during this mapping. Such a faulty mapping does not qualify as a functor. From a viewpoint of “sameness”, there is a certain “structural similarity” in subjective similarity characterized by category QX and QY in a sense that there exists a functor from one category to the other while richness of the structure is “different” as can be quantified from the number of arrows in each category. This nicely represents the situation of subjective experience of Ms X and Mr Y – they experience the objects with certain structural similarity and difference in richness in the structure.

In sum, functors can be used to compare structures between categories. In terms of categories of contents of consciousness or qualia Q, functors are not limited in the usage of comparison of qualia structures among individuals as in our examples. Functors can be used to compare structure of qualia category of color with qualia category of shape, qualia category of meaning, etc within individuals. This is one of the possible future directions of category theoretical research of qualia.

3.2 Category of categories, Cat, whose objects and arrows are categories and functors.

In a category of categories, Cat, objects and arrows are categories and functors (and this is the origin of the title for Section 3.1. This is an interesting characteristic of category theory; depending on the viewpoint, what we are talking about can be considered as objects in a certain category or arrows in another category. Further, a functor, which is considered as an arrow between two categories in this section, will be considered as an object in the later section! All of this comes from the fact that category theory is mathematics that focuses on “relationships”.

Let's consider a category of categories using an example of category of qualia Q. We ask Ms X and Mr Y to come back on the stage. If we consider category QX and QY as objects and functor F as arrows, the example in Figure 3 is an exemplar category of category Q.

4. What is a natural transformation?

4.1. A natural transformation is an arrow between functors (which are considered as objects).

A natural transformation is the central concept in category theory. Indeed, category theory was originally invented to introduce the concept of natural transformation. To explain natural transformation, the concept of functor was introduced. And to explain the concept of functor, the concept of category was introduced. Natural transformations are necessary to state and prove the Yoneda lemma. The Yoneda lemma and its inspiration for consciousness research is the central message of the paper: even if it is difficult or impossible to characterize a quale per se, we can do so by characterizing all arrows for the quale in category Q.

A natural transformation maps a functor to another functor while keeping the structure of the functor. The basic structure of category theory can be stated in terms of categories, functors and natural transformations. While there are more concepts (e.g., enriched categories in Appendix), solid understanding up to natural transformation is critical to handle category theory. In particular, it is important to understand and to develop intuition that a natural transformation is “a family of arrows” in the category that a functor maps the original category into (Figure 4).

Figure 4. What is a natural transformation? a) Category C (consisting of objects X and Y and an arrow f) is mapped into category D by two functors F and G . Each functor preserves the structure of category C in D (e.g., $F(f): F(X) \rightarrow F(Y)$, $G(f): G(X) \rightarrow G(Y)$). Natural transformation t maps functor F into G while preserving its structure. t can be considered as a family of arrows in category D , such as tX and tY . Each arrow is specified by an object in the original category C , such as tX , which maps an object $F(X)$ to $G(X)$ in category D . b) Visualizing t as a family of arrows in the destination category D .

Consider functors F and G that map category C to category D . When a mapping t from functor F to G satisfies the following two conditions, t is called a natural transformation and we write $t: F \Rightarrow G$.

1. Given an object X in category C , t gives an arrow $tX: F(X) \rightarrow G(X)$ in category D .
2. For any arrow $f: X \rightarrow Y$ in category C , $tX \circ G(f) = F(f) \circ tY$ holds in category D .

tX is called the X component of t .

4.1.1 Natural transformation in category L .

To understand the relevance of natural transformations in consciousness research, let's first consider the category of level of consciousness L and the functors. Suppose Ms F proposed a functor F as a consciousness meter, which maps all levels in preorder category L to preorder category D (objects and arrows are natural numbers and $\leq$). F maps all objects in L to numbers in D and $\leq$ in L into $\leq$ in D . On F 's monitor, level of consciousness under deep general anesthesia shows up as a number 10, while that under the arousal state is 100. Then Mr G also developed another functor G as a consciousness meter. On G 's meter, level under the deep general anesthesia is 20 and the arousal state is 200. In both cases, the arrow in category C is preserved as the arrow in category D as $\leq$. Now, we ask; what is the nature of the relationship between F 's and G 's consciousness meters?

Let us assume that there exists a natural transformation t between functors F and G . Then, according to condition 1 above, the natural transformation t gives us two arrows t_A and t_W for two objects that we considered in category L : A for anesthesia and W for wakefulness. t_A and t_W are two arrows in category D which satisfies the following: $t_A: F(A)=10 \rightarrow G(A)=20$, $t_W: F(W)=100 \rightarrow G(W)=200$. Interestingly, this arrow corresponds to $\leq$, the arrow in D .

Moreover, according to condition 2, with respect to two objects A and W and the arrow $f: A \rightarrow W$ in category L , we have $t_A \circ G(\leq) = F(\leq) \circ t_W$. The left hand side of this equation means that 1) we first evaluate the value for anesthesia in F ($=10$) to translate into the value in G ($=20$), then 2) we confirm the relation between A and W ($\leq$) in category L to be translated into the relation in G 's monitor ($20 \leq 200$). The right hand side of the equation means that 1) we first translate the relation $A \leq W$ in category L into category D ($F(A)=10 \leq F(W)=100$), then 2) we map the value for wakefulness in F 's monitor ($=100$) into G 's monitor ($=200$). In sum, this family of arrows in category D guarantees a lawful relationship between two functors, preserving the original relations in category L . (Note that we also have the two more relations with respect to the identity arrows for 1_A and 1_W in category L).

4.1.2 Natural transformation for category Q .

Next, let's consider a natural transformation for the category of qualia Q . We consider two categories based on Q introduced in Section 2.1.2 and make it a situation that can be empirically tested in a psychophysical experiment (Zelevnikow-Johnston, Fan, and Tsuchiya, in preparation). For simplicity, we consider two objects as subjective experience of a red Apple (A) and a red Berry (B) and arrows as a relationship for "nearly indistinguishable" in terms of color (Figure 5). Category C refers to objects and arrows in the Central visual field while category E refers to those in the Entire visual field, which includes the central, left and right visual field. A mapping R from C to E translates objects and arrows to those in the Right visual field. Likewise, a mapping L from C to E translates those in the Left visual field. The mappings L and R are likely to be functors in healthy subjects (which we confirmed in our preliminary results). In the following, we assume L and R satisfy the functor conditions.

Figure 5. Natural transformation for category Q. Objects of Q are Apples (A) and Berries (B). Arrows of Q are “nearly indistinguishable” in terms of color. Functors L and R map Central vision to Left and Right visual fields, respectively, preserving objects and arrows. If natural transformation t exists from functor L to functor R, then indistinguishability of Left Apple to Right Berry can be established by both through Left Berry ($=L(f) \circ tB$) or Right Apple ($=tA \circ R(f)$), that is, there is no path dependency. Note that tA is an arrow in category E, indicating that Apple in the Left is “nearly indistinguishable” from Apple in the Right. Same goes with tB for Berry in the Left and Right.

Natural transformation t can clarify the relationship between functors L and R. Let’s go through this example slowly as this is an important point. Assume an existence of a natural transformation t from functor L to functor R. Now, with respect to objects A (Apple) and B (Berry) as well as its arrow f ($f:A \rightarrow B$; A is nearly indistinguishable from B in terms of color) in category C for Central visual field, a natural transformation is a collection of arrows in category E for Entire visual field. These arrows include $tA: L(A) \rightarrow R(A)$ and $tB: L(B) \rightarrow R(B)$ (condition 1), both of which mean “nearly indistinguishable” in terms of color, which are indeed the arrows in category E. tA means that Apple is nearly indistinguishable in terms of color in the Left and Right visual field.

Condition 2 says for an arrow $f:A \rightarrow B$ in category C, we have $tA \circ R(f) = L(f) \circ tB$. The left side of this equation first confirms that Apple is indistinguishable between Left and Right visual fields, then next it confirms that Apple in Right is indistinguishable from Berry in Right. The right side of the equation first confirms that Apple is indistinguishable from Berry in Left, then Berry in Left is indistinguishable from Berry in Right. In other words, condition 2 means that the color of Left Apple is indistinguishable from the color of Right Berry, and it allows us to go in either way to prove this fact. The power of the natural transformation t is to handle a lot of arrows, such as tA and tB , all at once in a lawful manner.

4.2 Functor category, whose objects and arrows are functors and natural transformations.

Using the concept of natural transformation, we introduce another critical concept to prove the Yoneda lemma: functor category. Functor categories and “sameness” defined in this manner, we suspect, is something that is likely to be completely missing in current language in consciousness research, and it is very useful for future conceptual analyses. Functor category, $\text{Fun}(C,D)$, considers objects and arrows as functors from category C to category D and natural transformations between them.

From the viewpoint of qualia category Q , let's consider a functor category. We extend the example category C and E in Figure 5. Simply, treating functor L and R as objects and a natural transformation t as an arrow constructs a functor category $\text{Fun}(C,E)$.

4.2.1 Useful concepts of “sameness”: natural equivalence and categorical equivalence

What is the benefit of considering a functor category? One of the benefits is to allow us to introduce two concepts of “sameness”, called “natural equivalence” and “categorical equivalence”. Categorical equivalence is likely to be a critical target to establish in empirical studies of qualia structure within a framework of category theory.

Let's say, we want to examine if peripheral vision and central vision are “essentially the same” or not ([A. Haun, Tononi, Koch, & Tsuchiya, 2017](#); [Tyler, 2015](#)). Between these visual fields, objective resolutions are different, yet when the size of objects are matched for the resolution, essential visual phenomenology, including color experience, seems “essentially the same” ([Gordon & Abramov, 1977](#); [Hansen, Pracejus, & Gegenfurtner, 2009](#); [Webster, Halen, Meyers, Winkler, & Werner, 2010](#)). How can we proceed to extract and quantify the sameness of obviously different objects?

First, you might think of “isomorphism” between category C and E for Central and Entire visual fields, respectively (Remember Section 2.2). As we introduced in Section 3.2, C and E are objects of Cat (category of categories), where arrows are functors. Isomorphism in Cat , which is also called “categorical isomorphism”, requires that for functor $F: C \rightarrow E$, there exists functor $G: E \rightarrow C$, such that $1_C = F \circ G$ and $1_E = G \circ F$. This turns out to be a very strong requirement. This is because functor G collapses E into its subset C , thus F cannot recover the original E from C .

On the other hand, what occurs if we consider the “isomorphism” instead of equality in $1_C = F \circ G$ and in $1_E = G \circ F$? “Isomorphisms” here are nothing but invertible natural transformations, which is called natural equivalence, in functor categories, $\text{Fun}(C,C)$ and $\text{Fun}(E,E)$. In this case we obtain weaker, yet very proper and flexible, kind of “sameness” between two categories C and E that we were looking for, and potentially useful for consciousness research. The sameness, obtained through natural equivalence, is called “categorical equivalence” between category C and E .

In our example of Central and Entire visual field (Figure 5), C and E will not be “categorically isomorphic” which implies one-to-one relation on object and arrows, since functor from E to C should collapse the multiple objects or arrows into the same object or arrow. On the other hand, C and E can be “categorically equivalent”, since it allows such multiplicity up to isomorphisms. If this can be experimentally verified, then we can conclude that qualia structures are “essentially the same” across visual fields. If, however, something essential is lost within a part of the visual field, say, blindspot, scotoma ([Ramachandran & Blakeslee, 1998](#)) or colorblindness in a quadrant ([Gallant, Shoup, & Mazer, 2000](#)), then, categorical equivalence would not hold any more.

5 . Explaining the Yoneda lemma and its contents.

Finally, we are ready to introduce one of the most important results of category theory, the Yoneda lemma, to consciousness research. If we can apply the Yoneda lemma to consciousness research, we can define and characterize consciousness through its relationships with the others even if we cannot describe what consciousness is per se. To be more precise, the general conclusion of the Yoneda lemma is that characterization of an object in a category is completely determined up to isomorphism by its arrows to the other objects in that category. We believe this is a substantial change of perspective, especially in the context of consciousness research: properties of an object are essentially the same as how the object relates with the others.

Having said that, some may say that a concise and general direct definition of an object is vastly superior to the exhaustive characterizations of its relationships to other objects. However, as we noted in Section 1, the difficulty in dealing with consciousness is its difficulty to describe it in itself. How can I examine my definition of “redness” compared to another person’s definition of “redness” per se? An empirical approach is for me to characterize my structural relationship between my “redness” to the other qualia and to repeat the same for the other person’s qualia (see the discussion on “the intrinsic qualities of experiences themselves and the relational structures that holds among those experiences” in ([Palmer, 1999](#))). The Yoneda lemma provides a mathematical footing to this empirical approach: we can eventually establish the sameness of my and other person’s redness in this procedure.

5.1 Locally small categories and hom functors

Before applying the Yoneda lemma to consciousness categories, let's introduce only two more concepts. The first is the concept of a hom set. A collection of arrows from object X to object Y in category C is written as $\text{hom}_C(X, Y)$ ¹. The second concept is a hom functor. A

¹ Technically, $\text{hom}_C(X, Y)$ can be really large so that it cannot be defined as a set. If $\text{hom}_C(X, Y)$ qualifies as a set every time specific objects X and Y are picked, then category C is called “locally small”. The Yoneda lemma applies to locally small categories. If all arrows in a category can be considered as a set, then the category is regarded as “small”. All categories we proposed in this paper is “small” and we believe natural extension of consciousness categories will be at most

hom functor maps an object X in category C into a set, which is an object in category Set . A hom functor also maps an arrow f in category C into a function, which is an arrow in category Set . A function maps a set to another set in category Set .

Among hom functors, one that is obtained by fixing an object A in category C is called hom functor h_A . Let us go back to our familiar example: an object A for Apple. Then, hom functor h_A can be understood as “all relationships that Apple has with other objects in category C ”. Formally, h_A is a functor from category C to category Set . This is because h_A maps objects Berry (B) and Cherry (C) in category C into $hom_C(A, B)$ and $hom_C(A, C)$ in category Set , which can be understood as “how Apple relates to Berry in all possible ways” and “how Apple relates to Cherry in all possible ways”.

h_A also maps an arrow f to $hom_C(A, f)$, which is an arrow in category Set . Note that an arrow in Set is a function from a set to another set. For unfamiliar readers, $hom_C(A, f)$ is a little confusing. To understand this, recall that f is one of the arrows from object A to B in category C , which can mean “nearly indistinguishable”, for example. Let’s single out one of the arrows in $hom_C(A, B)$ as an arrow “ a ”, which means A is “nearly indistinguishable in its color” with B . Then, a can be composed with f to obtain $a \circ f$. This compositional situation means as follows: A is related to C (in a sense of “ $a \circ f$ ”), where C is related to B (in a sense of “ f ”), which is related to A as “nearly indistinguishable in its colour”. Note that, in categories where there are more than one arrows between objects, an arrow from A to C , which means “nearly indistinguishable in its color” may be a distinct arrow from $a \circ f$. $h_A(f)$ maps a particular arrow from A to B into the corresponding arrow from A to C . If you consider arrows as elements in a set, this means that $h_A(f)$ is a function between two sets of arrows, that is, $hom_C(A, B)$ and $hom_C(A, C)$.

a)

b)

$$h_A(B) = hom_C(A, B)$$

$$h_A(C) = hom_C(A, C)$$

c)

locally small. An example of a non-small category is a category of Set because a collection of sets cannot be considered as a set (but it is still locally small.)

Figure 6. What are hom sets and hom functors? a) A hom set, $\text{hom}_C(A, B)$, is a set of arrows from Apple (A) to Berry (B) in category C. $\text{hom}_C(A, B)$ includes our familiar arrow, “nearly indistinguishable in color”, but also many other different arrows (e.g., “indistinguishably tasty”, “indistinguishably sour”, etc). b) A hom functor, h_A , maps objects B and C in category C into objects $h_A(B)$ and $h_A(C)$ in category Set. h_A also maps an arrow f in category C into an arrow in category Set.

Note that until the last paragraph, we have always considered only one arrow from one object to the other, that is, we have simply considered categories of preorder in order to simplify various concepts in category theory. For us to use category theory in consciousness research in practice, however, we will need to consider many arrows from one object to another object. $\text{hom}_C(A, B)$ is a powerful conceptual tool to think about such situations. We can consider $\text{hom}_C(A, B)$ as a list of properties of Berry in terms of relationship from Apple.

Pushing this idea further, we can denote a list of properties of all objects in category C from the perspective of Apple as $\text{hom}_C(A, -) = h_A$. We use “-” to denote the argument in which any object in category C can be substituted. The dual notion can be represented as $\text{hom}_C(-, A) = h_A$, which can be considered as a list of Apple’s relationships in category C from the perspectives of the other objects. Both h_A and h_A are likely to be useful to consider the issue of “viewpoints” in consciousness research.

5.2 What can we say from the Yoneda lemma?

Due to the technicality, we will omit the proof of the Yoneda lemma here. The readers who grasped all concepts introduced so far can understand the proof in a standard textbook of category theory ([Awodey, 2010](#); [Leinster, 2014](#)). The important theorem that follows from the Yoneda lemma is as follows:

Theorem) For objects A and B in category C, $h_A \cong h_B$ is a necessary and sufficient condition for $A \cong B$.

Here, $A \cong B$ means that A and B are isomorphic objects in category C (Section 2.2). $h_A \cong h_B$ means that there exists a natural equivalence between h_A and h_B . A natural equivalence is a natural transformation from functor h_A to h_B , which is invertible. In Section 4.2, we introduced a functor category. Given that h_A and h_B are both functors from category C to category Set, they are objects of $\text{Fun}(C, \text{Set})$. Thus, “ $h_A \cong h_B$ ” means h_A and h_B are isomorphic objects in $\text{Fun}(C, \text{Set})$. In other words, there is an invertible natural transformation (=a natural equivalence) between them.

Let’s translate the theorem into English. h_A (and h_B) means all relationships that A has with all other objects in category C. Thus, $h_A \cong h_B$ means that all relationships for A and those for B are “naturally convertible” to each other. If A and B are isomorphic objects in the category C, then it is relatively easy to prove $h_A \cong h_B$. However, a nontrivial mathematical fact is that $h_A \cong h_B$ can prove $A \cong B$. Even if it is impossible to directly compare A and B, we can make a conclusion about it by examining how A and B relate to others, in terms of

meaning in semantics, environments in ecology, astronomical objects in cosmology, numbers in mathematics and qualia in consciousness research.

Some readers may think that $hA \cong hB$ is more difficult to test empirically than $A \cong B$. Surprisingly, however, there are many cases where it is overwhelmingly easier to check $hA \cong hB$ than $A \cong B$. There is no shortage of examples in mathematics. Next, we consider it in the context of consciousness research.

5.3 First application of the Yoneda lemma in consciousness research

Let's apply the above theorem for category of qualia Q . To gain some insights, we consider the Checkershadow illusion ([Adelson, 1995](#)) (Figure 7a). In this “illusion”, square A looks very dark while square B looks quite bright. In Figure 7b, we show square A^* and B^* without any background, which makes it easy to see that brightness of A^* and B^* are physically the same. The striking effect of this “illusion” does not reduce even if you try to cognitively convince yourself that it is just a 2D image or that the direction of the shadow in the image is contradictory with the lighting condition in your room. It exemplifies the fact that subjective experience of brightness of an object strongly depends on the context of the objects.

Are the brightness of A and B the same? Can we define subjective brightness of A and B? If there are individual differences in how we perceive brightness, is there any way to systematically examine and characterize subjective brightness?

Figure 7. Applying the Yoneda lemma to category Q . a) Checkershadow Illusion. The two squares labeled A and B look “different” in brightness, as in complementary squares in a checkerboard. b) Removing the surrounding of A and B to obtain A^* and B^* , whose brightness looks the “same”.

Let's apply the theorem derived from the Yoneda lemma in Figure 7a. Consider squares A and B. Within category Q , A and B are not isomorphic. Why? Let us think about this from a perspective of category theory. Here, we consider objects as squares denoted by A, B, i, ii, iii and iv and arrows as “nearly indistinguishable” as before.

First, consider if there is an arrow from A to i. Due to the clear boundary between A and i, they look quite different, thus there is no arrow. Meanwhile, there is an arrow from B to i, partly due to the fact that bottom right part of i is indeed the same as that of B. However, if you pay attention to the top left part of the i, it becomes clear that i has a gradation and that i and B are not that similar. Yet, the cylinder on the top right provides an explanation of this difference within i as an effect of shadow. Further, the overall configuration of the checkerboard also facilitates the sameness between B and i. In the end, we conclude an existence of an arrow from B to i. At this point, there is already a difference between a collection of relationships between A and the other objects and those between B and the others, thus the theorem concludes that A and B cannot be isomorphic. This prediction from the theorem coincides with our subjective phenomenology of “difference” between A and B, which provides a preliminary support of our framework of category Q, the application of the Yoneda lemma and its potential usefulness for the future consciousness research.

What has been said above applies similarly to i, ii, iii, and iv. According to our framework, this collection of arrows created a conscious experience where A and B do not look the same. Meanwhile, the theorem predicts that A and B should look the same if we erase the sources that generate the discordances between arrows from A and arrows from B (and this is the powerful aspects of the theorem: if $hA \cong hB$, then $A \cong B$). The sources of discordance may include the green column and its shadow, the regularity of the checkerboard pattern, etc. In fact, after painting square i, ii, iii, and iv onto a white background, one of the authors (HS) now feels that A and B look the same. The other author (NT) feels that A and B became much more similar than before, but they are still distinguishable. By removing further contextual cues, as in Figure 7b, A* and B* look the same to most people. At this point, some readers may find that the arrows that we have adopted so far, signifying “nearly indistinguishable”, are too coarse to be useful in this situation. To capture different degrees of perceived similarity, we believe that “enriched category” will be useful ([Fong & Spivak, 2019](#); [Lawvere, 1973](#); [Leinster & Meckes, 2017](#)) (See Appendix. The Yoneda lemma is also known to exist in enriched category ([Bonsangue, van Breugel, & Rutten, 1998](#))). In sum, the Yoneda lemma predicts that two objects A and B should look the same if we eliminate the discordances of relationships between A and the others vs. B and the others.

So, what new things this type of category theoretical analyses can bring into consciousness research? Normally, researchers consider properties of the external world, such as the amount of light, as objective and physical reality. And, we call our perception “veridical” under the situation, where our subjective experience matches with these objective properties as in Figure 7b. Whereas, we call our perception “illusory” when they disagree as in Figure 7a, which is the reason why it is called Checkersshadow “Illusion”. However, this framework assumes a lot on the world view, which is not necessarily endorsed by many consciousness researchers ([Velmans, 2009](#)). In fact, this type of attitude can lead to a denial of existence of subjective experience, which we believe is problematic for us to make real progress in consciousness research. Rather than imposing such a loaded view, category theory offers a more neutral view, potentially facilitating interdisciplinary investigation of the contextual effects combining experimental psychologists, neurophysiologists and computational neuroscientists, philosophers, mathematicians under the same hood.

Some readers may think that the Yoneda lemma is an overkill as A and B can be directly compared in Figure 7. Of course, this is a simple example to make a point. However, various controversies surrounding consciousness research are rooted in a question of whether qualia are essentially the same between two conditions, under which the direct comparisons are impossible. For example, with or without paying attention, are qualia essentially the same ([Block, 2007](#); [Cohen, Dennett, & Kanwisher, 2016](#); [A. Haun, Tononi, et al., 2017](#))? Are the foveal vision and peripheral vision essentially the same? What is, if any, the effect of expectation on qualia? In these situations, we just cannot compare objects in two conditions directly. The indirect approach of the Yoneda lemma is especially effective in these situations. What is particularly powerful is that it makes an explicit prediction about our phenomenology. And the prediction is possible to test empirically through experimentation. Category theory, indeed, is not only useful as a conceptual framework to summarize what we know already, but rather, its power originates from its predictive power to make further predictions.

6. Conclusion and outlook

In this paper, we introduced categories of consciousness in several forms and showed that the Yoneda lemma can be applied to them to obtain novel perspectives and predictions on consciousness. While the exact and concise definition of consciousness remains difficult, its characterization through indirect characterization of its relationships to others using convergent methods (see, for example, ([Velmans, 2009](#))) is actually a valid way as guaranteed by category theory, and in particular, the Yoneda lemma.

As a future prospect, we think it is critical to consider various categories of consciousness within a larger structure of category of qualia categories. Categorical equivalence offers a way to address hitherto philosophical problems, such as if my qualia are essentially the same as yours, if humans and animals experience the world in a similar way. Categorical equivalence between two qualia categories clarifies “in what sense” these consciousnesses are essentially the same. Here, the critical innovation resides in the choice of what aspects of consciousness to extract as objects and arrows. If these aspects contain much of self-consciousness, human consciousness categories won’t be equivalent to those of animals. In terms of visual consciousness, much of equivalence is likely to be found.

The importance and novelty of our proposal here can be better appreciated by the following criticism on similarity judgment in consciousness research by Pautz ([Pautz, 2019](#)).

“To see how Similarity-Congruence falls short, let us return to hypothetical case ... some unfamiliar sentient organism, Karl. You want to determine exactly what experiences Karl has... Karl is presented with three objects consecutively ... three T-shapes such that T1 is more like T2 than T3. Then, given Similarity-Congruence, you can deduce that Karl has some trio of experiences, E1, E2, and E3, such that E1 is more like E2 than E3. But, as a simple point of logic, Similarity-Congruence is not logically strong enough to tell us precisely what those experiences are... it doesn’t tell us whether they are colour

experiences of similar shades of red, or whether they are colour experiences of similar shades of green. In fact, it doesn't tell us whether they are experiences of colour or experiences of (say) smell. That is, it doesn't entail the specific, determinate qualitative contents of those experiences."

Our proposal, specifically with the Yoneda lemma, is all about the power of "a collection of arrows (or relationships)" that is necessary to determine the color experience of red or green, the smell of fish, the sound of a bell, etc. And the essential role of a collection of relations does make sense in light of the fact that there has been no clinical report of a patient who lost color of red, but not others. When brain damage causes color blindness, the reported cases are all about the entire loss of color experience (and associated concepts) in specific visual location ([Gallant et al., 2000](#)), but not a particular color or range of colors. Similar things can be said about visual motion ([Zihl, Von Cramon, & Mai, 1983](#)). Loss of a single category of perception, such as faces (e.g., prosopagnosia), objects, etc, have been linked to some specific brain lesions ([Kanwisher & Yovel, 2006](#); [Milner & Goodale, 1995](#)), but rarely on a particular face (Except for the loss of "familiar" faces and objects, called "Capgras Syndrome" ([Ramachandran & Blakeslee, 1998](#))). On the other hand, loss of a single linguistic concept (=forgetting) is common. Why can we lose a specific single concept in linguistic case but not for the qualia of color or motion, which is bound to the retinal locations? While anatomical localizations and functional mapping in neuroscience can give us a hint, they do not address this theoretical question.

We believe that one possible answer to this riddle is to do with the Yoneda lemma. We hypothesize that the essence of the color qualia resides in relations with other colors, that is, without the entire sets of relations, color qualia just cannot emerge. On the other hand, we hypothesize that the essence of linguistic conceptual qualia (if any ([Kemmerer, 2015](#); [McClelland & Bayne, 2016](#))) is its relational structure that allows one concept to be missing from the web of relations. In retrospect, therefore, the essential role of a collection of relations in consciousness may be obvious ([Palmer, 1999](#)). Yet, we are not aware if a collection of massive similarity ratings has been ever collected or considered in consciousness research. Replacing characterization of a quale with a collection of comparative descriptions with other qualia can have a mathematically solid foundation: the Yoneda lemma in category theory.

Consider the issue of animal consciousness. Comparative cognitive scientists have developed various behavioral paradigms, combining clever tricks in training, which allows us to discuss the structure of animal consciousness. In reality, however, many animal researchers are strongly influenced by behaviorism and are reluctant to investigate or even talk about animal consciousness. Some researchers even give up any hope of understanding animal consciousness altogether ([Nagel, 1974](#)). Combining the perspective of the Yoneda lemma with quantitative theories of consciousness, such as Integrated Information Theory ([Oizumi et al., 2014](#); [Tononi et al., 2016](#)), we should be able to establish a new research program into animal consciousness ([Tsuchiya, 2017](#)). Pioneering works in monkeys have introduced highly creative behavioral experimental tasks, such as binocular rivalry ([Leopold, Maier, & Logothetis, 2003](#)) and no-report paradigms ([Wilke, Mueller, & Leopold, 2009](#)), well beyond simple button press reports, which poorly characterize

consciousness even in humans (also see ([Matsuzawa, 1985](#))). These animal researchers have already brought various “arrows” in category Q for monkeys. This type of research can be extended to other animals, such as dolphins, birds, octopuses, mice and even flies. We have gained so much knowledge about arrows that are powerful to gain insight about animal consciousness over the last 30 years ([Boly et al., 2013](#)).

In this paper, we intentionally limited most of our examples into cases where up to one arrow can exist from one object to the other (preorder category). However, it is easy to extend our category Q to have multiple arrows between objects, such as “nearly indistinguishable in terms of X”, where X can be color, shape, size, location, etc. The richer the arrows, the better the structure of the category can be understood. Bringing richer sets of arrows will sharpen our understanding of “in what sense” category Q for human is similar to or not with those for animals.

In addition to what we introduced in this paper, there are many more tools in category theory which are likely to be illuminating in consciousness research. Some of these conceptual tools will clarify complex theoretical concepts about consciousness, which have been discussed by philosophers and psychologists for a long. Such conceptual clarification will inspire further theoretical and empirical research ideas to come in the near future.

Appendix: Enriched category of qualia

In the main text, we suggested an arrow for category Q as “nearly indistinguishable”. However, this concept is known to fall into Sortie’s paradox: imagine a graded redness of a sky in the sunset. Any given two close points can have an arrow in this situation, but two points that are distant are easily distinguishable. In this case, “composition” does not hold ([Pitt, 2018](#)). In our example, we avoided such difficulty by specifically selecting discrete entities as objects, where composition of arrows is designed to be true.

To properly deal with the continuous degrees of “similarity”, we propose a formulation through enriched category. Within enriched category theory, extended Yoneda lemma is known to exist (e.g., ([Bonsangue et al., 1998](#))), thus our arguments in this paper should apply to enriched category theory.

In enriched category, we replace “arrows” with something more flexible ([Fong & Spivak, 2019](#); [Lawvere, 1973](#); [Leinster & Meckes, 2017](#)). Mathematically speaking, we can obtain an enriched category from a locally small category by replacing a set of arrows into objects in a “monoidal category”. Thus, enriched category is a generalization of category. Monoidal category can be considered as a category where a coherent notion of “product structure” is defined among objects and among arrows. Mathematical examples of the monoidal category include the one whose objects are positive real numbers and arrows are $\leq$. In this case, product structure corresponds to “sum”. Enriching a category with such a monoidal category, we obtain “generalized metric space”, where there is “distance”-like concept is defined between objects.

Acknowledgement.

The authors thank Ariel Zeleznikow-Johnston and Steven Phillips for their comments on the earlier version of this manuscript.

Funding statement.

This work was supported by Australian Research Council (NT, grant number DP180104128, DP180100396); National Health Medical Research Council (NT, grant number APP1183280); and Japan Science and Technology Agency (HS, grant number JPMJCR17N2).

Conflicts of Interest.

None.

References

- Adelson, E. H. (1995). Checkershadow Illusion. Retrieved from <http://persci.mit.edu/gallery/checkershadow>
- Awodey, S. (2010). *Category Theory*: OUP.
- Balduzzi, D., & Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS Comput Biol*, 5(8), e1000462. doi:10.1371/journal.pcbi.1000462
- Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 2018(1), niy007.
- Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are There Levels of Consciousness? *Trends Cogn Sci*. doi:10.1016/j.tics.2016.03.009
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *BBS*, 30(5-6), 481-499.
- Boly, M., Seth, A., Wilke, M., Igmundson, P., Baars, B., Laureys, S., . . . Tsuchiya, N. (2013). Consciousness in humans and non-human animals: recent outstanding advances, and possible future directions. *Frontiers Psy*, 4(625), 1-20.
- Bonsangue, M. M., van Breugel, F., & Rutten, J. J. (1998). Generalized metric spaces: completion, topology, and powerdomains via the Yoneda embedding. *Theoretical Comp Sci*, 193(1-2), 1-51.
- Bradley, T.-D. (2017). The Yoneda Perspective. Retrieved from <https://www.math3ma.com/blog/the-yoneda-perspective>
- Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends Cogn Sci*, 23(9), 754-768. doi:10.1016/j.tics.2019.06.009
- Coecke, B., & Kissinger, A. (2017). *Picturing quantum processes*: Cambridge University Press.
- Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *TICS*, 20(5), 324-335. doi:10.1016/j.tics.2016.03.006

- Dennett, D. C. (1988). Quining qualia. In *Consciousness in modern science*: OUP.
- Engel, & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *TICS*, 5(1), 16-25.
- Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Conscious Cogn*, 20(3), 807-827. doi:10.1016/j.concog.2011.02.010
- Fong, B., & Spivak, D. I. (2019). *An Invitation to Applied Category Theory*: CUP.
- Frege, G. (1980). *The foundations of arithmetic: A logico-mathematical enquiry into the concept of number*. Northwestern University Press.
- Fuyama, M., & Saigo, H. (2018). Meanings, Metaphors, and Morphisms: Theory of Indeterminate Natural Transformation (TINT). *arXiv preprint arXiv:1801.10542*.
- Gallant, J. L., Shoup, R. E., & Mazer, J. A. (2000). A human extrastriate area functionally homologous to macaque V4. *Neuron*, 27(2), 227-235.
- Gordon, J., & Abramov, I. (1977). Color vision in the peripheral retina. II. Hue and saturation. *JOSA*, 67(2), 202-207.
- Graziano, M. S., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cogn Neurosci*, 2(2), 98-113. doi:10.1080/17588928.2011.565121
- Hansen, T., Pracejus, L., & Gegenfurtner, K. R. (2009). Color perception in the intermediate periphery of the visual field. *JOV*, 9(4), 26-26.
- Haun, A., Oizumi, M., Kawasaki, H., Kovach, C., Oya, H., Howard, M. A., . . . Tsuchiya, N. (2017). Conscious perception as integrated information patterns in human electrocorticography. *eNeuro*, doi: 10.1101/039032.
- Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, 21(12), 1160.
- Haun, A., Tononi, G., Koch, C., & Tsuchiya, N. (2017). Are we underestimating the richness of visual experience? *Neuro of Consc*, 1-4. doi:10.1093/nc/niw023
- Hoffman, W. C. (1966). The Lie algebra of visual perception. *Journal of mathematical Psychology*, 3(1), 65-98. doi:[https://doi.org/10.1016/0022-2496\(66\)90005-8](https://doi.org/10.1016/0022-2496(66)90005-8)
- Hoffman, W. C. (1980). Subjective geometry and geometric psychology. *Mathematical Modelling*, 1(4), 349-367. doi:[https://doi.org/10.1016/0270-0255\(80\)90045-7](https://doi.org/10.1016/0270-0255(80)90045-7)
- Hohwy, J. (2009). The neural correlates of consciousness: New experimental approaches needed? *Consciousness and Cognition*, 18(2), 428-438.
- Kanai, R., & Tsuchiya, N. (2012). Qualia. *Curr Biol*, 22(10), R392-396. doi:10.1016/j.cub.2012.03.033

- Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philos Trans R Soc Lond B Biol Sci*, 361(1476), 2109-2128.
- Kemmerer, D. (2015). Are we ever aware of concepts? A critical question for the Global Neuronal Workspace, Integrated Information, and Attended Intermediate-Level Representation theories of consciousness. *Neuroscience of Consciousness*, 2015(1).
- Kleiner, J. (2019). Models of consciousness. *arXiv preprint arXiv:1907.03223*.
- Kleiner, J., & Tull, S. (2020). The Mathematical Structure of Integrated Information Theory. *arXiv preprint arXiv:2002.07655*.
- Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat Rev Neuro*, 17(5), 307-321. doi:10.1038/nrn.2016.22
- Lamme, V. A. F. (2015). *The Crack of Dawn. Perceptual Functions and Neural Mechanisms that Mark the Transition from Unconscious Processing to Conscious Vision*: Barbara Wengeler Stiftung.
- Laureys, S. (2005). The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn Sci*, 9(12), 556-559.
- Lawvere, F. W. (1973). Metric spaces, generalized logic, and closed categories. *Reports of Math and Physics*, 43(1), 135-166.
- Lawvere, F. W., & Schanuel, S. H. (2009). *Conceptual mathematics*: CUP.
- Leinster, T. (2014). *Basic category theory* (Vol. 143): Cambridge University Press.
- Leinster, T., & Meckes, M. W. (2017). The magnitude of a metric space: from category theory to geometric measure theory. In *Measure Theory in Non-Smooth Spaces* (pp. 156-193).
- Leopold, D. A., Maier, A., & Logothetis, N. (2003). Measuring subjective visual perception in the nonhuman primate. *Journal of Consciousness Studies*, 10(9-10), 115-130.
- Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, 105(5), 776-798. doi:10.1016/j.neuron.2020.01.026
- Matsuzawa, T. (1985). Colour naming and classification in a chimpanzee (*Pan troglodytes*). *Journal of Human Evolution*, 14(3), 283-291.
- McClelland, T., & Bayne, T. (2016). Concepts, contents, and consciousness. *Neuroscience of Consciousness*, 2016(1).
- Milner, D. A., & Goodale, M. A. (1995). *The visual brain in action*. Oxford: Oxford University Press.
- Mormann, F., & Koch, C. (2007). Neural correlates of consciousness. *Scholarpedia*, 2(12), 1740.

- Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Reveiw*, 83(4), 435-450.
- Northoff, G., & Huang, Z. (2017). How do the brain's time and space mediate consciousness and its different dimensions? Temporo-spatial theory of consciousness (TTC). *Neurosci Biobehav Rev*, 80, 630-645. doi:10.1016/j.neubiorev.2017.07.013
- Northoff, G., Tsuchiya, N., & Saigo, H. (2019). Mathematics and the Brain: A Category Theoretical Approach to Go Beyond the Neural Correlates of Consciousness. *Entropy*, 21(12), 1234.
- Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comp*, 10(5), e1003588. doi:10.1371/journal.pcbi.1003588
- Palmer, S. E. (1999). Color, consciousness, and the isomorphism constraint. *BBS*, 22(6), 923-943.
- Pautz, A. (2019). What is the Integrated Information Theory of Consciousness?: A Catalogue of Questions. *Journal of Consciousness Studies*, 26(1-2), 188-215.
- Pitt, D. (2018). Phenomenal compositionality and context effects. *Inquiry*, 61(5-6), 494-498.
- Prentner, R. (2019). Consciousness and topologically structured phenomenal spaces. *Conscious Cogn*, 70, 25-38. doi:10.1016/j.concog.2019.02.002
- Ramachandran, V. S., & Blakeslee, S. (1998). *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William Morrow.
- Saigo, H., Naruse, M., Okamura, K., Hori, H., & Ojima, I. (2019). Analysis of soft robotics based on the concept of category of mobility. *Complexity*, 2019.
- Searle, J. R. (2000). Consciousness. *Annual review of neuroscience*, 23, 557-578. doi:10.1146/annurev.neuro.23.1.557
- Sloman, A. (1991). Developing concepts of consciousness. *Behavioral and Brain Sciences*, 14(4), 694-695.
- Spivak, D. I. (2014). *Category theory for the sciences*: MIT Press.
- Stanley, R. P. (1999). Qualia space. *Journal of Consciousness Studies*, 6(1), 49-60.
- Stanovich, K. E. (1991). Damn! There goes that ghost again! *Behavioral and Brain Sciences*, 14(4), 696-698.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci*, 5, 42.
- Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neuro*, 17(7), 450-461. doi:10.1038/nrn.2016.44

- Tsuchiya, N. (2017). What is it like to be a bat?" - a pathway to the answer from the Integrated Information Theory. *Philosophy Compass*, e12407. doi:DOI 10.1111/phc3.12407
- Tsuchiya, N., Taguchi, S., & Saigo, H. (2016). Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience research*. doi:<http://dx.doi.org/10.1016/j.neures.2015.12.007>
- Tull, S., & Kleiner, J. (2020). Integrated Information in Process Theories. *Arxiv*. doi: arXiv:2002.07654v1
- Tye, M. (2018). Qualia. *The Stanford Encyclopedia of Philosophy*.
- Tyler, C. W. (2015). Peripheral color demo. *i-Perception*, 6(6), 2041669515613671.
- Van Gulick, R. (2018). Consciousness. *The Stanford Encyclopedia of Philosophy*.
- Velmans, M. (2009). How to define consciousness: And how not to define consciousness. *J of Consciousness Studies*, 16(5), 139-156.
- Webster, M. A., Halen, K., Meyers, A. J., Winkler, P., & Werner, J. S. (2010). Colour appearance and compensation in the near periphery. *Proceedings of the Royal Society B: Biological Sciences*, 277(1689), 1817-1825.
- Wilke, M., Mueller, K. M., & Leopold, D. A. (2009). Neural activity in the visual thalamus reflects perceptual suppression. *Proc Natl Acad Sci U S A*, 106(23), 9465-9470. doi:0900714106 [pii] 10.1073/pnas.0900714106
- Yoshimi, J. (2011). Phenomenology and connectionism. *Front Psychol*, 2, 288. doi:10.3389/fpsyg.2011.00288
- Zihl, J., Von Cramon, D., & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain*, 106(2), 313-340.