G Model
NSR-3908; No. of Pages 7
ARTICLE IN PRESS

Neuroscience Research xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Neuroscience Research

journal homepage: www.elsevier.com/locate/neures

Perspective

Using category theory to assess the relationship between consciousness and integrated information theory

Naotsugu Tsuchiya[a,b,*,1], Shigeru Taguchi[c,2], Hayato Saigo[d,3]

[a] School of Psychological Sciences, Faculty of Biomedical and Psychological Sciences, Monash University, Australia
[b] Monash Institute of Cognitive and Clinical Neuroscience, Monash University, Australia
[c] Graduate School of Letters, Hokkaido University, Japan
[d] Department of Biosciences, Nagahama Institute of Bio-Science and Technology, Japan

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form
27 November 2015
Accepted 15 December 2015
Available online xxx

Keywords:
Consciousness
Qualia
Category theory
Integrated information theory
Phenomenology
Equivalence

ABSTRACT

One of the most mysterious phenomena in science is the nature of conscious experience. Due to its subjective nature, a reductionist approach is having a hard time in addressing some fundamental questions about consciousness. These questions are squarely and quantitatively tackled by a recently developed theoretical framework, called integrated information theory (IIT) of consciousness. In particular, IIT proposes that a *maximally irreducible conceptual structure* (MICS) is *identical* to conscious experience. However, there has been no principled way to assess the claimed identity. Here, we propose to apply a mathematical formalism, *category theory*, to assess the proposed identity and suggest that it is important to consider if there exists a proper *translation* between the domain of conscious experience and that of the MICS. If such translation exists, we postulate that questions in one domain can be answered in the other domain; very difficult questions in the domain of consciousness can be resolved in the domain of mathematics. We claim that it is possible to empirically test if such a functor exists, by using a combination of neuroscientific and computational approaches. Our general, principled and empirical framework allows us to assess the relationship between the domain of consciousness and the domain of mathematical structures, including those suggested by IIT.

1. Introduction

The material basis of subjective conscious phenomena remains one of the most difficult scientific questions (Chalmers, 1996). While it is impossible to doubt if the reader is consciously awake (as opposed to unconscious as in deep dreamless sleep) and visually conscious of this text at this moment (as opposed to blind and seeing nothing), it appears very difficult to be completely certain about conscious states of other persons, seems more difficult to infer consciousness in babies or animals, and looks even impossible to tell if artificial machines can ever achieve human-like consciousness.

Over the last 25 years, concerted neuroscientific approaches have established that consciousness arises from the interactions among some neurons in the thalamo-cortical systems (Boly et al., 2013; Dehaene and Changeux, 2011; Koch, 2004). Now the field has matured enough to result in a specific theory, called integrated information theory (IIT) of consciousness (Oizumi et al., 2014; Tononi, 2004, 2008, 2012, 2015). IIT has strong explanatory power in many observed neuroscientific enigmatic facts about consciousness and proposes a precise mathematical formalism that should be *identical* to consciousness.

However, it has been unclear whether there exists any principled and empirical ways to assess the proposed *identity*. And, it seems unclear what it means for some mathematical formalism and consciousness to be *identical.*

To address these issues, here, we propose that a fundamental mathematical formalism, called *category theory* can be a very powerful tool. In category theory, a category is defined as a collection of objects and arrows.[4] In its standard usage, a category refers

* Corresponding author at: Monash University, Room 145, Building 220, 770 Blackburn Rd, Clayton, 3168 VIC, Australia. Tel.: +61 399054564.
E-mail addresses: naotsugu.tsuchiya@monash.edu (N. Tsuchiya), tag@let.hokudai.ac.jp (S. Taguchi), h_saigoh@nagahama-i-bio.ac.jp (H. Saigo).
[1] Corresponding author is a member of the Japan Neuroscience Society (JNS).
[2] Address: Graduate School of Letters, Hokkaido University, Kita 10, Nishi 7, Kita-ku, Sapporo 060-0810, Japan.
[3] Address: Nagahama Institute of Bio-Science and Technology, Nagahama, Japan.

[4] Mathematically speaking, a category C consists of (1) a collection of objects, such as X and (2) a collection of arrows, which define relationship between any pair of objects, such as X and Y, such that (3) for every object X there is a self-referential arrow $1x$: $X \rightarrow X$, (4) any pair of arrows, such as f: $X \rightarrow Y$ and g: $Y \rightarrow Z$, are

http://dx.doi.org/10.1016/j.neures.2015.12.007

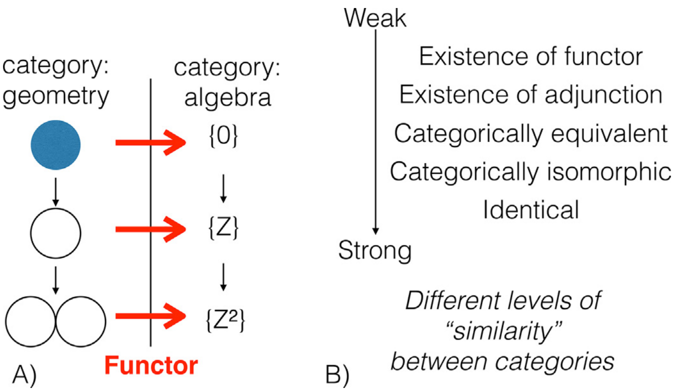

Fig. 1. Category theory. (A) Three objects in geometry and those in algebra can be mapped between them by preserving their structural relationships. This mapping is called a functor in category theory. (B) In category theory, *quality of similarity* can be precisely defined with graded levels. Existence of a functor is relatively weak, requiring relatively loose conditions. Yet, existence of a functor is already quite powerful to the extent that it can guarantee the translation of proof of the Brower's theorem between geometry and algebra. We conjecture that the domain of mathematical structures and the domain of consciousness or qualia can be shown to be *similar* up to "categorically equivalent".

to a class of things that share certain properties, as in categories of objects, humans and animals. The standard definition of category is consistent with, but less fundamental and abstract than, the definition of *category* in category theory.

This formalization of categories allows us to compare and relate two seemingly complete separate domains of knowledge, such as mathematical concepts and conscious experience. In the following, we first introduce category theory and then briefly review the two to-be linked worlds of consciousness and mathematical structures according to IIT. Next we outline what strategies need to be taken for this research program. Finally, we offer the future prospects, promising the dissolution of Hard problem.

2. Category theory

Category theory was introduced in 1945 by mathematicians Eilenberg and Mac Lane (Awodey, 2010; Mac Lane, 1998). Category theory is now considered as the foundation of mathematics, a position previously held by set theory. In category theory, everything is considered as either an object or an arrow that connects objects. Objects and arrows can include almost any concept. In Fig. 1A, we consider three objects in geometry (a disk, a ring and a double-ring) and three objects in algebra (a set that is composed of only zero {0}, a set of integer {Z}, and a set of 2 integers {Z^2}). Interestingly, when we relate these objects in each domain[5] of mathematics, their relationships can be proven to be *mathematically analogous*. More precisely, category theory says that there exists a *functor* between them. A functor is a structure-preserving map between categories.[6]

We focus on two important properties of category theory. First, category theory provides a mathematical framework for translating a relationship in one domain to a distinct and separate domain by use of a structure-preserving map, or a functor. Second, category theory brings a precise mathematical formalism to assess whether or not two separate domains of knowledge are *similar* and in what qualitative way they are similar.

As to the first point, we briefly describe a powerful example of a mathematical proof. In geometry, there is a fundamental theorem due to Brouwer (Fulton, 1995). This theorem, known as Brouwer's fixed-point theorem, is notoriously difficult to prove within the domain of geometry. Briefly, this theorem states that any continuous function that maps any point on a disk to another point on the same disk leaves at least a single point that does not change its position. For example, any rotation about the center of a disk would leave the position of the center of the rotation unchanged. While difficult to prove within the domain of geometry, by translating geometric objects over to the algebraic domain, it can be seen that the proof of this theory amounts to a proof that there is no isomorphic mapping from {0} into a set of integers, which is rather easy to prove. Today, many mathematicians go from one domain to another in order to prove theorems. In fact, a similar method has been used to prove Fermat's last theorem. Outside of the field of mathematics, category theory has bridged across different disciplines. Recent work, for example, have shown the analogy in the precise sense among quantum mechanics, topology, logic and computation (Baez and Stay, 2009), which can be used as a basis to translate the proofs in one domain to the others. Application of category theory into neuroscience and cognitive science has recently emerged (Ehresmann and Gomez-Ramirez, 2015; Phillips and Wilson, 2010).

As to the second point, category theory offers extremely useful tools to characterize similarity between the different domains. In category theory, the nature of similarity is precisely defined as different degrees of the structure-preservation through mathematical terms; requiring more and more conditions amounts to *strong* similarity. In this framework, the qualitative strength of *similarity* degrades from "identical", "categorically isomorphic", "categorically equivalent", "existence of adjunction" to "existence of functor". With these graded scales of similarity, we can precisely understand in what sense IIT's proposed mathematical structures and conscious experience are similar. Though it is weakest among the above list, "existence of functor" is sufficient to prove Brouwer's theorem mentioned above. We believe that finding a functor between IIT's mathematical structure and consciousness might be also sufficient to bring about many theoretical and empirical results, without requiring "identity" as claimed by the original theory. Before the invention of category theory, there was no systematic framework to characterize this kind of qualitatively graded levels of *similarity* (Mac Lane, 1998). From a category-theoretic point of view, strength of *similarity*, *analogy*, *metaphor*, and *relationship* that are used in many different scientific disciplines can be qualitatively characterized in a very precise manner. It might sound totally counter to some readers to see a claim such as a very "precise" and "qualitative" characterization, but this is indeed the core and general feature of category theory.

It is these two properties of category theory that are likely to be useful in considering the so-called "mind-body" problem which regards the nature of mapping between consciousness and brain. Decades of neuroscientific research have culminated to a suggestion that it is not the brain per se, but rather some type of mathematical structure that maps to the domain of consciousness (Oizumi et al., 2014; Tononi, 2004). While we focus on a particular mathematical structure, called a maximally irreducible conceptual structure (MICS) in the integrated information theory (IIT), our argument generalizes to any mathematical structures that can be derived from the brain.

In the next section, we explain what we mean by consciousness and mathematical structure, which are to be linked by a functor.

composable, that is, gf: $X \to Z$, (5) a self-referential arrow is both a left and right unit for composition, that is, if f: $X \to Y$, then $f1x = f = 1yf$, and (6) composition is associative, that is, $(hg)f = h(gf)$.

[5] Throughout this paper, we use "domain" to mean a slightly different concept in "domain" in mathematics. Our usage is more colloquial, referring to a set of highly related objects, concepts and phenomena.

[6] Mathematically speaking, it means that (1) any object m in M has a mapped object q in Q, that is, $F(m) = q$, (2) any arrow f in M has a mapped arrow g in Q, that is, $F(f) = g$, (3) F preserves identities, that is, for any object X in C, $F(1_x) = 1_{(F(X))}$ and (4) F preserves composition, that is, for any pair of arrows f: $X \to Y$ and g: $Y \to X$ in C, $F(gf) = F(g)F(f)$.

Please cite this article in press as: Tsuchiya, N., et al., Using category theory to assess the relationship between consciousness and integrated information theory. Neurosci. Res. (2016), http://dx.doi.org/10.1016/j.neures.2015.12.007