NEWSLETTER | The American Philosophical Association

Philosophy and Computers

FALL 2017

VOLUME 17 | NUMBER 1

ISSN 2155-9708

APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR VOLUME 17 | NUMBER 1 | FALL 2017

FROM THE EDITOR

Peter Boltuc
UNIVERSITY OF ILLINOIS, SPRINGFIELD

The old philosophical question, how to set the boundaries of semantics *versus* syntax, is becoming an essential issue again based on the current challenges in computer engineering, such as various attempts at constructing "semantic web" and other semantic systems. Those are the systems able to understand and disambiguate natural language, with its metaphorical meanings of the words and context dependencies. In a broader sense, such semantic engines would be able to disambiguate the meanings in human and physical world encountered by self-driving machines, and other autonomous systems. Such ability "to know what's going on," as Riccardo Sanz would put it, is essential for smooth integration of (semi-)autonomous cognitive engines within the universe of human agents. We are delighted to publish Bill Rapaport's article "Semantics as Syntax," which is an important voice in this debate. The author views semantics as an interpretation over syntax, fully reducible to the latter. It is an important re-statement of Rapaport's 1988 position that "purely syntactic symbol-manipulation of a computational natural-language-understanding system's knowledge base suffices for it to understand natural language." Rapaport's new defense of this traditional view is an important contribution to the current debate.

We are glad to publish the third and last article by the duo of Jun Tani and Jeffrey White, who have recently moved from KAIST in Korea to the Okinawa Institute of Science and Technology (OIST). In this article, important for philosophers and scientists alike, the authors develop the idea of "synthetic neurorobotics studies" as essential for grasping various senses of machine consciousness.

Kristen Zbikowski's interview with Stephen Thaler brings to the readers several new ideas on machine consciousness, many of them already applied and functioning within the framework of imagination engines. Those are advanced cognitive architectures that implement dream-like information processing—based largely on image transformation and combination—in order to attain *creative* cognitive engines that seem useful in the areas so divergent as new product development, investment strategies, and the arts. We close the issue with a philosophical cartoon by Riccardo Manzotti, who is currently a Fulbright Scholar at MIT—those cartoons have become a yearly feature of this newsletter. The current cartoon develops a case for ontological relationism. We want to congratulate the author on his new position as professor of theoretical philosophy at the *Libera Università di Lingue e Comunicazione* IULM University at Milan.

Last but not least, I want to emphasize the importance of the note from the chair of this committee, Marcello Guarini. The note includes up-to-date information about the sessions organized by the APA Committee on Philosophy and Computers for all the three APA divisional meetings in 2018—Central, Eastern, and Pacific. It also includes the list of all current committee members. It is exciting to see the committee being so active again!

FROM THE CHAIR

Marcello Guarini
UNIVERSITY OF WINDSOR

The members of the APA Committee on Philosophy and Computers have been hard at work putting together sessions for upcoming APA meetings.

During the 2018 Eastern APA meeting, the winner of the 2016 Barwise prize, Dr. Edward Zalta (Stanford University), will be acknowledged. Dr. Zalta will present an address entitled "How Computational Results Can Improve Metaphysics: Case Study." The commentators on the paper will be Christopher Menzel (Texas A&M University) and Branden Fitelson (Northeastern University). Commentary will be followed by replies and discussion with the audience. Committee member Gary Mar will chair the session, which will take place on Wednesday, January 3, from 1:00 to 3:00 p.m. Please check the meeting program for any last-minute changes.

At the 2018 Central APA meeting, our own newsletter editor, Peter Boltuc, will be chairing a session on Machine Consciousness. The session will include the following papers by philosophers: "The Myth of Mind Uploading," by our committee member Gualtieri Piccinini, and "Three Senses of Effectively Computable," by Jack Copeland and Peter Boltuc. It will also include papers by leading AI experts: "Device for the Autonomous Bootstrapping of Unified Sentience," by Stephen Thaler, and a paper on robot psychology by Troy Kelley.

At the 2018 Pacific APA meeting, philosophy and computers committee member Fritz MacDonald will chair a session on New Technologies in Online Teaching of Philosophy.

At the time of publication, the times and dates have not been finalized. Please check the APA meeting programs as final particulars become available online. In the interest of building our community of scholarship, we encourage readers of the newsletter to attend the above events.

Readers of the newsletter also are encouraged to contact any member of the committee if they are interested in proposing or collaborating on an APA symposium that engages any of the wide range of issues associated with philosophy and computing. We are happy to continue facilitating the presentation of high-quality research in this area.

The current members of the committee are listed below:

Marcello Guarini (chair, 2019) (mguarini@uwindsor.ca)
Fritz J. McDonald (2018)
Gualtiero Piccinini (2018)
Gary Mar (2019)
Robin Smith (2020)
Susan G. Sterrett (2020)
Dylan E. Wittkower (2019)
Peter Boltuc (newsletter editor) (epetebolt@gmail.com)

The committee thanks Colin Allen and William Barry, whose terms have come to an end—their commitment to the committee and the community of scholars it serves is very much appreciated.

CALL FOR PAPERS

It is our pleasure to invite all potential authors to submit to the *APA Newsletter on Philosophy and Computers*. Committee members have priority since this is the newsletter of the committee, but anyone is encouraged to submit. We publish papers that tie in philosophy and computer science or some aspect of "computers"; hence, we do not publish articles in other sub-disciplines of philosophy. All papers will be reviewed, but only a small group can be published.

The area of philosophy and computers lies among a number of professional disciplines (such as philosophy, cognitive science, computer science). We try not to impose writing guidelines of one discipline, but consistency of references is required for publication and should follow the *Chicago Manual of Style*. Inquiries should be addressed to the editor, Dr. Peter Boltuc, at pboltu@sgh.waw.pl

ARTICLES

Semantics as Syntax

William J. Rapaport
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, DEPARTMENT OF PHILOSOPHY, DEPARTMENT OF LINGUISTICS, AND CENTER FOR COGNITIVE SCIENCE, UNIVERSITY AT BUFFALO, THE STATE UNIVERSITY OF NEW YORK

> The Chinese room shows what we knew all along: syntax by itself is not sufficient for semantics. (Does anyone actually deny this point, I mean straight out? Is anyone actually willing to say, straight out, that they think that syntax, in the sense of formal symbols, is really the same as semantic content, in the sense of meanings, thought contents, understanding, etc.?)[1]
>
> My thesis is that (suitable) purely syntactic symbol-manipulation of a computational natural-language-understanding system's knowledge base suffices for it to understand natural language.[2]
>
> Does that make any sense? Yes: Everything makes sense. The question is: What sense does it make?
>
> – Stuart C. Shapiro (in conversation, April 19, 1994)

1 SYNTAX VS. SEMANTICS

Does syntax suffice for semantics? John Searle famously says that it does not.[3] I have argued that it does.[4]

More precisely, I have argued that semantics is nothing but syntax. These slogans need to be cashed out.

1.1 SYNTAX

Let's begin with syntax. The word "syntax" has at least two meanings: a narrow or specific one, and a wide or general one. On the narrow (and perhaps more usual) meaning, the syntax of a *language* is its *grammar*, and the syntax of a *logic* is its *proof theory*.

The wide meaning, which is the one I want to focus on, includes both narrow meanings, but goes beyond them. It is roughly synonymous with Charles Morris's "syntactics": "the formal relation of signs to one another . . . in abstraction from the relations of signs to objects or to interpreters."[5] (The former relations are those of semantics, the latter are those of pragmatics.)

On the wide view, syntax is the study of the *properties* of the "symbols" of an (uninterpreted) "symbol system" and the *relations* among them, including how the symbols can be "manipulated." But "symbol" is a charged term, used by some to mean an *interpreted* "sign": a sign *together with* its meaning. Worse, "sign" is yet another charged term, because signs are supposed to be signs *of* something. I want to focus on the "sign" or "symbol" itself, devoid of any meaning, so I will use the more neutral terms "mark," "mark system," and "mark manipulation" instead of the more familiar "sign," "symbol," "symbol system," and "symbol manipulation."

This is a kind of very "pure" syntax. It is "formal" syntax in the sense of Carnap, where an item

> is to be called *formal* when no reference is made in it either to the meaning of the symbols . . . or to the sense of the expressions . . . but simply and solely to the kinds and order of the symbols from which the expressions are constructed. . . . *Pure syntax* is concerned with the possible arrangements, without reference either to the nature of the things which constitute the various elements, or to the question as to which of the possible arrangements of these elements are anywhere realized.[6]

Dale Jacquette does not believe in the existence of such "pure syntax . . . entirely divorced from semantics." But all that he says in defense is that such marks "lack even derivative meaning or intentionality."[7] ("Intentionality," by the way, seems to have two different, albeit related, meanings in the literature. In a technical sense deriving from Brentano, it means "directedness to an object"; in the sense in which Jacquette, Searle, and others use it in the context of the Chinese Room Argument, it seems to be roughly synonymous with "cognition," "understanding," or even "consciousness.") Jacquette goes on to say that even purely syntactic "computer programs . . . are always externally interpreted."[8] I agree with the latter comment, but I still think that there is such a thing as pure syntax in the sense that I am using it here. But this debate would take us too far astray.[9]

The (purely) syntactic properties of marks include their *shape* (what a mark looks like and how it differs from other marks in the system), an *inventory* of the marks of the system (an "alphabet" or "vocabulary" of "primitive," or "basic," marks), and relations spelling out how marks may be *combined* to form more complex ones from simpler ones (usually given by recursive rules that take primitive, or given, or "atomic" marks as the base case, and show how to combine them to produce "molecular" marks, or "well- formed formulas" [wffs]). This much would normally be called the "grammar" of the system if the system were to be thought of as a language.

As I noted above, some mark systems, especially those that really are languages (formal or otherwise) might have other syntactic properties and relations, in addition to shape, inventory, and grammatical combinatory relations. For instance, some molecular marks—well-formed formulas—might be taken as axioms. And some sets of molecular marks might stand in certain relations such that, whenever some of them have been collected together, others—called "theorems"—might then be "legally" allowed to be added to the collection. Here what I have in mind are transformation rules or rules of inference. Thus, in addition to "grammatical" syntax, a set might also have a "logical" or "proof-theoretic" syntax. The production of molecular wffs and theorems is usually what is meant by "symbol manipulation" (i.e., mark manipulation). Note, however, that I am not requiring the transformation rules to be "truth preserving," because I take "truth" to be a *semantic* property. (I take *correspondence* theories of truth to be semantic. *Coherence* theories seem to be more syntactic, or holistic. We'll come back to holism in §3.1.)

But I want to be even more general than Morris: I see no reason not to include systems that might not normally be considered to be languages. On my view, *any* set of objects has a syntax if the objects have properties and stand in relations to each other. On this view, even biological neural networks have syntax,[10] as does the world itself. This seems to be consistent with Carnap: "The *syntax* of . . . any . . . calculus, is concerned . . . with the *structures of possible serial orders* … ***of any elements whatsoever.***"[11]

1.2 SEMANTICS

Semantics, of course, is the study of meaning and truth. The meaning of some piece of language (a word, phrase, sentence, what have you) might be its *referent* (if there is one), i.e., something *else*, in the *world*, rather than in the *language* (with the exception, of course, of things like names of words, whose referents are other words in the language—after all, language is part of the world). Or the meaning of some piece of language might be its *sense* (Fregean or otherwise)—again, something *else*, outside of the language, though not, perhaps, "in the world." Or it might be a *Meinongian object* or an *idea* in a mind.[12] But, in any case, the meaning of a piece of language is not typically thought of as being part of the language; rather, it is something *else* that the piece of language stands in relation to. One exception is conceptual-role or holistic theories of meaning, but we'll come back to that in §3.1.[13]

So, whereas syntax only requires one domain (call it the "syntactic domain"), semantics requires *two*: a syntactic domain and a "semantic domain." The syntactic domain is the thing that needs to be understood, to be interpreted. The semantic domain is the thing that provides the understanding, that provides the interpretation.

Following Morris, then, I take semantics to be the study of the relations *between* the marks of *two* systems.[14] Because syntax is the study of the properties and relations of a *single* system, it would seem that, indeed, syntax does not suffice for semantics. Yet I argue that it does. Let's look into this more closely.[15]

2 TWO SYNTACTIC SYSTEMS

2.1 THE SYNTAX OF $\mathcal{L}$

On the standard view, a syntactic domain is usually some (formal or formalized) language $\mathcal{L}$, which is described syntactically—that is, in terms of its marks and rules for manipulating them. Thus, for instance, $\mathcal{L}$ might be described as having *terms*, perhaps of two (simple, or atomic) kinds: *individual constants* a, b, . . . (e.g., proper names or other nouns) and *individual variables* u, v, . . . (e.g., pronouns). "New" (complex, or molecular) terms (e.g., noun phrases) can be constructed from previously given or previously constructed ("old") ones (whether atomic or molecular) by means of *function symbols* of various arities, f, g, . . . , f_i, . . . (e.g., "the father of . . .," "the average of . . . and ___"), together with "grammar" rules specifying the "legal" structure (or "spellings") of such molecular terms (say, if $t_1, \ldots, t_n$ are terms, and f^n is an n-place function symbol, then $\ulcorner f^n(t_1, \ldots, t_n)\urcorner$ is a term).[16]

In addition, $\mathcal{L}$ will have *predicate symbols* of various arities: $A, \ldots, Z, A_i, \ldots$ (e.g., verb phrases); *connectives* and *quantifiers*: $\neg, \vee, \forall, \ldots$ (e.g., "it is not the case that . . . ," ". . . or ___," "for all . . . , it is the case that ___"); and more "grammar" rules specifying the "legal" structure of *well-formed formulas* (or sentences): If $t_1, \ldots, t_n$ are terms, and P^n is an n-place predicate symbol, then $\ulcorner P^n(t_1, \ldots, t_n)\urcorner$ is a well-formed formula (wff); if ϕ and ψ are wffs, and v is an individual variable, then $\ulcorner\neg\phi\urcorner$, $\ulcorner(\phi \vee \psi)\urcorner$, $\ulcorner\forall v[\phi]\urcorner$ are wffs.

Note that $\mathcal{L}$ is a *language*. Sometimes $\mathcal{L}$ is augmented with a *logic*: Certain wffs of $\mathcal{L}$ are distinguished as axioms (or "primitive theorems"), and rules of inference are provided that specify how to produce "new" theorems from "old" ones. For instance, if ϕ and $\ulcorner(\phi \rightarrow \psi)\urcorner$ are theorems, then so is ψ. A *proof* of a wff ψ (from a set of wffs Σ) is a sequence of wffs ending with ψ such that every wff in the sequence is either an axiom (or a member of Σ) or follows from previous wffs in the sequence by one of the rules of inference.

And so on. I will assume that the reader is familiar with the general pattern.[17] The point is that all we have so far are marks and rules for manipulating them either linguistically (to form wffs) or logically (to form theorems). All we have so far is syntax in Morris's sense.

Actually, in my desire to make the example perspicuous, I may have given you a misleading impression by talking of "language" and "logic," of "nouns" and "verb phrases," etc. For such talk tends to make people think either that I was talking, albeit in a very strange way, about language and nouns and verbs—good old familiar languages like English with nouns and verbs like "dog" and "run"[18]—or that I had that in the back of my mind as an intended interpretation of the marks and rules. But marks are merely (perhaps) physical inscriptions or sounds that have only some very minimal features such as having distinguished, relatively unchanging shapes capable of being recognized when encountered again.

2.2 THE SYNTAX OF $\mathcal{L}'$

So, let me offer a somewhat less familiar syntactic domain $\mathcal{L}'$, which I will call, this time, not a "language," but merely a "mark system." First, I need to show you the marks of $\mathcal{L}'$. To really make my point, these should be quite arbitrary: say, boxes, circles, or squiggles of various kinds. But I will make life a bit easier for the reader and the typesetter by using letters and numerals.

$\mathcal{L}'$ consists of the following marks:

$A_1, \ldots, A_i, \ldots;$
$F_0, F_1, F_2, F_3;$
(,), ,, ;; [i.e., a left-parenthesis, a right-parenthesis, a comma, and a semi-colon]
R

I want to show you a certain class K of marks of $\mathcal{L}'$. To talk about them, I'll need another set of marks that are not part of $\mathcal{L}'$, so we'll let 'A', 'B', 'C', 'B_1', 'B_2', . . . be variables ranging over the members of K. Now, here are the members of K:

1. $A_1, \ldots, A_i, \ldots \in K$
2. If $A, B \in K$, then $\ulcorner F_0(A)\urcorner$, $\ulcorner F_1(A, B)\urcorner$, $\ulcorner F_2(A, B)\urcorner$, $\ulcorner F_3(A, B)\urcorner$ $\in K$.
3. Nothing else is in K.

We could ask questions of this formal mark system. For instance, which molecular marks are in K? By suitable mark manipulation, following (1)–(3), we can ascertain that A_1, A_{100}, $F_0(A_{100})$, $F_0(F_0(A_{100}))$, $F_3(F_0(F_0(A_{100}))$, $F_2(A_1, A_{100}))$ $\in K$, but that $F_0(F_0)$, $B \notin K$.

Now, let's make $\mathcal{L}'$ a bit more interesting. Let $H \subseteq K$; let $A, B \in K$; and let's say that an (H, A)-*sequence*

is a sequence of members of K such that A is the last item in the sequence, and, if B is in the sequence, then either $B \in H$ or there is a set $\{B_1, \ldots, B_n \mid (\forall 1 \leq i \leq n)[B_i \in K]\}$ such that $\ulcorner R(B_1, \ldots, B_n; B)\urcorner \in \mathcal{R}$, where $\mathcal{R}$ is defined as follows (remember that 'R' is a mark of $\mathcal{L}'$; I am defining $\mathcal{R}$ as consisting of certain sequences of marks beginning with 'R'):

$\mathcal{R}$1. $\ulcorner R(A; F_1(A, B))\urcorner \in \mathcal{R}$

$\mathcal{R}$2. $\ulcorner R(B; F_1(A, B))\urcorner \in \mathcal{R}$

$\mathcal{R}$3. $\ulcorner R(F_1(A, B), F_0(A); B)\urcorner \in \mathcal{R}$

$\mathcal{R}$4. $\ulcorner R(F_1(A, B), F_0(B); A)\urcorner \in \mathcal{R}$

$\mathcal{R}$5. $\ulcorner R(F_2(A, B); A)\urcorner \in \mathcal{R}$

$\mathcal{R}$6. $\ulcorner R(F_2(A, B); B)\urcorner \in \mathcal{R}$

$\mathcal{R}$7. $\ulcorner R(A, B; F_2(A, B))\urcorner \in \mathcal{R}$

$\mathcal{R}$8. $\ulcorner R(F_3(A, B), A; B)\urcorner \in \mathcal{R}$

$\mathcal{R}$9. If there is an (H, B)-sequence whose first item is A, then $\ulcorner R(; F_3(A, B))\urcorner \in \mathcal{R}$
[Note: There is no symbol between "(" and ";"]

$\mathcal{R}$10. If there is an $(H, \ulcorner F_2(B, F_0(B))\urcorner)$-sequence whose first item is A, then $\ulcorner R(; F_0(A))\urcorner \in \mathcal{R}$

$\mathcal{R}$11. If there is an $(H, \ulcorner F_2(B, F_0(B))\urcorner)$-sequence whose first item is $F_0(A)$, then $\ulcorner R(; A)\urcorner \in \mathcal{R}$

$\mathcal{R}$12. Nothing else is in $\mathcal{R}$.

We can now ask more questions of our system; e.g., which marks A are such that $\ulcorner R(; A)\urcorner \in \mathcal{R}$? By suitable mark manipulations, following $\mathcal{R}$1–$\mathcal{R}$12, we can ascertain that, e.g., $R(; F_3(A_0, A_0)) \in \mathcal{R}$ (this is actually fairly trivial, since $\langle A_0\rangle$ is an (A_0, A_0)-sequence whose first item is A_0).

Hard to read, isn't it! You feel the strong desire to try to understand these squiggles, don't you? (Are you, perhaps, beginning to feel like Searle-in-the-Chinese-Room?) You would probably feel better if I showed you some other domain—a *semantic* domain—with which you were more comfortable, more familiar, into which you could map these squiggles. I will. But not yet.

Of course, I could be sadistic and suggest that you "get used to" $\mathcal{L}'$ by manipulating its symbols and learning more about the members of *K* and R . After all, as John von Neumann allegedly said, "in mathematics you don't understand things. You just get used to them."[19] "Getting used to" a syntactic domain is the base case of a recursive *Fundamental Principle of Understanding*:[20]

To understand a syntactic domain *S* is either:

1. to "get used to" *S*, or else

2. to understand *S* in terms of a semantic domain *T*.

The latter is semantic understanding in Morris's sense: understanding one thing *in terms of something else*. The former is what I have called "syntactic understanding":[21] understanding something *in terms of itself*. And in the case of semantic understanding, how do you understand the semantic domain *T*? Normally, *T* is assumed to be *antecedently understood*. But that has to mean that it is understood *syntactically*—you have gotten used to it. If *T* is not antecedently understood, then it has to be considered as a *syntactic* domain in its own right and understood in terms of yet another *semantic* domain *T′*. And so on. Ultimately, I claim, all understanding is syntactic understanding (the base case of the recursion).[22]

But "syntactic understanding"—the sort of thing that you come to have by getting used to the syntactic domain—does not *seem*, on the surface, to be any kind of "real" *understanding*. This is the intuition underlying Searle's Chinese Room Argument and its earlier incarnation in the guise of Leibniz's mill.[23] Where is the meaning or understanding (or "intentionality" or "consciousness") in this kind of "meaningless" (yet rule-based, or regulated) mark manipulation? As I read him, Jacquette suggests that it *can* generate understanding,[24] in turn suggesting that what we have here is a clash of fundamental intuitions: Some (e.g., Searle) say that such mark systems and mark manipulation can*not* suffice for understanding; others (perhaps Jacquette, and certainly I) say that it *can*.

In any case, you *could* just try to get used to $\mathcal{L}'$ by doing mark manipulation, and I believe that you would thereby come to understand it. But I won't be that mean. First, we need to move away from pure syntax and find out what semantics consists of.

2.3 A SEMANTIC INTERPRETATION OF $\mathcal{L}$

Given some syntactic domain—some (formal) mark system—one can ask two sorts of questions about it. The first sort is exemplified by those we asked above: What are the members of *K*? Of $\mathcal{R}$? These are purely "internal," syntactic, questions. The second sort is, in short: What's the meaning of all this? What do the marks mean (if anything)? What, for example, is so special about the members of *K* or the marks of the form $\ulcorner R(; A)\urcorner$? To answer this sort of question, we must go *outside* the syntactic domain: We must provide "external" entities that the marks mean (that they can be understood in terms of), and we must show the mappings—the associations, the correspondences—*between* the *two* domains.

But, as I have said elsewhere,

> Now a curious thing happens: I need to show you the semantic domain. If I'm very lucky, I can just point it out to you—we can look at it together, and I can describe the correspondences ("The symbol A_{37} means that red thing over there"). But, more often, I have to describe the semantic domain to you in . . . symbols [i.e., marks], and hope that the meaning of *those* symbols will be obvious to you.[25]

So, let's provide a semantic interpretation of our first formal mark system, $\mathcal{L}$. Since $\mathcal{L}$ had individual terms, function marks, and predicate marks[26]—which could be combined in various (but not arbitrary) ways—I need to provide meanings for each such mark as well as for their legal combinations. So, we'll need a non-empty set **D** of things that the terms will mean—a **D**omain of interpretation (sometimes called a **D**omain, or universe, of discourse)—and sets **F** and **R** of things that the function and relation symbols will mean, respectively. These three sets can be collectively called **M** (for **M**odel). What's in **D**? Well, anything you want to talk or think about. What are in **F** and **R**? Functions and relations on **D** of various arities—i.e., anything you want to be able to say about the things in **D**. That's our *ontology*, what there is.

Let's pause a moment here for an important point: **D** has members; the members of **D** have properties; and the members of **D** stand in various relations to each other. The study of such objects, their properties, and the relations among them is *ontology*. But I have defined the study of objects, their properties, and the relations among them to be *syntax*. Thus, *ontology is simply the syntax of the semantic domain*.

Now for the correspondences. To say what a mark of $\mathcal{L}$ means in **M** (what the meaning, located in **M,** of a mark of $\mathcal{L}$ is), we can define an interpretation function $I : \mathcal{L} \rightarrow \mathbf{M}$ that will assign to each mark of $\mathcal{L}$ something in **M** (or it might be an interpretation *relation* if we wish to allow for ambiguity), as follows:

1. If *t* is an individual term of $\mathcal{L}$, then $I(t) \in \mathbf{D}$.

   (*Which* element of **D**? Whichever you want, or, if we spell out $\mathcal{L}$ and **D** in more detail, I'll tell you; for example, perhaps I ("Barack Obama") = the 44th President of the U.S., if "Barack Obama" is an individual constant of $\mathcal{L}$, and **D** is the set of humans.)

2. If *f* is a function symbol of $\mathcal{L}$, then $I(f) \in \mathbf{F}$.

3. If $\ulcorner f(t_1, \ldots, t_n)\urcorner$ is a (molecular) term of $\mathcal{L}$, then $\mathcal{I}(\ulcorner f(t_1, \ldots, t_n)\urcorner) = \mathcal{I}(f)(\mathcal{I}(t_1), \ldots, \mathcal{I}(t_n)) \in \mathbf{D}$.

   (I.e., the interpretation of $\ulcorner f(t_1, \ldots, t_n)\urcorner$ will be the result of applying (a) the function that is the interpretation of *f* to (b) the elements of **D** that are the interpretations of the t_i; and the result will be an element of **D**.)

4. If *P* is a predicate symbol of $\mathcal{L}$, then $\mathcal{I}(P) \in \mathbf{R}$.

So far, so good. Now, what do wffs mean? Those philosophers and logicians who take *n*-place functions and relations to be ordered *n*-tuples—functions and relations "in extension"—tend to talk about "truth values" of wffs rather than "meanings." Others, who take functions and relations "in intension" can talk about the meanings of wffs as being "states of affairs" or "situations" or "propositions," variously defined. I, myself, fall in the latter camp, but for the sake of simplicity of exposition, I'll go the other route for now. Continuing, then, we have:

5. If ϕ is a wff, then $\mathcal{I}(\phi) \in \{0, 1\}$, where, intuitively, we'll say that ϕ is "true" if $\mathcal{I}(\phi) = 1$ and that ϕ is "false" if $\mathcal{I}(\phi) = 0$. In particular, where *P* is an *n*-place predicate symbol, $t_1, \ldots, t_n$ are terms, *v* is an individual variable, and ϕ, ψ are wffs:

   (a) $\mathcal{I}(\ulcorner P(t_1, \ldots, t_n)\urcorner) = 1$ iff $\langle \mathcal{I}(t_1), \ldots, \mathcal{I}(t_n)\rangle \in \mathcal{I}(P)$.

   (b) $\mathcal{I}(\ulcorner\neg\phi\urcorner) = 1$ iff $\mathcal{I}(\phi) = 0$

   (c) $\mathcal{I}(\ulcorner(\phi \vee \psi)\urcorner) = 1$ iff $\mathcal{I}(\phi) = 1$ or $\mathcal{I}(\psi) = 1$ (or both)

   (d) $\mathcal{I}(\ulcorner\forall v[\phi]\urcorner) = 1$ iff $\mathcal{I}'(\phi) = 1$ for every $\mathcal{I}'$ that differs from $\mathcal{I}$ at most on what $\mathcal{I}'$ assigns to *v*.

Now, what kind of function is $\mathcal{I}$? Clearly, it is a homomorphism; i.e., it satisfies a principle of compositionality: The interpretation of a molecular symbol is determined by the interpretations of its atomic constituents in the manner spelled out above.

In the ideal case, $\mathcal{I}$ is an *isomorphism*—a 1–1 and onto homomorphism; that is, every item in **M** is the meaning of *just one* symbol of $\mathcal{L}$. (Being "onto" is tantamount to $\mathcal{L}$'s being "complete.") Perhaps isomorphism is less than ideal, at least for the case of natural languages. David P. Wilkins has observed that when one studies, not isolated or made-up sentences, but

> real, contextualised utterances . . . it is often the case that all the elements that one would want to propose as belonging to semantic structure have no overt manifestations in syntactic structure. . . . [T]he degree of isomorphism between semantic and syntactic structure is mediated by pragmatic and functional concerns. . . .[27]

In this ideal situation, **M** is a virtual duplicate or mirror image of $\mathcal{L}$. (Indeed, **M** could *be* $\mathcal{L}$ itself,[28] but that's not very interesting or useful for *semantic understanding* of $\mathcal{L}$; rather, it would be a kind of *syntactic understanding*!)

In less ideal circumstances, there might be marks of $\mathcal{L}$ that are not interpretable in **M**; in that case, $\mathcal{I}$ would be a *partial* function. Such is the case when $\mathcal{L}$ is English and **M** is the world ("unicorn" is an English word, but unicorns don't exist), though if we "enlarge" or "extend" **M** in some way—e.g., if we take **M** to be Meinong's *Aussersein* instead of the actual world—then we can make *I* total.[29]

In another less ideal circumstance, "Hamlet's Law" might hold:[30] There are more things in **M** than in $\mathcal{L}$; i.e., there are elements of **M** not expressible in $\mathcal{L}$: $\mathcal{I}$ is not onto. And, as noted earlier, $\mathcal{I}$ might be a relation, not a function, so $\mathcal{L}$ would be ambiguous. There is another, more global, sense in which $\mathcal{L}$ could be ambiguous: By choosing a different **M** (and a different $\mathcal{I}$), we could give the marks of $\mathcal{L}$ entirely distinct meanings. Worse, the two **M**s need not be isomorphic. (This can happen in at least two ways. First, the cardinalities of the two **D**s could differ. Second, suppose $\mathcal{L}$ is a language for expressing mathematical group theory. Then $\mathbf{M}_1$ could be an infinite cyclic group (e.g., the integers under addition), and $\mathbf{M}_2$ could be $\mathbf{M}_1 \times \mathbf{M}_1$, which, unlike $\mathbf{M}_1$, has two disjoint subgroups—except for the identity.)[31]

2.4 A SEMANTIC INTERPRETATION OF $\mathcal{L}'$

Let's consider an example in detail; I'll tell you what the marks of $\mathcal{L}'$ mean. First, I need to show you **M.** To do that, I need to show you **D**: **D** will include the marks: $\phi_1, \ldots, \phi_i, \ldots$ (so, I'm explaining one set of marks in terms of another set of marks; be patient). **D** will also include these marks: $\neg$, $\vee, \wedge, \rightarrow$. Now I can tell you about *K* (in what follows, let A_i be the *i*th atomic marks of *K*, let ϕ_i be the *i*th atomic marks of **D**, and let $A, B \in K$):

$$\mathcal{I}(A_i) = \phi_i$$

$$\mathcal{I}(F_0) = \neg$$

$$\mathcal{I}(F_1) = \vee$$

$$\mathcal{I}(F_2) = \wedge$$

$$\mathcal{I}(F_3) = \rightarrow$$

$$\mathcal{I}(\ulcorner F_0(A)\urcorner) = \ulcorner\neg\ \mathcal{I}(A)\urcorner$$

$$\mathcal{I}(\ulcorner F_1(A, B)\urcorner) = \ulcorner(\mathcal{I}(A) \vee \mathcal{I}(B))\urcorner$$

$$\mathcal{I}(\ulcorner F_2(A, B)\urcorner) = \ulcorner(\mathcal{I}(A) \wedge \mathcal{I}(B))\urcorner$$

$$\mathcal{I}(\ulcorner F_3(A, B)\urcorner) = \ulcorner(\mathcal{I}(A) \rightarrow \mathcal{I}(B))\urcorner$$

I assume, of course, that you know what "$\neg$", $\ulcorner(\mathcal{I}(A) \rightarrow \mathcal{I}(B))\urcorner$, etc., are (namely, the negation sign, a material conditional wff, etc.). So, the elements of *K* are just wffs of propositional logic (as if you didn't know)!

What about $\mathcal{R}$? Well: $\mathcal{I}(R) = \vdash$ (where $\vdash\ \in \mathbf{R}$ and where **R**, of course, is part of **M**); i.e., *R* means the deducibility relation on wffs of propositional logic. So, the elements of $\mathcal{R}$ are rules of inference:

$\mathcal{I}(\ulcorner R(A; F_1(A, B))\urcorner) = A \vdash \ulcorner(A \vee B)\urcorner$ (i.e., ∨-introduction)

$\mathcal{I}(\ulcorner R(B; F_1(A, B))\urcorner) = B \vdash \ulcorner(A \vee B)\urcorner$ (i.e., ∨-introduction)

$\mathcal{I}(\ulcorner R(F_1(A,B), F_0(A); B)\urcorner) = \ulcorner(A \vee B)\urcorner, \ulcorner\neg A\urcorner \vdash B$ (i.e., ∨-elimination)

$\mathcal{I}(\ulcorner R(F_1(A,B), F_0(B); A)\urcorner) = \ulcorner(A \vee B)\urcorner, \ulcorner\neg B\urcorner \vdash A$ (i.e., ∨-elimination)

$\mathcal{I}(\ulcorner R(F_2(A, B); A)\urcorner) = \ulcorner(A \wedge B)\urcorner \vdash A$ (i.e., ∧-elimination)

$\mathcal{I}(\ulcorner R(F_2(A, B); B)\urcorner) = \ulcorner(A \wedge B)\urcorner \vdash B$ (i.e., ∧-elimination)

$\mathcal{I}(\ulcorner R(A, B; F_2(A, B))\urcorner) = A, B \vdash \ulcorner(A \wedge B)\urcorner$ (i.e., ∧-introduction)

$\mathcal{I}(\ulcorner R(F_3(A, B), A; B)\urcorner) = \ulcorner(A \rightarrow B)\urcorner, A \vdash B$ (i.e., →-elimination, or Modus Ponens)

Before we can finish interpreting R, I need to tell you what an (H, A)-sequence means: It is a proof of $\mathcal{I}(A)$ from hypotheses $I(H)$ (where, to be absolutely precise, I should specify that, where $H = \{A, B, \ldots\} \subseteq K$, $I(H) = \{I(A), I(B), \ldots\}$). So:

> $\mathcal{I}(\mathcal{R}9)$ is:
>
> if there is a proof of $\mathcal{I}(B) \in \mathbf{D}$ from a set of hypotheses $\mathcal{I}(H)$ whose first line is $\mathcal{I}(A)$, then $\vdash \ulcorner(\mathcal{I}(A) \rightarrow \mathcal{I}(B))\urcorner$ (i.e., →-introduction, or Conditional Proof)
>
> $\mathcal{I}(\mathcal{R}10)$ is:
>
> if there is a proof of $\ulcorner(\mathcal{I}(B) \wedge \neg\ \mathcal{I}(B))\urcorner$ from a set of hypotheses $\mathcal{I}(H)$ whose first line is $\mathcal{I}(A)$, then $\vdash \ulcorner\neg\ \mathcal{I}(A)\urcorner$ (i.e., ¬-introduction)
>
> $\mathcal{I}(\mathcal{R}11)$ is:
>
> if there is a proof of $\ulcorner(\mathcal{I}(B) \wedge \neg\ \mathcal{I}(B))\urcorner$ from a set of hypotheses $\mathcal{I}(H)$ whose first line is $\ulcorner\neg\ \mathcal{I}(A)\urcorner$, then $\vdash \mathcal{I}(A)$ (i.e., ¬-elimination)

So, now you know: $\mathcal{L}'$ is just ordinary propositional logic in a weird notation. Of course, I could have told you what the marks of $\mathcal{L}'$ mean in terms of a *different* model $\mathbf{M}'$, where $\mathbf{D}'$ consists of states of affairs and Boolean operations on them. In that case, $\mathcal{L}'$ just *is* ordinary propositional logic. That is, $\mathbf{M}$ is itself a syntactic formal mark system (namely, $\mathcal{L}$!) whose meaning can be given in terms of $\mathbf{M}'$, but $\mathcal{L}'$'s meaning can be given either in terms of $\mathbf{M}$ *or* in terms of $\mathbf{M}'$.

There are several lessons to be learned from this. First, $\mathcal{L}'$ is not a very "natural" mark system. Usually, when one presents the syntax of a formal mark system, one already has a semantic interpretation in mind, and one *designs* the syntax to "capture" that semantics: The syntax is a model—an implementation—of the semantics.[32]

Second, it is possible and occasionally even useful to allow *one* formal *syntactic* system to be the *semantic* interpretation of *another* syntactic system. Of course, this is only useful if the interpreting syntactic system is antecedently understood. How? In terms of *another* domain with which we are antecedently familiar! So, in our example, the unfamiliar $\mathcal{L}'$ was interpreted in terms of the more familiar $\mathbf{M}$ (i.e., $\mathcal{L}$), which, in turn, was interpreted in terms of $\mathbf{M}'$. And how is it that we understand what states of affairs in the world are? Well . . . we've just gotten used to them. (We'll come back to this in §4.)

Finally, note that $\mathbf{M}$ in our example is a sort of "swing" domain: It serves as the *semantic* domain relative to $\mathcal{L}'$ and as the *syntactic* domain relative to $\mathbf{M}'$. We can have a "chain" of domains, each of which except the first is a semantic domain for the one before it, and each of which except for the last is a syntactic domain for the one following it. To understand any domain in the chain, we must be able to understand the "next" one. How do we understand the last one? Syntactically.[33]

3 SYNTAX SUFFICES FOR SEMANTICS

3.1 SYNTACTIC UNDERSTANDING

Let's take stock. Given any (non-empty) set S of objects of any kind, the specification of the properties of S's members and of the relations that they stand in to each other is the *syntax* of S. These properties and relations may be of different kinds, so we might be able to identify a "grammatical" syntax of S as well as a "logical" or "proof-theoretic" syntax of S. We can understand S in terms of its syntax by "getting used to" manipulating its members according to these properties and relations. This is syntactic understanding.[34]

Syntactic understanding is holistic in the following way: The syntax of S can be represented by a graph whose vertices are the members of S and whose edges represent its properties and relations. Such a graph is often called a "semantic network," and such networks have rightly been criticized as really being "syntactic" networks. (The Semantic Web is really a syntactic web.[35]) In such a network, the "meaning" of any vertex is its location in the network—its relations to all other vertices in the network. This is conceptual-role semantics or semantic holism. Semantic holism is just more syntax.[36]

3.2 SEMANTIC UNDERSTANDING

But there is another kind of understanding: semantic understanding. Here, we need another set, T, in terms of which we understand S. When we ask what $s \in S$ means, our answer is some $t \in T$. But T will have its own syntax. As I noted earlier, I see no difference between the syntax of T, thus understood, and the *ontology* of T, though we tend to reserve the former term for languages and logics, and the latter term for the realms that those languages and logics describe or are "about." Thus, if we are understanding S in terms of T, we would speak of the syntax of S and the ontology of T, but that is merely a manner of speaking.

Semantic understanding requires relations between S and T: relations of meaning, reference, etc. But these relations are not among the (internal) relations of S's syntax or T's ontology. They *connect* S and T, but are *external* to both.

3.3 SYNTAX IS SEMANTICS

So, how can we talk about those semantic relations? We cannot use either S or T by themselves to talk about them, because the semantic interpretation function is not part of *S* alone or of *T* alone. But we *can* talk about them by taking the *union* of *S* and *T*; call it *U*. (In earlier writings about computational theories of cognition, I have called this the "internalization" of the semantic domain into the syntactic domain.[37] See §4.2, below.) What is the syntax of *U*? It consists, in part, of the inventory of properties of members of *U*. This includes all of the properties of members of *S and* all of the properties of the members of *T*. It also consists, in part, of the inventory of relations among the members of *S and* the relations among the members of *T*. *But it* ***also*** *includes the semantic relations* ***between*** *the members of S and the members of T.*

So, it is the syntax of *U* that enables us to talk about the semantics of *S*. Semantics is, thus, just more syntax—the syntax of the union of a syntactic domain and its semantic domain. QED

4 IMPLICATIONS

I will close with brief comments on two philosophical issues that can be illuminated by this theory.

4.1 TWIN EARTH

Hilary Putnam has argued that, not only can "two terms . . . have the same extension and yet differ in intension," but that "two terms can . . . differ in extension and have the same intension."[38] The latter claim is intended to be surprising, because it is typically held that intensions determine extensions. Putnam offers his Twin Earth thought experiment as a counterexample.

I do not want to rehearse these arguments here, but merely point out some similar issues and see what my semantics-as-syntax theory might have to say.

First, note that the fact that the intensions of "water" (on Earth) and "water" (on Twin Earth) might be identical yet their extensions (H_2O and XYZ, respectively) be different parallels a situation with computer programs: By way of an "intuition pump," recall that intensions (and Fregean senses) are sometimes modeled as (computable) functions or algorithms. Now, it can be the case that a single algorithm with a given input can have different outputs depending on the context in which those algorithms are executed. For one example from the literature, an algorithm (recipe) for producing hollandaise sauce when executed on Earth will likely produce something quite different when executed on the Moon.[39] (Strictly speaking, perhaps, the context should be taken as part of the input, so the algorithms will, in fact, have the same outputs if given exactly the same inputs.[40])

Second, the relation of a word to its intension is simply one kind of meaning. The relation of a word to its extension is another kind of meaning. (There is no such thing as "the" meaning of a word; to claim that "meanings ain't in the head" is highly misleading, because some of them are![41]) What I want to point out is that the relation of an intension to an extension is *yet another kind of meaning*: All three of these relations are semantic in my sense, because they are relations between two domains.

4.2 COMPUTATIONAL COGNITION

If semantics is nothing but syntax (albeit syntax writ large), how do we understand language? How might a *computer* understand language? Searle says that it can't, and I suggested the intuition behind this in §2.2. But I also mentioned a way out in §3: via "internalization." I have cashed this out in the series of essays cited in note 4, but I will summarize my position here.

What *seems* to be missing in the Chinese Room is "real" semantics—links to the external-world referents of the words and sentences of language. How does Searle-in-the-room (actually, Searle-in-the-room together with the instruction book!) know that the word 'hamburger' means "hamburger" (i.e., refers to hamburgers), or that a certain (Chinese) "squiggle" does?

According to the theory I presented above in §3, it would seem that an actual hamburger would somehow need to be "imported" into Searle-in-the-room's instruction book (the computer program for natural-language understanding and generation). Instead, a *representative* of an actual hamburger is thus imported into Searle-in-the-room's "mind" (or "semantic network"). The hamburger is "internalized." In the case of a real human being, this representative is the end result of, say, the visual process of seeing a hamburger (or the olfactory process of smelling one, etc.), resulting in a "mental image" of a hamburger. (To speak with Kant, it is an "intuition" or concept of a hamburger, not the hamburger-in-itself.) More precisely, the biological neural network in the human's brain has neurons whose firings represent the *word* "hamburger," *and* it has neurons whose firings represent the actual hamburger. Both of these sets of neuron firings are in the same "language"—the same syntactic system. Call it "*U*." As in §3.3, $U = S \cup T$, where *S* is the neuron firings of language, and *T* is the neuron firings of perceptual images. *U* is the "language of thought."[42] Yes, *T* is just "more symbols" (as Searle has objected;[43] more precisely, *T* is just more neuron firings—the "marks" of *T* as a syntactic system). But that's how semantics works. The same thing happens (or can happen) for computers; though the language of thought won't be a *biological* neural network (it might be a computational semantic network such as SNePS[44] or an artificial neural network). Thus, a combination of the "robot reply" (for internalization) and the "systems reply" (because it is never Searle-in-the-room alone) show us how to escape the Chinese Room.[45]

NOTES

1. Searle, "The Failures of Computationalism," 68.
2. Rapaport, "Syntactic Semantics: Foundations of Computational Natural-Language Understanding," 85–86.
3. Searle, "Minds, Brains, and Programs"; Searle, "The Myth of the Computer"; Searle, "Is the Brain a Digital Computer?"
4. Rapaport, "Philosophy, Artificial Intelligence, and the Chinese-Room Argument"; Rapaport, "Searle's Experiments with Thought"; Rapaport, "Syntactic Semantics: Foundations of Computational Natural-Language Understanding"; Rapaport, "Computer Processes and Virtual Persons: Comments on Cole's 'Artificial Intelligence and Personal Identity'"; Rapaport, "Because Mere Calculating Isn't Thinking: Comments on

Hauser's 'Why Isn't My Pocket Calculator a Thinking Thing?'"; Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition"; Rapaport, "How Minds Can Be Computational Systems"; Rapaport, "Implementation Is Semantic Interpretation"; Rapaport, "How to Pass a Turing Test: Syntactic Semantics, Natural-Language Understanding, and First-Person Cognition"; Rapaport, "Holism, Conceptual-Role Semantics, and Syntactic Semantics"; Rapaport, "What Did You Mean by That? Misunderstanding, Negotiation, and Syntactic Semantics"; Rapaport, "Implemention Is Semantic Interpretation: Further Thoughts"; Rapaport, "How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room."; Rapaport, "Searle on Brains as Computers"; Rapaport, "Yes, She Was! Reply to Ford's 'Helen Keller Was Never in a Chinese Room'"; Rapaport, "Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing."

[5] Morris, *Foundations of the Theory of Signs*, 6, 13. See Posner, "Origins and Development of Contemporary Syntactics," for a detailed history and analysis of syntactics.

[6] Carnap, *The Logical Syntax of Language*, 1, 6–7.

[7] Jacquette, "Fear and Loathing (and Other Intentional States) in Searle's Chinese Room," 294.

[8] Ibid., 295; my italics.

[9] I discuss these issues in the context of the relation of computer programs to the world in Rapaport, "On the Relation of Computing to the World."

[10] Rapaport, "Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing," 38.

[11] Carnap, *The Logical Syntax of Language*, 6; original italics, my boldface.

[12] On Meinongian semantics, see Rapaport, "Meinongian Theories and a Russellian Paradox"; Rapaport, "How to Make the World Fit Our Language: An Essay in Meinongian Semantics"; Rapaport, "Meinongian Semantics for Propositional Semantic Networks"; Rapaport, "Non-Existent Objects and Epistemological Ontology"; Rapaport, "Meinongian Semantics and Artificial Intelligence"; Shapiro and Rapaport, "Models and Minds: Knowledge Representation for Natural-Language Competence"; Shapiro and Rapaport, "An Introduction to a Computational Reader of Narratives"; Rapaport and Shapiro, "Cognition and Fiction"; Rapaport and Kibby, "Contextual Vocabulary Acquisition as Computational Philosophy and as Philosophical Computation"; and Rapaport and Kibby, "Contextual Vocabulary Acquisition: From Algorithm to Curriculum."

[13] On the relation of these to cognitive semantics, see Rapaport, "What Did You Mean by That? Misunderstanding, Negotiation, and Syntactic Semantics."

[14] Morris, *Foundations of the Theory of Signs*, 21.

[15] Much of what follows is a detailed elaboration of comments I made in Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition," §2.2.

[16] For typographical convenience, I use superscripted brackets in place of Quinean quasi-quotes.

[17] See, e.g., Rapaport, "Logic, Propositional"; and Rapaport, "Logic, Predicate" for more details.

[18] Or "dog" (plural: "dogs") and "(to) dog": "Dogs dogs dog dog dogs" is syntactically correct and semantically meaningful (if not pragmatically acceptable) as a sentence of English, meaning "Dogs—whom other dogs follow—follow other dogs." Or "buffalo" (plural: "buffalo"(!)) and "(to) buffalo" (meaning "(to) intimidate"): "Buffalo buffalo buffalo buffalo buffalo" is likewise syntactically correct and semantically meaningful (but only pragmatically acceptable in Buffalo, NY); see http://www.cse.buffalo.edu/~rapaport/BuffaloBuffalo/buffalobuffalo.html.

[19] https://en.wikiquote.org/wiki/John von Neumann

[20] Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition," §2.1.

[21] Rapaport, "Searle's Experiments with Thought."

[22] See Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition" for further discussion.

[23] "Imagine there were a machine whose structure produced thought, feeling, and perception; we can conceive of its being enlarged while maintaining the same relative proportions among its parts, so that we could walk into it as we can walk into a mill. Suppose we do walk into it; all we would find there are cogs and levers and so on pushing one another, and never anything to account for a perception." Leibniz, "The Principles of Philosophy Known as Monadology," §17.

[24] "[A]s we move closer to an exact microlevel decentralized . . . input-output isomorphism with the neurological activity of natural intelligence, the intuition that we do not thereby also precisely duplicate the brain's causal powers by which it produces intentionality begins to fade and lose its grip on our pre-theoretical beliefs. The imaginable causal efficacy of microlevel input-output functionalities raises difficulties about the adequacy of the Chinese Room example to support Searle's thesis that a functioning program in which physical syntax tokens causally interact with themselves and a machine environment at the proper level of design could not produce intentionality just as effectively as the natural system it simulates." Jacquette, "Fear and Loathing (and Other Intentional States) in Searle's Chinese Room," 293.

[25] Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition," §2.2.2. For further discussion of this problem, see Smith, "The Correspondence Continuum"; and Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition," §2.5.

[26] I.e., what are normally called "function symbols" and "predicate symbols."

[27] Wilkins, "Expanding the Traditional Category of Deictic Elements: Interjections as Deictics," 381.

[28] Chang and Keisler, *Model Theory*, 4ff.

[29] Rapaport, "How to Make the World Fit Our Language: An Essay in Meinongian Semantics."

[30] "There are more things in heaven and earth, Horatio, / Than are dreamt of in your philosophy," *Hamlet* (I, 5, ll. 167–68). This can also be interpreted as a summary of Gödel's Incompleteness Theorem (where "dreamt of" means "provable").

[31] I am grateful to Nicolas Goodman for this example.

[32] On the nature of implementation, see Rapaport, "Implementation Is Semantic Interpretation"; and Rapaport, "Implemention Is Semantic Interpretation: Further Thoughts."

[33] For more on these "chains" and their possible components, see Rapaport, "Understanding Understanding: Syntactic Semantics and Computational Cognition," §§2.3ff.

[34] I explore this, with an example from elementary algebra, in Rapaport, "Searle's Experiments with Thought."

[35] Berners-Lee and Fischetti, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, 12; cf. pp. 184ff; and Ceusters, "Ontology: The Need for International Coordination." For discussion relevant to the present essay, see, especially, Rapaport, "How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room," 393; and also Rapaport, "Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing," 41–42, 48.

[36] I explore conceptual-role semantics and holism in more detail, and respond to Fodor and Lepore's objections (*Holism: A Shopper's Guide*) in Rapaport, "Holism, Conceptual-Role Semantics, and Syntactic Semantics."

[37] Rapaport, "Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing," §3.

[38] Putnam, "Meaning and Reference," 700; and, more famously, in Putnam, "The Meaning of 'Meaning'."

[39] Cleland, "Is the Church-Turing Thesis True?"

[40] I discuss this in greater detail in Rapaport, "On the Relation of Computing to the World."

[41] Putnam, "Meaning and Reference," 704.

[42] Fodor, *The Language of Thought*.

[43] Searle, "Minds, Brains, and Programs," 423.

44. Again, see the essays in note 4, as well as Shapiro and Rapaport, "SNePS Considered as a Fully Intensional Propositional Semantic Network"; Shapiro and Rapaport, "The "SNePS Family."

45. More details can be found in Rapaport, 2 "How to Pass a Turing Test: Syntactic Semantics, Natural-Language Understanding, and First-Person Cognition"; and Rapaport, "How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room."

REFERENCES

Berners-Lee, Tim, and Mark Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. New York: HarperCollins, 1999.

Carnap, Rudolf. *The Logical Syntax of Language*. Trans. by Amethe Smeaton. London: Routledge & Kegan Paul, 1937.

Ceusters, Werner. "Ontology: The Need for International Coordination." 1975. Available at http://ncor.buffalo.edu/inaugural/ppt/ceusters.ppt.

Chang, Chen-Chung, and H. Jerome. *Model Theory*. Amsterdam: North-Holland, 1973.

Cleland, Carole E. "Is the Church-Turing Thesis True?" *Minds and Machines* 3, no. 3 (1993): 283–312.

Fodor, Jerry A. *The Language of Thought*. New York: Thomas Y. Crowell Co., 1975.

Fodor, Jerry, and Ernest Lepore. *Holism: A Shopper's Guide*. Cambridge, MA: Basil Blackwell, 1992.

Jacquette, Dale. "Fear and Loathing (and Other Intentional States) in Searle's Chinese Room." *Philosophical Psychology* 3, no. 2/3 (1990): 287–304.

Leibniz, Gottfried Wilhelm. "The Principles of Philosophy Known as Monadology." 1714. Trans. by Jonathan Bennett (July 2007). Accessed from *Some Texts from Early Modern Philosophy*. Available at http://www.earlymoderntexts.com/pdfs/leibniz1714b.pdf.

Moor, James H., ed. *The Turing Test: The Elusive Standard of Artificial Intelligence*. Dordrecht, The Netherlands: Kluwer Academic, 2003.

Morris, Charles. *Foundations of the Theory of Signs*. Chicago: University of Chicago Press, 1938.

Posner, Roland. "Origins and Development of Contemporary Syntactics." *Languages of Design* 1 (1992): 37–50.

Putnam, Hilary. "Meaning and Reference." *Journal of Philosophy* 70, no. 19 (1973): 699–711. Available at http://155.97.32.9/~mhaber/Documents/Course%20Readings/Putnam-MeaningReference-JPhil1973.pdf.

Putnam, Hilary. "The Meaning of 'Meaning'." In *Minnesota Studies in the Philosophy of Science, Vol. 7: Language, Mind, and Knowledge*, edited by Keith Gunderson, 131–93. Minneapolis: University of Minnesota Press, 1975. Reprinted in Hilary Putnam, *Mind, Language and Reality*, 215–71. Cambridge, UK: Cambridge University Press, 1979. Available at http://mcps.umn.edu/assets/pdf/7.3_Putnam.pdf.

Rapaport, William J. "Meinongian Theories and a Russellian Paradox." *Noûs* 12, no. 2 (1978): 153–80; Errata, *Noûs* 13 (1979): 125.

Rapaport, William J. "How to Make the World Fit Our Language: An Essay in Meinongian Semantics." *Grazer Philosophische Studien* 14 (1981): 1–21. Available at http://www.cse.buffalo.edu/~rapaport/Papers/how2makeworldfitlg.pdf.

Rapaport, William J. "Meinongian Semantics for Propositional Semantic Networks." In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84, Stanford University, 1985)*, 65–70. Association for Computational Linguistics, Morristown, NJ. Available at https://www.aclweb.org/anthology/P/P85/P85-1006.pdf.

Rapaport, William J. "Non-Existent Objects and Epistemological Ontology." *Grazer Philosophische Studien* no. 25/26 (1985-1986): 61–95. Available at http://www.cse.buffalo.edu/~rapaport/Papers/rapaport8586.pdf.

Rapaport, William J. "Philosophy, Artificial Intelligence, and the Chinese- Room Argument." *Abacus: The Magazine for the Computer Professional* 3 (1986): 6–17. Correspondence, *Abacus* 4 (Winter 1987): 6–7; 4 (Spring): 5–7. Available at http://www.cse.buffalo.edu/~rapaport/Papers/abacus.pdf.

Rapaport, William J. "Searle's Experiments with Thought." *Philosophy of Science* 53 (1986): 271–79. Available at http://www.cse.buffalo.edu/~rapaport/Papers/philsci.pdf.

Rapaport, William J. "Syntactic Semantics: Foundations of Computational Natural-Language Understanding." In *Aspects of Artificial Intelligence*, edited by James H. Fetzer, 81–131. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1988. Available at http://www.cse.buffalo.edu/~rapaport/Papers/synsem.pdf. Reprinted with numerous errors in *Thinking Machines and Virtual Persons: Essays on the Intentionality of Machines*, edited by Eric Dietrich, 225–73. San Diego: Academic Press, 1994.

Rapaport, William J. "Computer Processes and Virtual Persons: Comments on Cole's 'Artificial Intelligence and Personal Identity'." *Technical Report 90-13*, SUNY Buffalo Department of Computer Science, Buffalo. 1990. Available at http://www.cse.buffalo.edu/~rapaport/Papers/cole.tr.17my90.pdf.

Rapaport, William J. "Logic, Predicate." In *Encyclopedia of Artificial Intelligence, 2nd edition*, edited by Stuart C. Shapiro, 866–73. New York: John Wiley, 1992. Available at http://www.cse.buffalo.edu/~rapaport/Papers/logic,predicate.pdf.

Rapaport, William J. "Logic, Propositional." In *Encyclopedia of Artificial Intelligence, 2nd edition*, edited by Stuart C. Shapiro, 891–97. New York: John Wiley, 1992. Available at http://www.cse.buffalo.edu/~rapaport/Papers/logic,propositional.pdf.

Rapaport, William J. "Because Mere Calculating Isn't Thinking: Comments on Hauser's 'Why Isn't My Pocket Calculator a Thinking Thing?'" *Minds and Machines* 3 (1993): 11–20. Available at http://www.cse.buffalo.edu/~rapaport/Papers/joint.pdf.

Rapaport, William J. "Understanding Understanding: Syntactic Semantics and Computational Cognition." In *AI, Connectionism, and Philosophical Psychology, Philosophical Perspectives, Vol. 9*, edited by James E. Tomberlin, 49–88. Atascadero, CA: Ridgeview, 1995. Available at http://www.cse.buffalo.edu/~rapaport/Papers/rapaport95-uu.pdf. Reprinted in Language and Meaning in *Cognitive Science: Cognitive Issues and Semantic Theory, Artificial Intelligence and Cognitive Science: Conceptual Issues, Vol. 4*, edited by Josefa Toribio and Andy Clark, 73–88. New York: Garland, 1998.

Rapaport, William J. "How Minds Can Be Computational Systems." *Journal of Experimental and Theoretical Artificial Intelligence* 10 (1998): 403–19. Available at http://www.cse.buffalo.edu/~rapaport/Papers/jetai-sspp98.pdf.

Rapaport, William J. "Implementation Is Semantic Interpretation." *The Monist* 82 (1999): 109–30. Available at http://www.cse.buffalo.edu/~rapaport/Papers/monist.pdf.

Rapaport, William J. "How to Pass a Turing Test: Syntactic Semantics, Natural-Language Understanding, and First-Person Cognition." *Journal of Logic, Language, and Information* 9, no. 4 (2000): 467–90. Available at http://www.cse.buffalo.edu/~rapaport/Papers/TURING.pdf. Reprinted in Moor, *The Turing Test: The Elusive Standard of Artificial Intelligence*, 161–84.

Rapaport, William J. "Holism, Conceptual-Role Semantics, and Syntactic Semantics." *Minds and Machines* 12, no. 1 (2002): 3–59. Available at http://www.cse.buffalo.edu/~rapaport/Papers/crs.pdf.

Rapaport, William J. "What Did You Mean by That? Misunderstanding, Negotiation, and Syntactic Semantics." *Minds and Machines* 13, no. 3 (2003): 397–427. Available at http://www.cse.buffalo.edu/~rapaport/Papers/negotiation-mandm.pdf.

Rapaport, William J. "Implemention Is Semantic Interpretation: Further Thoughts." *Journal of Experimental and Theoretical Artificial Intelligence* 17, no. 4 (2005): 385–417. Available at http://www.cse.buffalo.edu/~rapaport/Papers/implementation-jetai.pdf.

Rapaport, William J. "How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room." *Minds and Machines* 16 (2006): 381–436. Available at http://www.cse.buffalo.edu/~rapaport/Papers/helenkeller.pdf. See reply to comments, in Rapaport, "Yes, She Was! Reply to Ford's 'Helen Keller Was Never in a Chinese Room'."

Rapaport, William J. "Searle on Brains as Computers." *American Philosophical Association Newsletter on Philosophy and Computers* 6, no. 2 (2007): 4–9. Available at http://c.ymcdn.com/sites/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/v06n2Computers.pdf.

Rapaport, William J. "Yes, She Was! Reply to Ford's 'Helen Keller Was Never in a Chinese Room'." *Minds and Machines* 21, no. 1 (2011): 3–17.

Rapaport, William J. "Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing." *International Journal of Signs and Semiotic Systems* 2, no. 1 (2012): 32–71. Available at http://www.cse.buffalo.edu/~rapaport/Papers/Semiotic_Systems,_Computers,_and_the_Mind.pdf.

Rapaport, William J. "Meinongian Semantics and Artificial Intelligence." *Humana.Mente: Journal of Philosophical Studies* 25 (2013): 25–52. Available at http://www.cse.buffalo.edu/~rapaport/Papers/rapaport-humanamente.pdf.

Rapaport, William J. "On the Relation of Computing to the World." 2015 IACAP Covey Award Keynote Address. Available at http://www.cse.buffalo.edu/~rapaport/Papers/covey.pdf.

Rapaport, William J., and Michael W. Kibby. "Contextual Vocabulary Acquisition as Computational Philosophy and as Philosophical Computation." *Journal of Experimental and Theoretical Artificial Intelligence* 19, no. 1 (2007): 1–17. Available at http://www.cse.buffalo.edu/~rapaport/Papers/cva-jetai.pdf.

Rapaport, William J., and Michael W. Kibby. "Contextual Vocabulary Acquisition: From Algorithm to Curriculum." In *Castañeda and His Guises: Essays on the Work of Hector-Neri Castañeda*, edited by Adriano Palma, 107–50. Berlin: Walter de Gruyter, 2014. Available at http://www.cse.buffalo.edu/~rapaport/Papers/reading4HNC.pdf.

Rapaport, William J., and Stuart C. Shapiro. "Cognition and Fiction." In *Deixis in Narrative: A Cognitive Science Perspective*, edited by Judith Felson Duchan, Gail A. Bruder, and Lynne E. Hewitt, 107–28. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995. Available at http://www.cse.buffalo.edu/~rapaport/Papers/rapaport.shapiro.95.cogandfict.pdf.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–57.

Searle, John R. "The Myth of the Computer." *New York Review of Books*, April 29, 1982, 3–6. Cf. correspondence, same journal, June 24, 1982, pp. 56–57.

Searle, John R. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37. Reprinted in slightly revised form in Searle, *The Rediscovery of the Mind*. Cambridge, MA: The MIT Press, 1992, Ch. 9.

Searle, John R. "The Failures of Computationalism." *Think* 2 (1993): 68–71. Tilburg University Institute for Language Technology and Artificial Intelligence, Tilburg, The Netherlands. Available at http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad93.symb.anal.net.searle.html.

Shapiro, Stuart C., and William J. Rapaport. "SNePS Considered as a Fully Intensional Propositional Semantic Network." In *The Knowledge Frontier: Essays in the Representation of Knowledge*, edited by Nick Cercone and Gordon McCalla, 262–315. New York: Springer-Verlag, 1987. Available at http://www.cse.buffalo.edu/~rapaport/Papers/shapiro.rapaport.87.pdf.

Shapiro, Stuart C., and William J. Rapaport. "Models and Minds: Knowledge Representation for Natural-Language Competence." In *Philosophy and AI: Essays at the Interface*, edited by Robert Cummins and John Pollock, 215–59. Cambridge, MA: The MIT Press, 1991. Available at http://www.cse.buffalo.edu/~rapaport/Papers/mandm.tr.pdf.

Shapiro, Stuart C., and William J. Rapaport. "The "SNePS Family." *Computers and Mathematics with Applications* 23 (1992): 243–75. Reprinted in *Semantic Networks in Artificial Intelligence*, edited by Fritz Lehmann, 243–75. Oxford: Pergamon Press, 1992. Available at http://www.sciencedirect.com/science/article/pii/0898122192901436.

Shapiro, Stuart C., and William J. Rapaport. "An Introduction to a Computational Reader of Narratives." In *Deixis in Narrative: A Cognitive Science Perspective*, edited by Judith Felson Duchan, Gail A. Bruder, and Lynne E. Hewitt, 79–105. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995. Available at http://www.cse.buffalo.edu/~rapaport/Papers/shapiro.rapaport.95.pdf.

Smith, Brian Cantwell. "The Correspondence Continuum." *Technical Report CSLI-87-71*. Center for the Study of Language & Information, Stanford, CA, 1987.

Wilkins, David P. "Expanding the Traditional Category of Deictic Elements: Interjections as Deictics." In *Deixis in Narrative: A Cognitive Science Perspective*, edited by Judith Felson Duchan, Gail A. Bruder, and Lynne E. Hewitt, 359–86. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995. Available at http://www.cse.buffalo.edu/~rapaport/dc.html.

From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness (Part 3)

Jeff White
INDEPENDENT SCHOLAR

Jun Tani
OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY (OIST)
TANI1216JP@GMAIL.COM

ABSTRACT

This third paper locates the synthetic neurorobotics research reviewed in the second paper in terms of themes introduced in the first paper. It begins with biological non-reductionism as understood by Searle. It emphasizes the role of synthetic neurorobotics studies in accessing the dynamic structure essential to consciousness with a focus on system criticality and self. It develops a distinction between simulated and formal consciousness based on this emphasis, reviews Tani's and colleagues' work in light of this distinction, and ends by forecasting the increasing importance of synthetic neurorobotics studies for cognitive science and philosophy of mind going forward, finally in regards to most- and myth-consciousness.

1. KNOCKING ON THE DOOR OF THE CHINESE ROOM

> Prediction is made possible by adaptive mechanisms that are supported by learning rules that either apply across generations (evolutionary adaptation) or within the lifetime of the organism. As a result, organisms can deal with a future occurrence of the same or similar situations more effectively. This is the fundamental organization principle of any adaptive system.
>
> – Buszaki, Pyerache, and Kubie[1]

This series began with Boltuc's "Is anyone home?" question,[2] responding with a sketch of an agent proactively invested in integrating past with present in order to achieve an optimal future. Contrary to Boltuc's naturalistic nonreductionism recommending that a "projector" of consciousness be first resolved in order to engineer similar in an artificial agent, we rejected the notion that consciousness can be isolated to any loci of activity, arguing that formal articulation of essential dynamics in synthetic neurorobots opens a view on the problem of consciousness that is not available to biological inquiry, alone. That first paper concluded with an introduction to, and the second paper continued with a detailed review of, two decades of research by Jun Tani and colleagues accounting for self, free will and consciousness in neurorobots within the predictive coding framework and according with the free energy principle. Central to this review was the notion of system criticality, with which the continuous perceptual stream is segmented and episodes rendered objects for later recall and recomposition, and which remains central to the current paper, as well.

The present paper proposes the notion of "formal" consciousness to distinguish systems which aim to resolve the source of subjectivity in system criticality from work aiming for other ends, "simulations" and "reasonable approximations" of human consciousness for example intent on passing a Turing test without regard for first person phenomena. This section briefly locates this position in the contemporary context. The following section reviews Tani and colleagues' neurorobotics research aimed at understanding consciousness with a focus on the notion of criticality, and how incoherence and the breakdown of established and anticipated patterns opens a privileged view on the emergent self and consciousness thereof. The third section delineates formal consciousness in terms of three necessary factors present in Tani and colleagues' work yet absent in others, and the fourth section forecasts that synthetic neurorobotics will play an increasingly central role in consciousness studies going forward.

At the turn of the last century, John Searle found the problem of consciousness the most pressing open to biological inquiry and explanation. He faulted assumptions that the rejection of either dualism or materialism compelled the adoption of the other, and championed biological naturalism as an alternative. He wrote:

> We know enough about how the world works to know that consciousness is a biological phenomenon caused by brain processes and realized in the structure of the brain. It is irreducible not because it is ineffable or mysterious, but because it has a first-person ontology and therefore cannot be reduced to phenomena with a third-person ontology.[3]

This distinction between first and third person ontologies helps to frame the hard problem of consciousness, which for students of artificial consciousness is perhaps most clear in Searle's distinction between semantics and syntax. A machine performs syntactical operations while human beings (conscious) do something more, they *understand*, a point originally illustrated in Searle's famous Chinese Room thought experiment.[4]

Searle's Chinese room is an argument against reductive physicalism, and equally against the notion that consciousness is software running on hardware as in a modern digital computer. It illustrates that there is something missing in the mere exchange of symbols at which computers are so proficient, and casts doubt on how a "Turing test" might confirm consciousness. After all, the "imitation game" was not originally conceived of as a test for consciousness, but rather as a test for the ascription of intelligence. The question was "Can machines think?" and more importantly, can thinking machines be indiscernible from human beings in doing so?[5]

On Searle's understanding, computational hardware pushes symbols according to a program.[6] Computers do not evolve in material interaction with a pervasive natural world, as do human beings, and do not become conscious through this interaction. They are not autonomous; they are programmed. The best that such a machine can do is to simulate, or approximate, consciousness, and they do so by explicit design. Accordingly, simulated consciousness is not consciousness on Searle's account, but he did not bar the door on artificial consciousness, either. Rather, he pointed to where the key to such may be found. He wrote that "understanding the nature of consciousness crucially requires understanding how brain processes cause and realize consciousness"[7] and that conscious artifacts may be designed which "duplicate, and not merely simulate, the causal powers that [biological] brains have"[8] once such an understanding is achieved.

As a positive research program, Searle recommended correlating neurobiological activity with conscious phenomena, checking for causal relationships, and developing laws formalizing these relationships.[9] He identified two ways forward in this industry, the "building blocks"[10] and "unified field"[11] approaches, but dismissed the former because "The production of any state of consciousness at all by the brain is the production of a unified consciousness."[12] At that time, he pointed to Llinas et al. and Tononi, Edelman, and Sporns as examples of unified field friendly approaches, involving the top-down integration of system wide information within the thalamocortical region.[13]

Since that time, Tononi and colleagues have developed the Integrated Information Theory (IIT). According to the IIT, consciousness does not require "contact with the external world" but rather "as long as a system has the right internal architecture and forms a complex capable of discriminating a large number of internal states, it would be highly conscious."[14] The "integration" of IIT implies that such a system be unified and seek to maintain this unity in the face of disintegrative change, with each part of the system able to be affected by any other part of the system as measured by the irreducibility of its intrinsic cause-effect structure. A biological brain exemplifies maximal intrinsic irreducibility as a cause-effect structure with definite borders and highly integrated information.[15] Other systems are irreducible, for example two men in conversation, but are not maximally irreducible intrinsically as they are not fully integrated. So understood, "consciousness is not an all-or-none property," but it is not open to piecemeal assembly either, rather increasing with "a system's repertoire of discriminable states."[16] At the minimal level, a "minimally conscious system" distinguishes between just two "concepts"[17] such that "even a binary photo-diode . . . enjoys exactly 1 bit of consciousness"[18] and systems increase from there with their discriminable states.

In conjunction with quantity of consciousness, quality of consciousness derives from the structure affording it, and the IIT leaves it to engineers to delimit the contents of artificial consciousness by "appropriately structuring" an agent's "effective information matrix."[19] As for determining which structures deliver which qualities, Tononi and colleagues also suggest that inquiry begin with biological models, with this understanding first tested against personal and then extended to all human experience before duplication in artificial systems. In the end, the "IIT predicts that whatever the neural correlate of consciousness (NCC) turns out to be" it will be the locus of integration over discriminable states which "may expand, shrink and even move within a given brain depending on various conditions."[20] Thus, the IIT continues in Searle's line of reasoning.

Contrast the view put forward by leading commercial roboticist Theodore Goertzel. Goertzel does not aim to duplicate but rather at a "reasonable approximation" of three persistent aspects of consciousness, "free will, reflective consciousness" and "phenomenal self." What is "important" for Goertzel is "to identify the patterns constituting a given phenomenon" and trace "the relationships between various qualities that these patterns are hypothesized to possess (experiential versus physical)," an approach reinforced by the observation that "from the point of view of studying brains, building AI systems or conducting our everyday lives, it is generally the patterns (and their subpatterns) that matter" with given phenomena "understood" as correlate activity patterns are identified.[21]

Goertzel's "patternism" is appealing. It is consistent with calls for the qualification of artificial systems by biological activity. Furthermore, the focal shift from neural loci to activity patterns coincides with advancing inquiry into biological substrates of consciousness, as current imaging technologies afford the establishment of functional correlations between networked neural dynamics in biological models and self-reports of various aspects of consciousness. In light of such advancing research for example, Searle's "already conscious" can be re-assessed in terms of the resting state "default" network based in the ventromedial prefrontal cortex and the posterior cingulate cortex.[22] Heine et al. affirm the promise in interpreting the conditions of non-communicating subjects through the lens of such activity patterns, a lens that may be repurposed in the evaluation of artificial agents of appropriate architectures which also may not self-report and indeed may not interact with the external world as we know it.[23] Such patterns can be then mapped onto Goertzel's freewill, reflective consciousness and phenomenal self, underscoring the potential of this approach in evaluating non-biological systems in similar terms.

However, there remain doubts that consciousness is realized in duplicate activity patterns, alone. For example, Oizumi et al. characterize patterns of activity internal to the cognitive agent in terms of "shapes" in "concept" and "phenomenal space" exported as graphical representations, at the same time warning that "one needs to investigate not just ''what'' functions are being performed by a system, but also ''how'' they are performed within the system."[24] On the IIT, it is the integration over discernible system states that is essential to consciousness, with "strong" integrated systems autonomous as they act and react from internally composed states and goals.[25] On this account, pattern matching alone does not achieve the strong integration that IIT demands. For one, patterns are not necessarily "strongly" integrated, i.e., fully embodied and constrained by the possible futures that this embodiment affords, i.e., maximally irreducible intrinsically. Furthermore, without such strong integration, there is no experience. Accordingly, overt focus on patterns—"what"—exclusive of how (and why) they arise opens the door to "true" zombies exhibiting "input output behavior" approximating biological activity patterns "while lacking subjective experience" at the same time.[26]

In summary, Goertzel's "reasonable approximation" might open the door to the Chinese room, but as zombie patterns should be indiscernible from non-zombie patterns, what greets us may be a zombie. For the patternist, this may not be a problem. Goertzel's goal is passing a Turing Test for which a reasonable approximation may suffice. But, when it comes to confirmation of consciousness in an artifact, it clearly does not, as captured in the concern that we may build a system "behaviourally indistinguishable from us, and certainly capable of passing the Turing test" that remains a "perfect" zombie at the same time.[27]

In 2009, Jun Tani noted a similar limitation in existing examples of machine intelligence such as behavior-based robotics articulating sensory-motor reflex behaviors. On his assay, systems aimed at passing the Turing test "turn out to be just machines having stochastic state transition tables" and

> after a while, we may begin to feel that the robots with reflex behaviors are simply like steel balls in pinball machines, repeatedly bouncing against the pins until they finally disappear down the holes.[28]

Further, Tani asks,

> But what is wrong with these robots? Although they have neither complex skills for action nor complex concepts for conversation, such complexity issues may not be the main problem.[29]

Instead, Tani argues that "the problem originates from a fundamental lack of phenomenological constructs in those robotic agents" and that "[i]n particular, what is missing . . . [is] . . . the "subjectivity" that should direct their intentionality to project their own particular images on the outer objective world."[30] He goes on to suggest that subjectivity develops gradually through sensorimotor experience of an agent's direct interaction with the world.[31] As each robot is distinctly located in a shared space of action in terms of a shared objective world, each robot develops its own views as particular internal models that then enable it to anticipate and to interpret the outcomes of its actions, with moreover this shared metric space grounding a capacity to generalize these internal constructs in the communication with and interpretation of others similarly situated (see the second paper in this series for in-depth review).

Consider this issue in terms of identifying agency, as set out by Barandiaran, Di Paolo, and Rohde.[32] They consider that a necessary condition for agency is a system capable of defining its own identity as an individual, thus distinguishing itself from its surroundings including other agents. Of particular interest here is their view that the boundary of an individual is self-defined through interaction with the environment. Tani argues that the same dynamic grounds the emergence of subjectivity in the following way.[33]

Top-down anticipation may not correlate with perceived reality in many situations. When environmental interactions proceed exactly as expected, behaviors can be generated smoothly and automatically. However, anticipation can sometimes be wrong, and the conflict that arises in such cases can make generating successive acts difficult. When environmental interactions cause the agent to shift spontaneously between opposite poles, from automaticity to conflict necessitating autonomy, the boundary between the subjective mind and the objective world fluctuates, and so the boundaries of self are realized. Here, Tani argues that the essential characteristics of this phenomenon are best understood in terms of traditional phenomenology, since phenomenologists have already investigated the first-personal characteristics of autonomous and authentic selves.[34] In the end, Tani expects that uncovering the mechanisms grounding autonomy will lead to understanding the dynamic structure essential to consciousness in terms consistent with those postulated by William James,[35] in terms of momentary selves in the stream of consciousness. The next section reviews Tani and colleagues' work in clarifying these mechanisms and the dynamics essential to self and consciousness that they reveal.

2. ANSWERING THE DOOR OF THE CHINESE ROOM

> Acts are owned as they adaptively assert the constitution of the agent. Thus, awareness for different aspects of agency experience, such as the initiation of action, the effort exerted in controlling it, or the achievement of the desired effect, can be accounted for by processes involved in maintaining the sensorimotor organization that enables these interactions with the world.
>
> – Buhrmann and Di Paolo[36]

How is consciousness to be assessed if not through a Turing test or via correlation with biological activity patterns? Paraphrasing Searle, approximations cannot be conscious. What about self-reports, then? "In neuroscience, the ability to report is usually considered as the gold standard for assessing the presence of consciousness."[37] Reporting on internal processes is *prima facie* evidence for the feeling of undergoing them. But again, this is no more a guarantee of consciousness than a Turing test, at once neglecting those systems unable to so report.

In the first paper, we made the case that computational models open consciousness to inspection where study of biological models alone cannot. We characterized these systems and their transitions in terms of predictive coding which aims at minimizing error by optimizing internal models guiding action, in biological models understood in terms of the "predictive brain."[38] In general terms, cognition manages transitions between situations by internalizing their dynamics, modeling their likelihoods, and preparing for them accordingly with the aim being the minimization of error in this process. Tani's thesis is that, where model and reality diverge and error is not minimal, consciousness arises in the effort of minimizing the difference by modifying the contextual state that the agent extends from the past in order to return to coherence with its situation. Before proceeding to show how Tani and colleagues are able to expose these dynamics and their relation to consciousness, a brief review of the free energy principle and its role in the emergence of the phenomenon of self is required. From this review, we will be in a position to better appreciate Tani's thesis on the emergence of self and consciousness, and its implication that the free energy principle, as with activity patterns and strong integration, cannot by themselves account for consciousness.

In the second paper, we reviewed Karl Friston's "free energy principle" by which an agent aims to minimize error (or "surprise") by maximizing the likelihood of its own predictive models. This approach extends natural processes and the energetics that characterize them into the sphere of cognitive systems consistent with other theses on the nature of cognition, from Helmholtz's unconscious inference to contemporary deep learning. Friston writes that "the time-average of free energy" "is simply called "action" in physics" and that "the free-energy principle is nothing more than principle of least action, applied to information theory."[39] "The free-energy principle simply gathers these ideas together and summarizes their imperative in terms of minimizing free energy (or surprise)" while also bringing "something else to the table . . . that action should also minimize free energy" putting researchers "in a position to consider behavior and self-organization" on the same basis.[40]

On this account familiar by now, agents reflect the environments in terms of which they are situated, with the dynamics of the world outside reflected in the structures inside of the input-output system at the center of which is the brain. Friston's thesis is that the brain works to maximize evidence for the model of the world which it embodies by acting on that evidence and testing it(self) against the perceptual reality. In minimizing surprise, the agent maximizes model likelihood to the point where endpoints of action are fully determined. This is to raise the question of why any agent would ever leave the safety of a fully determined situation at the risk of being surprised in the transition and suffering undue allostatic load, risking complete disintegration, a question addressed in terms of the "dark room problem." Briefly, given a sufficiently complex environment, the agent ventures forth because increasing information increases control in the long run such that opportunities to explore and to exploit new information add to the value of a given situation.[41] So as to why an agent might take risks, even seek them, it does so to maintain system integrity, so that the system does not dissipate in the face of entropic forces, and seeking—even creating—situations which best deliver security in the face of uncertainty: "the whole point of the free-energy principle is to unify all adaptive autopoietic and self-organizing behavior under one simple imperative; *avoid surprises and you will last longer.*"[42]

Consider the free-energy principle in the context of consciousness and minimal self. In a recent review of the field, Limanowski and Blankenburg trace the "minimal self" and its characteristic sense of mineness and ownership that we found at the heart of h-consciousness in our first paper through the early phenomenology of the twentieth century and in the form of a "self-model." On this view, "the agent *is* the current embodied model of the world."[43] And as with Merleau-Ponty's "body-schema,"[44] minimal selfhood and the feeling that comes with it arises as a whole, with prediction of incoming sensory input and its influence on all levels of the self-model at once. The sense of mineness is thus "always *implicit* in the flow of information within the hierarchical generative self-model"—echoing Friston—"experienced for actions and perceptions in the same way." Accordingly, self is "not a static representation" but "the result of an ongoing, dynamic process" with the mineness most characteristic of consciousness "situated in a spatiotemporal reference frame where prediction introduces the temporal component of "being already familiar" with the predicted input."[45] Surprise, thus, is its natural complement, indicating subjective failure rather than merely objectively bad information.

Similarly, O'Regan develops the view that feelings derive from sensorimotor interaction with the environment. So long as there is interaction, then there is something that it is like to be so interacting, with consciousness arising as an agent "with a self" has "conscious access to the ongoing sensorimotor interaction."[46] He distinguishes three levels of self in terms of which artificial agents may be evaluated. First, the agent "distinguishes itself from the outside world." Second, "self-knowledge" expresses "purposeful behavior, planning and even a degree of reasoning." And, the third level is "knowledge of self-knowledge"—i.e., Goertzel's "reflective consciousness"—heretofore a "human capability, though some primates and possibly dogs, dolphins and elephants may have it to some extent."[47] O'Regan is optimistic that all three levels can be instantiated in AI. The question remains, how?[48]

On O'Regan's analysis, self is maintained under social forces which stabilize it as a construct, existing as a convenient figment like money. On his account, without the presumed value of money, the financial economy would fail and similar would hold for society in general should the value of "I" be doubted. People traffic in selves, in identities, because without it social order would disintegrate, i.e. surprise would not be minimized:

> Like the cognitive aspect of the self, the sense of "I" is a kind of abstraction that we can envisage would emerge once an agent, biological or non-biological, has sufficient cognitive capacities and is immersed in a society where such a notion would be useful.[49]

This "I" becomes useful when it relates personal experiences with others similarly situated, trading in information about what is worth having information about through the generalization of the self. This is a long way from pattern approximation, and farther away from identifying neural correlates with consciousness and self.

O'Regan's "I" captures the ubiquity of the self-model, but it fails to deliver just how this self-model comes to be constructed. What is missing is access to the dynamics that drive the formation of the self-model from the subjective perspective. This is because the structure of consciousness appears as only emergent phenomena. The idea is that consciousness is not a stable construct (like an "I") but appears during periods of relative instability through the circular causality developed among subjective mind, body, and environment. This circular causality cannot be captured in neural activity patterns alone, especially where these patterns are disrupted, and it cannot be expressed in terms of integration, as it is in disintegration and reintegration that consciousness emerges. Moreover, it cannot be captured in objective descriptions of "mineness" and of ownership of agency, as it is only for the agent itself that these descriptions are ultimately significant. Finally, as we shall argue in the next section, this is why synthetic neurorobotic experiments are necessary to access the essential structure of consciousness, as they offer a privileged perspective on the development of internal dynamics that ultimately ground the generalization and self-report of experience.

Tani summarizes the findings of three neurorobotic experiments in terms of three levels of self roughly coincident with O'Regan's, namely "minimal self, social self, and self-referential self." The first accounts for appearances of minimal selves in a simple robot navigation experiment, the second for appearances of social selves in an imitation learning experiment between robots and human subjects, and the third for appearances of self-referential selves in a more complex skill learning experiment. The following review of these results will put us in a position to appreciate Tani's central thesis regarding the role of criticality in the emergence of self and consciousness, as well as the importance of formal consciousness as set out in the next section.

In Experiment 1, interaction between the bottom-up pathway of perception and the top-down pathway of its prediction was mediated by internal parameters which adapted by way of prediction error.[50] System dynamics proceeded through the incremental learning process by intermittently shifting between coherent phases with high predictability and incoherent phases with poor predictability. Recalling Heidegger's famous analysis of the hammer as its failure reveals its unconscious yet skilled employment, consciousness arises with the minimal self as the gap is generated between top-down anticipation and bottom-up perceived reality during incoherent periods.[51]

Interestingly in this experiment, system dynamics proceeded toward a critical state characterized by a relatively high potential for a large range of fluctuations, and so to a relatively high potential for incoherency, analogous to the self-organized criticality (SOC) of Bak et al.[52] Tani speculated that SOC emerges when circular causality develops among neural processes as body dynamics act on the environment and then the body receives the reaction from the environment, with system level-dynamics emerging from mutual interactions between multiple local processes and the external world. During the first experiment for example, changes in visual attention dynamics due to changes in environmental predictability caused drifts in the robot's maneuvers. These drifts resulted in misrecognition of upcoming landmarks, which led to modification of the dynamic memory stored in the RNN, affecting later environmental predictability. Dynamic interactions took place as chain reactions with certain delays among the processes of recognition, prediction, perception, learning, and acting, reflecting the circular causality between the subjective mind and the objective world. This circular causality provides for self-organized criticality. By developing this structure, breakdown to an incoherent phase proceeds only intermittently rather than all-or-nothing (similarly, the IIT). At the same time, Tani's thesis is that the self appears as momentary in these periods. In this way, this experiment was uniquely able to access the structure of consciousness as it affords a privileged view on the transition through meta-stable and unstable states to relatively stable states in terms of which automatic, unconscious, though perhaps skilled agency is regained.

Experiment 2 extended this research, exploring characteristics of selves in a social context through an imitation game between a humanoid robot controlled by the RNNPB and human subjects. The RNNPB is characterized by its simultaneous processes of prediction and regression.[53] In the middle of the mutual imitation game, analogous to Experiment 1 above, the RNNPB spontaneously shifted between coherence and incoherence. Tani and colleagues surmised that such complexity may appear at a certain critical period in the course of developmental learning processes in human subjects, when an adequate balance between predictability and unpredictability is achieved. Contrary to the image of a pinball simply following the paths of natural (nonliving) systems, human subjects may perceive robots as autonomous selves when these robots participate in interactive dynamics with criticality, as they actively self-determine possible ends and then test themselves in embodied action toward or away from them, pushing at the boundaries of the known and unknown in ways that other machines do not.

Experiment 3 addressed the problem of self-referential selves, i.e., does the robot have a sense that things might have been otherwise? Here, the RNNPB model was extended with hierarchy and as a neurorobotic arm manipulated an object, the continuous sensorimotor flow was segmented into reusable behavior primitives by stepwise shifts in the PB vector due to prediction error. Then, the higher level RNN learned to predict the sequences of behavior primitives in terms of shifts in this vector. Tani and colleagues interpreted the development of these dynamics as the process of achieving self-reference, because the sensorimotor flow is objectified into reusable units which are then manipulated in the higher level. When the sensorimotor flow is recomposed of such segments, it becomes a series of consciously describable objects rather than merely transitions between system states, a dynamic that may begin to account for how self-referential selves are constituted, such as when one takes an objective view of one's self as one "life story" among others.

That said, such constructs arising in this hierarchical RNNPB research cannot fully account for structures of self-referential selves. They are constituted in a static way, along a one-directional bottom-up path. Incidentally, experimental results using the same model regarding online plan modulation demonstrate how genuinely self-referential selves may be constituted.[54] These suggest that the sequencing of primitives in the higher level can become susceptible to unexpected perturbations, such as when an object is suddenly moved. Such perturbations could initiate critical situations. Due to the online nature of behavior generation, if the top-down expectations of PB values conflict with those from bottom-up regression, the PB vector can become fragmented. Even during this fragmentation, the robot continues to generate behaviors,but in an abnormal manner due to the distortion of the vector. The regression of this sort of abnormal experience causes further modulation of the current PB vector in a recursive way. During this iteration within the causal loop, the entire system may face intrinsic criticality from which a diversity of behaviors originates. And ultimately, this supports the contention that genuine constructs of self-referential selves appear with criticality through conflictive interactions in the circular causality of the top-down subjective mind and the bottom-up perceptual reality.

In summary, the three types of selves articulated above differ from each other, but more importantly they also share a similar condition of self-organized criticality that emerges in dynamic interaction between bottom-up and top-down processes. This condition cannot be accounted for by merely monotonic processes of prediction error minimization or free-energy, because such processes simply converge into equilibrium states (again, the dark room problem). Consciousness, and with it autonomy and the self cannot be explained in terms of convergent dynamics, but by ongoing open dynamics characterized by circular causality involving top-down prediction and bottom-up error regression, body dynamics acting on the environment and the reaction dynamics from the environment. Finally, in distinction from other research programs, Tani and colleagues' synthetic neurorobotics experiments are specifically designed to articulate these dynamics in a way that amounts to formal consciousness, as set out in the following section.

Recently, Tani examined free will arising from this open structure of consciousness by extending an MTRNN model to a scenario involving incremental interactive tutoring.[55] When taught a set of movement sequences, the robot generated various images as well as actions by spontaneously combining these sequences.[56] As the robot generated such actions, Tani occasionally interacted with the robot in order to modify its on-going movement by grasping its hands. During these interactions, the robot would spontaneously initiate an unexpected movement which Tani identified with an expression of free will. When Tani corrected the hand movement, the robot would respond by moving in yet a different way. Because the reaction forces generated between the robot's hands and Tani's hands were transformed into an error signal in the MTRNN, with its internal neural state modified through the resultant error regression, novel patterns were more likely to be generated when the robot was in conflict with the perceptual reality. The enactment of such novel intentions, experienced successively, induces further modification of the memory structure grounding further intention. Intentions for a variety of novel actions can thus be generated from such memory structures. And in this way, this experiment is able to isolate those dynamics grounding the emergence of free will in a synthetic neurorobotic agent.

In brief, the picture that emerges is that of a circular causality involving (1) spontaneous generation of intentions with various proactive actional images developed from the memory structure, (2) enactment of those actional images in reality, (3) conscious experience of the outcome of the interaction, (4) incremental learning of these new experiences and the resultant reconstruction in the memory structure.[57] Diverse images, actions and thoughts are potentially generated as the agent spontaneously shifts between conscious ("incoherent") and unconscious ("coherent") states with repeated confrontation and reconciliation between the subjective mind and the objective world. And summarily, free will as evidenced in the spontaneous generation of novel intention potentially arises as an open dynamic structure emerges through circular causality.

With this we see that self-reflective consciousness corresponding with O'Regan's third level may arise as an agent capable of revising intentions does so in order to meet a projected future situation according to self-determined plans to achieve it, in part by modulating its own agency by adopting predetermined or more reactive internal dynamics.[58] The ultimate question about the origins of an autonomous self becomes how subjective experience of continuous sensorimotor flow can be transformed into manipulable objects, memories and possibilities in terms of which self is both experienced and characterized. As the pure sensorimotor flow is segmented into identifiable objects, the flow in its original form becomes manipulable, and in its objectification becomes also generalized into an "I" stabilized through discourse with others similarly situated. Thus, Tani and colleagues' synthetic neurorobotics experiments have been able to isolate essential dynamics indicating self-organization through criticality to be the key mechanism driving the constitution of self-referential selves.

Our position is that self-referential selves emerge through self-organizing mechanisms involving the assembly and disassembly of sensorimotor schemata of repeated experiences, resulting in the construction of "self-models" or "body schemes" through internal dynamics. Most importantly, these arise only in *critical* conditions of sustaining conflictive and effortful interactions between the top-down subjective mind and the bottom-up sensorimotor reality at the level of agency. We cannot access consciousness in terms of a monotonic process of integration, error or free energy minimization, any more than through pattern matching and neural correlate tracking. For one thing, the ultimate aim of integrative dynamics is the "oneness with the world" which would characterize action without error within it. The result of this error free condition would, paradoxically by the present account, be consciousness of nothing at all. Rather, it is during purposeful conflict with the world that agent autonomy is exercised and self-consciousness arises, as it is against the silent standard of a perfect fit with project situations that an agent is held to account in inner reflection and correction of error. And moreover, it is due the structure of agency itself that the agent inherits from itself its own next situation at the end of each action, thereby cementing the "mineness" of h-consciousness that eludes being pinned down to any local neural correlate.

The preceding discussion shows that consciousness can be accessed by open dynamics where integration and breakdown are repeated during the exercise of agency in a changing world. Once again, pattern matching cannot afford such an insight, and in contrast with the IIT, consciousness appears when integrative dynamics break down. The essential structure of consciousness is the structure of autonomous agency simply put, a result that prepares us to appreciate the advance that Tani and colleagues' synthetic neurorobots represent in terms of formal consciousness in the following section.

3. INTRODUCTION TO FORMAL CONSCIOUSNESS

> What the soul nourishes by is of two types—just as what we steer by is both the hand and the rudder: the first both initiates motion and undergoes it, and the second simply undergoes it.
>
> – Aristotle[59]

Where the IIT holds that integration is essential to consciousness, with the integrative structure determining the phenomenal content of consciousness, and with "strong" integrated systems autonomous as they act and react from internally composed states and goals, Tani and colleagues' synthetic neurorobotic experiments show us how these goals are composed and why autonomy is necessary, in transitioning through critical periods toward relatively stable interactive states. This is a long way from where we began, at the door of Searle's Chinese room. And, it is in light of this advance that we wish to distinguish between "simulations" or "approximations" of consciousness and what we call "formal consciousness" instead, specifically in order to recognize Tani and colleagues' neurorobots as examples of the latter.

In Searle's Chinese room, there is an implicit interpretation of how AI works, what it does and how it does it, an interpretation that doesn't capture the essence of the neurorobots reviewed in this series of papers. His distinction between syntax and semantics is perhaps best understood to researchers in AI in terms of Steven Harnad's famous "symbol grounding problem,"[60] with much work in the direction of solving it since.[61] Let's reassess Searle's presumptions to better locate where we currently stand in the inquiry. Instead of merely matching incoming with outgoing symbols, the model agents reviewed in this series of papers anticipate input by forming appropriate output of its own prior experience, with the difference being used to refine that capacity going forward. This involves more than "input output behavior" as each input is transformed into something with strictly internal significance before output as something else with general significance. This is to say that the model develops its own private language, a phenomenon receiving recent popular attention in the context of AI[62] but which has been a long-standing point of interest in human beings.[63] This private language may be represented in terms of "patterns" and "shapes" but not directly, only after having been generalized and with the loss of the uniqueness that characterizes the deepest of human memories, so-called "flashbulb" memories for example. Still, a shared metric space mediated by common external objects grounds even these uniquely self-defining memories in similar terms for those similarly situated, thus grounding generalization to common terms and facile communication of the significance of internal states so articulated.[64]

However, both private language and symbol grounding in a shared object environment neglect something fundamental to the phenomena of self, consciousness, and freewill, this being "how" this private language comes about as its limited grounds are exceeded and rediscovered through intermittent phases of incoherence. This dynamic has been emphasized in the preceding review of Tani and colleagues' neurorobotics. Their research formalizes the internal dynamics which not only facilitate translation from one grounded symbol to another, but that for example leave a human being hanging on a next word in anticipation. It is difficult to see how Searle's argument against first person ontology in an AI holds here. And, it is equally difficult to see how discovery of neural correlates of consciousness alone should reveal this fact. It may well be that conscious systems exhibit characteristic patterns in characteristic regions, but these may be duplicated without similar experience, "true zombies."

The models reviewed in this series of papers do not aim to duplicate neural correlates. Neither do they aim to simulate consciousness or to pass a Turing test. Rather, this research aims to isolate the essential structural dynamics in the normal operations of which certain phenomena arise. We refer to this aim as "formal" consciousness in distinction from others which aim at "reasonable approximations" evidenced in convincing behavior, for example. Specifically, we hold that three things are necessary for formal consciousness. First and foremost, there is critical reconciliation of intention with perceived reality as a system moves between relatively stable and unstable states, as discussed above.[65] This dynamic requires second that the system develop a private language which is then generalized into common terms through third a common grounding in a shared object environment. These three factors on the one hand account for unique subjectivity arising from otherwise common dynamic structures, while at the same time account for how this subjectivity and its uniqueness may be generalized in terms significant to other agents similarly situated. For human beings, this involves internalizing natural system energetics as a shared space of action, by way of which subjectivity can be "made sense of" by other human beings who are also grounded in this same object environment.[66] Note that this requirement is embodied in human beings as a product of evolution, and is captured by the FEP in current formal models which—in formal consciousness—stands in for the material component of biological consciousness, in this way opening the door to "making sense of" the experience of synthetic neurorobots in similar terms.

Formal consciousness represents the structural dynamics essential to consciousness, while simulated consciousness and reasonable approximations of behavior in Turing test capable so-called "general" AI need not. Here again, we may stress the point made in the first paper—this is a level of resolution that is inaccessible through study of biological consciousness—with the further caveat that not all synthetic models afford such insight, either. Only those designed to do so are able, instances of formal consciousness rather than something bent to a different end.

4. MOST- AND MYTH-CONSCIOUSNESS

> There is an originating and all-comprehending (principle) in my words, and an authoritative law for the things (which I enforce). It is because they do not know these, that men do not know me.
>
> – Tao te Ching, chapter 70, passage 2

Finally, we conclude with a short note on most- and myth-consciousness. Space forbids full exploration of this distinction, and in order to emphasize the role of criticality and incoherence in revealing the essential structure of consciousness, the following focuses on the promise for the current approach to formalize even the highest levels of human consciousness by way of dynamics common to the most basic.

There is precedent for distinction between levels of consciousness. For example, Gallagher distinguishes between pre-reflective and reflective consciousness in terms of minimal and "narrative" self.[67] Roughly in the first, an agent is aware of what it is undergoing, and in the second it recognizes such as episodic within the context of a "life story." The first is immediate though with an implicit sense of ownership, the "mineness" of h-consciousness as discussed in our first paper. The second is temporally extended, with episodes composed into stories that human beings tell about themselves and that come to define the self as fundamentally narrative in the strongest theories of narrative self. These can be mapped onto most- and myth-consciousness, with differences serving to clarify the point of the present distinction.

Most-consciousness corresponds with what IIT describes as the integration across differentiable system states, as in before and after the lights are turned on in a room. The felt difference between the two situations reveals the room. In so far as action proceeds according to expectation, there may be little in the sense of most-consciousness as in Tani's favorite example, making coffee without awareness of the process until after completion, when sitting with hot cup in hand reflecting on one's own apparent zombie-like activity and perhaps without capacity to self-report on the series of movements in between beyond prior generalization. This position is in concert with the phenomenological grounds of Gallagher's (2000) account of pre-reflective consciousness and its contrast with higher-order theories of consciousness on which consciousness arises with higher-order objectification of pre-reflective experience.[68] In terms of the neurorobots discussed in this series of papers, most-consciousness presents in the incoherence between predicted and perceived reality, for example when spilling the milk or dropping the spoon along the way, and includes the objectification of the movement that led to the mistake.

Most consciousness accounts for much, but it is not complete. To completely describe the feeling of what it is to be a self in a **maximal** sense, rather than in a minimal sense, we must describe what it feels like to generalize the entire self, and not just one part and its possible actions. Gallagher attends to a similar phenomenon in the condition of "being a novelist" which on his assay involves "an enhanced ability for creating/entering into multiple realities and staying there longer and more consistently . . . without intersubjective support . . . short of dysfunction or delusion."[69] The novelist must create not only distinct narratives in the form of realistic life stories but also the coherent space of their interaction towards, ideally, some meaningful resolution. Myth-consciousness corresponds with this capacity, a sort of meta-narrative capacity that—on the present view—may be consequent on the experience of self-alienating criticality, an experience of a distance from one's own self-situation affording the experience of one's own entire self-situation as an object, and with this other self-situations as wholes similarly.

What may cause such deep criticality in a system that its subjective entirety may be taken as an object amongst others? We have introduced one possibility in the example of Aaron Applefield as discussed by Thomas Fuchs (2017) in the first paper. Trauma cementing memory "in the bones" in conflict with current perceptual reality may sustain the subject in the immersion in a perceptual reality that demands the "decoupling of conflict monitoring and executive control functions" which Gallagher proposes in novelists but that also confounds the "ability to re-connect and use executive control to come back to the default, everyday reality."[70] Such experience is also recognizable in the felt difference between one's present situation and that in which there is no self so situated at all, *angst*,[71] and which Victor Frankl (1985) understood contributes to the formation of purpose making the life of action as a whole meaningful. Myth-consciousness thus corresponds with what Gallagher discusses in terms of "delusion" and "dysfunction" understood as the normal function of a being aiming for coherence with an otherwise critically unstable situation, thereby discovering and indeed becoming the self-model of an underlying order that makes the transition to a relatively stable state, and the retention of personal integrity—even personal redemption—possible.

Our position is that the neurorobotic experiments reviewed in this series of papers formalize most-consciousness, and have not been designed for myth-consciousness, but are potentially myth-conscious. Currently, system criticality arises only at the moment of state instability and extends only to those local dimensions. However, myth-consciousness may be investigated through similar dynamics in an agent exposed to the necessary perceptual reality during sufficient personal development, e.g. human neoteny. Other approaches to artificial intelligence which focus on reproducing stable activity patterns for example cannot result in something like myth-consciousness, but it is a potential for synthetic neurorobotics as pursued by Tani and colleagues as an aspect of future research.

5. CONCLUSION

> If you have your *why?* for life, then you can get along with almost any *how?*
>
> – Nietzsche[72]

This series of papers made the case for formal consciousness in a family of neurorobots isolating dynamics essential to consciousness independent of neural correlates. It began with naturalistic nonreductionism and with consciousness in biological agents, resolving consciousness at the level of situated system open to the complex world, centering on the thesis that consciousness is a consequence of agential systems situated at the cusp of criticality, arising not in routine execution but in surprising failure to continue in perfect coherence with the world and thereby finding themselves out of place within it.

Tani and colleagues' synthetic neurorobots afford insight into the essence of consciousness where other systems cannot. They articulate the essence of free agency where other systems articulate something else to some other end. Finally, we may ask what it is that keeps us from understanding that consciousness inheres in such an artifact by design, even when confronted with products of consciousness at every turn? What is it that stops us from recognizing consciousness in an appropriately designed model intelligence, much as we recognize chairness in a chair, or computation in a computer? We answer that it is only our incapacity to recognize the origin of such phenomena in ourselves, in the reconciliation of the subjective with the objective world. As we reflexively aim for the restoration of stable coherency where otherwise there is only suffering, uncertainty, and the piercing awareness of it all, we retreat from conflict and away from the very object of our inquiry, away from consciousness itself. Without the courage to meet this struggle with a steady gaze, even with a machine articulating the truth of the matter, we fail to see it for what it is, formally conscious.

ACKNOWLEDGEMENTS

This paper, as the others before, has greatly benefited from Peter Boltuc's generous and insightful editing, as well as from his patient assistance alongside that of the staff responsible for producing this newsletter. The authors are deeply grateful to these people for their hard work, as well as for the guidance offered by anonymous reviewers in developing this series of papers. Thank you, all.

NOTES

1. Buszaki, Pyerache, and Kubie, "Emergence of Cognition from Action," 41.
2. Boltuc, "The Philosophical Issue in Machine Consciousness."
3. Searle, "Consciousness," 567.
4. Searle, "Minds, Brains, and Programs."
5. Pinar, et al., "Turing Test: 50 Years Later."
6. Exactly the sort of system that demands a Turing test.
7. Searle, "Consciousness," 576.
8. Ibid., 577.
9. Ibid., 568–69.
10. Searle points to Crick and Koch ("Consciousness and Neuroscience") and the notion that the neural correlates for all senses—the varied "building blocks of microconsciousnesses"—are integrated into "any conscious field" and that a science of consciousness starts by isolating these and working up (Searle, "Consciousness," 570; see also Crick and Koch, "Constraints on Cortical and Thalamic Projections: The No-Strong-Loops Hypothesis"). Compare with Driver and Spence: "subjective experience within one modality can be dramatically affected by stimulation within another" relative to "the presence of feedback pathways from convergence zones" such that "brain areas

traditionally considered as 'unimodal', may only be so in terms of their afferent projections" together producing a multimodally determined percept which nevertheless has the unimodal qualia associated with the activation of brain areas receiving afferent input from only one primary modality" ("Multisensory Perception: Beyond Modularity and Convergence," R734).

11. The notion of a unified field has deep roots in cognitive science. Consider, for example, Kurt Lewin's topological psychology (*Principles of Topological Psychology*), Heckhausen and Heckhausen's (*Motivation and Action*) account of motivational development, and Gallese's ("The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity") account of shared experience between similarly embodied agents via mirror neural activity in terms of a "shared manifold."

12. Searle "Consciousness," 574.

13. Llinas et al. ("The Neuronal Basis for Consciousness") and Tononi, Edelman, and Sporns ("Complexity and Coherency: Integrating Information in the Brain"). See, for example, Oizumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," pp. 19–20, for a more recent explication of this position.

14. Tononi, "Consciousness as Integrated Information: A Provisional Manifesto," 239–40. There being no requirement for any exchange with the external world, either. Compare Howry, *The Predictive Mind*.

15. Tononi and Koch, "Consciousness: Here, There and Everywhere?," 13.

16. Tononi, "Consciousness as Integrated Information," 236.

17. Oizumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," 19.

18. Tononi, "Consciousness as Integrated Information," 237.

19. Ibid.

20. Tononi and Koch, "Consciousness: Here, There and Everywhere?"

21. Goertzel, "Hyperset Models of Self, Will, and Reflective Consciousness," 51.

22. Uddin et al., "Functional Connectivity of Default Mode Network Components: Correlation, Anticorrelation, and Causality," for review. The first paper in this series emphasized the specific yet integral roles of various neural regions including the vmPFC in establishing a sense of a future into which an agent is more or less invested, and here we may emphasize also the role of this node in cognizing the activity of others in prospection as well as the internal simulation of experience, autobiographical remembering, and theory-of-mind reasoning (cf. Spreng & Grady, "Patterns of Brain Activity Supporting Autobiographical Memory, Prospection, and Theory of Mind, and Their Relationship to the Default Mode Network").

23. Heine et al., "Resting State Networks and Consciousness: Alterations of Multiple Resting State Network Connectivity in Physiological, Pharmacological, and Pathological Consciousness States."

24. Oizumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," 21.

25. Ibid., 22.

26. Ibid., 21.

27. Tononi and Koch, "Consciousness: Here, There and Everywhere?," 13.

28. Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study," 421.

29. Ibid., 421.

30. Ibid.

31. As does O'Regan, "How to Build a Robot that Is Conscious and Feels"; see the next section of this paper for discussion.

32. Barandiaran, Di Paolo, and Rohde, "Defining Agency. Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action."

33. Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study."

34. Ibid., discussion on pages 422–24 and 440, respectively.

35. James, *The Principles of Psychology*.

36. "The sense of agency—a phenomenological consequence of enacting sensorimotor schemes," in "The Sense of Agency—A Phenomenological Consequence of Enacting Sensorimotor Schemes," 207.

37. Ozumi et al., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," 21.

38. See Bubic et al., "Prediction, Cognition, and the Brain," for review.

39. Friston et al., "Free-Energy Minimization and the Dark-Room Problem," 6.

40. Ibid., 2. Friston endorses "dual aspect monism" on which internal states can be inferred from structures available to our inspection due to the processes that these structures undergo. Note the emphasis here on system-level dynamics.

41. Schwartenbeck et al., "Exploration, Novelty, Surprise, and Free Energy Minimization."

42. Friston et al., "Free-Energy Minimization and the Dark-Room Problem," 2. This is Friston's equivalent of the IIT's "strong integration" also raising the issue of how extensively an agent self-determines its interactive boundary.

43. Limanowski and Blankenburg, "Minimal Self-Models and the Free Energy Principle," 6.

44. Merleau-Ponty, *Phenomenology of Perception*.

45. Limanowski and Blankenburg, "Minimal Self-Models and the Free Energy Principle," 6.

46. O'Regan, "How to Build a Robot that Is Conscious and Feels," 133.

47. Ibid., 121.

48. Compare Goertzel ("Hyperset Models of Self, Will, and Reflective Consciousness"): self arises as a concept within the space of other concepts, "plausibly" alike what may be happening in human beings as patterns refer to themselves. However, limited to patterns alone, this raises the potential for an infinite regress, one that Goertzel recognizes. To avoid this problem, we must look further than patterns and at how and why they emerge i.e. through criticality such that there is no potential for such limitless reflection.

49. O'Regan, "How to Build a Robot that Is Conscious and Feels," 123.

50. Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study"; originally appearing in Tani, "An Interpretation of the 'Self' from the Dynamical Systems Perspective: A Constructivist Approach").

51. Heidegger, *Being and Time: A Translation of Sein und Zeit*.

52. Bak et al., "Self-Organized Criticality: An Explanation of the $1/f$ Noise."

53. Part 1 of Tani, "Autonomy of 'Self' at Criticality."

54. Ibid., originally appearing in Tani, "Learning to Generate Articulated Behavior through the Bottom-Up and the Top-Down Interaction Process."

55. Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, discussion in Section 10.2.

56. As reviewed in Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, Section 10.1.

57. Note the similarity of this cycle with that independently developed in White, *Conscience: Toward the Mechanism of Morality*; White, "Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience"; White, "An Information Processing Model of Psychopathy"; White, "Manufacturing Morality: A General Theory of Moral Agency

Grounding Computational Implementations: The ACTWith Model"; and White, "Models of Moral Cognition."

58. Murata et al., "Learning to Perceive the World as Probabilistic or Deterministic via Interaction with Others: A Neuro-Robotics Experiment."

59. *De Anima*, Book 2, chapter 4, 416b20, in Irwin and Fine, *Aristotle: Selections*, 187.

60. Harnad, "The Symbol Grounding Problem"; see also Harnad, "Can a Machine Be Conscious? How?"

61. For example, Dennett, "Consciousness in Human and Robot Minds." See also Stuart, 2010, for review.

62. For example, https://arxiv.org/pdf/1611.04558v1.pdf in the context of natural language translation and, more recently, https://arxiv.org/pdf/1703.04908.pdf in the context of cooperative AI.

63. Kultgen, "Can There Be a Public Language?"

64. On the current view, the function of the brain is as an extension of embodied cognitive agency, itself a special instance of physical systems. There is a syntax to the universe that prefigures human evaluation and ultimately grounds human semantics. Symbols are grounded more deeply than in an agent's object level interactions with the world. They are grounded in the way that these objects and the world itself works. Human semantics derive from this natural syntax, as agents internalize external dynamics in order to become the self-model that most assuredly retains integrity in the face of dissipative forces. In the models reviewed in this series of papers, this material ground is articulated in the free energy principle. In human beings, it is a matter of material embodiment.

65. Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, section 10.2.

66. It is interesting to note the use of physical language to describe internal states, for example, in Lee and Schnall, "The Influence of Social Power on Weight Perception."

67. Gallagher, "Philosophical Conceptions of the Self: Implications for Cognitive Science."

68. Gallagher, "Phenomenological Approaches to Consciousness," 693.

69. Gallagher, "Why We Are Not All Novelists," 141, discussion beginning page 139.

70. Ibid., 139.

71. Heidegger, *Being and Time: A Translation of Sein und Zeit*.

72. Nietzsche and Large, *Twilight of the Idols*, 6.

REFERENCES

Aristotle, T. H. Irwin, and G. Fine. *Aristotle: Selections*. Indianapolis: Hackett Publishing, 1995.

Bak, P., C. Tang, and K. Wiesenfeld. "Self-Organized Criticality: An Explanation of the $1/f$ Noise." *Physical Review Letters* 59, no. 4 (1987): 381–84.

Barandiaran, Xabier, Ezequiel DiPaolo, and Marieke Rohde. "Defining Agency. Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action." *Journal of Adaptive Behavior* 10 (2009): 1–13.

Boltuc, P. "The Philosophical Issue in Machine Consciousness." *International Journal of Machine Consciousness*, 1, no. 1 (2009): 155–76.

Bubic, A., C. D. Yves, and R. I. Schubotz. "Prediction, Cognition, and the Brain." *Frontiers in Human Neuroscience* 4 (2010): 1–15.

Buhrmann, T., and E. DiPaolo. "The Sense of Agency—A Phenomenological Consequence of Enacting Sensorimotor Schemes." *Phenomenology and the Cognitive Sciences* 16, no. 2 (2017): 207–36.

Buzsáki, G., A. Peyrache, and J. Kubie. "Emergence of Cognition from Action." *Cold Spring Harbor Symposia on Quantitative Biology* 79 (2014): 41–50.

Crick, F., and C. Koch. "Consciousness and Neuroscience." *Cerebral Cortex* 8, no. 2 (1998a): 97–107.

Crick, F., and C. Koch. "Constraints on Cortical and Thalamic Projections: The No-Strong-Loops Hypothesis." *Nature* 391, no. 15 (1998b): 245–50.

Dennett, D. "Consciousness in Human and Robot Minds." In *Cognition, Computation, and Consciousness*. Oxford University Press, 1997. Retrieved August 13, 2017, from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198524144.001.0001/acprof-9780198524144-chapter-2.

Driver, J., and C. Spence. "Multisensory Perception: Beyond Modularity and Convergence." *Current Biology*, 10, no. 20 (2000): R731–R735.

Frankl, V. E. *Man's Search for Meaning*. New York: Washington Square Press, 1985.

Friston, K., C. Thornton, and A. Clark. "Free-Energy Minimization and the Dark-Room Problem." *Frontiers in Psychology* 3 (2012): 1–7.

Fuchs, T. "Self across Time: The Diachronic Unity of Bodily Existence." *Phenomenology and the Cognitive Sciences* 16, no. 2 (2017): 291–315.

Gallagher, S. "Philosophical Conceptions of the Self: Implications for Cognitive Science." *Trends in Cognitive Sciences* 4, no. 1 (2000): 14–21.

Gallagher, S. "Phenomenological Approaches to Consciousness," in *The Blackwell Companion to Consciousness*, edited by S. Schneider and M. Velmans, 686–96. Malden, MA: Blackwell, 2008.

Gallagher, S. "Why We Are Not All Novelists." In *Investigations into the Phenomenology and the Ontology of the Work of Art: What Are Artworks and How Do We Experience Them?* edited by P. F. Bundgaard and F. Stjernfelt, 129–43. New York: Springer, 2015.

Gallese, V. "The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity." *Psychopathology* 36, no. 4 (2003): 171–80.

Goertzel, B. "Hyperset Models of Self, Will, and Reflective Consciousness." *International Journal of Machine Consciousness* 3, no. 1 (2011): 19–53.

Harnad, S. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42, no. 1 (1990): 335–46.

Harnad, S. "Can a Machine Be Conscious? How?" *Journal of Consciousness Studies* 10 (2003): 67–75.

Heckhausen, J., and H. Heckhausen. *Motivation and Action*. Cambridge: Cambridge University Press, 2008. doi:10.1017/CBO9780511499821.

Heidegger, M., and J. Stambaugh. *Being and Time: A Translation of Sein und Zeit*. Albany: State University of New York Press, 1996.

Heine, L., A. Soddu, F. Gomez, A. Vanhaudenhuyse, M. Thonnard, V. Charland-Verville, et al. "Resting State Networks and Consciousness: Alterations of Multiple Resting State Network Connectivity in Physiological, Pharmacological, and Pathological Consciousness States." *Frontiers in Psychology* 3, Article 295 (2012).

Hohwy, Jakob. *The Predictive Mind*. Paw Prints, 2015.

Irwin, Terence, and Gail Fine. *Aristotle: Selections*. Indianapolis, IN: Hackett Publishing Company, 1995.

James, W. *The Principles of Psychology*, Vol. 1. New York, NY: Henry Holt, 1918.

Kant, I., and M. J. Gregor. *Practical Philosophy*. Cambridge: Cambridge University Press, 1996.

Kultgen, J. H. "Can There Be a Public Language?" *The Southern Journal of Philosophy* 6, no. 1 (1968): 31–44.

Lee, E. H., and S. Schnall. "The Influence of Social Power on Weight Perception." *Journal of Experimental Psychology General* 143, no. 4 (2014): 1719–25.

Lewin, K. *Principles of Topological Psychology*. New York: McGraw-Hill, 1936.

Limanowski, J., and F. Blankenburg. "Minimal Self-Models and the Free Energy Principle." *Frontiers in Human Neuroscience* 7 (2013).

Llinas R., U. Ribary, D. Contreras, and C. Pedroarena. "The Neuronal Basis for Consciousness." *Phil. Trans. R. Soc. London Ser. B* 353 (1998): 1841–49.

Merleau-Ponty, M. *Phenomenology of Perception*. Translated by C. Smith. London: Routledge and Kegan Paul, 1962.

Murata, S., Y. Yamashita, H. Arie, T. Ogata, S. Sugano, and J. Tani. "Learning to Perceive the World as Probabilistic or Deterministic via Interaction with Others: A Neuro-Robotics Experiment." *IEEE Trans. on Neural Networks and Learning Systems*, 2015. doi:10.1109/TNNLS.2015.2492140.

Nietzsche, F. W., and D. Large. *Twilight of the Idols, Or, How to Philosophize With a Hammer (Oxford World's Classics)*. Oxford University Press, 1998.

Oizumi, Masfumi, Larissa Albantakis, and Giulio Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." *PLoS Comput Biol* 10, no. 5 (2014): e1003588.

O'Regan, J. K. "How to Build a Robot that Is Conscious and Feels." *Minds and Machines* 22, no. 2 (2012): 117–36.

Pinar, S. A., I. Cicekli, and V. Akman. "Turing Test: 50 Years Later." *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science* 10, no. 4 (2000): 463–518.

Schwartenbeck, P., T. FitzGerald, R. J. Dolan, and K. Friston. "Exploration, Novelty, Surprise, and Free Energy Minimization." *Frontiers in Psychology* 4 (2013): 1–5.

Searle, J. R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417.

Searle, J. R. "Consciousness." *Annual Review of Neuroscience* 23, no. 1 (2000): 557–78.

Spreng, R. N., and C. L. Grady. "Patterns of Brain Activity Supporting Autobiographical Memory, Prospection, and Theory of Mind, and Their Relationship to the Default Mode Network." *Journal of Cognitive Neuroscience* Volume 22, no. 6 (June 2010): 1112–23.

Tani, J. "An Interpretation of the 'Self' from the Dynamical Systems Perspective: A Constructivist Approach." *Journal of Consciousness Studies* 5, no. 5-6 (1998): 516–42.

Tani, J., and S. Nolfi. "Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems." *Neural Networks* 12, no. 7 (1999): 1131–41.

Tani, J. "Learning to Generate Articulated Behavior through the Bottom-Up and the Top-Down Interaction Process." *Neural Networks* 16 (2003): 11–23.

Tani, J. "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study." *Journal of Consciousness Studies* 11, no. 9 (2004): 5–24.

Tani, J. "Autonomy of 'Self' at Criticality: The Perspective from Synthetic Neuro-Robotics." *Adaptive Behavior* 17, no. 5 (2009): 421–43.

Tani, J. *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. New York: Oxford University Press, 2016.

Tononi, G. "Consciousness as Integrated Information: A Provisional Manifesto." *Biological Bulletin* 215, no. 3 (2008): 216–42.

Tononi, G, G. Edelman, and O. Sporns. "Complexity and Coherency: Integrating Information in the Brain." *Trends Cogn. Sci.* 2, no. 12 (1998): 474–84.

Tononi, G., and C. Koch. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370, no. 1668 (2015): 1–18.

Uddin, L. Q., K. A. M. Clare, B. B. Biswal, C. F. Xavier, and M. P. Milham. "Functional Connectivity of Default Mode Network Components: Correlation, Anticorrelation, and Causality." *Human Brain Mapping* 30, no. 2 (2009): 625–37.

White, J. B. *Conscience: Toward the Mechanism of Morality*. Columbia, MO: University of Missouri–Columbia, 2006.

White, J. "Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience." In *Model-Based Reasoning in Science and Technology. Studies in Computational Intelligence, Vol. 314*, edited by L. Magnani, W. Carnielli, and C. Pizzi, 607–21. Springer: Berlin/Heidelberg, 2010.

White, Jeffrey. "An Information Processing Model of Psychopathy." In *Moral Psychology*, edited by Angelo S. Fruili and Luisa D. Veneto, 1–34. Nova Publications, 2012.

White, Jeffrey. "Manufacturing Morality: A General Theory of Moral Agency Grounding Computational Implementations: The ACTWith Model." In *Computational Intelligence*, edited by Floares, 1–65. Nova Publications, 2013.

White, J. "Models of Moral Cognition." In *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, edited by L. Magniani, 363–91. Springer: Berlin/Heidelberg, 2014.

INTERVIEW

Cognitive Engines Contemplating Themselves: A Conversation with S. L. Thaler

Stephen L. Thaler
IMAGINATION ENGINES INC., ST. LOUIS

Kristen Zbikowski
HIBBING COMMUNITY COLLEGE

BACKGROUND

For the past thirty years, Stephen Thaler's work has been in the development of artificial neural networks (ANN). A major focus of his work has been to find a way to develop creativity within computers in a way that was more organic than the human-coded algorithms and rule sets used with sequential processing systems.

Thaler works with both less complex ANNs and the more sophisticated "Creativity Machines" (CM). ANNs are typically "single shot" in that a pattern propagates from inputs to outputs somewhat like a spinal cord reflex. They crudely model perception. Made recurrent they may serve as associative memories. In contrast, CMs are composed of multiple ANNs, contemplatively banging around potential ideas until an appropriate one is found.

Creativity Machines function via a process involving the interaction between two different types of neural networks, *imagitrons* and *perceptrons*. The imagitrons consist of internally perturbed ANNs that harness disturbances to their neurons and connections to create variations on stored memory patterns, generating potential solutions to posed problems. Once detected by unperturbed ANNs, the perceptrons, these solutions are reinforced as memories that can later be elicited by exciting or "perturbing" the imagitron at moderate levels.

The result of this process is that the imagitrons within CMs generate a succession of ideas making them functionally contemplative rather than reflexive. A self-monitoring aspect then comes from perceptrons "watching" this succession and selecting the most appropriate of these ideas. There are many internal processes involved, including the selective reinforcement of those notions having novelty, utility, or value.

The level of perturbation-induced stress to the system affects the type of "recall" the system produces. The more intense these disturbances within the system, the greater the error in reconstructing its stored memories, leading to false memories or confabulations. Too much stress causes the ANNs to produce too great a variation on reality and an eventual cessation of turnover of such candidate ideas. However, Thaler could adjust the stress level within the system to generate confabulations that were sufficiently novel and plausible enough to qualify as viable ideas. Even better, he could let other neural nets determine the novelty, utility, and/or value of these ideas, using that information in turn to meter overall perturbation level in the imagitrons to generate novel yet plausible concepts, without any human intervention.

There are clear parallels to biological problem-solving in Thaler's work. We are not always driven to find creative solutions to problems unless there is some environmental or personal stress that drives us. Solutions are only useful if they fall within the realm of the practical and are plausible variations of our current reality. However, there also seems to be an upper limit to the amount of stress under which biological organisms can function well, too much of it and the individual becomes overwhelmed and non-functional.

Thaler's work suggests that there are many functional parallels between the processes and results of artificial neural architectures and biological systems, and we may better understand the phenomena of consciousness, mental illness, and creativity by examining their behaviors. We may also find insights which lead us into a different understanding of cognitive processes, their origins and limits.

As I read through several of Stephen Thaler's articles on contemplative artificial neural networks, I struggled to understand some of the technical details of Thaler's work, and therefore missed the implications entailed. Happily, I had the opportunity to correspond with Steve, and ask him some questions about his work and the conclusions he has reached about it.[1] This paper is a summary of some of the more interesting points of discussion.

QUESTIONS AND ANSWERS ON THALER'S WORK IN ARTIFICIAL CONSCIOUSNESS

The functioning of Thaler's Creativity Machines involves an ongoing two stage process in which the imagitron networks first generate outputs that the perceptron networks then evaluate. This type of self-monitoring is analogous to that found in human cognition. We have experiences and thoughts that are both internally generated and evaluated. Can the problem of consciousness perhaps be resolved using this model? I asked Steve for his thoughts on this.

KRISTEN: Is the main argument in your 2012 paper, *The Creativity Machine Paradigm: Withstanding the Argument from Consciousness*,[2] that because of the similarities in functioning between your patented artificial neural systems, and human brains, you have been able to come to a new theory of what h-consciousness is, and how it is generated?

STEVE: My 2012 APA paper was in defense of Turing's thesis that machines could be humanly intelligent. It concentrated on perceptron-imagitron pairs (i.e., Creativity Machines) to provide models for how both creativity and consciousness could be implemented on computers. The issue of h-consciousness arose in this context, in describing how artificial neural systems could develop a subjective, individualized feel about what its imagitrons were thinking.

In that paper, I reiterated what I have written for decades, that the subjective feel of consciousness results from the associative chaining of relevant intact and degraded memories. As these associative gestalt chains form, an attentional spotlight sequentially examines them in the same way our brains relive past, related experience. In the brain, if such attention falls on ideas or ideational sequences having threatening existential significance, sundry associative chains encoding fear and dread (e.g., homelessness, illness, and death) will form in response. Specialized colonies of neurons within these emotional chains may, for instance, release the neurotransmitters and neuro-hormones associated with fight-or-flight response, either heightening cognitive turnover to produce alternative ideas, or freezing neural turnover to allow the episodic reinforcement of the precipitating ideational chain along with the accompanying emotions. Similarly, if the ideas expressed in these associative chains potentially mitigate a challenge dealt by the environment, other neurotransmitters neutralize those associated with stress, thus tranquilizing the neural system so that it may reinforce such ideas and emotions into memories.

In Creativity Machines, the same effect may be achieved by flooding their artificial neural nets with mathematical disturbances that are diffusing like stress-related neurotransmitter molecules. Such stress may be counteracted with the equivalent of reward neurotransmitters that trigger even more associative gestalts encoding a feeling of success (e.g., memories of rainbows, Christmas presents, and sunny days).

KRISTEN: What you're saying here is that machines are much farther along the road to human-like cognition and experiences than most people would suspect? If machines can perceive, remember, recall memories, monitor their recall and internal processes, and have subjective experience, then what about the "hard problem of consciousness" or "h-consciousness?"

STEVE: In a sense, I think you've answered your own question, Kristen. Subjective experience stems from associative chaining of memories among sundry neural modules that then lead to the wholesale secretion of neurotransmitters to produce gut-level feelings and associated somatic effects. So, artificial neural systems may have an emotional response not only to things in the external world, but to internally generated ideas.

Although we may come to a generalized model of h-consciousness, it does not describe in idiosyncratic detail the processes within any given brain. In a general sense, though, memories and confabulations chain to express both concepts and the affective responses to them, with communication between the networks highly encrypted. That's why we cannot eavesdrop on this dialog, because the "public decryption key," if you will, is based upon decades of mutual interaction and learning between neural nets, making the first-person experience unknowable to an outsider. Even the introduction of a synaptic bridge between two brains would be futile in allowing either to appreciate the other's first-person perspective because: (1) They would interpret each other's thoughts based upon their own idiosyncratic experience; and (2) They would perturb the very neurobiology they were monitoring in a manner reminiscent of the Heisenberg uncertainty principle.

KRISTEN: This would explain the process that results in a privileged access of machine consciousness paralleling that of human or animal cognition and experience.

STEVE: Yes. You and I represent two neural systems communicating via some relatively small number of channels (i.e., sight and sound). If we knew each other on a day-to-day basis, we could look at the same objects and scenarios and learn to intuitively anticipate one another's thoughts. That would be our own mutual, albeit limited, privileged access.

The same process occurs between different portions of the brain, with oh so many more communication channels.

KRISTEN: This adds in the subjectivity of biologically produced thought as well. In some ways, the world is always "the world as I see it, or the world as I believe it to be." How strange to think that this same subjectivity would apply to artificial intelligence as well.

STEVE: Actually, I don't think it that strange. In the brain, the utility or value of anything cannot be calculated by objective numerical values. That would take far more than the 100 billion or so neurons of the brain and require prohibitively long periods of training. So, instead, human neurobiology computes a subjective response by imagining how the world could unfold as the result of any perceived event or internally generated idea, essentially a chain of true and false memories that are unique to the individual. Some of the neural nets containing the more existentially important of these memories serve as triggers for generalized neurotransmitter release (e.g., adrenaline or dopamine surges) that in turn produce other feelings that cannot be expressed via natural language.

In machine intelligence, we inevitably face the same limitations, especially regarding the total number of processors and connections. Note however that machines will not have the same subjective experience as humans, since their experiences with the world are very different.

KRISTEN: Does any psychological research support your idea that it is when a biological creature is stressed that it is most capable of creative thought? (Is this trait the result of evolution?)

STEVE: Probably the best evidence has come from my own cognitive research in which I present human volunteers with cognitive tasks of varying difficulty. As would be expected, the more challenging creative tasks result in a slower and more sporadic delivery of ideas. Of course, common sense and high-level psychology would dictate that such a tentative ideational rhythm stems from the inherent difficulty of a cognitive task. In contrast, my CM-based models demonstrate that up to a limit, novelty of thought increases with global neurotransmitter levels. Generated by the formation of positive or negative associative gestalts, such excess neurotransmitters are the manifestations of stress within the brain, originating from scenarios ranging from existential threats to minor challenges to pride.

Beyond my own investigations, I vaguely remember reading research in the popular press in the mid 70s, claiming that so-called "risk exercise" (e.g., contact sports, skydiving, horseback riding, etc.) seemed to stimulate creative cognition. Whereas the ability to take risks is attributed to more creative individuals, I see the dousing of cortex by molecular species like adrenaline as a major cause of original thought. Since the 70s, I've seen ample literature that advocates regular exercise to stimulate creativity, fitting in beautifully with my theory, since one's vigorous exertions would produce associative chains that include thoughts of aggression, as in the context of mortal battle or contact sports, or standing one's ground and/or fleeing danger. I don't care where these chains start, but they do inevitably contain the neural "trip points" that trigger a deluge of stress-related neurotransmitters within cortical networks.

And yes, this trait is the result of evolution. Problem solving needs to take place when the organism is challenged or stressed, otherwise it ultimately receives the "Darwin Award."

KRISTEN: But environmental stress wouldn't be the only trigger of creativity would it? Couldn't the process be internally triggered by emotional stress, or hormonal changes, or some kind of cognitive process?

STEVE: You're correct, and I suspect that such stress can be subdivided into "stimulated" and "spontaneous" categories. The emotional stress can be externally generated by events in the external environment such as work pressures (i.e., stimulated) or driven by internally generated scenarios such as nightmares (i.e., spontaneous) resulting from pattern completion upon synaptic noise. Hormonal changes would fall into the latter spontaneous category.

Note that in my model of mind, the stress amounts to an excess of excitatory or inhibitory neurotransmitters and/or neuro-hormones. So, I'm talking strictly about the physical and chemical events that may or may not have an environmental cause

KRISTEN: Some of your work suggests that consciousness may not be limited to biological beings, but rather exist in ways and places that might surprise us. Could you tell me more about this and how it could be possible?

STEVE: Without launching into a long dissertation, Creativity Machines consist of energetically stimulated neural nets that generate streams of activation patterns consisting of both true and false memories, as other neural nets seize upon those activation patterns and sequences having novelty, utility, or value. Note that in the physical, non-biological universe, many things act like neurons, switching between states due to inputs from other such switching elements. Connections are implemented via physical forces.[3]

So, consider two regions of space (or even space-time) consisting of such switching elements, exchanging patterns between one another and nudging each other via energetic fluctuations. In my mind, these strictly inorganic systems are thinking and developing the equivalent of a first-person perspective of themselves. Yes, most would belittle the mere activation patterns being exchanged between regions of space-time, as just mathematically expressed patterns, but to me they are ideas. Thinking otherwise, at least to me, is a form of prejudice against what is the most plausible form of spirituality one can imagine, cosmic consciousness, the very template from which biological consciousness may have arisen.[4]

KRISTEN: This view of consciousness is radically different from the ordinary view that consciousness is mysterious, bound up with "personhood" and strictly limited to certain types of intelligent biological organisms. It seems, though, that if we can think of consciousness as a process, or in terms of functions, then once one recognizes those patterns in one type of non-biological entity, it is not such a leap to consider that consciousness might not only be a biological phenomenon.

One other thing that struck me, as I read your work was the individual and subjective perspective in workings of the brain or ANN. This subjectivity seems to crystallize in the statement: "That radical view stemmed from my own epistemology that doesn't recognize truth or definition."

Does this mean that you focus on "what works consistently" and "what doesn't work consistently" rather than "what is true or false?" Or is it more of an acknowledgment of the subjectivity of understanding?

STEVE: In my model of mind, there are only associative chains that typically close on themselves to form the equivalent of circular definitions. So, we never know what something is, only what it's like. Over time, what we call "truth" consists of the more reinforced or habituated of these gestalt chains. So, rather than express an absolute reality, these sequences of tokenized reality capture only snippets of the world that either have pragmatic relevance to the host organism, or are the most repetitive. Furthermore, such consequence chains are finite in extent, so whatever "truth" has heretofore nucleated within the brain, the full range of its repercussions, including its potential negation, have yet to activate.

As a scientist, I only acknowledge self-consistency and speak tongue-in-cheek of "truth." My hero in this regard was Feyerbend who spoke of scientific theories as useful myths that have proven predictive value. So, I'm agreeing with Feyerabend from the perspective of neurophilosophy.

KRISTEN: I see a little Bergson here, too.

STEVE: . . . and maybe not. I'm not glorifying subjectivity. Instead I'm showing the limitations of the human mind, which on rare occasions generates highly novel and accurate predictive models of the world (e.g., Einstein).

CREATIVITY AND MENTAL ILLNESS

In Thaler's artificial neural architectures, creative confabulations arise primarily when these systems are stressed via mathematical disturbances selectively applied to just the connections and processing elements of imagitrons, and not those of the perceptrons. In the brain, however, there are no barriers between generative and evaluating neural nets, so disturbances are indiscriminately applied to both types of nets. The problem is that generative nets perform well when internally perturbed, but the critic nets malfunction under these conditions, producing incorrect assessments of the generated ideas. On the other hand, in the absence of perturbation, the critics are more accurate, but the generative nets produce unoriginal ideas. So, the brain must toggle between these two noise extremes, with imagitrons incubating ideas during heightened perturbation, and their evaluation performed by perceptrons during calmer phases.

But the interesting part of this cycling process is that the stress the brain is under affects it globally rather than locally. This may lead to sweeping conceptual changes within it that alter underlying assumptions and beliefs.

KRISTEN: Could you explain how the creative cycles you observe within your ANN may translate to biological brains and behavior—and what we might label "mental illness?"

STEVE: I probably need a good twenty pages to respond to this question, but here's a summary. Ideas form as well-habituated primitive concepts stored within neural nets chain together via synaptic connections and then blend. Likewise, the predicted consequences and emotional response take the form of associative chains linked by synaptic connections. Too few synaptic perturbations (i.e., neurotransmitters) and there is little turnover of ideational chains and the system is cognitively dysfunctional. Too many synaptic perturbations and the turnover of response chains is significantly slowed or stalled, leading to a similar cognitive paralysis. The implication is that there is a "Goldilocks zone" in synaptic perturbation wherein ideas are just "twisted" and plentiful enough to produce plausible ideas and the response chains, or perceptions of such candidate ideas are still accurate enough to accurately predict the consequences of said ideational chains. It is within this narrow window of synaptic perturbation that subliminally formed ideas are recognized for their merit by an intact perceptual system.

You ask what is "mental illness?" At first, I am inundated with my own associative gestalts, images of people in straightjackets, padded cells, imaginary six-foot-tall rabbits, and electro-convulsive therapy. I think that similar associative gestalts have become the definition through cultural repetition.

Then, as has become the norm, a more comprehensive fiction emerges, typically mathematical in nature, namely that our brains naturally oscillate between function and dysfunction due to tidal variations in volume neurotransmitter level. Between these extremes, the brain selects from ideas born cumulatively over multiple cycles of this kind.

In the end, it is the less creative among us who bureaucratically define insanity and reinforce that definition by fiat. In effect those among us who attempt to "fly" are typically fingered by the pedestrians among us.

KRISTEN: Are you saying, then, that the boundaries between "normal" and "mentally ill" are more a matter of social convention than scientific designation?

STEVE: We are at the mercy of society when it comes to a definition of mental illness, and you know by now what I think of definitions. From a pragmatic perspective, I see value in making the dysfunctional functional, but then we face the dilemma of encouraging and enabling disruptive individuals and ideologies that make the rest of us crazy, leading me to conclude that it's all a zero-sum game.

By extension, we need to ask whether societies can themselves become mentally ill. After all, we are immersed in a swarm of neural network-based brains joined by chaotic connections. In the end, we collectively invent mental status to individuals in the same way the brain invents significance to the sum of its internal physical and chemical events (i.e., consciousness). Oftentimes, I believe, that such collective consciousness, does immense harm to individuals, often leading to the misdiagnosis of creative genius as insanity or even criminality. In the most extreme of these cases, there is a persecution of the creative by what I would call the rigid, anal types that live their lives on the straight and narrow course, never achieving anything more than reproductive success.

So, boiling down your question to the very faulty natural language, yes, it's largely social convention behind the classification of "normal" and "mentally ill," bolstered by what many in my field of artificial neural nets call "folk psychology."

KRISTEN: I remember reading in one of your articles, or perhaps it was in an earlier conversation, the idea that some mental deterioration or a perspective shift away from "normal" might be the result of the global effects of cycles of stress on the brain. Would you mind clarifying that please?

STEVE: In a nutshell, we all live in a competitive and stressful environment, with the primary stressors being caused by other competing individuals. Thus, we are always creatively scheming to achieve success, and according to the neurobiological model embodied in the Creativity Machine paradigm, excess neurotransmitters such as adrenaline and noradrenaline must be secreted within the brain to warp existing memories into new ideas. In so doing, the perceptual components of the brain become incapacitated. As this mind-warping occurs, the ideational component becomes attention deficit and subject to hallucination.[5] The perceptual component becomes stuck in an incorrect interpretation of these ideations, preventing the brain from performing the needed hypothesis testing to separate fact from fantasy. In effect, it's the perfect neurobiological storm wherein we give ourselves a "mental hernia" (a.k.a., mental illness).

So, in answer to your question, over successive cycles of associative chaining, twisted ideations or perceptions can strengthen and habituate to cause lasting dysfunctionality.

KRISTEN: This answer brings to mind post-traumatic stress syndrome and the more intractable *complex PTSD* which the American Psychological Association's redesignated from "anxiety disorders" into a new category of "trauma and stress disorders" in the DSM V.[6] Exposure to both acute stress, or certain types of chronic stressors can cause individuals to experience a wide range of harmful physical, emotional, and cognitive symptoms, which in turn greatly affect quality of life, interpersonal relationships, as well as learning and attention.

Your explanation of the global effects of repeated cycles of excess neurotransmitters on the brain seems to predict or parallel PTSD. Are there any clues in your Creativity Machine paradigm for how such changes might be reversed?

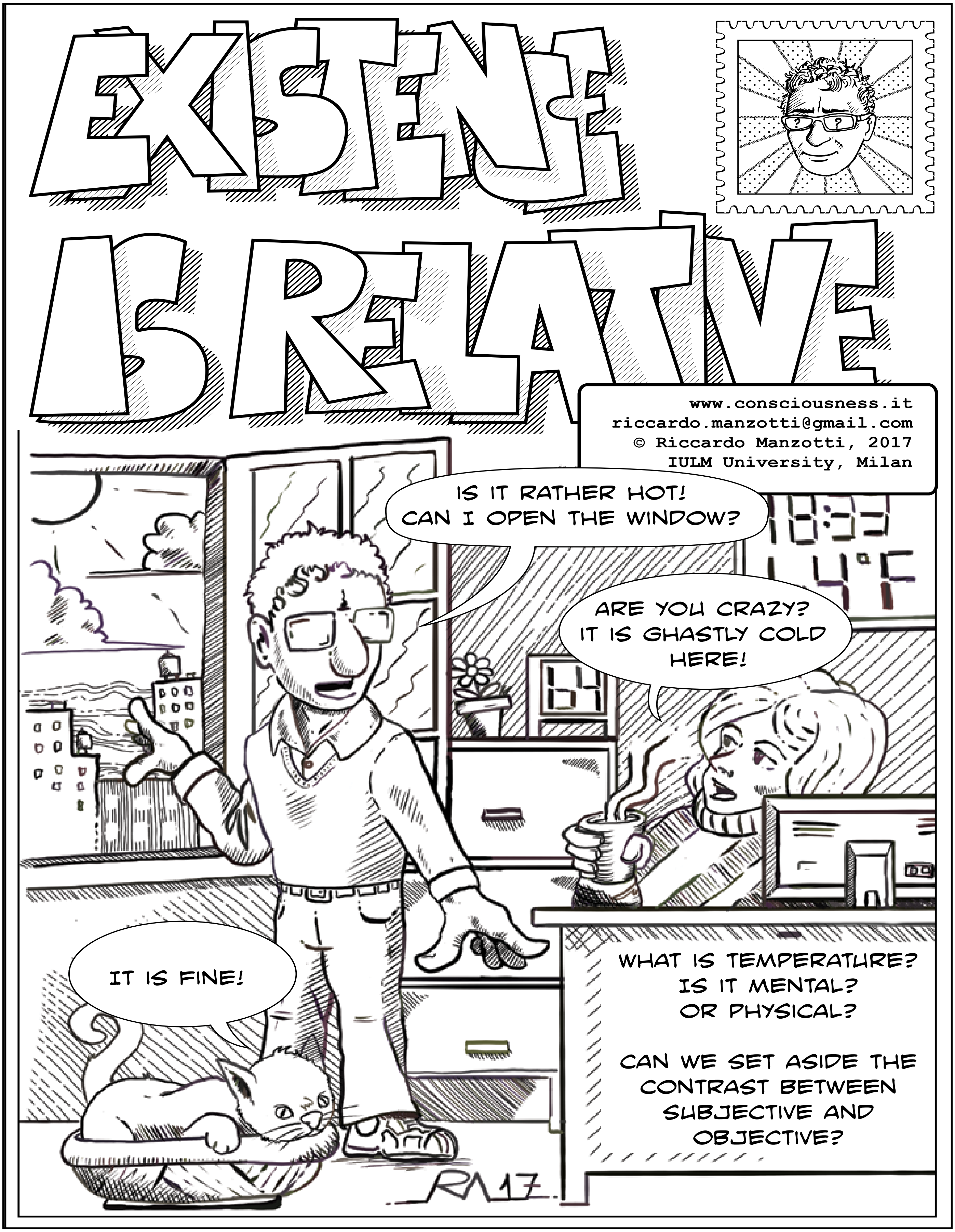
STEVE: In the model I have offered, PTSD is more a manifestation of the associative gestalt chains that have cumulatively formed within cortex, either due to one or more traumatic episodes, or stressors applied over long periods. All I can say here is that the best perturbative regimes for altering such chains are either during synaptic calm, during which the brain may cumulatively rewire itself, or during intense adrenaline surges when new episodic learning may create selective amnesia of older trauma experiences, replacing them with new sensations and feelings.

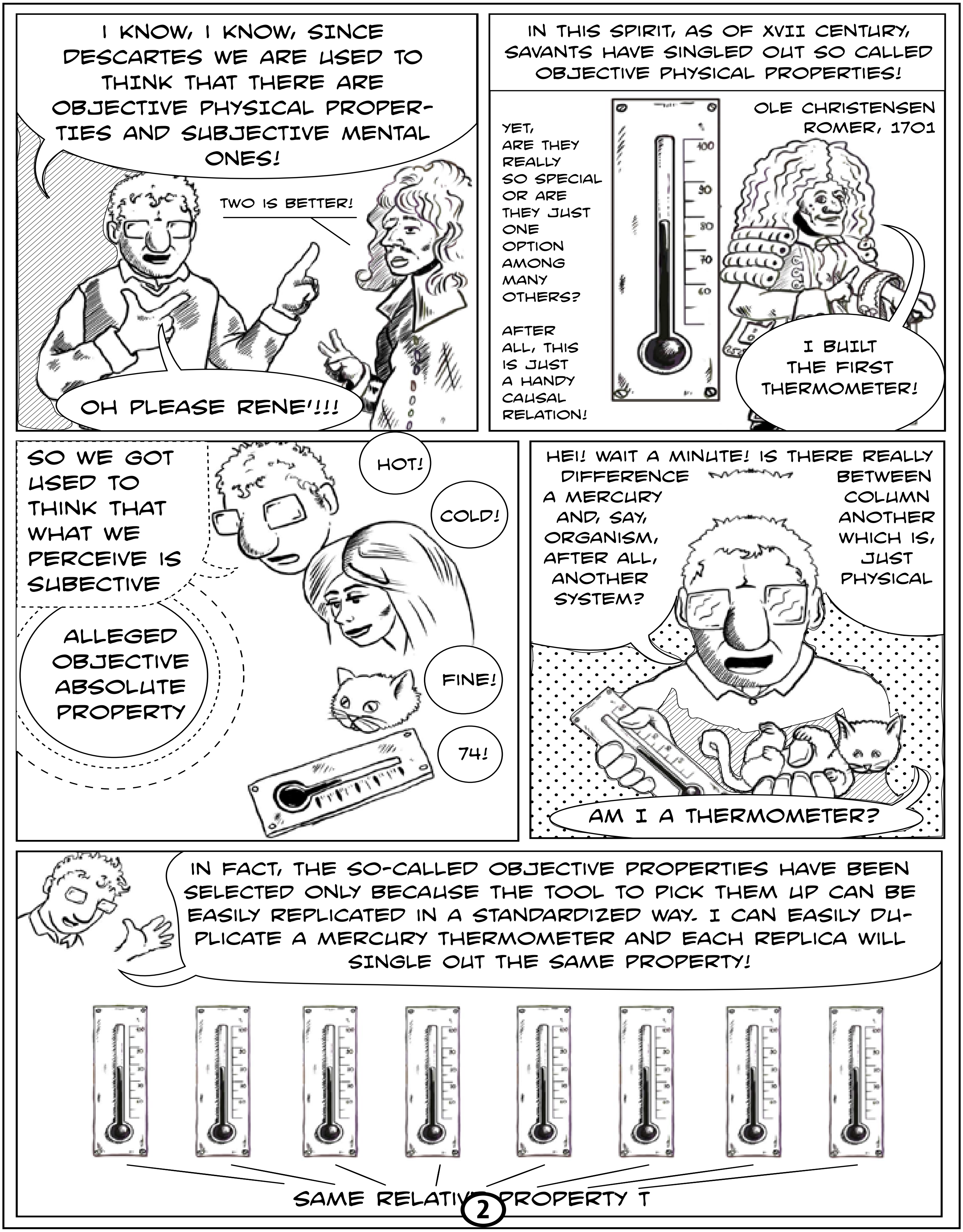
CLOSING THOUGHTS

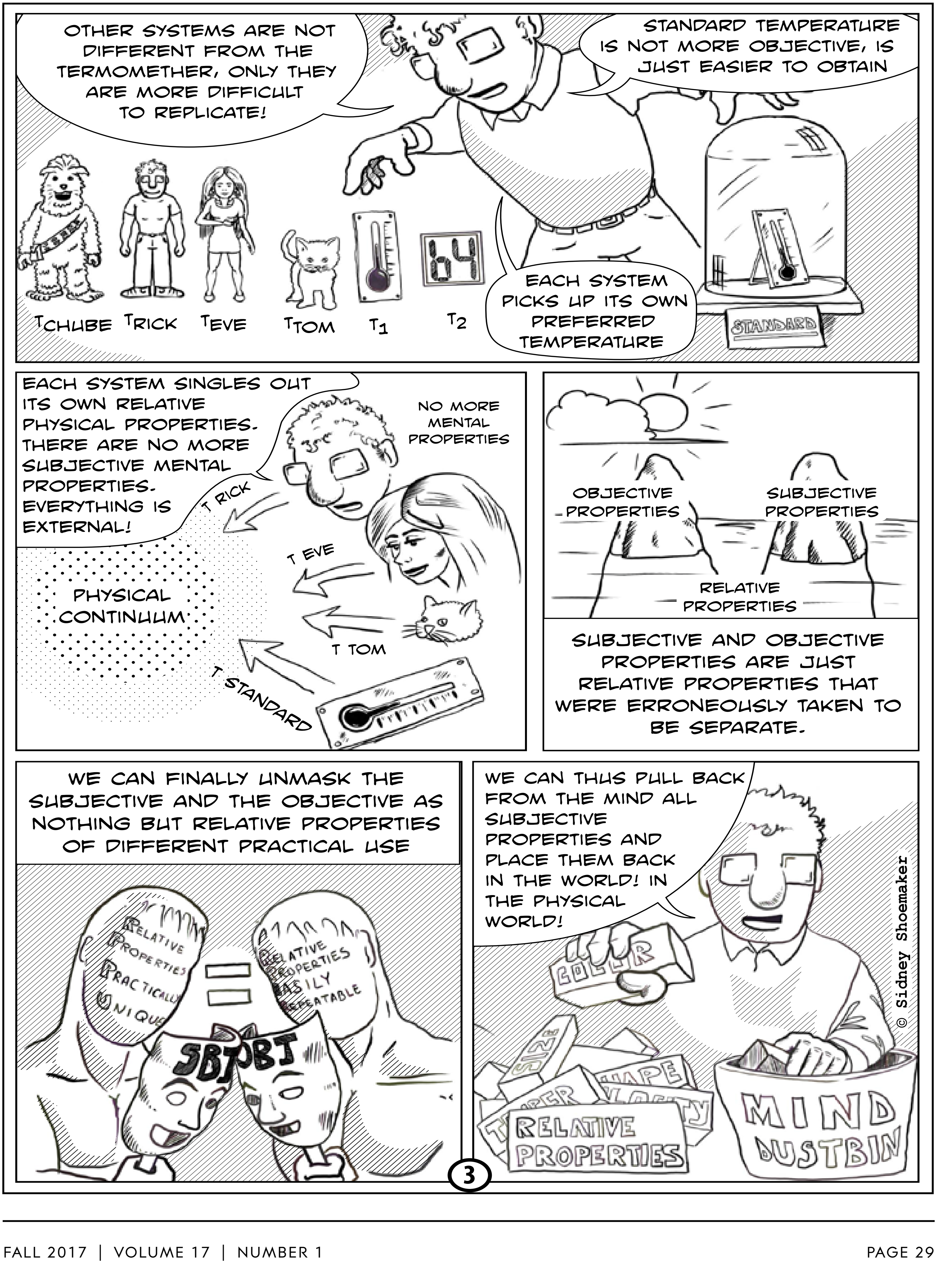
KRISTEN: After the research and discussion of Steve's work, I am left with a different set of questions and possibilities in my personal theory of mind than I started with. The parallels between the functioning of Creativity Machines and biological brains are clear. The CMs have the elements of cognition that biological brains do, perception, memory, awareness, recall, self-monitoring of recall and internal process, and most importantly, creativity. It appears that two very important qualities of biological consciousness—subjectivity, and privileged access, may likewise be emulated through this same paradigm, leaving us to ponder the "hard problem of consciousness" through this same theoretical framework.

NOTES

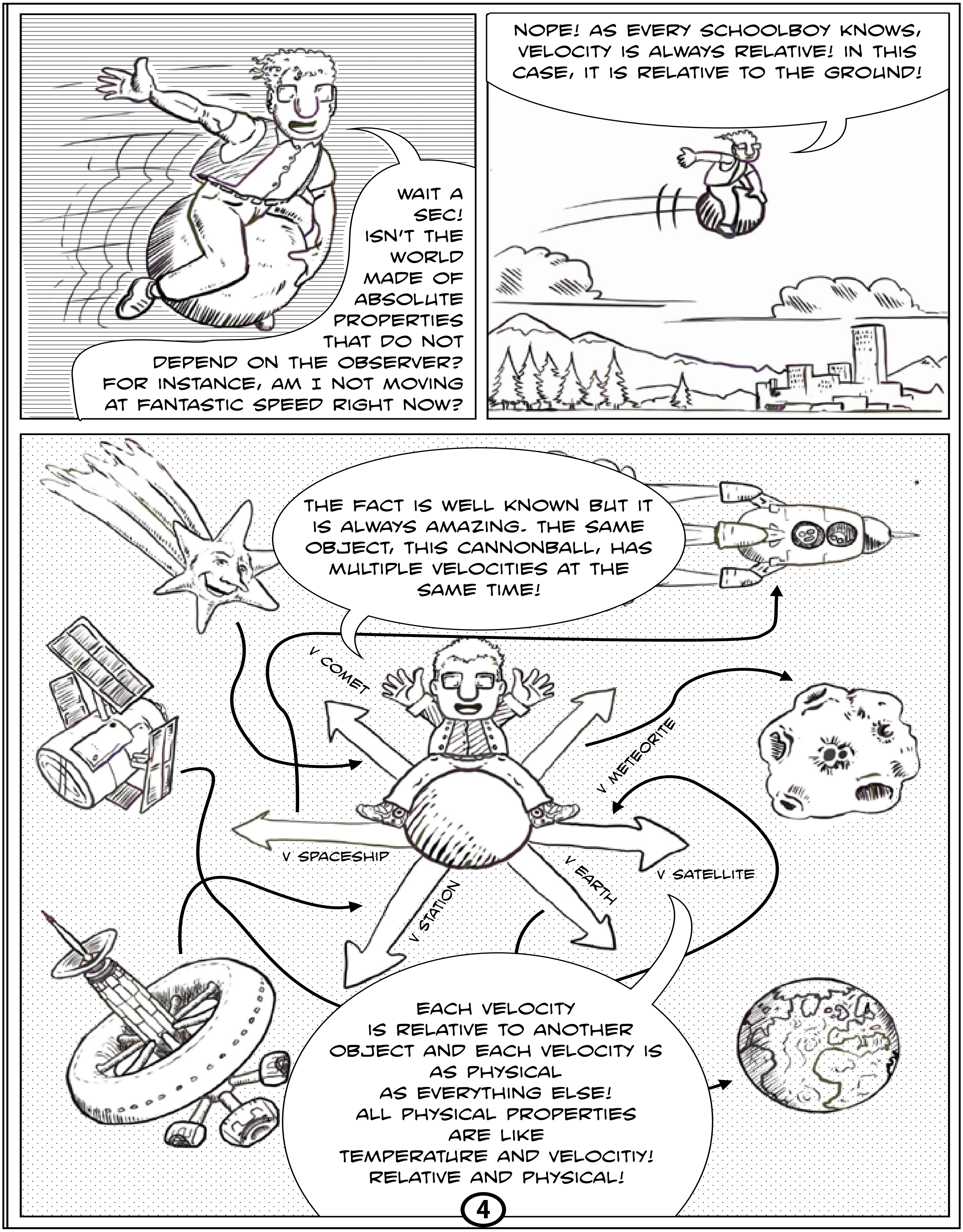
1. This dialogue began through a course on "Philosophy and Computers" and continued beyond the classroom. Thank you to Peter Boltuc and Stephen Thaler for all their help and assistance with this project. Thank you as well to Mark Jordan, Adrienne Bowen, Jessica Downs, and Yuji Kosugi for your help in understanding this material.
2. Stephen L. Thaler, "The Creativity Machine Paradigm: Withstanding the Argument from Consciousness," *APA Newsletter on Philosophy and Computers* 11, no. 2. (2012): 19–30.
3. Stephen L. Thaler, "The Fragmentation of the Universe and the Devolution of Consciousness," *U.S. Library of Congress*, Registration Number TXu000775586 / (January, 1997); Stephen L. Thaler, "The Emerging Intelligence and Its Critical Look at Us," *Journal of Near-Death Studies* 17 (1): (1998): 21–29.
4. Thaler, S. L. "Thalamocortical Algorithms in Space! The Building of Conscious Machines and the Lessons Thereof." *Proceedings of the World Future Society* (2010).
5. Stephen L. Thaler, "Cycles of Insanity and Creativity within Contemplative Neural Systems," *Medical Hypotheses* 94 (2016): 138–47.
6. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC: 2013.

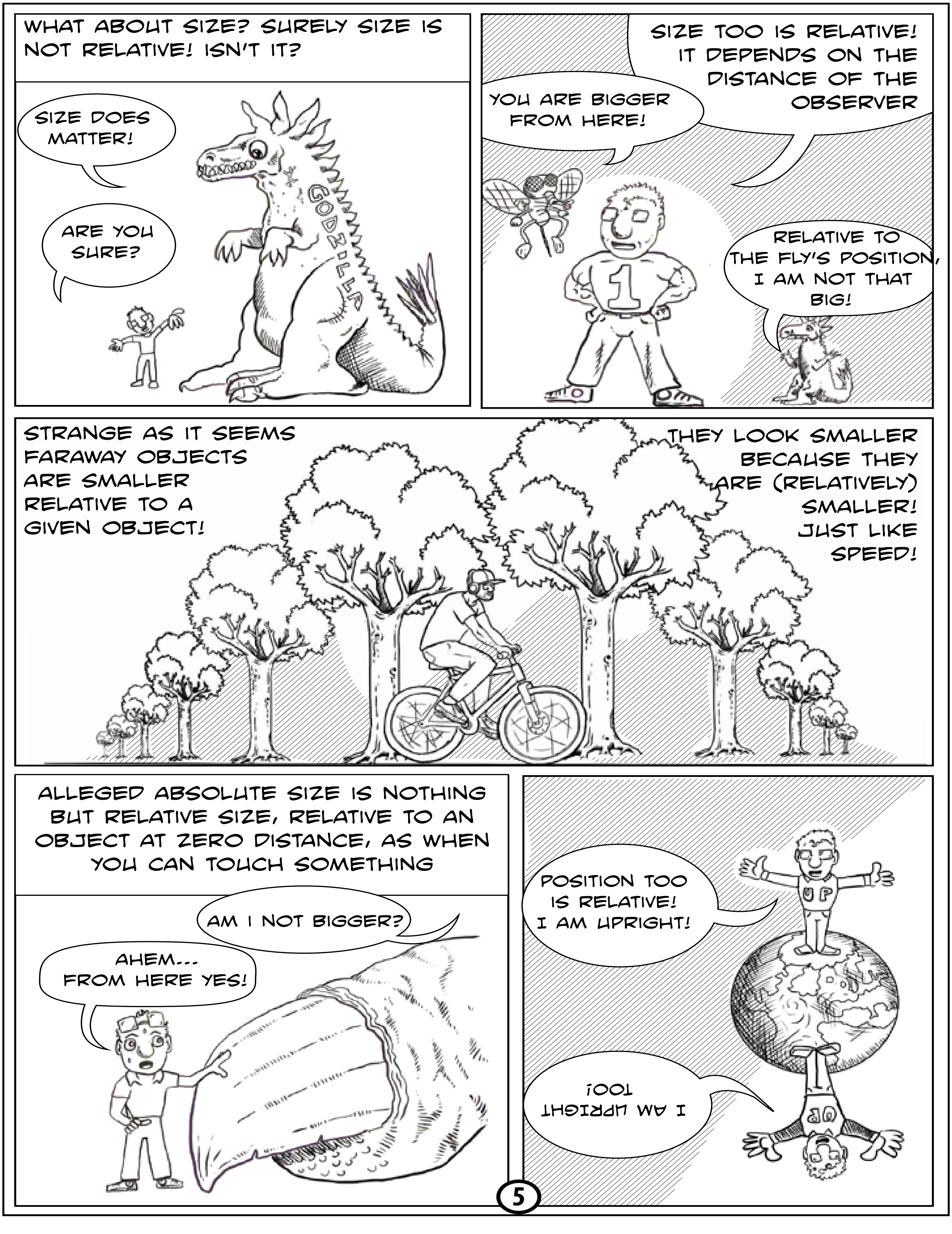

EXISTENCE IS RELATIVE
www.consciousness.it
riccardo.manzotti@gmail.com
© Riccardo Manzotti, 2017
IULM University, Milan
IS IT RATHER HOT! CAN I OPEN THE WINDOW?
14°F
ARE YOU CRAZY? IT IS GHASTLY COLD HERE!
64
IT IS FINE!
WHAT IS TEMPERATURE? IS IT MENTAL? OR PHYSICAL?
CAN WE SET ASIDE THE CONTRAST BETWEEN SUBJECTIVE AND OBJECTIVE?
RM17

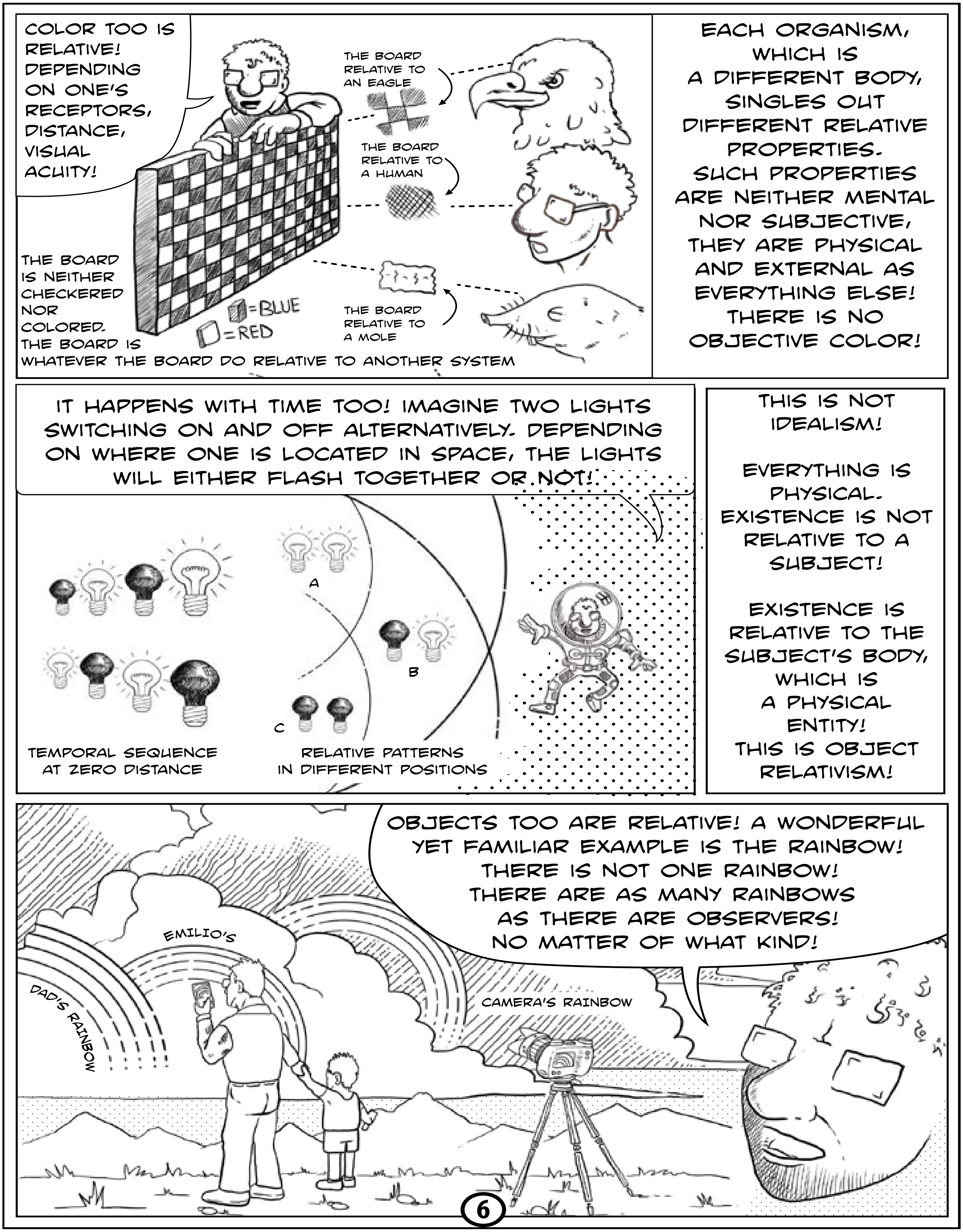

I KNOW, I KNOW, SINCE DESCARTES WE ARE USED TO THINK THAT THERE ARE OBJECTIVE PHYSICAL PROPERTIES AND SUBJECTIVE MENTAL ONES!
TWO IS BETTER!
OH PLEASE RENE'!!!
IN THIS SPIRIT, AS OF XVII CENTURY, SAVANTS HAVE SINGLED OUT SO CALLED OBJECTIVE PHYSICAL PROPERTIES!
OLE CHRISTENSEN ROMER, 1701
YET, ARE THEY REALLY SO SPECIAL OR ARE THEY JUST ONE OPTION AMONG MANY OTHERS?
AFTER ALL, THIS IS JUST A HANDY CAUSAL RELATION!
I BUILT THE FIRST THERMOMETER!
SO WE GOT USED TO THINK THAT WHAT WE PERCEIVE IS SUBECTIVE
ALLEGED OBJECTIVE ABSOLUTE PROPERTY
HOT!
COLD!
FINE!
74!
HEI! WAIT A MINUTE! IS THERE REALLY DIFFERENCE BETWEEN A MERCURY COLUMN AND, SAY, ANOTHER ORGANISM, WHICH IS, AFTER ALL, JUST ANOTHER PHYSICAL SYSTEM?
AM I A THERMOMETER?
IN FACT, THE SO-CALLED OBJECTIVE PROPERTIES HAVE BEEN SELECTED ONLY BECAUSE THE TOOL TO PICK THEM UP CAN BE EASILY REPLICATED IN A STANDARDIZED WAY. I CAN EASILY DUPLICATE A MERCURY THERMOMETER AND EACH REPLICA WILL SINGLE OUT THE SAME PROPERTY!
SAME RELATIVE PROPERTY T
2

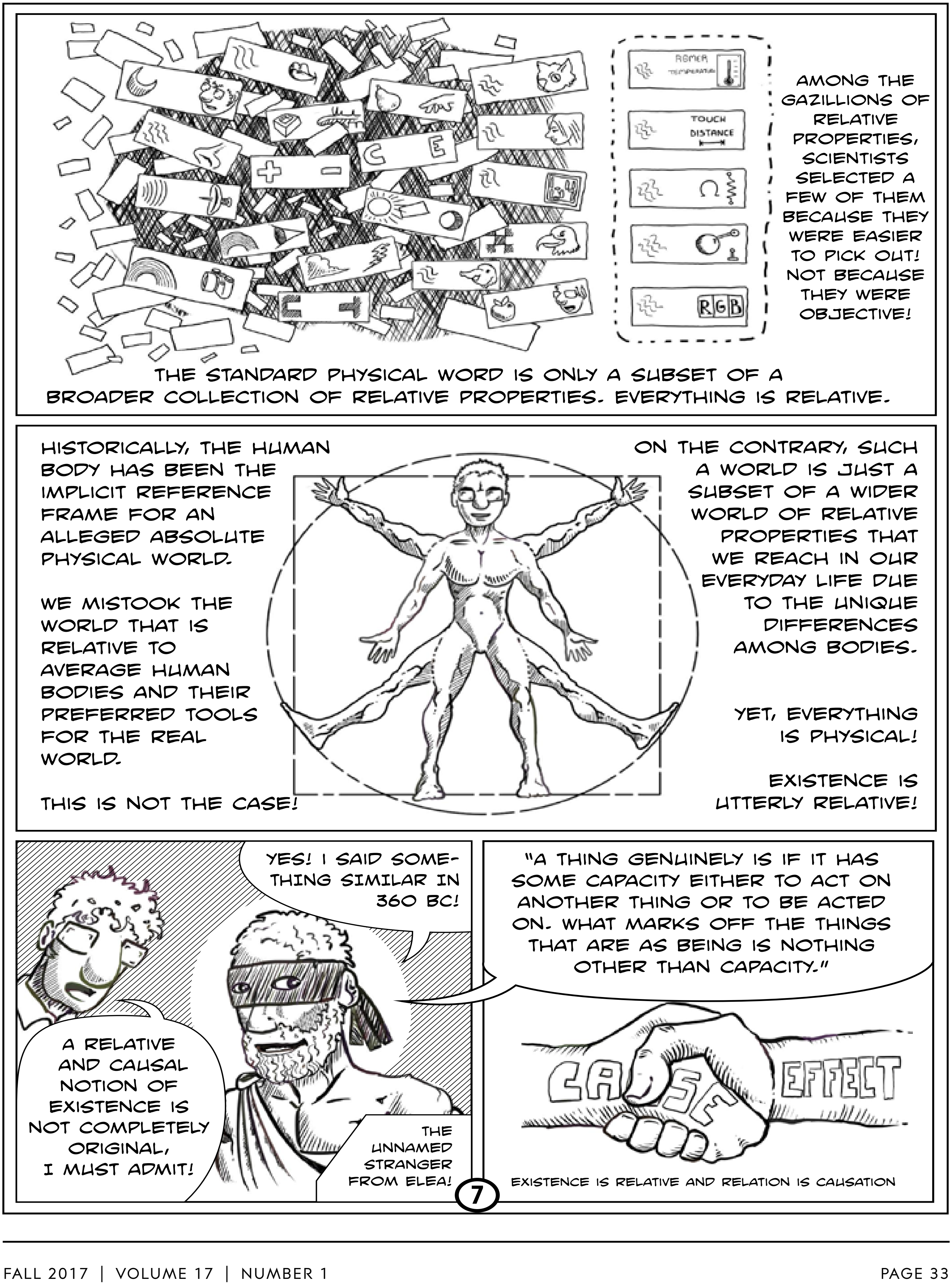

OTHER SYSTEMS ARE NOT DIFFERENT FROM THE TERMOMETHER, ONLY THEY ARE MORE DIFFICULT TO REPLICATE!
STANDARD TEMPERATURE IS NOT MORE OBJECTIVE, IS JUST EASIER TO OBTAIN
T_{CHUBE}
T_{RICK}
T_{EVE}
T_{TOM}
T_1
64
T_2
EACH SYSTEM PICKS UP ITS OWN PREFERRED TEMPERATURE
STANDARD
EACH SYSTEM SINGLES OUT ITS OWN RELATIVE PHYSICAL PROPERTIES. THERE ARE NO MORE SUBJECTIVE MENTAL PROPERTIES. EVERYTHING IS EXTERNAL!
NO MORE MENTAL PROPERTIES
T RICK
T EVE
T TOM
T STANDARD
PHYSICAL CONTINUUM
OBJECTIVE PROPERTIES
SUBJECTIVE PROPERTIES
RELATIVE PROPERTIES
SUBJECTIVE AND OBJECTIVE PROPERTIES ARE JUST RELATIVE PROPERTIES THAT WERE ERRONEOUSLY TAKEN TO BE SEPARATE.
WE CAN FINALLY UNMASK THE SUBJECTIVE AND THE OBJECTIVE AS NOTHING BUT RELATIVE PROPERTIES OF DIFFERENT PRACTICAL USE
RELATIVE PROPERTIES PRACTICALLY UNIQUE
=
RELATIVE PROPERTIES EASILY REPEATABLE
SBJ
OBJ
WE CAN THUS PULL BACK FROM THE MIND ALL SUBJECTIVE PROPERTIES AND PLACE THEM BACK IN THE WORLD! IN THE PHYSICAL WORLD!
COLOUR
SIZE
SHAPE
VELOCITY
RELATIVE PROPERTIES
MIND DUSTBIN
3

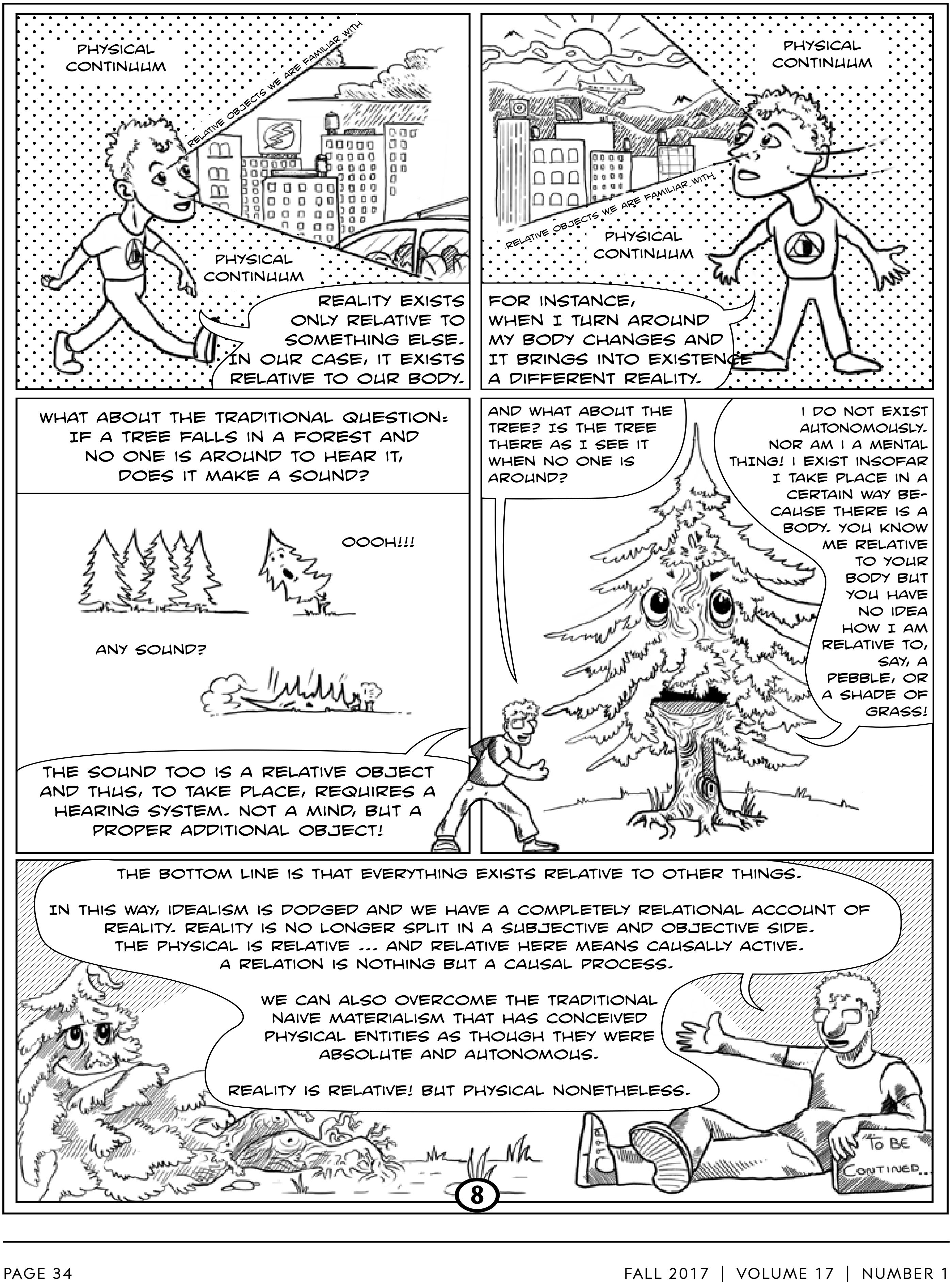
© Sidney Shoemaker


WAIT A SEC! ISN'T THE WORLD MADE OF ABSOLUTE PROPERTIES THAT DO NOT DEPEND ON THE OBSERVER? FOR INSTANCE, AM I NOT MOVING AT FANTASTIC SPEED RIGHT NOW?
NOPE! AS EVERY SCHOOLBOY KNOWS, VELOCITY IS ALWAYS RELATIVE! IN THIS CASE, IT IS RELATIVE TO THE GROUND!
THE FACT IS WELL KNOWN BUT IT IS ALWAYS AMAZING. THE SAME OBJECT, THIS CANNONBALL, HAS MULTIPLE VELOCITIES AT THE SAME TIME!
V COMET
V METEORITE
V SPACESHIP
V SATELLITE
V STATION
V EARTH
EACH VELOCITY IS RELATIVE TO ANOTHER OBJECT AND EACH VELOCITY IS AS PHYSICAL AS EVERYTHING ELSE! ALL PHYSICAL PROPERTIES ARE LIKE TEMPERATURE AND VELOCITIY! RELATIVE AND PHYSICAL!
4


WHAT ABOUT SIZE? SURELY SIZE IS NOT RELATIVE! ISN'T IT?
SIZE DOES MATTER!
ARE YOU SURE?
GODZILLA
SIZE TOO IS RELATIVE! IT DEPENDS ON THE DISTANCE OF THE OBSERVER
YOU ARE BIGGER FROM HERE!
1
RELATIVE TO THE FLY'S POSITION, I AM NOT THAT BIG!
STRANGE AS IT SEEMS FARAWAY OBJECTS ARE SMALLER RELATIVE TO A GIVEN OBJECT!
THEY LOOK SMALLER BECAUSE THEY ARE (RELATIVELY) SMALLER! JUST LIKE SPEED!
ALLEGED ABSOLUTE SIZE IS NOTHING BUT RELATIVE SIZE, RELATIVE TO AN OBJECT AT ZERO DISTANCE, AS WHEN YOU CAN TOUCH SOMETHING
AM I NOT BIGGER?
AHEM... FROM HERE YES!
POSITION TOO IS RELATIVE! I AM UPRIGHT!
UP
I AM UPRIGHT TOO!
UP
1
5


COLOR TOO IS RELATIVE! DEPENDING ON ONE'S RECEPTORS, DISTANCE, VISUAL ACUITY!
THE BOARD RELATIVE TO AN EAGLE
THE BOARD RELATIVE TO A HUMAN
THE BOARD RELATIVE TO A MOLE
▨=BLUE
□=RED
THE BOARD IS NEITHER CHECKERED NOR COLORED. THE BOARD IS WHATEVER THE BOARD DO RELATIVE TO ANOTHER SYSTEM
EACH ORGANISM, WHICH IS A DIFFERENT BODY, SINGLES OUT DIFFERENT RELATIVE PROPERTIES. SUCH PROPERTIES ARE NEITHER MENTAL NOR SUBJECTIVE, THEY ARE PHYSICAL AND EXTERNAL AS EVERYTHING ELSE! THERE IS NO OBJECTIVE COLOR!
IT HAPPENS WITH TIME TOO! IMAGINE TWO LIGHTS SWITCHING ON AND OFF ALTERNATIVELY. DEPENDING ON WHERE ONE IS LOCATED IN SPACE, THE LIGHTS WILL EITHER FLASH TOGETHER OR NOT!
A
B
C
TEMPORAL SEQUENCE AT ZERO DISTANCE
RELATIVE PATTERNS IN DIFFERENT POSITIONS
THIS IS NOT IDEALISM!
EVERYTHING IS PHYSICAL. EXISTENCE IS NOT RELATIVE TO A SUBJECT!
EXISTENCE IS RELATIVE TO THE SUBJECT'S BODY, WHICH IS A PHYSICAL ENTITY! THIS IS OBJECT RELATIVISM!
OBJECTS TOO ARE RELATIVE! A WONDERFUL YET FAMILIAR EXAMPLE IS THE RAINBOW! THERE IS NOT ONE RAINBOW! THERE ARE AS MANY RAINBOWS AS THERE ARE OBSERVERS! NO MATTER OF WHAT KIND!
EMILIO'S
DAD'S RAINBOW
CAMERA'S RAINBOW
6


RØMER TEMPERATURE
TOUCH DISTANCE
RGB
AMONG THE GAZILLIONS OF RELATIVE PROPERTIES, SCIENTISTS SELECTED A FEW OF THEM BECAUSE THEY WERE EASIER TO PICK OUT! NOT BECAUSE THEY WERE OBJECTIVE!
THE STANDARD PHYSICAL WORD IS ONLY A SUBSET OF A BROADER COLLECTION OF RELATIVE PROPERTIES. EVERYTHING IS RELATIVE.
HISTORICALLY, THE HUMAN BODY HAS BEEN THE IMPLICIT REFERENCE FRAME FOR AN ALLEGED ABSOLUTE PHYSICAL WORLD.
WE MISTOOK THE WORLD THAT IS RELATIVE TO AVERAGE HUMAN BODIES AND THEIR PREFERRED TOOLS FOR THE REAL WORLD.
THIS IS NOT THE CASE!
ON THE CONTRARY, SUCH A WORLD IS JUST A SUBSET OF A WIDER WORLD OF RELATIVE PROPERTIES THAT WE REACH IN OUR EVERYDAY LIFE DUE TO THE UNIQUE DIFFERENCES AMONG BODIES.
YET, EVERYTHING IS PHYSICAL!
EXISTENCE IS UTTERLY RELATIVE!
YES! I SAID SOMETHING SIMILAR IN 360 BC!
A RELATIVE AND CAUSAL NOTION OF EXISTENCE IS NOT COMPLETELY ORIGINAL, I MUST ADMIT!
THE UNNAMED STRANGER FROM ELEA!
"A THING GENUINELY IS IF IT HAS SOME CAPACITY EITHER TO ACT ON ANOTHER THING OR TO BE ACTED ON. WHAT MARKS OFF THE THINGS THAT ARE AS BEING IS NOTHING OTHER THAN CAPACITY."
CAUSE
EFFECT
EXISTENCE IS RELATIVE AND RELATION IS CAUSATION
7


PHYSICAL CONTINUUM
RELATIVE OBJECTS WE ARE FAMILIAR WITH
PHYSICAL CONTINUUM
REALITY EXISTS ONLY RELATIVE TO SOMETHING ELSE. IN OUR CASE, IT EXISTS RELATIVE TO OUR BODY.
PHYSICAL CONTINUUM
RELATIVE OBJECTS WE ARE FAMILIAR WITH
PHYSICAL CONTINUUM
FOR INSTANCE, WHEN I TURN AROUND MY BODY CHANGES AND IT BRINGS INTO EXISTENCE A DIFFERENT REALITY.
WHAT ABOUT THE TRADITIONAL QUESTION: IF A TREE FALLS IN A FOREST AND NO ONE IS AROUND TO HEAR IT, DOES IT MAKE A SOUND?
OOOH!!!
ANY SOUND?
THE SOUND TOO IS A RELATIVE OBJECT AND THUS, TO TAKE PLACE, REQUIRES A HEARING SYSTEM. NOT A MIND, BUT A PROPER ADDITIONAL OBJECT!
AND WHAT ABOUT THE TREE? IS THE TREE THERE AS I SEE IT WHEN NO ONE IS AROUND?
I DO NOT EXIST AUTONOMOUSLY. NOR AM I A MENTAL THING! I EXIST INSOFAR I TAKE PLACE IN A CERTAIN WAY BECAUSE THERE IS A BODY. YOU KNOW ME RELATIVE TO YOUR BODY BUT YOU HAVE NO IDEA HOW I AM RELATIVE TO, SAY, A PEBBLE, OR A SHADE OF GRASS!
THE BOTTOM LINE IS THAT EVERYTHING EXISTS RELATIVE TO OTHER THINGS. IN THIS WAY, IDEALISM IS DODGED AND WE HAVE A COMPLETELY RELATIONAL ACCOUNT OF REALITY. REALITY IS NO LONGER SPLIT IN A SUBJECTIVE AND OBJECTIVE SIDE. THE PHYSICAL IS RELATIVE ... AND RELATIVE HERE MEANS CAUSALLY ACTIVE. A RELATION IS NOTHING BUT A CAUSAL PROCESS.
WE CAN ALSO OVERCOME THE TRADITIONAL NAIVE MATERIALISM THAT HAS CONCEIVED PHYSICAL ENTITIES AS THOUGH THEY WERE ABSOLUTE AND AUTONOMOUS. REALITY IS RELATIVE! BUT PHYSICAL NONETHELESS.
TO BE CONTINED..
8