APA Newsletters

NEWSLETTER ON PHILOSOPHY AND COMPUTERS

Volume 11, Number 2 Spring 2012

apa
The American Philosophical Association
University of Delaware
Newark, DE 19716

 ISSN 2155-9708

APA Newsletter on

Philosophy and Computers

Piotr Bołtuć, Editor | Spring 2012 | Volume 11, Number 2

From the Editor

The APA ad hoc committee on philosophy and computers started as largely a group advocating the use of computers and the web among philosophers, and by the APA. While today philosophical issues pertaining to computers are becoming more and more important, we may have failed in some way since problems that have been plaguing the APA's website for about the last year have put us all back, unnecessarily. This also pertains to the *Newsletter*; not only did we lose positioning in the web-search engines but the *Newsletter* reverted to just PDFs. The good news is that archival issues are successively coming back. I remember the advice that David Chalmers gave to the *Newsletter* upon receiving the Barwise Prize a few years ago, to either become a regular journal or, if we stay open access, to use much more of blog-style communications. It is my hope that one day the latter option may become more realistic.

Let me change gears a bit and restart on a somewhat personal note. My first philosophy tutor was my mother; among other things she taught me that philosophy is the theory of the general theories of all the sciences. I still like this definition. My first philosophy tutor also warned me that philosophy should not become overly preoccupied with just one theory, at one stage of its development, which has been Spencer's predicament. Consistent with this advice, when I was starting my own philosophical thinking I was always puzzled that few philosophers drew sufficient conclusions from Einstein's relativity theory, in particular its direct implications for Newtonian and Kantian understanding of time and space. Today it seems that more and more philosophers focus on the philosophical implications of quantum physics, and in particular the issue of quantum pairs. Therefore, I was very interested in Terry Bynum's paper, when I heard its earlier version at the 2011 CAP conference in Aarhus, Denmark. I am very glad that Terry accepted my invitation so that his interesting article is featured in the current issue. Of course, the question who is able to avoid excessive reliance on the current state of science and who falls into the Spencer-trap is always hard to answer without a longer historical perspective. I am also glad that Luciano Floridi responds to Terry's paper in this issue with an important historical outlook. More responses are expected and encouraged for submission to the next issue.

In his provocative article Tony Beavers argues that it may be morally required to build a machine that would make human beings more moral. I think the paper is an important contribution to the recently booming area of robot ethics. Alexandre Monnin contributes to the set of articles pertaining to ontology of the web that started with a paper by Harry Halpin. In his tightly argued work, originally written in French, Alexandre shows why URIs are philosophically interesting, not only for philosophers of computers but also for the more traditional colleagues interested in philosophy of language. In the next paper Stephen Thaler talks about creativity machines. While some philosophers may still not be sure whether and by what standards machines can be creative, Thaler designed, patented, and prepared for useful applications some such machines so the proof seems to be in the pudding, and some of the proof can also be found in this interesting article. We end with a cartoon by Richardo Manzotti; this time it is on an ontological topic. As always cartoons tend to be overly persuasive for philosophical discussion; yet, they serve as a good tool for putting forth the author's ideas.

I am sure the chair of the committee would want to mention the very successful session on machine consciousness at the Central APA meeting. The session brought together papers by Terry Horgan, Robert van Gullick, and Ned Block (who was unable to come due to illness), as well as by two members of this committee, David Anderson and myself. The session was very well attended, so that some people had to sit on the floor or in the doorway. I do hope to have more on this committee's activities in the next issue.

Articles

On Rethinking the Foundations of Philosophy in the Information Age[*]

Terrell Ward Bynum
Southern Connecticut State University

1. Introduction: physics and the information revolution

It is commonplace today to hear people say that we are "living in the Age of Information" and that an "Information Revolution" is sweeping across the globe, changing everything from banking to warfare, medicine to education, entertainment to government, and on and on. But why are these dramatic changes taking place? How is it possible for information technology (IT) to transform our world so quickly and so fundamentally? Scholars in the field of computer ethics are familiar with James Moor's suggested answer; namely, that IT is revolutionary because it is *logically malleable*, making IT one of the most powerful and flexible technologies ever created. IT is a nearly universal tool, Moor said, that can be adjusted and fine tuned to carry out almost any task. The limits of IT, he noted, are basically the limits of our imagination. Moor's influential analysis of the Information Revolution (including associated concepts like *policy vacuums, conceptual muddles,* and *informationalization*) has shown itself to be practical and insightful (see Moor 1998).

Today, recent developments in physics, especially in quantum theory and cosmology, suggest an additional—almost metaphysical—answer to explain why IT is so effective in transforming the world. During the past two decades, many physicists have come to believe that the universe is made of information; that is, that *our world is a vast "ocean" of quantum bits* ("qubits") *and every object or process in this ocean of information (including human beings) can be seen as a constantly changing data structure comprised of qubits.* (See, for example, Lloyd 2006 and Vedral 2010.) If everything in the world is made of information, and IT provides knowledge and tools for analyzing and manipulating information, then we have an impressive explanation of the transformative power of IT *based upon the fundamental nature of the universe*!

It is not surprising that important developments in science can have major philosophical import. Since the time of ancient Greece, profound scientific developments have inspired significant rethinking of "bedrock" ideas in philosophy. Indeed, scientists working on the cutting edges of their field often engage in thinking that is borderline metaphysical. Occasionally, the scientists and philosophers have been the very same people, as illustrated by Aristotle, who created physics and biology and, at the same time, made related contributions to metaphysics, logic, epistemology, and other branches of philosophy. Or consider Descartes and Leibniz, both of whom were excellent scientists and world-class mathematicians as well as great philosophers. Sometimes, thinkers who were primarily scientists—for example, Copernicus, Galileo, and Newton—inspired others who were primarily philosophers—for example, Hobbes, Locke, and Kant. Later, revolutionary scientific contributions of Darwin, Einstein, Bohr, Schrödinger, and others significantly influenced philosophers like Spencer, Russell, Whitehead, Popper, and many more.

Today, in the early years of the twenty-first century, cosmology and quantum physics appear likely to alter significantly our scientific understanding of the universe, of life, and of human nature. These developments in physics, it seems to me, are sure to lead to important new contributions to philosophy. Among contemporary philosophers, Luciano Floridi—with his pioneering efforts in the *philosophy of information, informational realism,* and *information ethics* (all his terms)—has been leading the way in demonstrating the importance of the concept of information in philosophy. (See, for example, his book *The Philosophy of Information*, 2011.) Given the above-mentioned developments in physics, it is not surprising that Floridi was the first philosopher ever (in 2008-2009) to hold the prestigious post of Gauss Professor at the Göttingen Academy of Sciences in Germany (previous Gauss Professors had been physicists or mathematicians). Floridi's theory of *informational realism,* though, focuses primarily upon *Platonic information* that is not subject to the laws of physics. A materialist philosopher, perhaps, would be more inclined to focus instead upon qubits, which *are* physical in nature. Whether one takes Floridi's Platonic approach or a materialistic perspective, I believe that recent developments in philosophy and physics with regard to *the central importance of information will encourage philosophers to rethink the bedrock concepts of their field*.

2. "It from bit"

It is my view that a related materialist "information revolution" in philosophy began in the mid 1940s when philosopher/scientist Norbert Wiener triumphantly announced to his students and colleagues at MIT that "entropy is information." He realized that information is physical and, therefore, it obeys the laws of physics. As a result, in 1948 in his book *Cybernetics*, Wiener made this important claim about philosophical materialism:

> Information is information, not matter or energy. No materialism which does not admit this can survive at the present day. (p. 132)

According to Wiener, therefore, every physical being can be viewed as an informational entity. This is true even of human beings; and, in 1954, in the second edition of his book *The Human Use of Human Beings*, Wiener noted that the essential nature of a person depends, not upon the particular atoms that happen to comprise one's body at any given moment, but rather upon the informational pattern encoded within the body:

> We are but whirlpools in a river of ever-flowing water. We are not stuff that abides, but patterns that perpetuate themselves. (p. 96)

> The individuality of the body is that of a flame . . . of a form rather than a bit of substance." (p. 102)

In that same book, Wiener presented a remarkable thought experiment to show that, if one could encode, in a telegraph message, for example, the entire exquisitely complex information pattern of a person's body, and then use that encoded pattern to reconstitute the person's body from appropriate atoms at the receiving end of a message, people could travel instantly from place to place via telegraph. Wiener noted that this idea raises knotty philosophical questions regarding not only personal identity, but also "forking" from one person into two, "split" personalities, survival of the self after the death of one's body, and a number of others (Wiener 1950, Ch. VI; 1954, Ch. V).

Decades later, in 1990, physicist John Archibald Wheeler introduced his famous phrase "it from bit" in an influential paper (Wheeler 1990), and he thereby gave a major impetus to an information revolution in physics. In that paper, Wheeler declared that "all things physical are information theoretic in origin"—that "every physical entity, every it, derives from bits"—that "every particle, every field of force, even the spacetime continuum itself . . . derives its function, its meaning, its very existence" from bits. He predicted that "Tomorrow we will have learned to understand and express *all* of physics in the language of information" (emphasis in the original).

Since 1990, a number of physicists—some of them inspired by Wheeler—have made great strides toward fulfilling his "it-from-bit" prediction. In 2006, for example, in his book *Programming the Universe*, Seth Lloyd presented impressive evidence supporting the view that the universe is not only a vast ocean of qubits, it is actually a gigantic quantum computer:

> The conventional view is that the universe is nothing but elementary particles. That is true, but it is equally true that the universe is nothing but bits—or rather, nothing but qubits. Mindful that if it walks like a duck and it quacks like a duck then it's a duck . . . since the universe registers and processes information like a quantum computer, and is observationally indistinguishable from a quantum computer, then it *is* a quantum computer. (p. 154, emphasis in the original)

More recently, in 2011, three physicists used axioms from information processing to derive the mathematical framework of quantum mechanics (Chiribella et al. 2011). These are only two of a growing number of achievements that have begun to fulfill Wheeler's "it from bit" prediction.

The present essay explores some philosophical implications of Wheeler's view that every physical entity—every particle, every field of force, even space-time—derives its very existence from qubits. But if, as Wheeler has said, qubits are responsible for the very existence of every particle and every field of force,

then qubits were, in some sense, prior to every other physical thing that exists. Qubits, therefore, must have been part of the Big Bang! As Seth Lloyd has said, "*The Big Bang was also a Bit Bang*" (Lloyd 2006, 46).

Unlike traditional bits, such as those processed in today's computing devices, qubits have quantum features, such as *genuine randomness, superposition*, and *entanglement*—features that Einstein and other scientists considered "spooky" and "weird." As explained below, these scientifically verified quantum phenomena raise important questions about traditional bedrock philosophical concepts.

3. To be is to be a quantum data structure

In most computers today, each bit can only be in one or the other of two specific states, 0 or 1. Such a "classical" bit cannot be both 0 and 1 at the same time. A qubit, on the other hand, *can* simultaneously be 0 and 1, and indeed it can even be in an infinite number of different states between 0 and 1. As Vlatko Vedral noted, in his book *Decoding Reality: the Universe as Quantum Information* (2010),

> we are permitted to have a zero and a one at the same time present in one physical system. In fact, we are permitted to have an infinite range of states between zero and one—which we call a qubit. (p. 137)

This remarkable feature of qubits is not just a theoretical possibility. It is *real,* in the sense that it is governed by the laws of physics, and it enables quantum computers to calculate far more efficiently than a traditional computer using classical bits (see below).

If every physical thing in the universe consists of qubits—in keeping with Wheeler's "it from bit" hypothesis—then one would expect that *any* physical entity could be in many different states at once, depending on the many states of the qubits of which it is composed. Indeed, quantum physicists have found that, under the right circumstances, "All objects in the universe are capable of being in all possible states" (Vedral 2010, 122). This means that objects can be in many different places at once, that a particle could be both positive and negative at the same time, or simultaneously spinning clockwise and counter clockwise around the same axis. It means that living things—like Schrödinger's famous cat or a human being—could be both alive and dead at the same time, and at least some things can be teleported from place to place instantly over long distances faster than the speed of light without passing through the space in between. Finally, it also means that, at the deepest level of reality, *the universe is both digital and analogue at the same time*. These are not mere speculations, but requirements of quantum mechanics, which is the most tested and most strongly confirmed scientific theory in history. So, philosophers, it seems, will have to rethink many fundamental philosophical concepts, like *being* and *non-being*, *real* and *unreal*, *actual* and *potential*, *cause* and *effect*, *consistent* and *contradictory, knowledge* and *thinking*, and many more (see below).

4. Coming into existence in the classical universe: information and decoherence

A familiar "double-slit experiment," which is often performed today in high school physics classes and undergraduate laboratories, illustrates the ability of different kinds of objects to be in many different states at once. In such an experiment, particles or larger objects are fired, one at a time, by a "particle gun" toward a screen designed to detect them. The particles or objects in the experiment, can be, for example, photons, or electrons, or single atoms, or much larger objects, such as "buckeyballs" (composed of sixty carbon atoms comprised of 1,080 subatomic particles), or even larger objects.

To begin a double-slit experiment, a metal plate with two parallel vertical slits is inserted between the gun and the detection screen. The gun then fires individual particles or objects—one at a time—at the double-slit plate. If the particles or objects were to act like classical objects, some of them would go through the right slit and strike the detection screen behind that slit, while others would go through the left slit and strike the detection screen behind that slit. But this is not what happens. Instead, surprisingly, *a single particle or object goes through both slits simultaneously*, and when a sufficient number of individual particles or objects has been fired, a *wave-interference pattern* is created on the detection screen from the individual spots where the particles or objects landed. In such an experiment, an individual particle or object travels toward the double-slit plate *as a wave*; and then, on the other side of the double-slit plate, it travels toward the detection screen *as two waves interfering with each other*. When the two interfering waves arrive at the detection screen, however, a classical particle or object suddenly appears on the screen *at a specific location* which could not have been known in advance, even in principle.

In summary, then, in a double-slit experiment, single particles or objects behave also like waves—even like two waves creating an interference pattern. How is a philosopher to interpret these results? Perhaps we could try to make sense of this "weird" behavior by adopting a distinction much like Aristotle's distinction between the *potential* and the *actual*. When a child is born, for example, Aristotle would say that the child is *potentially* a language speaker, but not *actually* a language speaker. The potential of the child to speak a language is, for Aristotle, *something real* that is included in the very nature of the child. In contrast, a stone or a chunk of wood, for example, does not have the potential ever to become a language speaker. For Aristotle, *the potential and the actual are both real in the sense that both are part of the nature of a being; and the potential of a being becomes actualized through interactions with already actualized things in the environment.* A child, for example, becomes an actual language speaker by interacting appropriately with people in the community who are actual language speakers. And, similarly, an unlit candle, which potentially has a flame at the top, becomes a candle with an actual flame when it interacts appropriately with some actual fire in the environment.

If we adopt a distinction that is very similar to Aristotle's, we could say, perhaps, that the waves in a double-slit experiment consist of *potential paths* that the particle or object could follow on its way to the detection screen. Indeed, this is an interpretation that many quantum scientists accept. The potential paths, then, are *real* entities that travel through space-time together as a wave or "packet of possibilities" between the gun and the screen. But where is the *actual (*that is, *classical)* particle or object while its *packet of possibilities* is traveling to the screen? Has the classical particle or object itself disappeared? Or does it exist *as* a packet of possibilities? And how could it be *an actual particle or object* when it is still in the gun, or when it strikes the screen, but then *only be a wave of possibilities* while traveling between the two? Typical philosophical ideas about real and unreal, cause and effect, potential and actual don't seem to fit this case. Nevertheless, double-slit experiments are regularly performed in high school classrooms and undergraduate labs around the world—and always with the same "weird" results. Indeed, quantum mechanics requires that every object in the universe, no matter how large, would behave the same way under the right circumstances!

In quantum mechanics, the possibilities that form the wave are said to be "superposed" upon each other, and so together they are called the "*superpositions*" of the particle or

object. Some quantum scientists would say that the particle or object exists *everywhere at once within the wave*. Other scientists would say that *no actual particle or object exists within the wave*, and it is illegitimate even to ask for its specific location. In any case, when a wave of possibilities interacts appropriately with another physical entity in its environment *by sharing a bit of information with another physical entity*, all the "superposed" possibilities—*except one*—suddenly disappear and *one* actualized classical particle or object instantly appears randomly at a specific location. Quantum physicists call this phenomenon, in which a wave of possibilities gets converted into an actualized classical object, *decoherence*.

Decoherence, then, is a remarkable phenomenon! It is what brings into existence actualized classical objects—located at specific places and with specific properties that can be observed and measured. Decoherence somehow "extracts" or "creates" classical objects out of an infinite set of possibilities within our universe; and this "extraction" process is *genuinely random*. As Anton Zeilinger explains,

> The world as it is right now in this very moment does not determine uniquely the world in a few years, in a few minutes, or even in the next second. The world is open. We can give only probabilities for individual events to happen. And it is not just our ignorance. Many people believe that this kind of randomness is limited to the microscopic world, but this is not true, as the [random] measurement result itself can have macroscopic consequences. (Zeilinger 2010, 265)

Random or not, *being or existing* in our universe has two different varieties:

1. One is *quantum existence* as a wave of superposed possibilities, while the other is
2. *Classical existence* as a specific object located at a specific place in space-time with classical properties which can be observed and measured.

In our universe, the quantum realm and the classical realm exist together and interact with each other. The ultimate source of physical being is the constantly expanding ocean of qubits, which establish what is physically possible by generating—or being?—an infinite set of superposed possibilities. From this infinite, always expanding, set of possibilities, the sharing of specific information (decoherence) generates the everyday classical objects of our world in specific locations with observable and measurable properties. *Information, then, combined with the process of sharing information, is the ultimate source of everything physical in our universe. It from bit!*

5. Additional quantum puzzles for philosophy

Similar philosophical challenges arise from other quantum phenomena, such as *entanglement, "spooky action at a distance," teleportation*, and *quantum computing.* Each of these phenomena is briefly discussed below along with some of the philosophical questions that arise from them.

Entanglement and "Spooky Action at a Distance" — As indicated above, a quantum entity can be indefinite in the sense that its properties can be superposed possibilities that have not yet been actualized. For example, an electron could be spinning clockwise and counterclockwise around the same axis at the same time. When one observes or measures that electron (or when it interacts with another physical entity in the environment), its spin—instantly and randomly—becomes definitely clockwise or definitely counterclockwise. This happens because of decoherence in which the electron shares information about itself with the measurer (or something else in the environment).

If two electrons (or other quantum entities) are close together and interact appropriately, instead of acting like two separate entities, each with its own superposed possibilities, the two electrons *share their superpositions and begin to act like a single quantum entity*. This phenomenon is called *entanglement.* Thus, the spins of two entangled electrons, both of which are spinning simultaneously clockwise and counterclockwise, depend upon each other in such a way that if one of the electrons is measured or observed, thereby randomly making it spin definitely clockwise or definitely counterclockwise, the other electron's spin instantly becomes the opposite of the spin of the first one. The amazing and puzzling (Einstein said "spooky") thing is that such entanglement can continue to exist even if the electrons are separated by huge distances. For example, if one entangled electron is on Earth and the other one is sent to Mars, they still can be entangled. So if someone measures the electron on Earth yielding, at random, a definite clockwise spin for the Earth-bound electron, then the other electron—the one on Mars—must instantly spin definitely counterclockwise! *This instant result occurs no matter how far away the other electron is,* and it violates the speed of light requirement of relativity theory. That is why Einstein considered it to be "spooky action at a distance."

How is a philosopher to interpret these phenomena, which do not fit well with the usual philosophical accounts of cause and effect? Apparently, philosophers need to become creative—perhaps even daring—by questioning old, familiar foundational concepts that have formed the metaphysical bedrock of philosophy for centuries. For example, given the growing belief among physicists that the universe is an ocean of quantum information, and given Seth Lloyd's view that the universe behaves like a gigantic quantum computer, perhaps we could interpret superpositions as entities much like subroutines stored within the quantum computer/universe and waiting to be run. When the computer/universe randomly sends a bit of information to one of its subroutines, *that* subroutine is the one that gets run, while the others get erased or taken "off line." This would be the phenomenon called *decoherence,* which randomly "extracts" classical reality from an infinite source of possibilities generated by the underlying quantum computer/universe.

Given this suggested story, the *entanglement* of two quantum entities could be interpreted as the establishment of something very like a *hyperlink* connecting subroutines within the cosmic quantum computer. The "classical" world, including all physical objects and processes—perhaps even space-time and gravity—could be a projection or "virtual reality" generated by the cosmic quantum computer. The "laws of nature" of the classical world—such as Einstein's speed of light requirement—would then be part of the virtual reality projection; while "spooky action at a distance" would be the result of a "hyperlink" inside of the cosmic quantum computer—that is, *inside the underlying ocean of qubits which create our classical world through the process of decoherence*. In such a situation, there would be no need—and no way—to unite relativity and quantum mechanics, because they would exist in different worlds (or different parts of the same world). This is only one metaphysical speculation (my own) regarding the ultimate nature of the universe in our "Age of Information." Creative philosophers need to come up with many more stories until we find one that can be scientifically confirmed. *Metaphysicians, start your engines!*

Teleportation — Another quantum phenomenon that presents a challenge to traditional philosophy is called "teleportation," a process in which the quantum properties of one object are transferred instantly to another object by means of entanglement and measurement. Because the transfer of

properties takes place via entanglement, it occurs instantly no matter how far apart the objects might be in the classical world, and without the need to travel through space-time. The object which acquires the quantum properties of the original is rendered *identical to the original*, and the original is destroyed by measurement. (In some cases, some classical information also must be sent to the receiving station, using a traditional communication channel, to make adjustments in the recipient of the teleported properties and thereby assure that the recipient is identical to the original.) It is important to note that in teleportation it is *quantum information* that gets transferred*, not the matter/energy of the original object.* The recipient of the teleported quantum properties contains matter/energy that is not the original matter/energy of the donor object, but *the recipient is otherwise absolutely identical to the original*.

How should philosophers interpret these results? Is *the original entity* teleported, or merely an exact copy of it? If we agree with Norbert Wiener that all physical objects and processes are continually changing data structures, and not the matter/energy that happens to encode the data at a given moment (Bynum 2010), then *the teleported entity is actually the original data structure, and not merely a copy*. On the other hand, if Wiener's view is rejected, what is a better interpretation of quantum teleportation?

Quantum Computing — Because qubits can simultaneously be in many different states between 0 and 1, and because of the phenomenon of entanglement, quantum computers are able to perform numerous computing tasks at the very same time. As Vlatko Vedral explains,

> *any problem* in Nature can be reduced to a search for the correct answer amongst several (or a few million) incorrect answers. . . . [and] unlike a conventional computer which checks each possibility one at a time, quantum physics allows us to check multiple possibilities simultaneously. (Vedral 2010, 138, emphasis in the original)

Once we have learned to make quantum computers with significantly more than 14 qubits of input—which is the current state of the art—quantum computing will provide remarkable efficiency and amazing computing power! As Seth Lloyd has explained,

> A quantum computer given 10 input qubits can do 1,024 things at once. A quantum computer given 20 qubits can do 1,048,576 things at once. One with 300 qubits of input can do more things at once than there are elementary particles in the universe. (Lloyd 2006, 138-139)

For philosophy, such remarkable computer power has major implications for concepts such as *knowledge*, *thinking*, and *intelligence*—and, by extension, *artificial intelligence*. Imagine an artificially intelligent robot whose "brain" includes a quantum computer with 300 qubits. The "brain" of such a robot could do more things simultaneously than all the elementary particles in the universe! Compare that to the problem-solving abilities of a typical human brain. Or consider the case of so-called human "idiot savants"—who can solve tremendously challenging math problems "in their heads" instantly, or remember every waking moment in their lives, or remember, via a "photographic memory," every word on every page they have ever read. Perhaps such "savants" have quantum entanglements in their brains which function like quantum computers. Perhaps *consciousness* itself is an entanglement phenomenon. The implications for epistemology and the philosophy of mind are staggering!

6. The need to rethink the foundations of philosophy

In the June 2011 issue of *Scientific American*, Vlatko Vedral made a convincing case for the view that *quantum properties are not confined to tiny subatomic particles* (Vedral 2011). Most people, he noted, including even many physicists, make the mistake of dividing the world into two kinds of entity: on the one hand, tiny particles which are quantum in nature; and on the other hand, larger "macro" objects, which obey the classical laws of physics, including relativity.

> Yet this convenient partitioning of the world is a myth. Few modern physicists think that classical physics has equal status with quantum mechanics; it is but a useful approximation of a world that is quantum at all scales. (Vedral 2011, 38 and 40)

Vedral went on to discuss a number of "macro" objects which apparently have exhibited quantum properties, including, for example, (1) entanglement within a piece of lithium fluoride made from trillions of atoms, (2) entanglement within European robins who use it to guide their yearly migrations of 13,000 kilometers between Europe and central Africa, and (3) entanglement within plants that use it to bring about photosynthesis.

Given what has been said above, and given all the important developments in the information revolution that is happening within physics today, *it is time for philosophers to awaken from their metaphysical slumbers and join the Information Age!*

*An earlier version of this paper was the 2011 Preston Covey Address at the IACAP2011 conference in Aarhus, Denmark.

References

Bynum, Terrell Ward. 2011. The historical roots of information and computer ethics. In *The Cambridge Handbook of Information and Computer Ethics,* ed. Luciano Floridi. Cambridge University Press.

Chiribella, Giuli; D'Ariano, Giacomo; Perinotti, Paolo. July 2011. Informational derivation of quantum theory. *Physical Review A.* 84.

Floridi, Luciano. 2011. *The Philosophy of Information*. Oxford University Press.

Lloyd, Seth. 2006. *Programming the Universe: A Quantum Computer Scientist Takes on the Universe*. Alfred A. Knopf.

Moor, James H. 1998. Reason, relativity and responsibility in computer ethics. *Computers and Society* 28(1):14-21.

Vedral, Vlatko. 2010. *Decoding Reality: The Universe as Quantum Information*. Oxford University Press.

Vedral, Vlatko. 2011. Living in a quantum world. *Scientific American* June:38-43.

Wheeler, John A. 1990. Information, physics, quantum: the search for links. In *Complexity, Entropy, and the Physics of Information,* ed. W. Zurek. Addison-Wesley.

Wiener, Norbert. 1948. *Cybernetics: or Control and Communication in the Animal and the Machine*. MIT Press.

Wiener, Norbert. 1950, 1954. *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin, First Edition; Doubleday Anchor Books, Second Edition Revised.

Zeilinger, Anton. 2010. *Dance of the Photons: From Einstein to Teleportation*. Farrar, Straus, and Giroux.

Hyperhistory and the Philosophy of Information Policies

Luciano Floridi
*University of Hertfordshire and University of Oxford**

1. Preface

I am hugely indebted to Terry Bynum's work. Not merely for

his kind and generous acknowledgement of my efforts to establish a philosophy of information but, way more seriously and significantly, because of his ground-breaking work, which opened new research paths to philosophers of my generation, especially, but not only, in computer ethics.

I suppose the best way to honor his work is probably by trying to contribute to it. In this short article, I shall attempt to do so by taking seriously two important points made in Bynum's article. One is his question: "How is it possible for information technology (IT) to transform our world so quickly and so fundamentally?" The other is his exhortation: "we need to bring philosophy into the Information Age […]. We need to rethink the bedrock foundations of philosophy that were laid down hundreds of years ago by philosophers like Hobbes, Locke, Hume, and Kant. Central philosophical concepts should be re-examined […]." I shall accept Bynum's exhortation. And I shall try to contribute an answer to his question by calling the reader's attention to the need to reconsider our philosophy of politics, our philosophy of law, and our philosophy of economics, in short, to the need of developing a philosophy of information policies for our time. The space is of course limited, so I hope the reader will forgive me for some simplifications and sweeping remarks that will deserve much more careful analysis in a different context.

2. Hyperhistory

More people are alive today than ever before in the evolution of humanity. And more of us live longer and better today than ever before. To a large measure, we owe this to our technologies, at least insofar as we develop and use them intelligently, peacefully, and sustainably.

Sometimes, we may forget how much we owe to flakes and wheels, to sparks and ploughs, to engines and satellites. We are reminded of such deep technological debt when we divide human life into prehistory and history. That significant threshold is there to acknowledge that it was the invention and development of information and communication technologies (ICTs) that made all the difference between who we were and who we are. It is only when the lessons learnt by past generations began to evolve in a Lamarckian rather than a Darwinian way that humanity entered into history.

History has lasted six thousand years, since it began with the invention of writing in the fourth millennium BC. During this relatively short time, ICTs have provided the *recording* and *transmitting* infrastructure that made the escalation of other technologies possible. ICTs became mature in the few centuries between Guttenberg and Turing. Today, we are experiencing a radical transformation in our ICTs that could prove equally significant, for we have started drawing a new threshold between history and a new age, which may be aptly called *hyperhistory*. Let me explain.

Prehistory and history work like adverbs: they tell us *how* people live, not *when* or *where*. From this perspective, human societies currently stretch across three ages, as ways of living. According to reports about an unspecified number of uncontacted tribes in the Amazonian region, there are still some societies that live prehistorically, without ICTs or at least without recorded documents. If one day such tribes disappear, the end of the first chapter of our evolutionary book will have been written. The greatest majority of people today still live historically, in societies that rely on ICTs to *record* and *transmit* data of all kinds. In such historical societies, ICTs have not yet overtaken other technologies, especially energy-related ones, in terms of their vital importance. Then there are some people around the world who are already living hyperhistorically, in societies or environments where ICTs and their data *processing* capabilities are the necessary condition for the maintenance and any further development of societal welfare, personal well-being, as well as intellectual flourishing. The nature of conflicts provides a sad test for the reliability of this tripartite interpretation of human evolution. Only a society that lives hyperhistorically can be vitally threatened informationally, by a cyber attack. Only those who live by the digit may die by the digit.

To summarize, human evolution may be visualized as a three-stage rocket: in prehistory, there are no ICTs; in history, there are ICTs, they record and transmit data, but human societies depend mainly on other kinds of technologies concerning primary resources and energy; in hyperhistory, there are ICTs, they record, transmit, and, above all, process data, and human societies become vitally dependent on them and on information as a fundamental resource.

If all this is even approximately correct, the emergence from its historical age represents one of the most significant steps taken by humanity for a very long time. It certainly opens up a vast horizon of opportunities, all essentially driven by the recording, transmitting, and processing powers of ICTs. From synthetic biochemistry to neuroscience, from the Internet of things to unmanned planetary explorations, from green technologies to new medical treatments, from social media to digital games, our activities of discovery, invention, design, control, education, work, socialization, entertainment, and so forth would be not only unfeasible but unthinkable in a purely mechanical, historical context.

It follows that we are witnessing the outlining of a macroscopic scenario in which an exponential growth of new inventions, applications, and solutions in ICTs are quickly detaching future generations from ours. Of course, this is not to say that there is no continuity, both backward and forward. Backward, because it is often the case that the deeper a transformation is, the longer and more widely rooted its causes are. It is only because many different forces have been building the pressure for a very long time that radical changes may happen all of a sudden, perhaps unexpectedly. It is not the last snowflake that breaks the branch of the tree. In our case, it is certainly history that begets hyperhistory. There is no ASCII without the alphabet. Forward, because it is most plausible that historical societies will survive for a long time in the future, not unlike the Amazonian tribes mentioned above. Despite globalization, human societies do not parade uniformly forward, in synchronic steps.

3. The philosophy of information policies

Given the unprecedented novelties that the dawn of hyperhistory is causing, it is not surprising that many of our fundamental philosophical views, so entrenched in history, may need to be upgraded, if not entirely replaced. Perhaps not yet in academia, think tanks, research centers, or R&D offices, but clearly in the streets and online, there is an atmosphere of confused expectancy, of exciting, sometimes naïve, bottom-up changes in our views about (i) the world, (ii) ourselves, (iii) our interactions with the world, and (iv) among ourselves.

These four focus points are not the result of research programs, or the impact of successful grant applications. Much more realistically and powerfully, but also more confusedly and tentatively, the changes in our *Weltanschauung* are the result of our daily adjustments, intellectually and behaviorally, to a reality that is fluidly changing in front of our eyes and under our feet, exponentially, relentlessly. We are finding our new balance by shaping and adapting to hyperhistorical conditions that have not yet sedimented into a mature age, in which novelties are no longer disruptive but finally stable patterns of "more of

approximately the same" (think, for example, of the car or the book industry, and the stability they have provided).

It is for this reason that the following terminology is probably inadequate to capture the intellectual novelty that we are facing. As Bynum rightly stressed, our very conceptual vocabulary and our ways of making sense of the world (our semanticising processes and practices) need to be reconsidered and redesigned in order to provide us with a better grasp of our hyperhistorical age, and hence a better chance to shape and deal with it. With this proviso in mind, it seems clear that a new philosophy of history, which tries to makes sense of our age as the end of history and the beginning of hyperhistory, invites the development of (see the four points above) (i) a new philosophy of nature, (ii) a new philosophical anthropology, (iii) a synthetic e-nvironmentalism as a bridge between us and the world, and (iv) a new philosophy of politics among us.

In other contexts, I have argued that such an invitation amounts to a request for a new philosophy of information that can work at 360 degrees on our hyperhistorical condition (Floridi 2011). I have sought to develop a philosophy of nature in terms of a philosophy of the infosphere (Floridi 2003), and a philosophical anthropology in terms of a fourth revolution in our self-understanding—after the Copernican, the Darwinian, and Freudian ones—that re-interprets humans as informational organisms living and interacting with other informational agents in the infosphere (Floridi 2008; 2010). Finally, I have suggested that an expansion of environmental ethics to all environments—including those that are artificial, digital, or synthetic—should be based on an information ethics for the whole infosphere (Floridi forthcoming). What I have not done but I believe to be overly due is to outline a philosophy of information policies consistent with such initial steps, one that can reconsider our philosophical views of economics, law, and politics in the proper context of the hyperhistorical condition and the information society.

4. Conclusion

Six thousand years ago, a generation of humans witnessed the invention of writing and the emergence of the State. This is not accidental. Prehistoric societies are both ICT-less and stateless. The State is a typical historical phenomenon. It emerges when human groups stop living in small communities a hand-to-mouth existence and begin to live a mouth-to-hand one, in which large communities become political societies, with division of labor and specialized roles, organized under some form of government, which manages resources through the control of ICTs. From taxes to legislation, from the administration of justice to military force, from census to social infrastructure, the State is the ultimate information agent and so history is the age of the State.

Almost halfway between the beginning of history and now, Plato was still trying to make sense of both radical changes: the encoding of memories through written symbols and the symbiotic interactions between individual and *polis*-State. In fifty years, our grandchildren may look at us as the last of the historical, State-run generations, not so differently from the way we look at the Amazonian tribes, as the last of the prehistorical, stateless societies. It may take a long while before we shall come to understand in full such transformations, but it is time to start working on it. Bynum's invitation to "bring philosophy into the Information Age" is most welcome.

* Research Chair in Philosophy of Information, and UNESCO Chair in Information and Computer Ethics, University of Hertfordshire; Faculty of Philosophy and Department of Computer Science, University of Oxford.

Address for correspondence: Department of Philosophy, University of Hertfordshire, de Havilland Campus, Hatfield, Hertfordshire AL10 9AB, UK; l.floridi@herts.ac.uk

References

Floridi, L. 2003. On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology* 4(4):287-304.

Floridi, L. 2008. Artificial intelligence's new frontier: artificial companions and the fourth revolution. *Metaphilosophy* 39(4/5):651-55.

Floridi, L. 2010. *Information - a Very Short Introduction*. Oxford, Oxford University Press.

Floridi, L. 2011. *The Philosophy of Information*. Oxford, Oxford University Press.

Floridi, L. Forthcoming. *Information Ethics*. Oxford, Oxford University Press.

Is Ethics Headed for Moral Behaviorism and Should We Care?

Anthony F. Beavers
The University of Evansville

> *The righteous are responsible for evil before anyone else is. They are responsible because they have not been righteous enough to make their justice spread and abolish injustice: it is the fiasco of the best which leaves the coast clear for the worst.*
>
> Levinas (1976/1990, 186), paraphrasing the prophet Ezekiel

A Provocation

I start with a premise that may appear at first as a moral imperative: if it is within our power to build a machine that can make human beings more moral, both individually and collectively, then we have a *prima facie* moral obligation to build it. Objections to this claim are, of course, tenable, though they may assume particular conceptions of ethics that have historically carried great credibility, but whose credibility we might have new reason to doubt. Some of these objections are apparent if we substitute the word "nation" with "machine" and claim that if it is within our power to build a *nation* that can make human beings more moral, then we have a *prima facie* obligation to build *it*. While this claim, too, may at first seem intuitively correct, it could prove objectionable if the most direct way to build such a state requires totalitarianism or, minimally, an overly-coercive state that punishes moral (and not merely legal) wrongdoers. We thus find ourselves at the nexus of several inter-related issues, including not only how to determine in a precise way what is morally correct, but also the role that freedom plays in moral culpability. If a total nation-state holds individuals at gun point and demands that they act morally under pain of death, their actions are no more deserving of reward than they would be deserving of punishment if at gun point they were made to act immorally.

Indeed, it is a common ethical assumption, in the West at least, that someone can be morally praised or blamed (that is, culpable) only for actions that are in their power to do or refrain from doing. Thus, a good character in virtue ethics is only worthy of respect because it is in the power of individuals to sculpt their own characters, and in Kantian ethics, moral praise and blame can only be attributed to creatures that are free. Such an assumption, however, itself becomes problematic if we rearrange our initial premise a bit and suggest that if it is in our power to design human beings genetically to be moral, then we have a *prima facie* obligation to do so. In this case, humans might still choose the right course of action with the same feeling of freedom that we do, but only because they are engineered to do so. That some among us would object to such a course of action is readily apparent in the fact that many find Huxley's *Brave New World* a piece of dystopian, and not utopian, fiction. Furthermore, the theological among us might worry that if it is morally imperative to engineer moral human

beings, then God must have made a tragic mistake in the first place by making us the way he did.

New possibilities from research in computational machinery and bio-engineering are raising a daring question: Are we not morally required to engineer a moral world, whether by deference to moral machines, social engineering, or taking control over our biology? When we consider the great lengths we go to in training a child by nurturing guilt and a sense of shame (scolding, for instance), fighting, even killing, in (so called) moral wars, punishing and rewarding wrongdoers accordingly, sanctioning acceptable conduct in our institutions through mechanisms of law, etc., such a question does not seem misplaced. It is as if we want to create a moral world, but in the most difficult, unproductive, and possibly even immoral way possible. History itself bears testimony to our failure: witness the fact that the U.S. is quickly approaching involvement in the longest war in its history contrasted against the fact that most Americans are barely aware that we are fighting at all and seem to have lost any interest in seeing it come to an end. Furthermore, even if this war were to end, we collectively characterize war in general as inevitable, which means also that we have accepted it as unavoidable. Arriving at this point is simply to have given up on the matter. But, to be fair to ethics, this fatalism (or indifference) must itself be seen as a serious moral transgression—one that is only apparently, but not actually, banal—if there is in fact something we can do to fix the situation. Should we, at this point in history, start to think seriously about putting an end to our moral indecency? Might Huxley's *Brave New World* or some variant thereof be utopian after all? What should the world look like morally, given that technology is slowly giving us the power to shape it as we wish, and would it be worth the cost if developing a moral world meant abandoning several cherished assumptions about ethics?

The goal of ethics is to make itself obsolete, hopefully, though, by fulfillment in moral community and not by just defining it out of existence. Yet, current trends in technology and, more broadly, in society seem to be leaning toward the latter. Ethics, traditionally conceived, is under attack on several fronts. Yet, given its historical failure, we must wonder whether it is worth saving. I'm beginning to think not. The goal of the rest of this essay is to say why.

Honestly, Is Honesty a Virtue?

Temperance, courage, wisdom, and justice made it into Plato's list of virtues in the *Republic,* but, ironically, the author of the cave allegory did not include honesty. Yet, as his text clearly shows, this was no oversight, since honesty is necessary for avoiding self-deception and is thus necessary for the named virtues as well. Self-deception is quite hard to avoid, even in matters of epistemology and especially in ethics. In this spirit, Dennett says of the frame problem that it is not merely "an annoying technical embarrassment in robotics," but "on the contrary, that it is a new, deep epistemological problem—accessible in principle but unnoticed by generations of philosophers—brought to light by the novel methods of AI, and still far from being solved" (1984, 130). More recently, he remarked that AI "makes philosophy honest" (2006). In a similar vein after citing this last quote from Dennett, Anderson and Anderson observe that "ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas" (2007, 16). In this light, it is common among machine ethicists to think that research in computational ethics extends beyond building moral machinery because it helps us better understand ethics in the case of human beings. This is because of what we must know about ethics in general to build machines that operate within normative parameters. Unclear intuitions are unworkable where engineering specifications are required.

Implicit in this observation is the notion that *ought implies implementability*. Admittedly, this claim looks counter-intuitive at first blush, but it is a logical extension of the Kantian notion that *ought implies can* properly situated by the possibility of moral machinery. "Can" in this context means that one must have the ability to *x*, before we can claim that one ought to *x*. This, in turn, implies that the behavioral recommendations of any moral theory must fall within the power of an agent to perform, or, in other words, that the theory itself must be able to be implemented, whether in wetware or hardware. Consequently, computational ethics sets a criterion for evaluating the tenability of moral theories. If it can be shown that a particular theory cannot be physically implemented, whether for logical or empirical reasons, we are justified in claiming that that theory insofar as it is a *moral* theory is untenable.

Initially, this might sound well and good if it weren't for the fact that such a criterion poses serious problems for Kantian deontology and classical utilitarianism, because they both run into moral variants of the frame problem and are therefore not implementable. (For further discussion on Kant, see Beavers 2009.) Without rehearsing the full arguments here, a quick sketch might be sufficient to get the point across.

Kant's universalization formula of the categorical imperative says "Act as if the maxim of your action were to become through your will a universal law of nature" (1785/1994, 30), where a maxim is defined as "the subjective principle of acting." It is the rule that I employ as a subject when acting individually, and it is moral if and only if I can at the same time permit any agent in the same situation to employ the same maxim. The problem here is that the possibility of universalization depends on the scope I set for the maxim. If the subject is defined as a class of one (i.e., anyone exactly like me in exactly my particular situation), any maxim will universalize, and thus every action could be morally permissible. To avoid this conclusion one must find a non-arbitrary way to establish the legitimate scope of a maxim that should be taken into account. The prospects for doing so objectively seem poor without simultaneously begging the question.

Similarly, Mill runs into problems with the principle of utility where "actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness" (1861/1979, 7). As is commonly known, Mill does not mean the promotion of my happiness and the reduction of my private pains. He means those of the (global?) community as a whole. Because the success of an action hangs on future states that are wholly unknown to the agent, the principle of utility is computationally intractable. Without some specification of the scope, it is impossible to know whether any particular action promotes or impedes happiness across the whole. The worst atrocities might, over time, turn out to maximize happiness, while the kindest gestures to some could lead to tragic consequences for others.

Utilitarianism might be salvageable by modifying it into some computationally tractable form . . . maybe. It is too soon to say, but I have my doubts about Kant, *pace* Powers, who has made a worthy attempt to save him by treating the categorical imperative in its various forms as heuristics for behavior rather than strict rules (2006). This approach, I worry, leads to problems of its own, such as losing the objective criterion for determining precisely when a behavior is moral which the categorical imperative was meant to provide. (If the categorical imperative is a heuristic, what is the algorithm for which it provides the short cut?) But I have deeper worries about Kant that I have presented elsewhere (2009 & 2011b) and that are appropriate to repeat here.

For reasons that should be clear from the above, *ought cannot imply must*. That is, if it is impossible for me to refrain from an action, then the notion of ought does not apply. (This is why angels and animals are not moral agents in Kant's moral architecture.) Said in other words, *ought implies might not*. However, if so, then we are heading for an uncomfortable situation that I have identified as "the paradox of automated moral agency" or P-AMA (2011b). In brief, it starts with a few definitions, followed by a question and then an argument. The definitions are intended to avoid starting with question-begging biases. Thus,

> {*def MA*} any agent that does the right thing morally, however determined.

In stating the definition in this way, we do not imply any moral evaluation or theory of moral behavior. We do so in order to clear room for the question just intimated. Having defined an MA neutrally, we can now distinguish between responsible moral agents (RMAs) and artificial moral agents (AMAs). In turn, the notion of an RMA is intentionally morally loaded to fit traditional assumptions about what it means for an agent to be worthy of moral praise or blame for its actions.

> {*def RMA*} an MA that is fully responsible and accountable for its actions.

It can decide things for itself and so may do or refrain from doing something using its own discretion. Because it is the cause of its own behavior it can be morally culpable. Finally, to return to a more neutral definition:

> {*def AMA*} a manufactured MA that may or may not be an RMA.

Regardless of the technical possibilities of current research in artificial moral agency and whether we are disposed to think that an RMA can be the only genuine kind of MA, we can now ask the important question, *should an AMA be an RMA*, assuming it possible for us to make one so. If we cling to the notion of responsibility assumed thus far, the answer would seem to be no.

Given that the need to make a machine an MA in the first place stems from the fact that such machines are autonomous, that is, they are self-guided, rather than act by remote control, we run into a paradox, P-AMA, which says:

1) If we are to build autonomous machines, we have a *prima facie* moral obligation to make them RMAs, that is, agents that are responsible and able to be held responsible for their actions.
2) For an RMA to be responsible and able to be held responsible for its actions, it must be capable of both succeeding and failing in its moral obligations.
3) An AMA that is also an RMA must therefore be *designed* to be capable of both succeeding and failing in its moral obligations.
4) It would be a moral failure to unleash upon the world machines that are capable of failing in their moral obligations.
5) Therefore, we have a moral obligation to build AMAs that are not also RMAs.

P-AMA might be escapable as a paradox by simply denying premise 1, but doing so might not be as easy as it first appears, mostly because of the technical aspects involved with "autonomy" as it applies to machinery. A full discussion of the point exceeds the scope of this paper, but the problem can quickly be summarized by noting that as the world becomes increasingly automated, machines are being left to "decide" things on their own. Internet routers and the switches on the U.S. power grid do so to help with load balancing, the automatic braking system on my car does, and even my dishwasher and dryer do, since neither stop until they sense that the job is done. Such machines interact with environmental cues that may in certain circumstances lead to dire consequences. More pressingly, advances in auto-generative programming allow machines to write their own code, often producing innovative and unpredictable results. To set such machines free on the world without building in moral constraints would simply be irresponsible on the part of their designers, but to anticipate every contingency is not possible either. So these constraints themselves have to autonomously decide things as well. In short, they must be able to evaluate situations and use some procedure to act in morally acceptable ways.

The issue is pulled into greater focus when we address the question of who is to blame when such machines fail. If they are autonomous and left to their own devices, blaming their creators would seem to be cruel and no more justified than blaming parents for the moral failures of their children or God, for that matter, for the failures of the free creatures that he unleashes on the world. We could, of course, argue that the creators of such machines should not make them autonomous in the first place, but this is tantamount to arguing that parents should not have children or that God should not have made his creatures autonomous either.

The real issue with the paradox here points, I believe, to a problem with our traditional notion of moral responsibility. To be consistent, if we cannot morally want machines to be RMAs as opposed to non-responsible MAs, we cannot want humans to be either. Moral responsibility in this light appears to be a solution of last resort for "fallen creatures." Since I am not theistically inclined, I have no stake in either exonerating or indicting God, but the matter does speak to the point that responsibility and accountability, when they carry the weight of moral praise and blame that we attach to them, are necessarily correlated with the notion that we, humans, are morally broken. If we can repair the situation, we ought to; seriously . . . we physicians ought to heal ourselves . . . if we can.

Non-Responsible Moral Agents . . . Really?

The notion of a non-responsible moral agent is not coherent if we assume conventional conceptions of responsibility or see it as a necessary part of the moral enterprise. But it seems that the definition of moral responsibility is being reduced to causal responsibility by challenges on several fronts. This is to say that *x* is responsible for *y* means only that *x* is the precipitating cause of *y*. This shift of focus in matters of morals is visible in the conflation between ethics and codes of conduct that we see in several of our institutions, in the notion that immoral behavior results from neurological deficit embraced by several neuroscientists (and sometimes by our courts), and in the advent of moral machinery. The bottom line, it seems, is not the need to have agents to blame, but the need to have immoral behavior cease. In other words, the social problem of ethics is to create (or encourage) agents, whether human or otherwise, to behave morally. The coercion of moral behavior, whether by the promise of rewards or punishments, is but one means to this end (and one, we must admit, that is sometimes effective and sometimes not).

In 2011a, 2011b, and 2011c, I advanced what I called the "sufficiency argument." It is intimated here already. The argument maintains that the kind of moral interiority necessary for an agent to be an RMA is a sufficient though not necessary condition for being an MA. Therefore, moral interiority is not essential for moral agency. One corollary of the argument is that there are other (and perhaps more effective) ways to be an

MA that do not require the internal psychological components involved in conscience, guilt, shame, etc. Advancing this position seriously is really to do nothing other than pinpoint the direction that ethics is already heading: the general focus of our moral regard is no longer the salvation of the individual soul, but individual behavior, properly contextualized, insofar as it has a moral impact on our social situation. To have come this far, however, is already to have wreaked havoc on the historical foundations of ethics, (again) at least in the West.

To make this clear, in 2011c, I invited the reader to consider the headline "First Robot Awarded Congressional Medal of Honor for Incredible Acts of Courage on the Battlefield." I then asked, "What must we assume in the background for such a headline to make sense without profaning a nation's highest award of valor? Minimally, fortitude and discipline, intention to act while undergoing the experience of fear, some notion of sacrifice with regard to one's own life, and so forth, for what is courage without these things? That a robot might simulate them is surely not enough to warrant the attribution of virtue, unless we change the meaning of some terms." At the time of that writing, I was worried that we, as a species (meaning irrespective of the concerns of professional ethicists), were in the midst of an inevitable entry into a post-ethical age. In a sense, I still think we are, but it might be better to put this in Nietzschean terms and say that we are tacitly in the process of revaluing value. The ethical landscape is transforming at its very roots as we are forced by new technological possibilities and life in a highly connected world to recognize a plurality of lifestyle choices, religious (and non-religious!) commitments, and political ideologies. Whether this leads to relativism is besides the point; the problem we must face is whether we can find a way to work together to solve some very pressing problems that the species is just beginning to confront without destroying ourselves in the process. This change of moral focus from the individual soul to the common good now seems to me to be a positive step in the right direction, even if it amounts to a *no-fault ethics*. Indeed, this is what I mean by non-responsible moral agency; pointing fingers gets us nowhere when there is serious work to be done.

Fortunately, information ethics (IE), as advanced by Floridi, starts in the right direction with a macro-ethics that might best be described as an eco-informational environmentalism. Floridi's views are spread across several papers and will soon be released as a book, *Information Ethics,* the second volume of a quadrilogy on the philosophy of information, which will comprise part of an intricate system of philosophical overhaul. Thus, a detailed treatment is not possible here. To paint the picture in broad strokes though, Floridi advocates following the lead of environmental ethics by shifting our focus from the agent in a moral situation to the patient. This move is in direct contrast to virtue ethics, which focuses its attention on the character of the subject, but it is also in contrast to utilitarianism, deontology and contractarianism, which, though relational, tend to "treat the relata, i.e., the individual agent and the individual patient, as secondary importance" (1999, 41), by putting their focus on the action itself. Additionally, they (including virtue ethics here) are also anthropocentric in the sense that they view ethics primarily as a matter of managing relations between human beings. This contrasts strongly with Land Ethics, where the environment itself can become a patient worthy of our moral regard because it is intrinsically valuable and not just valuable *for us*. Following this lead, Floridi advocates an "object-oriented and ontocentric theory" (1999, 43) that extends our moral concern to "anything that exists."

While I must confess that, on first encountering this view, my moral sensibilities were offended by a theory that seems not to be able to distinguish between persons and things, I have come to appreciate what is going on at a deeper level: by broadening our moral regard to include non-human, indeed, non-living, things, we also broaden the concept of "harm" to that of "damage" (Floridi 2002). This view squares well with the no-fault ethics mentioned above insofar as harm invites compensation whereas damage invites repair. In traditional views, if we harm a person, justice demands compensation, but "harming" a painting only makes sense by extension of metaphor. We cannot pay recompense to a painting for its pain and suffering. We can, however, see to its repair. This shift of focus from harm to damage invites us to fix problems rather than place blame. It is in this spirit that moral behaviorism starts to make sense.

Setting aside the motives, drives, and desires of moral agents to focus on the damage that they do and the repairs that they (or others) can make gets us to what really matters in ethics. Once again, the point of ethics is not grounded in the need to have agents to blame, but in the need to make immoral behavior cease. The "whys" and "what fors" are beside the point, though, for those who wish to preserve them, they may do so with limited concession, as I shall demonstrate momentarily. Indeed, I regard the possibility of their preservation as one of the benefits of moral behaviorism.

Getting Practical about Moral Philosophy

In their book *Moral Machines: Teaching Robots Right from Wrong*, Wallach and Allen call attention to a problem that morally demands a change of perspective from traditional ethics to something more along the lines of the above. This demand is forced by new possibilities regarding emerging technologies, though in some sense it might always have been in the waiting. They write:

> Companies developing AI are concerned that they may be open to lawsuits even when their systems enhance human safety. Peter Norvig of Google offers the example of cars driven by advanced technology rather than humans. Imagine that half the cars on U.S. highways are driven by (ro)bots, and the death toll decreases from roughly forty-two thousand a year to thirty-one thousand a year. Will the companies selling those cars be rewarded? Or will they be confronted with ten thousand lawsuits for deaths blamed on the (ro)bot drivers? (207)

Given our current ethical and legal climate, companies are right to be concerned that their technologies to improve our world may shift the burden of responsibility from others to themselves. Yet, from a patient-centered point of view, this demonstrates precisely what is wrong with approaching ethics from a traditional, agent-oriented perspective, since it should be clear that if we can save ten thousand lives by employing autonomous vehicles we ought to do so, regardless of where this places responsibility and accountability. Some forgiveness here is in order. In cases such as this, the traditional, fault-oriented perspective gets in the way of doing the right thing. As more technologies with possible positive ethical consequences emerge, this problem will inevitably become a greater concern we will have to address.

There is room to be concerned as well about what happens to individual responsibility and accountability if we fail to defer appropriately to certain machines. In 2011b, I put forth a thought experiment involving MorMach, an all knowing moral machine, the ultimate oracle in all matters concerning ethics, in order to illustrate the emerging possibility that we might one day transcend our faulty neural wiring and hormone control systems by deference to a machine that is better at ethics than we are.

If such a machine were to exist, would not ethics itself require our deference, even in cases where our conscience, an affective component of our frail biology after all, might disagree? Suppose MorMach were widely employed across every sector of society, including, for instance, the medical profession. Where should we place the blame if a physician were to follow his conscience against the advice of MorMach and end up engaged in an action with serious negative consequences? On a traditional approach to ethics, it would seem that fault in this case would fall to the physician who should have let the AMA do the moral work for him. Speculating about the future is dangerous business, but I suspect that if MorMach were a reality, the courts would inevitably agree. In this light, we may wonder whether one day moral failures will be indistinguishable from other kinds of failures, like, for instance, not prescribing a medication according to the advice of established medical practice or failing to follow an owner's manual regarding warnings when using various tools.

Practically speaking, these examples suggest that ethics requires us to acknowledge human limitations when confronting moral matters. Being able to be morally successful, and therefore worthy of praise, only because it is possible for us to be immoral, is not, as Kant thought, a sign of the dignity of the human being, but the sign of an ethics that assumes human beings to be broken from the start. In this light, we should take care to see that ethics becomes behavior-oriented.

Finally, to deliver on the promise made in the last paragraph of the previous section, the sufficiency argument allows us to approach moral behaviorism without entirely dismissing the several motivations that come from inherited ethical and religious tradition. To remind the reader, the sufficiency argument maintains that the kind of moral interiority necessary for an agent to be an RMA is a sufficient though not necessary condition for being an MA. Therefore, moral interiority is not essential for moral agency. It is not *essential*, but this is not to say that it is not helpful, particularly for beings constituted like us. Of course, what is true for sufficient conditions in general is also true for this one. This is to say that there may be (and are, I believe) a number of sufficient conditions that will lead one to being an MA; several existing moral beliefs and systems are, no doubt, among them. All are fine and acceptable, as long as the necessary condition for being an MA is met, and this is, straightforwardly, moral behavior. Used in this way, the sufficiency argument permits a plurality of paths to moral objectives based on a singular necessary condition. Perhaps this pluralism of motivation can get us all on the same page regarding moral behavior without having to reach agreement about incidentals that often clutter ethical debate. Perhaps this is what we need in a quickly globalizing moral community.

References

Anderson, M., and Anderson, S. 2007. Machine ethics: creating an ethical intelligent agent. *AI Magazine* 28(4):15-26.

Beavers, A. 2009, March. Between angels and animals: the question of robot ethics, or is Kantian moral agency desirable? Association for Practical and Professional Ethics, Eighteenth Annual Meeting, Cincinnati, Ohio.

Beavers, A. 2010. Editorial to *Robot ethics and human ethics*. Special issue of *Ethics and Information Technology* 12(3):207-208.

Beavers, A. 2011a, July. Is ethics computable, or what other than *can* does *ought* imply? Presidential Address at the Annual International Association for Computing and Philosophy Conference, Aarhus University, Aarhus, Denmark.

Beavers, A. 2011b, October. Could and should the ought disappear from ethics? International Symposium on Digital Ethics, Loyola University, Chicago, Illinois.

Beavers, A. 2011c. Moral machines and the threat of ethical nihilism. In *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. Lin, P., Bekey, G., and Abney, K., 333-344. Cambridge, MA: MIT Press.

Dennett, D. 1984. Cognitive wheels: the frame problem of AI. In *Minds, Machines and Evolution*, ed. Hookway, C., 129-151. Cambridge, UK: Cambridge University Press.

Dennett, D. 2006, May. Computers as prostheses for the imagination. The International Computers and Philosophy Conference. Laval, France.

Floridi, L. 1999. Information ethics: on the philosophical foundation of computer ethics. *Ethics and Information Technology* 1:37-56.

Floridi, L. 2002. On the intrinsic value of objects and the infosphere. *Ethics and Information Technology* 4:287-304.

Kant, I. 1785/1994. *Grounding for the Metaphysics of Morals*, ed. and trans. Ellington, J. Indianapolis: Hackett Publishing Company.

Levinas, E. 1976/1990. Damages due to fire. In *Nine Talmudic Readings by Emmanuel Levinas*, ed. Aronowicz, A., 178-197. Bloomington, IN: Indiana University Press.

Mill, J. S. 1861/1979. *Utilitarianism*. Indianapolis: Hackett Publishing Company.

Powers, T. 2006. Prospects for a Kantian machine. *IEEE Intelligent Systems* 1541-1672:46-51.

Wallach, W. and Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford, UK: Oxford University Press.

The Artifactualization of Reference and "Substances" on the Web: Why (HTTP) URIs Do Not (Always) Refer nor Resources Hold by Themselves

Alexandre Monnin

Université Paris 1 Panthéon-Sorbonne (PhiCo, EXeCo), Institut de Recherche et d'Innovation (IRI) du Centre Pompidou, INRIA (Wimmics), CNAM (DICEN)

"we now have to pay our way in order to subsist"[1]

(B. Latour)

Introduction

From an architectural point of view, the Web can be conceived as an information space full of URIs—Web identifiers. Contrary to popular belief it is *not* a traditional hypertext linking documents or "pages" to one another. Indeed, to account for all the situations encountered on the Web (Web services, dynamic pages, applications, feeds, content negotiation, etc.), a more encompassing theory was needed. According to the latter (the REST style of architecture), Web identifiers have to be treated as *dereferenceable proper names*—URIs (Uniform resource Identifiers), instead of the more well-known URLs (Uniform Resource Locators).

URIs are especially interesting for philosophers. Like proper names, a concept central both to the philosophy of language and metaphysics, they seem to refer to an object. If the architecture of the Web retains some of their characteristics, then philosophers are no longer facing a *terra incognita* but rather a familiar landscape. Unlike proper names, however, URIs also give access to Web contents. As such, they betoken an important change, from a symbolical dimension, where proper names are bestowed certain functions and used to solve philosophical conundrums regarding identity, to a technological one, to quote the late German media theorist Friedrich Kittler, where they earn new functionalities and act as the pillar of a world-wide information system.[2]

This shift is what we call *artifactualization*,[3] the becoming-artifact of philosophical concepts. Our first goal in this paper is to show that reference, the frail symbolic relation between a sign and its referent, is turned into something entirely different on the Web, the space *between* referent and reference, the relation itself, being *adjusted* so as to warrant that reference doesn't fail.

Our second goal is to deal in the same movement with the correlate of URIs, "resources." About ten years after the birth of the Web, it was understood/decided, after careful analysis, that its architecture was a *resource*-oriented one. A very paradoxical move inasmuch as resources are *not* accessible *per se*. But a most important one since it provided the URIs a means to identify "anything at all." Things on the Web, outside of the Web, chairs, people, rates, square circles, etc. The introduction of resources can be seen as a potent way to reopen the ontological question afresh.

Yet, it must also be understood that while resource can be anything, they also share very specific characteristics which have not been properly identified. Drawing from Kittler once again, we could say that the concept of an object for philosophers from Goclenius, Lohardus, and Suarez to Kant to Bretano, Twardowki, and Meinong, belonged to the symbolic realm while the very notion of a resource belongs to the technical realm as well, born as it was out of an effort to restore consistency to a technical project.

As the Web is spreading and becoming more ubiquitous day after day, we witness an interesting change whereby objects are becoming resources. From an online document to a person or an RDFID-enhanced product or device, they are everywhere—or *everyware*, to borrow designer Adam Greenfield's portmanteau word.

Interestingly, on the surface resources share many aspects with what used to be the dominant ontological conception of objects for centuries: substance. However, unlike substances, the category of resource is no longer a natural one. The function of substances was to explain how things like people, organisms, or artifacts persisted over time. Without such an ontological background, the issue remains open. We will see that on the Web, resource persistence has a *cost* which has to be assumed by a publisher and depends on protocols and standards. Overall, this will lead to a completely different ontological framework. One that is gaining more and more traction insofar as the network expands.

I. From Web pages to resources

It has been said that the new digital continent opened new perspective for ontology. "Not since the first work of fiction was produced have philosophers been confronted with such an impressive and so totally unexplored new realm of ontological inquiry as is presented by cyberspace," says David Koepsell in the opening pages of his book, *The Ontology of Cyberspace*. In a similar vein, Luciano Floridi prefers to speak of a process of re-ontologization[4] but the idea is roughly the same.

The issue is that on specific questions such as "What exactly is a Web page?" philosophers—except for a few exceptions worth mentioning like Harry Halpin—haven't taken into account the work of Web architects. Thus, up until now, a lot more has been done to understand the fundamentals of the Web inside standardization bodies like the W3C.[5] Koepsell, for instance, in the already quoted book, explains the "retrieval" of a Web page the following way:

> Web pages are just another form of software. Again, they consist of data in the form of bits which reside on some storage medium. Just as with my word processor, my web page resides in a specific place and occupies a certain space on a hard drvie [sic] in Amherst, New York. When you "point" your browser to http://wings.buffalo.edu/~koepsell, you are sending a message across the Internet which instructs my web page's host computer (a Unix machine at the university of Buffalo) to send a copy of the contents of my personal directory, specifically, a HTML file called "index.html," to your computer. That file is copied into your computer's memory and "viewed" by your browser. The version you view disappears from your computer's memory when you no longer view it, or if cached, when your cache is cleaned. You may also choose to save my web page to your hard drive in which case you will have a copy of my index.html file. My index.html file remains, throughout the browsing and afterward, intact and fixed.[6]

While the default view of the Web is conform to the paragraph quoted, a more general theory was needed to account for cases not covered in this picture:

- The dynamic Web which is also, incidentally, becoming the default Web (services,[7] constantly changing "pages" like newspapers homepages, blogs, etc.)
- "Content negotiation" (abbreviated as "conneg"). A feature of the HTTP protocol accounting for the fact that users may specify the form of the information they get access to according to such criteria as languages, accessibility, formats, etc. This means that it is not possible to generalize on the basis of a single case that of retrieving a single HTML page on a server. After all, what gets sent to a browser may take many different forms. It may even be generated on the fly and thus nowhere to be found "on a server" before a request is even sent. In which cases, what is identified by a URI can simply no longer be a single (HTML) file.
- URIs without addressable content (temporarily or not).[8]
- The lack of a file versioning system[9] (WebDAV could be used as a counter-example but it never really scaled).

Further examination of the intricate history of Web identifiers is needed to understand why the naïve picture of how the Web works is no longer tenable. Before the creation of the W3C, the Web's implementation and principles were not thoroughly distinguished. The Web existed in the guise of programming libraries, software, and the likes, but no agreed upon standards defined the very principles to which these libraries had to stick. This led to many a conceptual difficulty when the first Web standards were devised around 1994-1995.

The latter had to do both with the nature of the objects available on the Web and their identifiers. At first, the notion of a document (or page) seemed to prevail. The obvious conclusion was that Web identifiers had to be addresses (URLs for *Uniform Resource Locators*) allowing for document retrieval in a hypertextual environment. Pages evolving over time (even in the so-called web 1.0—forums being a good example of the latter), the identification of stable entities as exemplified through library identifiers like ISBNs for books or ISSNs for journals, was transferred to URNs (for *Uniform Resource Names*)—proper names referring to objects not available on the Web. The only problem of these identifiers is that the Web's main feature is to provide information about a range of entities, whatever status ("inside" or "outside" of the Web) they have. URNs no longer giving access to anything, their value became disputable. The contradiction regarding addressing, on the other hand, became flagrant in one official document, RFC[10] 1736[11]:

> Locators may apply to resources that are not always or not ever network accessible. Examples of the latter include human beings and physical objects that have no electronic instantiation.

This is no mere contradiction, rather the renegotiation, *in media res*, of the most fundamental features of a technical project. It is

precisely this non-sense that was corrected three years later, in 1998, when the notion of a resource first appeared (elsewhere than in acronyms such as "URIs," "URLs," "URNs," or "URCs"). Merely as a correlate of URIs, the latter being established as the new Web identifier after having been sundered in URNs and URLs. URIs are peculiar inasmuch as they add a technical dimension to identification, namely, access.[12] They have the status of "dereferenceable proper names" for this reason; being, in other words, proper names that *identify* a resource and *give access* to its representations.

Why resources instead of Web pages, a concept everyone is acquainted with? Simply said, because what is aimed at here is a stable entity whose "representation" can nevertheless vary over time or at a given moment (with conneg). The homepage of the newspaper *The Guardian* I access at time *t* is different from the same homepage I access at *t'*. Likewise, accessing it from a mobile phone or a textual browser will yield different results. These various representations are subject to "synchronic" and "diachronic" modifications.[13] Albeit not the least identical to one another, they must be somehow "faithful" to a given resource (*The Guardian* homepage, not accessible *per se*). Such a notion is especially important with regards to the fact that it allows reference not only to documents ("page") but also services, physical objects, etc. Overall, it is of paramount importance to restore the Web's coherence as a technological project beyond the technical changes it underwent with the evolution from a Web of documents to a Web of data or things (also known as the "Semantic Web").

II. The ontological status of resources

Up until now with the resource/representation duality, paralleled in the identification/access one, the debate mainly focused on URIs and their referring prowess. Here, one needs to distinguish between the URI *minter*—the person or institution that create a single URI based of the possession of a domain name—and the *service provider*, the person, institution (generally a company) that will work towards maintaining the access to a given set of representations. The issue at stake was to understand how URIs refer to resources and to find a suitable explanation accounting for their referential stability.

It looks like we currently lack an explanation of the URI/resource binding. But this seems to us to be a profoundly misguided way of begging the question. Indeed, at first sight, it seems to be the case that URIs, save for access, behave like philosophical proper names. However, the remainder of this paper will be dedicated to showing that after close examination such an assumption cannot be taken for granted.

Resources as abstract artifacts

Let us begin by reminding the reader how Web architecture defines resources. A resource, says RFC 2396, can be just about anything: the homepage of *The Guardian*, Tim Berners-Lee, founder of the Web, the number of married people in the U.S., a square circle, etc.

The URI directly identifies a resource which, in turn, can be (among many things):

- The rigid designation of "the Moon" (Kripkean style)
- The current satellite of Earth (Russellian definite description)
- One of the entities to which we have no present access or knowledge about (an "indefinite description" to quote Pierre Livet[14]).

Each time, we find a different way to identify or pick up an object. We follow eminent Web architect Roy Fielding, who states that identification, picking up an entity, does in fact give the true account of the resource. A resource, in Husserlian terms, is the intentional act of picking up something, and by doing so, aiming at an object. It has a content (the object identified) and a form (the action of identifying something). The distinction at stake is reminiscent of the hulé/morphé distinction in Husserl's *Ideen*. Unfortunately, the Husserlian vocabulary is tied to a somewhat mentalistic approach to the mind that is not entirely suitable to explain a *socio-technical system* like the Web (something outside of the scope of Husserl's phenomenological investigation up until his later books, particularly the *Krisis* and the *Origin of geometry*).

Another way of putting things would be to conceive of a resource as a *rule for identification*. It presents the advantage of allowing for different ways of identifying an object. In the example above, the rule can be of a Russellian nature ("the current King of France" is relatively similar to "the homepage of the Guardian" yielding different—including the possibility of no—results over time) or a Kripkean one ("the Moon"[15])—among infinite possibilities. Anyone is entitled to choosing any rule. When the standards explain that a resource can be anything, this is precisely what they mean: this choice is completely free. We're led back to Roy Fielding's definition, undoubtedly the most precise ever given. The Web Architectural style REST (for Representational State Transfer) he authored[16] indeed defines "a resource to be the semantics of what the author intends to identify, rather than the value corresponding to those semantics at the time the reference is created."[17] Precisely what gives enough room to distinguish between a rule and the result of its *application(s)* at a specific time (here, the date a resource was created).

This is not to say resources have no properties of their own: they're also arguably *abstract*, the way a concept is. This is central to Fielding's account of resources in REST. REST provided the Web its *post hoc* theory. Its influence on Web standards is not just a known fact but what made possible the transition from the first standards of 1994-1995 to those of 1998 where resources were first defined. Fielding's definition is the first hint that resources have got some specific properties distinct from those of its representations. This is generally not well understood and standards bear the mark of this difficulty. Especially when the existence of "physical resources" such as chairs, rocks, or even online documents (collections of bits) is assessed. How, then, are we to make sense of this dual notion, torn between conflicting requirements?

An entity is identified in contrast with some features of resources. If we are to distinguish between two sets of properties, those of the entity identified and those of the resource, Edward Zalta's account of fictions[18] might prove immensely useful. He indeed distinguishes between properties that are *exemplified* or *encoded* by a fiction.

For instance, Sherlock Holmes is as much a well-known drug addict whose genius elder brother, Mycroft, works for the Queen, as he is the creation of Conan Doyle, mentioned (according to Wikipedia) in four stories authored by Doyle: "The Greek Interpreter," "The Final Problem," "The Empty House," and "The Bruce-Partington Plans."

Similarly, it seems plausible to split resources in two comparable (which is not to say identical) sets of properties, as illustrated in Table 1.

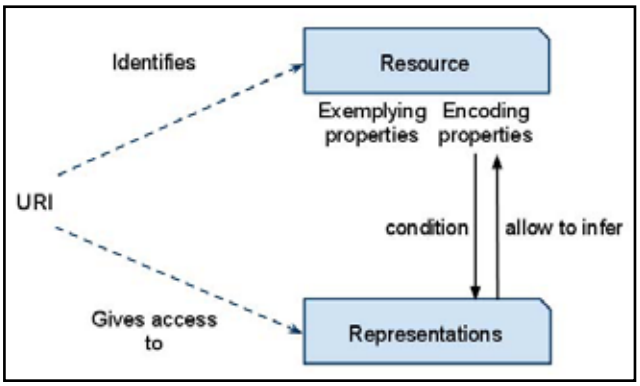
To better distinguish between the two, Zalta has chosen to vary the direction of predication. *Fx* designates exemplifying properties while *xF* is used for encoding one. This helps to ground the data/metadata distinction in ontology and the more intuitive notion that while my resource may be Tim Berners-Lee and may thus encode the property of being a man, English, creator of the Web, etc. (*xF*), Tim Berners-Lee himself is no resource (*Fx*). Like a fiction, a resource can be

Table 1.

Exemplification	**Encoding**
Type properties shared by all resources: - Abstract - Published at a given date - Published by a given person, institution, etc.	**Type properties** of a rule: The way an entity is identified: through rigid designation, definite descriptions, etc. The result will change according to the chosen rule.
Token properties of the above types: What distinguishes sets of exemplifying properties of resources from each other's.	**Token properties** of a token resource: The elements according to which an entity is identified. Properties linked to the content can be used to formulate any rule of identification.

anything *and* an abstract artifact (to borrow an expression from contemporary metaphysician Amie Thomasson[19]) at the same time. We may now make more sense of the little conundrum to which standards had no answers for: all resource are abstract (type exemplifying properties) and yet a given resource may be physical as well (token encoding property).

Figure 1. Once the two sets of properties of a resource are clearly separated, it becomes easy to expend the Webarch picture with additional ontological details.



III. The demise of reference under mutual adjustment between identification and access

I hereafter understand ontology as being what I call the *science of reference* or a *theory of objects*. It is always useful to mention that the word "ontology" appeared in the seventeenth century, thanks to Jacobus Lohardus and Rudulphus Goclenius. Two thousand years had passed since Aristotle's *Metaphysics*. The shift is perceptible, especially with Francisco Suarez,[20] from an understanding of Being dominated by the divine, to a more general definition of Being as *aliquid*, something in general, that needs not be actual. Actuality being no longer a necessary condition, Being came to be understood as the possibility of an object, its conceivability on the mind. Between actual objects and mere fictions (*ens rationis*) a new realm was discovered: metaphysics thus turned into the *a priori* science of the possible (*ens reale*). The problem is that for objects not to fall into fictions, they had to be properly distinguished. Suarez's answer relied on the principle of non-contradiction: for a truly thinkable object to merely be, its very concept had to be free of any contradiction. This is exactly what later led to Kant's critique of logical possibilities, understood in terms of compatibility of concepts, in the name of his own solution that dealt with the transcendental possibility of things through the synthesis of the knowing mind.

Later, the question known as the "problem of representations without objects," raised by Bernard Bolzano (How do you distinguish between the square circle and the golden mountain?), had a tremendous influence on thinkers such as Brentano, Twardowski, Meinong, and Husserl. After all, it created a breach between what is thinkable, and the possibility of an object. Meinong's answer was to explore those very possibilities which Suarez had avoided: to include fictions and impossible objects (impossible to conceive) in order to somehow free objectivity from conceivability. This became known as the theory of (non-existent) objects (*Gegenstandstheorie*). Ontology itself was transcendended by something broader than a description of the furniture of our world, even broader than a science of the possible. Something that encompasses the impossible itself: a theory of objects. And what those objects lacked in reality and possibility they made up in identity. How? Because it was still possible to *refer* to such objects as they came to be though as correlates of naming. What is remarkable is that objects were thus free from actuality and possibility. Naming, positing a referent, was acknowledged as the most fundamental (and trickiest!) ontological operation. Husserl later reframed the whole issue but that is outside of the scope of the present study.

The Web can be perceived as a technological equivalent of a *general system of reference*, owing to the paramount importance of URIs. This is also why the ontological question of the status of the referent is so prevalent. But unlike traditional ontology, where objects were substances, and theories of objects, which had no regard to technology, we will see that the Web deals with *non-natural,* even *technical* objects, with the surprising consequence that reference (and thus naming) is no longer a suitable concept in this context.

a) The ontology of resources as a theory of *non-natural* objects

Resources, it has been said, can be "anything at all." This is strongly reminiscent of the ontological notion of an object: "something in general" (*aliquid*). The difference here, and this is why taking into account the technological aspect is very important, is that objects (resources) on the Web are no longer *mental*. They have been *artifactualized*. One can consider them intentional objects, but only the way books by contrast to mental states are intentional since books are *published* and thus follow a very complicated chain comprising events, people, institutions, book shops, etc. Hence, instead of intentional objects we prefer to speak of *institutional objects*.

Resources are also context-dependent, not unlike facts in Latour's account (see his *Science in Action*, for instance). We

can establish a very good analogy between both: to produce a fact, as science does, requires that a statement be reused by as many people as possible while, at the same time, remaining stable. The same goes for resources. Either through tagging and hyperlinking or with RDF (the knowledge representation language of the Semantic Web), resources are just nodes in a graph whose identity depends on their ever-changing position inside that same graph since anyone can say anything about anything. As referent, they're *not* context independent—and the important point is precisely here to mix ontology and epistemology, otherwise one would end up with some form of nominalism as in Saussure's *arbitraire du signe*, whereby signifiers are distinguished only through differences and the positions they entails: semiotics always had a dissolving effect on ontology. If we don't want to lose the referent and end up with mere signs or representations because only the latter can traditionally be treated as context-dependent, we must attribute other identity conditions to referent (resources) *themselves*.

Finally, there is the importance of *trust*: the Semantic Web is no mechanism to help one decide how to link words to the world (reference!) or how to end up with a well-founded theory of truth. After all, it is dominated by *trust*, the highest layer on the famous Semantic Web "cake."[21] Take Dbpedia,[22] for example, the Sematic Web's most successful application that semantize Wikipedia and treats its entries as entities about which we have knowledge. Entities found in the semantic version of Wikipedia are always the result of the contributions of thousands of users all around the planet. The latter is subject to peer-review, contradiction, improvement, re-writing, etc. It's no longer "given." Only once the activities that support our ontology are negated and forgotten can it appear natural (in a very paradoxical way: here it is the artifactual activity of machines that produces naturalness). This is partly what currently happens with Dbpedia: human discussions as well as machine extractions disappear from the final product, giving it an uncanny pristine appearance that doesn't correspond to the reality of the complex and muddled lifecycle that made it possible. These entities or resources are always the result of collective choices and evaluations.

Recently, Halpin, Hayes, McCusker, McGuinness, and Thompson[23] have shown that many resources that seemed identical, already distinguished entities that only need to be picked-up, are in fact different under close examination (of course, using our framework, we could always distinguish resources by their *token exemplifying properties* but let's just assume this was not possible for the sake of the argument). For instance, the sodium element both on the CYC[24] knowledge-base and on Dbpedia appear to genuinely correspond to the same entity or referrent. However, they do not share the same definition (*token encoding properties*).[25] Sometimes the problem will seem circumscribed with regards to a common reality beneath these "representations" and distinct from them, well-established enough to warrant that dissimilarities are only symptomatic of a lack of precision. Two and a half millennia of philosophy should temper such optimism—also because this so-called underneath level, call it nature or society depending on your discipline, has always proved quite hard to find.

Does it mean that all we need to get rid of all the problems we're facing on the Web is a good social epistemology? We'd rather keep up with the ontological inquiry, not lose the worlds, and go back to the Latin language for whom, as French historian of philosophy Jean-François Courtine puts it, *substantia* (the translation of the Greek *ousia*) was the answer to the question *an sit?* "How about the reality of a given fact?" "Can we make something out of it?" "Is there something sure and solid (*res certa et solida*)?" What Aristotle, after Plato named in terms of presence (*ousia*), became the topic of debate, a moot point, a fact in dire need to be established. Latin rhetoric explicitly aims at convincing (*facere fidem*—to produce trust), stabilizing, what is always given as a matter of doubt (*Res dubia*) at first.[26] Following Bruno Latour we should say that realities (instead of only representations) are fallible and need to be constantly adjusted.[27]

b) The cost of maintaining a resource: why reference once artifactualized is no longer truly reference

Substances used to hold by themselves

"A 'resource'," writes Tim Berners-Lee, "is a conceptual entity (a little like a Platonic ideal)."[28] More than a clarification *per se*, this sentence indicated that its exact status remains to be investigated. Berners-Lee never really explains this reference to Plato. Furthermore, one would be in dire pain to find a proper theory of Ideas in Plato's writings. That is why we will now turn to Jules Vuillemin's *a priori* deduction of philosophical systems,[29] including Plato's idealist position, from what he calls the "scheme of pure predication" ("linguistic universals" in opposition to "perception universals" as in Aristotelianism).

In nominal sentences such as "Humility is a virtue," sentences that assert conceptual truths, explains Vuillemin, two linguistic universals are associated, one positioned as an argument, the other one as a function to use Frege's terminology. In both cases this kind of predication (but can we really speak of predication and at the same time mobilize Frege's function/argument dichotomy?) is not liable to change—such states of affairs will not vary since their arguments are neither located in space nor time and as such are completely intangible. Such predicative functions then hold or don't independently from circumstances. For the same reason, nothing is displayed to the sense, not even the referent. To borrow from French linguist François Rastier, the Idea belongs to the distal anthropic zone. In the same fashion, the Web might be seen as a technical device aiming at bringing the distal (the identified resource), what is never accessible, under the guise of the proximal (the accessed representation). Since no translation is available that would make these entities break out from their respective zones, doing so makes it mandatory to pick up entities from a zone we can have access to, the proximal one. The gap between the resource and its representation is thus never overcome but mediated through makeshifts (descriptions, information realizations, etc.). From this point of view, Berners-Lee was not so wrong to identify resources to a Platonic idea rather than an Aristotelian substance, despite the rather obvious analogy between resources and substances on the one side and accidents and http-representations on the other side.

Yet, the issue remains to determine what ties the core-content to what gets attached to it (substance and accident with regards to the canonical objects of the Western metaphysical tradition, resources and representations in Webarch)—the difficulties are of a very different nature in both cases. A) How, on the one hand, is the rule, in other words the content of a resource, being communicated to those who ignore what a URI is supposed to identify and only have a *de facto* access to potentially constantly changing representations?[30] B) On the other hand, as long as the resource doesn't change, it nevertheless undergoes modifications, but internal ones. How then are we to define a rule and all its applications *ab initio*? The stability of URIs is directly tied to this capacity. Apache Software Foundation President Justin Erenkrantz[31] is right to equate resources to "network continuity" in his Ph.D. thesis, though it doesn't explain *how* this continuity of subsistence is produced insofar as we chose to abstain from adopting substances as an explanation—a notion whose value resides precisely in its capacity to beg the issue at stake (as a metaphysical concept,

its specific "agency" instead of making things hold together like ontological cement might simply be to prevent the real problem from coming under scrutiny).

Nelson Goodman's paradoxes of predicate projectibility described in *Facts, Fiction, and Forecast* here come to mind. To infer the content of a resource from a finite number of representations (or what we may call local "projections" on a screen) is not very different from the proverbial problem of induction. Projectibility paradoxes were later reactivated from a Wittgensteinian perspective as rule paradoxes by Saul Kripke (*Wittgenstein on Rules and Private Language*). This undoubtedly makes him the philosopher whose works were the most influential on philosophical engineering[32]—and thus on the Web. From the idea of "baptism" to the paradox of rules (rigidity is another matter since the Web was not conceived with possible worlds in mind but rather from a universalist point of view that is reminiscent of Frege and the early twentieth-century logicians as opposed to algebraists like De Morgan, Boole, or Schröder who put forward the notion of a universe of discourse, prompting such thinkers as Jean van Heijenoort and later Jaakko Hintikka to draw a distinction between "logic as language" and "logic as calculus").

Many Web architecture discussions borrow from Kripke's idea of baptism but only in order to underline the fact that the publisher of a resource is to decide on its "content." This, in a sense, is very true. After all, *access* is what first and foremost distinguishes URIs from philosophical proper names. Apart from URIs understood as common names (the string of character considered as meaningful as in http://www.w3.org/People/Berners-Lee/) you only get a glimpse of a resource by being acquainted to its representations. Users can only fathom the meaning of a resource from its representations. This opens the door to Kripke's skeptic paradox as exemplified in his famous "quaddition" argument: one can infer that a resource is X from its various representations until one gets a result that no longer seems consistent with the rule first postulated (and for which there is no standard way on the Web to make it explicit).

Rules and resources as virtual trajectories

As a consequence, the issue at stake concerns not only the synthesis which binds resources and representations, but also the way it is realized and its cost. This includes understanding how the rule/resource constrains the supply of representations and, in turn, how obstacles to the application of the said rule, or the fact that it may be put on hold, lead to modifying it in return.

The framework recently proposed by Pierre Livet and Frédéric Nef for the analysis of social beings[33] offers a number of conceptual tools to think about such syntheses. Its starts as ontology of *processes* bearing on a coupling of the actual and the virtual—the startup stage of a process going from the actual to the virtual, to end up with an actualization of the virtual and a virtualization of the actual.

In addition to processes, the two authors introduce the notion of *quality*, whose peculiarity is to rest at the crossroads between the epistemic and the ontological; hence the impossibility to describe it through a single process. Qualification is indeed defined as the articulation of two processes, the second being the qualifying one: "The first corresponds to what is always actual whereas the second one is what binds this actual to the virtual."[34] The first is the actualization of the qualification expected from the second one: "In whatever sense we take things, a qualifying process requires the following coupling: the introduction of an expectation, a virtuality, and the accomplishment of this virtuality by another process so that coexists in actuality the process of expectation or initial reception of another process and the accomplishment of this virtuality by the second process."[35]

The expectation, here, is due to the identified rule/resource, a rule whose application constrains representations to be accurate. The process with which this second "qualifying process" (noted "Re" for "resource" regarding its virtual aspect and "StRe" for "state of the resource" with regards to its actual aspect) is assembled is one that allows to generate representations of the resource (noted "Dref," for "dereferencing"). Representations derive their quality from the resource, understood *as a requirement of the virtual bearing on the actual*. Qualification results from an actualization of the virtual and a concomitant virtualization of the actual. It entails a tight *coupling* between the two processes that will prove crucial for our investigation.

Following the notation[36] used by Livet and Nef to represent the coupling of processes, we would get:

$$\begin{array}{ll} \text{Dref } V_1^1 & \text{Dref } A_2^1 \\ \textbf{Dref } \mathbf{A_1} \rightarrow (\text{Dref } V_2^1) & \textit{StRe } A_1^2 \\ & \text{StRe } A_2^2 \rightarrow \text{Re } V_1^2 \end{array}$$

Such an approach sheds light on the canonical examples of the W3C Technical Architecture Group, used to explain the difference between a resource and its representations:

> While planning a trip to Mexico, Nadia reads "Oaxaca weather information: 'http://weather.example.com/oaxaca'" in a glossy travel magazine. Nadia has enough experience with the Web to recognize that "http://weather.example.com/oaxaca" is a URI and that she is likely to be able to retrieve associated information with her Web browser.
>
> (…)
>
> Dirk would like to add a link from his Web site to the Oaxaca weather site. He uses the URI http://weather.example.com/oaxaca and labels his link "report on weather in Oaxaca on 1 August 2004." Nadia points out to Dirk that he is setting misleading expectations for the URI he has used. The Oaxaca weather site policy is that the URI in question identifies a report on the current weather in Oaxaca—on any given day—and not the weather on 1 August. Of course, on the first of August in 2004, Dirk's link will be correct, but the rest of the time he will be misleading readers. Nadia points out to Dirk that the managers of the Oaxaca weather site do make available a different URI permanently assigned to a resource reporting on the weather on 1 August 2004.[37]

The first resource being a daily report of the weather in Oaxaca in contrast with a report of the weather in Oaxaca on 1 August 2004, Nadia's expectations are very different: she knows that actual representations are constrained on a virtual level by the resource. Any two apparently similar representations, if states of different representations, are in fact *moments belonging to heterogeneous virtual trajectories* that are part of their respective identities. Hence, despite the actual outward similarity between the two, they are in fact completely unalike—this becomes obvious once our gaze is no longer solely focused on the actual (similar remarks could be made with regards to a range of cases, including mirror representations hosted on a server with a different domain name for instance).

The (ontological and technical) coupling between resources and representations means URIs do not refer

Such a coupling also makes it possible to conceive of the dependency between the rule and the results of its applications. Widespread (but not in any way less troubling) phenomena, often shunned in standards owing to their normative grasp of

the Web, may finally get an explanation, starting with "non-cool URIs," in other words, URIs that do change contrary to the stability requirement imposed on "cool-URIs"[38]—a somewhat inappropriate name since, all things considered, URIs never change unless they disappear or are discarded in favor of new ones. Only the URI/resource pairing may evolve over time,[39] for the simple reason that the resource (rule) identified by a URI will have undergone modification first. As Livet and Nef put it, it is indeed impossible

> to foresee all the obstacles to the application of a rule and anticipate all the subsequent right revisions they entail. The spirit of the rule only makes sense in retrospect, when an obstacles lead to revising the rule in a satisfying way, albeit not one that complies with previous trajectories. Prospectively, we cannot pretend to know in advance how to follow a rule for every new situation that presents obstacles. It is because such forward-looking knowledge is out of reach that rules seem overhanging. But this, in a sense, is a mere appearance for the application of the rule rests upon past successes, thus only in retrospect.[40]

Once seen as a rule, and analyzed from the perspective of a coupling of processes associating the actual and virtual dimensions, resources can no longer be conceived outside of their representations: the latter only make sense against the rule that gave them birth, as well as they may modify it according to the borderline cases that force to renegotiate the application of the rule. Many a creator of conference websites knows that after the first installment, it may become necessary to turn the original resource (for instance PhiloWeb 2010 symposium homepage) into something more generic (PhiloWeb symposiums homepage), should it prove somewhat successful. The "ontologization" of the rule/resource that an analysis in terms of actual and virtual modalities warrants offers an explanation that escapes the Kripkean skeptic paradoxes and could help to find a more careful treatment of such phenomena as non-Cool URIs and mutable resources.

Let us add that for users, and probably for a majority of the institutions that publish a resource, representations provide (with connotative URIs, they function not only as proper names but also common ones though there is no consensus regarding the way such URIs should be written to avoid obsolescence[41]) the most efficient means to infer its content. This, as we have seen, has a cost. As Brian Cantwell Smith[42] once said, *"reference to succeed doesn't need adjustment to its target."* On the Web we witness exactly the contrary. Only URIs with no dereferenceable function may be said to refer although one may rather consider that they simply do not identify a resource, nor give access to its representation—which is admittedly quite different. We may call the aforementioned adjustment the *editorial commitment* made by the publisher of a resource to ensure "network" (and service[43]) "continuity," to borrow Justin Erenkrantz's definition of the resource.[44]

Nowhere do we need to ask ourselves whether or not URIs refer to something permanently and how. URIs do *not* refer for the aforementioned reasons. We rather publish Web resource identified by the latter and, if given enough resources (in the traditional meaning of the word), then we maintain a positive feedback loop between Web resources and accessible representations; all in all, a very different story.

URIs are undoubtedly the result of the artifactualization[45] of proper names. "Web proper names," to quote Thompson and Halpin,[46] are no longer conceptual, philosophical, or semiotic objects, but rather *technical* ones. The consequences of this simple yet decisive truth are yet to be properly measured. While regular (philosophical) proper names may possess the distinctive feature to refer if used accordingly, neither identification nor access on the Web have to do with reference. In other words, the artifactualization of proper names is tantamount to replacing reference with other (technical) processes, coupled to one another. Thus the explanation of the binding between URI and resources rests upon entirely new principles. The issue at stake is no longer to *point* at or *designate*, but rather to *maintain* a coupling between two kinds of processes through socio-technical means.[47]

Conclusion

Despite the above reasons, according to the title of this paper, URIs do not *always* refer. Indeed, the Semantic Web foundational language, RDF, a knowledge representation syntax, functions as an additional layer to the existing pile of standards that govern the Web. According to RDF and its semantics,[48] URIs *are* indeed proper names, interchangeable props or tags with no meaning whatsoever. As a corollary, URIs once moved to this context do indeed revert to the perennial definition of proper names in logic (and, we may add, in philosophy, though the emphasis then is less on finding an "interpretations controlled by the pure semantic power of the axioms that use them"—see below). Therefore, it can be argued that URIs keep referring since from the perspective of RDF all the previously observed intricate details just vanish into the background. What we must now think are the different semantic (and ontological) commitments across the layers[49] formed by the heap of accumulated standard, formal languages, logics, that characterizes the Web. Equally necessary is a theory describing how such layers *assemble* and how the properties of objects shift from layer to layer, from logical proper name to genuine Web proper names—and back, for instance. No one better described the situation with regards to names and URIs than Patrick J. Hayes in a keynote where he advanced the idea of Web logic or "Blogic." Let us quote him at length:

> *Names* are central in blogic. They are *global* in scope. They have *structure*. They *link* blogical content to other meaningful things, including other blogical content. They *embody* human/social meanings as well as being conduits and route maps for information transfer. In many ways, the Web is *constituted* by the links which are the blogic names, and the logical content which we write using those names is only one component, perhaps a minor one, of the whole social and technical structure which determines their meanings. And yet seen from the perspective of the logic, these IRIs are merely "logical names," elements of an arbitrary set of meaningless character strings. In AI/KR, we teach our students that the names are irrelevant, because one can replace them all with gensyms *without changing the logical meaning*.
>
> Clearly, there is something unsatisfactory about this picture, a serious disconnect between the classical logical view of names as simply uninterpreted strings waiting in a kind of blank innocence to have their possible interpretations controlled by the pure semantic power of the axioms that use them, and the reality of the almost unrestricted referential power that these names actually have in the dynamics of the Web. *Think of the concern and attention that is devoted to their choice, who owns them, who is responsible for maintaining and controlling them, and the ways they are decomposed and used in the planet-wide machinery called the Internet, none of which has very much at all to do with logical assertions. Another way*

*to put it: IRIs are *identifiers*, not mere logical names. Unfortunately, nobody seems to be able to say what in God's name that can possibly mean.* [our emphasis]

In a sense, this paper can be construed as an attempt to shed some light on this conundrum by looking directly at the "the whole social and technical structure which determines [URIs] meanings."

Endnotes

1. Latour et al. 2007, 110.
2. We argue elsewhere that concepts, as semiotic constructs, are also technical tools. Artifactualization is thus always a re-artifactualization.
3. Monnin 2009.
4. Floridi 2005.
5. See in particular the "Technical Architecture Group" (TAG). Cf. Thompson 2007 for a presentation. Our guess is that while they constitute a pre-requisite for any philosophy of the Web, we should not on the other hand accept non-critically every observation from Web architects.
6. Koepsell 2003.
7. "Second, there exist many addresses that corresponded to service rather than a document—authors may be intending to direct readers to that service, rather than to any specific result from a prior access of that service." Fielding & Taylor 2002.
8. "Finally, there exist addresses that do not correspond to a document at some periods of time, as when the document does not yet exist or when the address is being used solely for naming, rather than locating, information." *Ibidem*.
9. "First, it suggests that the author is identifying the content transferred, which would imply that the identifier should change whenever the content changes." *Ibid*.
10. RFCs (Requests for Comments) are documents of the IETF (Internet Engineering Task Force) where most Internet standards are consigned. In spite of the creation of the W3C, where Web recommendations are produced, IETF continues to publish standards for Web identifiers.
11. Kunze 1995.
12. See Hayes and Halpin 2008, Halpin 2009, Halpin and Presutti 2009.
13. Monnin 2011.
14. "Web ontologies as renewal of classical ontology," to be published in a forthcoming special issue of *Metaphilosophy* dedicated to the Philosophy of the Web.
15. Though, properly speaking, in Web architecture there is no room for possible worlds. This, however, is not sufficient to prevent the institution that mints URIs from referring in a way that presupposes possible worlds.
16. Fielding 2000 and Fielding & Taylor 2002.
17. Fielding & Taylor 2002, 135.
18. Zalta 2003.
19. Thomasson 1999.
20. Courtine 1990.
21. Many versions of it have been proposed, the canonical one now seems to be: http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24).
22. http://dbpedia.org/About
23. Halpin et al. 2010.
24. CYC is a longstanding AI project run by Douglas Lenat first at MCC then at Cycorp, whose goal is to construct "a foundation of basic "common sense" knowledge—a semantic substratum of terms, rules, and relations—that will enable a variety of knowledge-intensive products and services" (see: http://www.cyc.com/cyc/technology/whatiscyc).
25. *Ibidem*: "In particular, this issue comes into play when different agents describe the world at different levels of granularity. For example, different sources of Linked Data may make subtley different claims about some common-sense term like 'sodium.' This occurs in the case of the concept of sodium in DBPedia, which has a sameAs link to the concept of sodium in OpenCyc. The OpenCyc ontology says that an element is the set of all pieces of the pure element, so that sodium in Cyc has a member which is a lump of pure metallic sodium with exactly twenty-three neutrons. On the other hand, sodium as defined by DBPedia includes all isotopes, which have different number of neutrons than 'standard' sodium, and in this particular case are unstable. So, one should not state the number of neutrons in DBPedia's use of sodium, but one can with OpenCyc. At least in web settings with little inference or reliance on detailed structures, it is unlikely that most deployers of Linked Data actually check whether or not all the properties and their associated inferences are shared amongst linked data-sets."
26. The remainder of this paragraph is adapted from Courtine (2003): "Les traductions latines d'OYΣIA et la compréhension romano-stoïcienne de l'être."
27. Latour 2007.
28. Berners-Lee 1996.
29. Vuillemin 1986.
30. See Booth 2007b and Halpin 2008.
31. Erenkrantz 2009.
32. See, on this notion, Halpin 2008b and, in French, Monnin 2012.
33. Livet & Nef 2009.
34. *Ibidem*.
35. *Ibid*.
36. The arrow stands for the coupling typical of a process; the characters between parentheses, the aspect that is going to be replaced; in italics, the aspect that will take its place; in bold, the process qualified; the subscript indicate the initial (1) or final (2) aspect of a process; the superscript, process number 1 or 2. "V" stands for Virtual, "A" for Actual.
37. Jacobs & Walsh 2004.
38. Berners-Lee 1998, Sauermann & Cyganiak 2008.
39. In a recent communication (Arwe 2011), John Arwe mentions four of them:
    1. Needs evolve: successful proof-of-concept systems are pressed into wider use; department-level systems grow into corporate systems with different quality of service needs that require different deployments.
    2. Domain name ownership changes. Acquired organizations' names are retired, and there is a (small but non-zero) economic incentive to release unneeded domain names. Sometimes there is a legal requirement to do so.
    3. Some URLs are poorly constructed to begin with, in that they include components that lead people to want to change them over time. They include brand names, or version numbers, that are really mutable properties of the identified resource. Some organizations are simply used to being able to change URLs because their current consumers are human-attended user agents rather than fully automated/autonomous processes.
    4. Organizations out-source control of their network environments; exerting control of the sort required to control DNS entries or issue 301 redirects essentially becomes a legal process, ill-suited to solving technical problems.
40. Livet & Nef 2009, 201.
41. It is even quite the contrary as a result of the much discussed "axiom of opacity," which states that the "meaning" of a URI should only be inferred from dereferenced representations instead of any connotative aspect of the string of characters that constitutes the URI itself.
42. "Reference and Identity on the Internet," presentation given at the "Philosophy of the Web" seminar organized the author and Harry Halpin at La Sorbonne on January 28, 2012.

43. Ensuring a continuity of service constitutes no less than a necessary condition of the dereferencing process. Often monitored by third parties (instead of the institution that published the resource and guarantees that compliance of the representations is implemented in the long run), thus adding an additional line of expenditure when it comes to summarizing the efforts required to maintain the resource over time—or, to be more precise, the coupling between processes of dereferencing and qualification that we have previously analyzed.
44. The lack of substances brings this very issue to the foreground: "Why do things subsist? Once [enduring] substance has been excluded, subsistence comes to the fore, and then the big question is how many ways there are for the entities to graze their subsistence in the green pastures." Latour et al. 2011, 48.
45. For an introduction to this concept, see Monnin 2009. More recently, Luciano Floridi has been using a similar line of argument.
46. Halpin & Thompson 2005.
47. The fact that reference is no longer the issue sits well with the Web's reluctance to deal with the notion of truth. As already said, the epistemology of the Web is one of trust. Content providers, including resource publishers, must thus ensure that the definition they give of a resource (its encoding properties) are *trustful*. See also Henry Thompson, member of the TAG, who has dedicated a lot of thought to the analysis of URI persistence: "persistent identifier efforts can and should save *huge* amounts of fuss by focussing (sic) on the non-technology substrate issues involved in producing persistence" (Thompson 2007).
48. http://www.w3.org/TR/rdf-mt/
49. Specifically referring to RDF, Patrick Hayes called this problem "Death by layering," thus summarizing the issue at stake in a most fitting way: "names have a different logical status at different levels." What's true within the framework of the Semantic Web is all the more true within the broader framework of the Web itself here examined.

References

Arwe, J. 2011. Coping with un-cool URIs in the web of linked data. http://www.w3.org/2011/09/LinkedData/ledp2011_submission_5.pdf (January 31, 2012).

Berners-Lee, T. 1998. Cool URIs don't change. http://www.w3.org/Provider/Style/URI (April 23, 2011).

Courtine, J.-F. 1990. *Suarez et le système de la métaphysique*. Presses Universitaires de France - PUF.

Courtine, J.-F. 2003. *Les catégories de l'être: Etudes de philosophie ancienne et médiévale. Presses Universitaires de France*, Paris, PUF.

Erenkrantz, J. R. 2009. Computational REST: A New Model for Decentralized, Internet-Scale Applications. Ph.D. Thesis. University of California, Irvine.

Fielding, R. T. 2000. Architectural Styles and the Design of Network-based Software Architectures. Ph.D. Thesis. University of California, Irvine.

Fielding, R. T. & Taylor R. N. 2002. Principled design of the modern Web architecture. *ACM Transactions on Internet Technology (TOIT)* 2(2):115-50.

Floridi, L. 2005. The ontological interpretation of informational privacy. *Ethics and Information Technology* 7(4):185-200.

Halpin, H. 2008. The principle of self-description: identity through linking. *Proceedings of the 1st IRSW2008 International Workshop on Identity and Reference on the Semantic Web*, ed. Paolo Bouquet et al. Tenerife, Spain. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-422/irsw2008-submission-13.pdf.

Halpin, H. 2008b. Philosophical engineering: towards a philosophy of the web. *APA Newsletter on Philosophy and Computers* 7(2).

Halpin, H. 2009. *Sense and Reference on the Web*, Ph.D. Thesis. Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh. http://www.ibiblio.org/hhalpin/homepage/thesis/.

Halpin, H. & Presutti, V. 2009. An ontology of resources: solving the identity crisis. In *The Semantic Web: Research and Applications*, ed. Aroyo et al. Springer-Verlag Berlin, Heidelberg.

Halpin, H. & Thompson, H. S. 2005. Web Proper Names: Naming Referents on the Web. Chiba, Japan. http://www.instsec.org/2005ws/papers/halpin.pdf (May 25, 2009).

Halpin, H. et al. 2010. When owl:sameAs isn't the same: an analysis of identity in linked data. *The Semantic Web–ISWC 2010*:305-20.

Hayes, P. J., (ed.) 2004. RDF Semantics (W3C Recommendation 10 February 2004). http://www.w3.org/TR/rdf-mt/ (February 20, 2012).

Hayes, P. J. 2009. BLOGIC or Now What's in a Link? http://videolectures.net/iswc09_hayes_blogic/ (February 20, 2012).

Hayes, P. J., and H. Halpin. 2008. In defense of ambiguity. *International Journal on Semantic Web & Information Systems* 4(2):1-18.

Husserl, E., *passim*.

Jacobs, I. & Walsh N. 2004. Architecture of the World Wide Web, Volume One (W3C Recommendation 15 December 2004). http://www.w3.org/TR/webarch/#formats (February 1, 2009).

Koepsell, D. R. 2003. *The Ontology of Cyberspace: Philosophy, Law, and the Future of Intellectual Property*. New edition. Open Court Publishing Co, U.S.

Kunze, J. 1995. RFC 1736 - Functional Recommendations for Internet Resource Locators. http://www.rfc-editor.org/rfc/rfc1736.txt (February 20, 2012).

Latour, B. 2007. Quel cosmos? Quelle cosmopolitiques. In *L'émergence des cosmopolitiques,* ed. Lolive, J. & Soubeyran, O., 69-84. Colloque de Cerisy, Recherches, La Découverte, Paris.

Latour, B., Harman, G., Erdelyi, P. 2011. *The Prince and the Wolf: Latour and Harman at the Lse*. Zero Books.

Livet, P. & Nef F. 2009. *Les êtres sociaux : Processus et virtualité*. Hermann.

Monnin, A. 2009. Artifactualization: Introducing a new concept. Southampton, United Kingdom. http://hal-paris1.archives-ouvertes.fr/hal-00404715_v1/ (September 21, 2009).

Monnin, A. 2011. La resource et l'ontologie du Web (to be published in *Intellectica*). http://hal-paris1.archives-ouvertes.fr/hal-00610652 (February 5, 2012).

Monnin, A. 2012. L'ingénierie philosophique comme design ontologique. In Archéologie des nouvelles technologies, *Réel-Virtuel: enjeux du numérique*, 3.

Sauermann, L., & Cyganiak, R. 2008. Cool URIs for the Semantic Web (W3C Interest Group Note 03 December 2008). http://www.w3.org/TR/cooluris/ (February 1, 2009).

Thomasson, A.L. 1999. *Fiction and Metaphysics.* Cambridge University Press.

Thompson, H. S. 2007. URIs and Persistence: How long is forever? http://www.ltg.ed.ac.uk/~ht/UKOLN_talk_20070405.html (May 31, 2010).

Thompson, H. S. 2012. An introduction to naming and reference on the Web. http://www.ltg.ed.ac.uk/~ht/PhilWeb_2012/ (February 5, 2012).

Vuillemin, J. 1986. *What are Philosophical Systems?* Cambridge University Press.

Zalta, E. N. 2003. Referring to fictional characters. *Dialectica* 57(2):243-54.

The Creativity Machine Paradigm: Withstanding the Argument from Consciousness

Stephen L. Thaler
Imagination Engines, Inc.

Abstract

In Alan Turing's landmark paper, "Computing Machinery and Intelligence," the famous cyberneticist takes the position that machines will inevitably think, supplied adequate storage,

processor speed, and an appropriate program. Herein we propose the solution to the latter prerequisite for contemplative machine intelligence, the required algorithm, illustrating how it weathers the criticism well anticipated by Turing that a computational system can never attain consciousness.

1. Introduction. In his 1950 article in *Mind*, entitled "Computing Machinery and Intelligence," Alan Turing anticipated nine objections to his conjecture that machines would one day think, and that they could succeed at the so-called "imitation game." The foremost of these objections, in my mind, was the so-called "argument from consciousness" in which machines are denied full contemplative status on the basis of their lack of emotion, in particular the feelings they have about their own thinking. Appropriately, Turing quotes Professor Jefferson's Lister Oration from 1949 to drive home the dissenting point of view, "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only to write it but know that it had written it. No mechanism could feel pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, or depressed when it cannot get what it wants."

Recently, a new direction in artificial intelligence technology, called the "Creativity Machine Paradigm," allows the generation of new ideas and plans of action without the "chance fall of symbols," accelerated toward its goals by what is tantamount to the subjective pleasure or frustration felt by the human mind as it originates seminal concepts. Whereas this connectionist principle has not written poems in iambic pentameter, it has proven itself capable of both generating and interpreting natural language, to the extent of autonomously fomenting controversy over its self-originated commentary (Hesman 2004). While not generating a concerto, it has achieved the equivalent by spontaneously authoring an album of original musical tunes (Thaler 2007) that are capable of passing the equivalent of a "musical Turing test," after being mentored not by "if-then-else" heuristics or tedious statistical studies, but by the detection of the raw emotions on its audience's face. In military projects, battlefield robots have bootstrapped impressive tabula rasa behaviors, spontaneously developing improvised reactions to unexpected scenarios, and displaying socially conscious gestures of cooperative planning and mutual protection within a swarm (Hambling 2006). In all three of these examples, the system was well aware of the consequences of its generated concepts before unleashing them upon the world. With its only "valves" being transistors, and its reproductive tendencies limited to software-based object instantiation, I will argue that it experiences the gamut of emotions to both its external environment and its own imaginings, ranging from frustration, to panic, to elation. These feelings then govern the generation, acceptance, and savoring of its own ruminations.

2. Background. To properly relate the concept of a Creativity Machine it is important to review several underlying building blocks that contribute to the paradigm's ability to achieve not only thought, but also self-regulating meta-thought. These key principles include the perceptron and what I have coined the "imagitron."

2.1 Perceptrons. To most readers, the more familiar component of a Creativity Machine is the *perceptron*, a specialized neural network that emulates the non-contemplative aspects of cognition wherein raw numerical patterns, representing both interoceptive and exteroceptive inputs to the brain, are mapped to associated memories. Just as within neurobiology, the creation of such mappings is achieved through the adjustment of synaptic connection strengths, via simple learning algorithms, as numerous exemplary input-output pair patterns are applied to the network. Having attained such mappings, two very important learning processes have taken place within the perceptron: (1) distributed colonies of neurons have synaptically bound themselves into token representations of frequently encountered features within the body of input training exemplars, and (2) additional synapses have acquired strengths that reflect the intrinsic relationships between such features in generating an associated pattern-based memory at the net's output layer.

Introspectively relating this learning process to human cognition, the world observable to the brain is automatically carved up into its dominant themes, consisting of repeating entities and scenarios in the external world. As such themes appear within the outer reality, the token representations thereof, once again consisting of distributed colonies of neurons, activate, thereafter driving the subsequent excitation of associated memories within downstream neuron layers. During such forward propagation of patterns, no contemplative processes are at work. Instead, the net reflexively and instinctively generates a stored memory in response to a sensed pattern originating from the environment. Therefore, the process emulates the brain's inherent ability to generate immediate and hopefully useful associations when the time-intensive luxury of understanding is detrimental to the host organism.

One skilled in both artificial neural networks and the workings of the brain realizes that while the perceptron epitomizes non-contemplative perception via pattern association, neurobiological perception involves hierarchical cascades of neural assemblies and not just a single, monolithic neural network. Within these compound neural architectures, an individual perceptron may activate into a particular memory. A subsequent perceptron accepting the output memory of the first will then activate into a related memory, and so forth and so on. In this manner, multiple perceptrons are recruited into associative chains, often terminating upon themselves to form closed loops. The topology of such chains may be dynamic due to their inclusion of specialized neurons capable of triggering the secretion of synapse-altering agents (i.e., neurotransmitters and neurohormones). Because of such weight plasticity, memory linkages will not only be constantly rerouting themselves, but the experiences stored therein will be deforming themselves to various degrees. The net result, if you will, is that the co-activating neural patterns will consist of a mixture of intact and degraded experience.

To any given brain, such complex patterns of neural activations will be idiosyncratic in that all neural modules cumulatively habituate to one another, in what amounts to a highly encrypted communications scheme. Such subjective experience cannot be shared with other neural networks from another brain, since these "outsider nets" do not possess the hard earned encryption key that has been attained through cumulative, joint exposure of the resident nets to sensory patterns. In lieu of such joint training, we as humans employ very slow and inefficient schemes such as symbolic language to convey these jointly activating memories, the result being the wholesale loss of information contributing to an overall picture that falls short of the synaptic reality.

Even if supplied such synaptic detail, we would find that interconnected memories are severely lacking in detail and fidelity, with receiving networks filling in features as does the visual cortex in supplying multiple draft guesses as to information within the retinal blind spot (Dennett 1991). Due to the accumulated guesswork within such transiently linked neural modules, any semblance of reality degrades as in a

child's game of telephone. Accordingly, the story contained within such cascaded memories is much greater than their sum. For this reason, I call such chained memories an "associative gestalt." Because of their intangibility, resistance to high-level description, circularity, self-driven evolution, and their subjective interpretation of an objective reality, I identify such "associative gestalts" with emotions and feelings.

With no loss of generality, the monolithic perceptron can likewise carry out the pattern association that many generations of human beings regard simply as subjective experience. Henceforth I will at times represent the complex associative chains and loops as the flattened output pattern of a single perceptron, that may potentially represent a sublime recollection, past physical pleasure, or, if need be, a non-descript buzzing sensation. To achieve these mappings, both input and output patterns presented to the network by the environment will need to somehow correlate spatially and/or temporally as synapses adjust their strengths to learn the association. Henceforth we would be lax in our language to claim that the input and output patterns are truly related to one another, when in reality all we can say with confidence is that the patterns are associated in a strictly mathematical "mapping" sense.

2.2 Imagitrons. If the synaptic connection weights of a trained perceptron are subjected to time varying disturbances, two very important things happen to make it a pattern generator rather than the pattern associator: (1) Synaptic disturbances serve as a succession of pseudo-inputs from the environment driving the activation turnover of downstream processing units; and (2) The same connective disruptions continually reshape the attractor landscape of the network so as to create new features therein. Because of such internal noise, the net's activation trajectory is over an unstable and dynamic attractor landscape as it activates into patterns it has never before encountered within its environment. Appropriately, I call such generative perceptrons, imagitrons.

Realizing that the synaptic organization of the perceptron implicitly contains the rules binding neurons into tokenized representations of the external world, as well as the intrinsic heuristics interrelating such features, variations upon such connection weights are the only means by which to force the perceptron turned imagitron to exit its absorbed conceptual space and to generate other than the mapping it has gleaned through training. Introducing such weight deviations, the repeating features of the input space are transmogrified and their interrelationships softened or broken. The overall result is that the network fails to activate into its learned output exemplars at its terminal layer. In effect, the net is then generating false memories or *confabulations* as its synaptic connections are continually perturbed (Thaler 1998).

In Figure 1, we present a general result for a multilayer perceptron (MLP) based associative memory[1] that has learned a mapping and is then subjected to increasing levels of synaptic perturbation, plotting the probability, P_{mem}, of generating an intact output memory as the mean level of synaptic disturbance, $<\Delta w>$, is increased. Typically as such mean perturbation level rises the network predominantly outputs, to within a small error, the training exemplars it has already been exposed to, tantamount to the network's memories (Thaler 1995b). However, beyond a critical threshold of perturbation, near the end of what is called the "regime of graceful degradation," (near $<\Delta w>_c$) the network now begins to output slightly defective memories or novel patterns that are mathematically distinct from what the network has directly experienced (Thaler 1997a). Increasing the noise level even more, the hard-earned connection weights become randomized, thereby destroying the absorbed constraint relationships that capture the essence of the learned conceptual space. As a result the network tends to output nonsensical patterns.

Based upon the veracity and utility of output patterns produced by the imagitron as mean synaptic noise levels increase, I have identified three distinct regimes (Thaler 1996) that are called out in Figure 1:

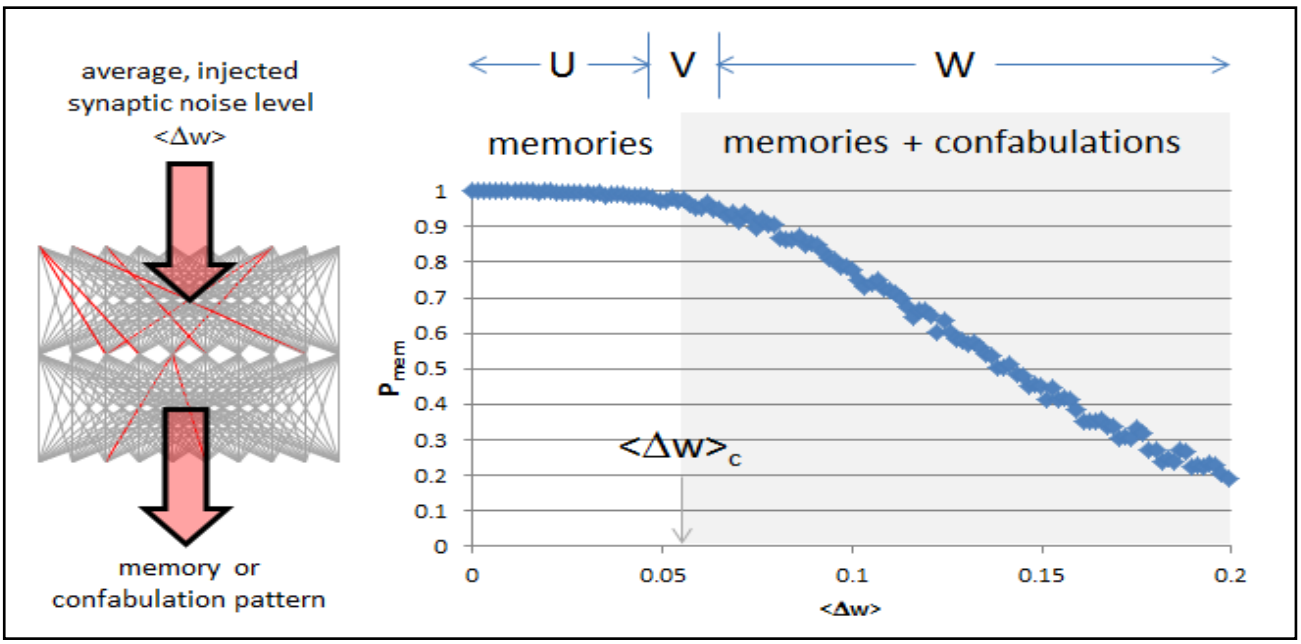


Figure 1. Confabulation Generation within a Synaptically Perturbed Perceptron. Shown here is a representative plot of the probability of activating a memory, P_{mem} versus increasing levels of mean synaptic perturbation, $<\Delta w>$, within a trained perceptron. The so-called "U regime" is characterized by intact memory generation, and the "W regime" marked by unconstrained, nonsensical output patterns. The narrow "V regime" near the critical point, $<\Delta w>_c$, produces novel patterns largely qualifying as members of the learned conceptual space. The left inset depicts this transiently perturbed network's weights in red.[2]

- U-Mode - Generally, U represents an imagitron into which minimal noise has been introduced (**<**Δw**>** < **<**Δw**>**$_c$), thus driving it to visit a series of rote memories that have been drawn from the network's previous training experience, its universe, if you will.
- V-Mode - Imagitrons operating at the critical noise level, near **<**Δw**>**$_c$, are depicted as V, suggesting that they are producing virtual memories of potential things and scenarios that could be part of the net's external environment, but hitherto have not been directly experienced by it through learning.
- W-Mode - Finally, W denotes an imagitron driven by noise levels in excess of those injected in the critical regime, (**<**Δw**>** > **<**Δw**>**$_c$). As a result, most of the constraint relationships characteristic of the conceptual space have been destroyed, leading to the generation of predominantly meaningless noise, in a manner reminiscent of the *blind watchman* allegory.

When enlisting an imagitron to search for solution patterns it should be apparent that the U-mode is only useful when purposely selecting among the network's finite recollections. On the other hand, W-mode represents an imagitron sufficiently "battered" so as to dissolve the hard-earned constraints and thereby generate an enormous search space littered predominantly with nonsensical patterns. It should make sense that the intermediate V-regime, what has been called the "multi-stage" regime (Rowe and Partridge 1993) in rule-based, computational creativity research, offers the best chance at producing a pattern that is novel in comparison with the network's memories, yet qualifying as a potential thing or action representative of the conceptual space learned by the net. In mathematical terms, the V regime produces output patterns that largely satisfy constraint relationships implicit within the imagitron's training patterns, thus qualifying them as potential and novel members of the learned conceptual space.

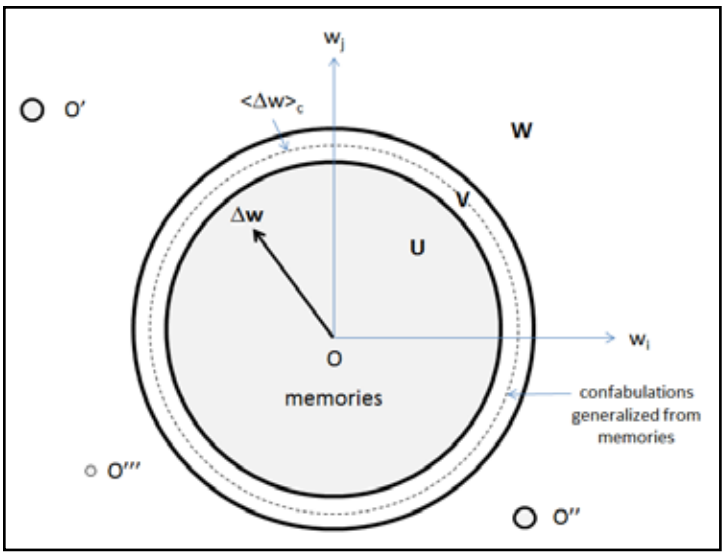


Figure 2. A Synaptically Perturbed Perceptron's Exit from Its Learned Conceptual Space. Illustrated here is a two-dimensional slice from weight space of a perceptron, depicting its weight solution, O. Other neighboring solutions are also shown. Progressively increasing the mean synaptic perturbation level allows the network output patterns to exit the original conceptual space, to produce potentially useful novel patterns such as those encountered at V. Increased perturbation levels generated totally unconstrained patterns represented by W. (Other potential weight space solutions, O′, O″, and O‴, are shown projected from a third weight dimension.)

I sometimes speak of the U, V, and W modes using diagrams like that depicted in Figure 2, where the origin represents a weight space solution for a trained perceptron, here simplified to two weight dimensions. Any pattern of synaptic perturbation to this perceptron may be represented as the vectorial deviation of the connection weights from their train-in values. Therefore, for a constant level of root-mean-square (RMS) weight fluctuations, the succession of perturbed vectors should randomly sweep out spherical hyper-surfaces that project down to a circle in the 2-D weight space shown. Below an RMS level corresponding to <Δw>$_c$ in Figure 1, the synaptic perturbation vector, randomly moves within a circular domain representing the U regime wherein the perturbations are seeding the formation of rote memories. Approaching the <Δw>$_c$ "membrane," in the V regime, the weight perturbation nucleate confabulatory patterns that are slight twists upon the net's absorbed memories. Finally, as perturbation vectors extend into the W-regime the synaptic tumult generates a stream of largely nonsensical, unconstrained patterns.

With imagitron function described in this geometrical fashion, the escape from the conceptual space stored within an imagitron, to produce new notions distinguishable from direct experience, is literally represented by the weight change vector, Δ**w**, departing from the U domain and penetrating into the thin V shell. Just before this U-V boundary traversal, the net is generating intact memories at an optimal rate, what might be likened to frenzy. With the slightest "thump" to mean synaptic perturbation level, the network may catastrophically transition to confabulation generation wherein notions generalized from, but distinct from those of the U domain are formed.

In producing biological intelligence, the brain's ability to exit a conceptual space by simply graduating the RMS synaptic noise level seems advantageous. Effectively, brains can live on a cusp, so to speak, and in response to environmental stress, bathe their neurobiology in slightly increased synaptic perturbation levels so as to drive them through a bifurcation separating mundane and improvised thought. It is at such times that there is the most need for new and viable strategies to preserve the host organism.

That the fidelity of a neural network's activation patterns to its learned reality is most sensitive to synaptic disturbances should make sense: Even within artificial perceptrons, the number of connection weights scale roughly with the square of the processing units therein offering the highest capture cross section for randomly distributed disordering effects. By far, the most numerous "trip points" for signal transmission in the brain are the chemical synapses, outnumbering neurons 10,000:1. With communication through these neuron gaps achieved with minimally small packets of neurotransmitter molecules, it would seem that unintelligent evolutionary forces could easily discover the selective advantage of secreting ever so slightly increased levels of perturbations (i.e., diffusing chemical species) so as to think that which had not previously been thought.

2.3 Perceptron-Imagitron Assemblies (Creativity Machines). When a noise-driven, pattern-generation network, an imagitron, is coupled with a pattern-recognition network such as a perceptron, those confabulatory outputs generated by the former net may be either objectively or subjectively evaluated by the latter so as to selectively filter for those true or false memories offering utility or value. Any such numerical figure of merit generated by the perceptron may be exploited to modulate noise injected into the imagitron's synaptic system. The permanent or transient combination of at least two such neural assemblies is called a "Creativity Machine" and the principle, applicable

to any computational platform, the "Creativity Machine Paradigm." Within the patent literature both the architecture and the paradigm are known as "Device for the Autonomous Generation of Useful Information" (Thaler 1997b) or "Device for the Autonomous Bootstrapping of Useful Information" (Thaler 2008). These two generations of inventive neural systems are therefore regarded as "DAGUIs" and "DABUIs," respectively.

If we construct a specialized DAGUI such that its perceptron generates a numerically based figure of merit proportional to the rate at which it is witnessing satisfactory pattern solutions from the imagitron, the networks equilibrate, with synaptic noise level automatically moving into the V regime (Thaler 1997c). This equilibrium arises due to the inherent insufficiency of novel, problem-solving patterns within the U domain and the sparseness of coherent patterns in the W regime.

Equating the imagitron with the brain's neo-cortex, I conjecture that the brain resides largely within the vicinity of the V regime of synaptic perturbation, essentially riding a cusp separating rote and novel pattern generation. As noted above, brain modality can thereby shift catastrophically from mundane stream of consciousness to more inventive ideation purely through the adjustment of the statistical average of synaptic perturbation, $<\Delta w>$. In neurobiology and the interconnected endocrine system, environmental stress can result in the secretion of appropriate neurotransmitters to alter long term potentiation, allowing us to consider that which has not been directly experienced or pondered before. In other words, the ability to rapidly bifurcate into false memory generation is favored by Darwin so as to allow effective strategy generation under traumatic, life-threatening circumstances. What we would consider convergence toward a viable solution would be marked by the subsidence of stress-related neurotransmitters such as adrenaline, as they are swamped out by less perturbative molecular agents such as serotonin and dopamine.

Depending upon the synaptic noise level within the imagitron, this neural cascade may interact with its environment in three fundamentally different ways. Referring to Figure 3, imagitrons and perceptrons may operate at very low noise levels, making them most attuned to the environment. The imagitron may serve as an associative memory, comparing any input environmental pattern, **E**, against the memories stored within it. Any patterns deemed novel through this comparison process (via reconstruction error, δ, Thaler 2000) may be selectively passed to the perceptron to access the value, utility, or threat thereof.

As mean synaptic noise level is raised into the U and V regime, the imagitron may either straightforwardly or creatively interpret the input stimulus, **E**, by activating into several rival memories or confabulations that are alternating due to synaptic disturbances. A context-aware perceptron (connections to environment not shown) may then maintain such noise so as to juggle these competing E-interpretations until the perceptron's "understanding" of the environment pattern is consistent with the overarching circumstances. At that time, the perceptron stage modulates the synaptic noise toward zero, effectively freezing in the environmental pattern's most favored interpretation.

Given sufficiently high levels of synaptic fluctuations, the imagitron is vastly more sensitive to internal disturbances than to the succession of environmental patterns, **E**, appearing at the network's inputs. It is within these V and sometimes W mode imagitrons that the equivalent of "eyes-shut" discovery takes place, with ideas synthesized from the combination of either intact or degraded token representations of world features.

Obviously, Creativity Machines may become much more complex than just the canonical, two network system described above. To facilitate their description and function, whether

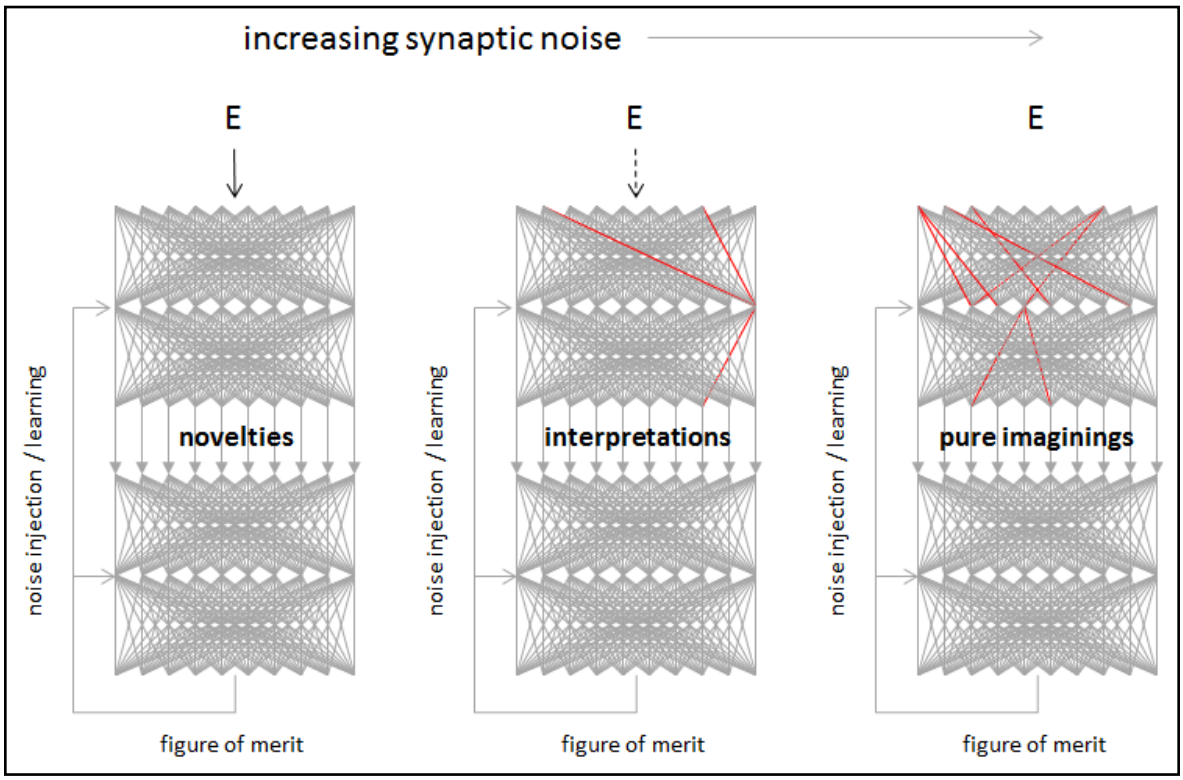


Figure 3. Changing Function of a Creativity Machine with Increasing Synaptic Noise Levels. As the perceptron injects increasing levels of synaptic noise (red weights) the system transitions from recognizing environmental patterns of interest, to inventive interpretation of things and events in the environment. With even more noise, the network becomes "attention deficit," freely imagining based upon a mixture of stored memories and derivative confabulations.

synthetic or biological, I have used a symbolism of my own making (Thaler 1996) that represents observing perceptrons by the letter "O." For instance, the cognitive feat of disambiguating some environmental pattern is describable as an E-U=O process and the "eyes shut" brand of creativity is denoted as V=O, with the equal sign conveying the reciprocal dialog between the V and O neural agencies.

More ambitious forms of discovery involving the identification of multiple imagitron assemblies simultaneously activating into juxtapositional concepts may be denoted as "U_iV_j=O" discovery[3] wherein any number of memories (U_i) and confabulations (V_j) may link into new combinations of tokenized entities or actions that are all collectively "judged and nudged" via a perceptron, O. Such juxtapositional discoveries can span the range of cognitive tasks that include the pragmatic combination, for example, of box, wheel, and axle memories to produce the epiphanal pattern of a wheeled vehicle, or the association of a deductive conclusion from combined predicates.

In demonstrating that a Creativity Machine can have thoughts about its thoughts, the O stage is critical because it is responsible for not only recognizing useful memories or confabulatory patterns, but also elevating synaptic perturbation until it is satisfied with the imagitron's output. Typically, the activation level of one or more output neurons, representing some figure of merit, can modulate the noise levels injected into the imagitron. In the simplest of cases, the perceptron could conceivably incorporate just one output neuron, continuously activating from a value of 0, symbolizing satisfaction, to an excitation of 1, representing utter discontent. That single output could in turn be tied with the effectiveness of any past ideas upon the environment, as learned through cumulative training.

Whereas such a simple perceptron would not lead to a complex chain of associations I have spoken of as a gestalt, it does produce a parade of memories and potential ideas in what might be considered to humans as frenzy. Having found a useful solution pattern, the perceptron could utilize its near zero output to modulate the imagitron's noise proportionately, thereby latching onto the currently activating pattern in a process tantamount to satisfaction and perhaps even ecstasy. More complex Creativity Machine designs are capable of producing the complex associative gestalts that "tag" neural assemblies capable of taking charge of the imagitron's injected noise level. As these specialized networks squeeze off the equivalent of adrenalin or serotonin, they are simultaneously activating into an evolving chain of associations. That these are not the kinds of associations humans experience is irrelevant. They are pattern-based associations none the less.

Recent improvements to the fundamental Creativity Machine architecture involve both perceptrons and imagitrons that are capable of adaptation (Thaler 2008), as symbolized through an asterisk. So, in the example of the passive V=O architecture, the V*=O* variation allows the perceptron stage to trigger reinforcement learning of confabulations deemed promising through the perceptron's opinion formation process. In this way, novel patterns deemed useful through the perceptron's associative gestalt are reinforced as memories within the imagitron. Simultaneously, the mapping between imagitron output patterns and the perceptron's predicted figure of merit is likewise perfected through additional training cycles. Implicit in this architecture are actuators fed by imagitrons to effect the environment, and sensors feeding perceptron outputs to assess the effect of such concepts or strategies upon the environment or the neural system itself.

The operation of this newest form of Creativity Machine (DABUI) should make introspective sense: In one instant we may have a brilliant idea, but in the next the revelation becomes only a memory. From a dynamical perspective, a perceptron may "take a liking" to an imagitron's activation state represented by a mountain top in the attractor landscape, thereafter transforming this same feature into a deep attractor basin through reinforcement learning. Subsequently, such new attractors, representing advantageous concepts or strategies, may be further mutated and merged into even better ideas through continuing cycles of synaptic perturbation and reinforcement learning. The overall effect is that DABUI operation, although initially stochastically seeded, becomes progressively more systematic as the perceptron intelligently triggers the storage and recombination of memories within a dialog of ever-growing sophistication.

In the latest and most ambitious DABUIs, the core perceptron-imagitron pair is able to instantiate additional neural modules that are gradually annexed to create vast brain-like pathways. In this application of the paradigm, confabulatory patterns represent candidate dimensioning and positioning strategies for these auxiliary nets, with the perceptron stages sensing the "wisdom" of the tentative architecture based upon the performance of other self-recommended architectures. Such performance may be gained through human mentorship or through the system's own self-defined objectives.

All in all, DABUIs represent a vastly generalized and even more rigorous and quantitative version of Baars' (1997) Global Workspace Theory (GWT) in which telephone numbers may be rehearsed in U-mode imagitron function. Speech may be formulated or visual art conceived at V levels of synaptic noise. Within the "theater of mind" originating such ideation, imagitrons serve as stage actors and perceptrons, the audience.

Aside from the vast utility and power in modeling GWT-style cognition, which I and others (Boltuc 2007, 2009) differentiate from consciousness, I point out a subtle process taking place within the DABUI that may have a significant consequence upon the subsequent discussion. As the former net nucleates a candidate concept or strategy upon injected synaptic noise, both nets simultaneously observe both the outgoing stimulation of and the incoming response from the environment. The imagitron component preferentially learns those stimulus patterns whose environmental response satisfies the perceptron while the perceptron stage perfects its mapping between said stimulus and response. In the process, a language is automatically built up understandable only to the networks involved in what is tantamount to a first-person perspective, involving otherwise indecipherable activation patterns that the philosophy of consciousness regards as *qualia*.

2.4 Creativity Machines and Consciousness. Heretofore I have mostly spoken of the Creativity Machine primarily in a pragmatic sense, as a simple and canonical neural architecture for invention and discovery, but I envision it as a model of so much more, namely, consciousness itself and how to implement machines that have thoughts about their thoughts.

Peering into the brain as scientists engaged in the process of free inquiry, all we see are evolving patterns of neural activation. However, querying the human test subject undergoing the functional brain scan, one hears a very subjective account of the overall conscious experience dominated by two very salient features: (1) the inexorable parade of memories, ideas, and sensations that seem to originate from nothingness, a stream of consciousness, so to speak, and (2) a reaction to that parade via emotions and what many have called the intrinsic "buzz" of consciousness that we associate with the hard problem (Chalmers 1995). The primary question then becomes one of how to resolve these diametrically opposed perspectives.

Just for a moment, allow me to pessimistically conjecture that consciousness isn't what it's hyped to be and that

intrinsically, it is just the evolving pattern of neural activations. If that's all there really is, then some creative process is required to relate a mechanism to what most of the human race considers mystical, profound, and inimitable. As I have already demonstrated, the Creativity Machine Paradigm is the fundamental neural architecture for achieving this end, especially when the apparatus involved, the brain, functionally consists of only neurons, synaptic interconnects, and a form of long range chemical connectionism represented, for example, by the endocrine system.

Let us assume that the Creativity Machine is at the heart of consciousness, not the kind related to attentional awareness, but to our inner mental experience and the so-called "subjective feel." After all, one may place a test subject into a sensory deprivation chamber, blocking visual or auditory input, allowing more visceral sensations such as warmth and wetness to habituate into nothingness. At this point, the stream of thoughts and the reactive associations are modeled by the inattentive Creativity Machine appearing in the right panel of Figure 3, wherein the turnover of memories and confabulations is primarily governed by the random noise fluctuations introduced into the imagitron. The succession of thoughts (a.k.a., thinking) trigger output patterns within the perceptron that are tantamount to the associative gestalts we have about such meandering thoughts.

Figure 4 summarizes what at this point is still a hypothetical model of how consciousness can arise in the brain via Creativity Machine Paradigm. Ubiquitous, energetic fluctuations (noise) drive a succession of memories and confabulations tantamount to thought, with absolutely no qualification that they be accurate or productive in nature. By virtue of connections to another neural assembly, associated patterns form, chain, and often loop in response to the evolution of faux things and events in the former neural assembly. Imagery of scenarios in the former assembly may evoke a chain of associated memories in the latter, all of which have been formed via the known sensory channels. That is why when we have feelings, we express them as though they are like something else. Such analogy chains form up, decay into others, and that is essentially the feelings we have of any thought. It is certainly true that there is no particular perceptron in the brain that has "good" and "bad" output nodes. Nevertheless, when we have an idea that is favorable to our being or livelihood, the associative chains formed include pleasant experience, including virtual, physical sensations. Having thoughts related to threat or adversity, the associative gestalt may include memories of physical pain that may trigger stress related neurotransmitters that keep the imagitron stage churning out progressively twisted notions until the perceptrons are satisfied.

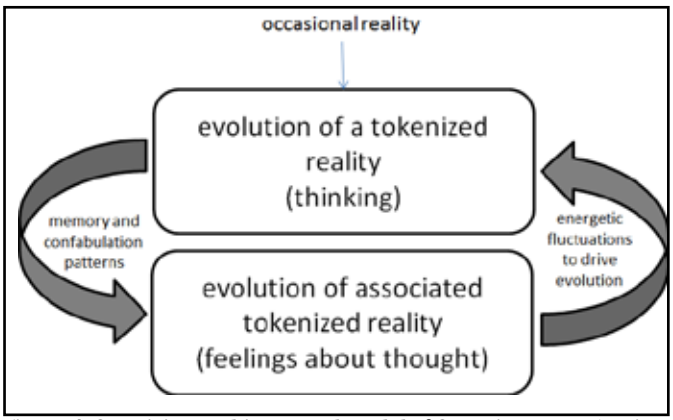


Figure 4. Creativity Machine Based Model of Consciousness. A noise-driven stream of tokenized world features activate within imagitrons, emulating so-called stream of consciousness or thought. Associated thoughts, known as feelings, nucleate in response to the imagitrons' stream of consciousness. They consist of chains and loops of memories gleaned from the sensory channels related to sights, sounds, and sensations such as physical pain and pleasure.

A salient aspect of Figure 4 is that consciousness is a loop from which the only escape is death or brain injury. There is no monitoring mechanism therein that can allow the brain to understand itself at the level of its synaptic organization and the momentary disturbances to such connections. Of nearly equal saliency is the fact that everything about this process is for all intents and purposes, bogus: The upper, imagitron stage is harnessing energetic disturbances to create a succession of entities and scenarios, none of which is real. Similarly, the lower, perceptron stage is producing likewise counterfeit impressions of this virtual reality, through associative chains and loops connecting memories and confabulations drawn from prior sensory experience. In effect, the entire process is an illusion, but the overall advantage is very real, namely, to preserve the life of the host organism and to provide survival advantage over other organisms.

Though the process may be an illusion, it may operate in a wealth of modalities that represents all aspects of inner mental life (Figure 5), again tied to one essential system feature, the mean synaptic fluctuation, $<\Delta w>$ within the imagitron stages.

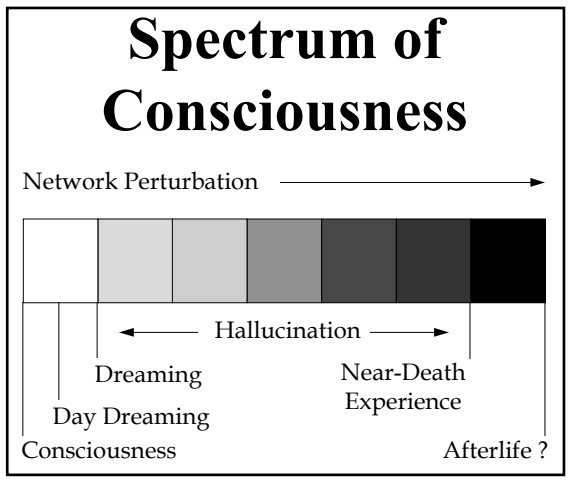


Figure 5. The Single Parameter underlying the Full Gamut of Conscious Experience, Synaptic Perturbation Level (from Thaler 1997c). Here "network perturbation" includes all synaptic and circuit-equivalent perturbations within neurons. However, because of the preponderance of connections over processing units, most disturbances are expected to be synaptic in nature.

For instance, in normal waking consciousness, imagitron assemblies and perceptrons are bathed in minimal noise, allowing them to lucidly detect anomalies in the environment (see Figure 3, left panel) as well as opportunities and threats therein. In daydreaming, heightened noise levels, at least within the cortical imagitrons, lead to attention deficit as internal activation turnover dominates over activations seeded by external events. In the resulting reverie, the noise level is sufficient to produce confabulatory entities and scenarios representing potential, alternative realities.

Effectively cut off from sensory input and the mean level of synaptic perturbation increased, the Creativity Machine

architecture can dream. Such synaptic fluctuations are likely essential to the transmogrification of entities and the intrinsic and nonsensical discontinuities within reported dream sequences. So, whereas ponto-geniculo-occipital (PGO) waves originating from the diencephalon (Hobson 1993) may seed the image in visual cortex of a tiger charging us, the resulting adrenaline rush can suddenly transform the big cat into a dove.

Within trauma or drug-induced hallucination, both imagitron and perceptron stages are subjected to intense synaptic fluctuations, leading to not only transmogrify of absorbed features, but also misinterpretation by perceptrons of noise-seeded entities and scenarios simulated within imagitrons.

Finally, within near-death experiences (NDE), it is plausible to assume that the entire gamut of noise levels is visited, beginning with stress-induced neurotransmitter release that overwhelms the sensory channels with an internally generated succession of memories tantamount to life review. Thereafter, cell apoptosis effectively nullifies synapses in an irreversible form of perturbation, wherein memories and then confabulations nucleate upon patterns of what appear to the surviving portions of the network as resting state (i.e., zeroed) neurons (Thaler 1995). It is my suspicion that: (1) perceptron modules dedicated to distinguishing reality from mental imagery become less adept at doing so, and (2) that other perceptrons, sensing a growing cascade of virtual events, mistakenly perceive that they are experiencing eternity. All then fades to black with a torrent of illusion, a fitting finale for a life of cognition based upon the same (Thaler 1993, 1995, 2010) in what may be described as a virtual brand of afterlife denoted with a question mark, "?", in Figure 5.

All aspects and life stages of human cognition can be imitated using the fundamental $U_iV_j=O_k$ architecture wherein multiple imagitrons, in both U and V modes, are under supervisory control by many perceptrons, such governance being exercised through average synaptic perturbation level. Throughout all these conscious modalities, imagitrons and perceptrons are mutually learning from one another to create a private and evolving language exercised between them that I identify as the first-person, subjective experience at the core of so-called "h-consciousness" (Boltuc 2011). The same adaptive encryption scheme may be achieved in machines based upon perceptron-imagitron ensembles, clearing the way for the engineering of machine consciousness.

In demonstrating the equivalence between human and machine intelligence, Turing relied upon gedanken experiments in which machines were remotely interrogated via natural language. To this great visionary, imitation of human behavior was sufficient to demonstrate equivalence. Currently we need not bother with the exchange of words to appraise the consciousness of machine intelligence muted by design. Instead, we may watch and compare the operation of both a Creativity Machine and a brain, side by side, with DABUIs monitored through graphical user interfaces and brains observed via the latest functional brain scan techniques. Within each of these systems we observe an evolution of activation patterns with one pattern ostensibly triggering the next. With causality smeared through this inherently cyclic process reminiscent of Figure 4, it should make perfect sense that ideas and feelings about such ideas become one and the same, simply an endless chain of patterns spawning other such patterns.

Now, through thirty-seven years cumulative experience with both the Creativity Machine and "easy chair neurobiology," I feel that we may now emulate all modalities and life-cycle aspects of this ostensibly complex and conscious computing scheme through the adjustment of just one simple parameter, the mean synaptic perturbation.

3.0 Dealing with the Other Objections. Having proposed a neural architecture that may implement the core phenomena of consciousness, it seems that the other objections to the very notion of thinking machines fall into place:

3.1 The Theological Objection. Like Turing, I am not impressed with the theological position that thinking is a function of man's immortal soul. However, in contrast to Turing's view, it is the Creativity Machine, and not generic AI, that is effectively a "mansion" for what many perceive as an immaterial spirit integral to the brain (Thaler 2010). Depending upon the experience of one's perceptrons, the system occupant can be that which defies definition such as supernatural entities, or, as in my case, a statistically definable average of energetic fluctuations among the synapses and neurons of a neural network.

When I was very young I entertained the former concept. Later in life my interpretation toggled to the latter view point, with my perceptrons appropriately biased through many cumulative experiments in synthetic psychology.

3.2 The "Head in the Sand" Objection. Turing accurately predicted one of Hollywood's principle money-making themes, that the consequences of machines thinking would be too dreadful (i.e., "Terminator" and "The Matrix"). Whereas these are theatrical scenarios involving human extermination or enslavement, there are less severe possibilities in store for humanity involving the mere intellectual humiliation of the species. In this vein Turing makes an extraordinarily perceptive observation that this objection would "likely be quite strong in intellectual people, since the value of the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power."

I resonate with Turing's observation on a daily basis wherein I interact with very knowledgeable individuals who are specialists within various problem domains. All but a few are nonplussed by my ability to rapidly absorb their chosen area of expertise into brainstorming assemblies of imagitrons and perceptrons to solve the problems they themselves have deemed top priority. Often denial and rejection, rather than glowing acceptance, is the result as rationales against the Creativity Machine methodology are stimulated via adrenaline rush. Later, through patient inquiry, I often discover the revulsion caused by such a simple model of human ingenuity. Even more intense emotions erupt with their own revelation that their very consciousness may be reduced to that of a neural net bombarded by noise to create a stream of consciousness as another net develops an attitude thereof.

Looking into the future, I see this objection continuing, with humanity producing more reasons why such thinking machines, most notably Creativity Machines, aren't really thinking. Ironically, though, they will be harnessing Creativity Machine Paradigm within their own brains to generate such oppositional sentiments.

3.3 The Mathematical Objection. Citing Gödel's incompleteness theorems, Turing correctly predicts that many would reject machine intelligence based upon its inherent limitations, namely, the generation of statements by a logical machine whose veracity could not always be verified by the same closed set of rules by which said machine operates. He quickly dismisses this objection based upon the observation that human intellect likewise has its limitations and that oftentimes we may know that a notion is true, but are incapable of analytically proving so. Under pressure to seek such proof, we must creatively transcend the rules or principles exploited in idea synthesis and essentially find validation via another logical system, either discrete or fuzzy.

I would have to say that both the mind and the Creativity Machine share the same pathology wherein pattern generation can outpace pattern analysis. In effect, perceptrons may recognize the effectiveness or validity of a confabulatory pattern but, because of their non-contemplative function, can only intuit such utility. It is only after skeletonization of such perceptrons to comprehend the logic captured therein that the underlying logical schema are revealed, at least to humans or some externalized neural assembly.

In spite of not possessing such an onboard explanation facility, I would claim that the cognitive weakness of a Creativity Machine is also its strength, an imagitron's ability to err toward creative possibilities harnessing unintelligent noise, while monitoring neural nets instinctively select the best of these candidate notions. My suspicion is that this is the initial stage of great ideas and that through multiple drafts (Dennett 1997) the formal logic, mathematical symbolism, and explanatory narrative become just the icing on the cake.

3.4 Arguments from Various Disabilities. "…but you will never be able to make one do X" is another objection, intimating that a machine must possess the diversity of behaviors typical of a human. Whereas Turing points out that in his time, most would use logical induction to infer that narrowly focused machines of his time could not attain the flexibility characteristic of the human brain.

I, on the other hand would claim that both the cognitive and conscious aspects of the brain have been intensively, rather than extensively captured via Creativity Machine. That is to say that smaller implementations of the paradigm recreate narrowly focused cognition and a consciousness less rich than that allowed by the human experience. Simple scaling of the paradigm, adding a sensor suite far more extensive and capable than the human sensoria, and an actuator ensemble more adept than human hands, fingers, and feet, and we are now in the regime that should genuinely concern the "head in the sand" faction who might then themselves be regarded as disabled.

3.5 Lady Lovelace's Objection. In effect, the Creativity Machine is the epitome of generative artificial intelligence, perhaps forming the ultimate response to Lady Lovelace's Objection that state machines like Babbage's Analytical Engine" are incapable of originating any ideas on their own, or "taking us by surprise," as Turing himself semantically fine-tuned the Lovelace's critique.

Certainly, the Creativity Machine has produced concepts that have taken many by surprise, beginning with the generation of natural language, wherein a perceptron-imagitron pair exposed to sundry Christmas carols generated the controversial lyric (at the granularity of letters and words), "In the end all men go to good earth in one eternal silent night." Sales figures serve as testament to the paradigm shift product designs formulated by the architecture. Many have marveled at the ability of totally untrained neural models interconnected as perceptron-imagitron teams to develop totally unanticipated and sometimes unfathomable robotic behaviors to deal with newly arising scenarios on the battlefield or the factory floor.

Then again, critics have charged that the concepts generated by the Paradigm aren't that powerful, and its artistic creations not that moving. But then again, isn't that the case for any human originating within any conceptual space, surrounded by critics with all manner of perceptual biases and hidden agendas? Are we to then claim that such brains are not capable of thought?

In taking a strictly analytical view of the model of seminal cognition offered by the Creativity Machine, what really counts is that the monitoring perceptrons are taken by surprise with confabulatory outputs they have never before experienced, sometimes associating utility or value to such false memories. Anticipating that human brains are Creativity Machine based then such perceptron-imagitron assemblies operative in the low noise regime may first sense novelty to these freshly generated concepts, thereafter raising synaptic noise level to interpret them and evaluate their utility or value. If the consensus among societies of such neurobiological systems is favorable, in terms of novelty and utility, the concept becomes by popular fiat, an example of historical or H-creativity (Boden 2004). If, later, an archeological expedition finds evidence that the idea is an ancient one, or contact with a well advanced extraterrestrial civilization is made, attribution and perhaps historical status *may* change.

As Turing amplified, Lovelace viewed state machines of her time as capable of doing only what they were told. "Inject" an idea, representable as a pattern, into the machine and it will generate a response, in the form of another pattern, effectively responding, but then dropping into a state of quiescence. Essentially such "one-shot" operation represents that of the perceptron, which distinguishes itself from all other such mathematical transformations in that it crafts itself, using simple rules (i.e., Hebbian learning or back-propagation) that even "unintelligent" nature can supply given that the environment is providing ample input-output examples.

Turing further draws an analogy between mind and an atomic pile, noting that both can operate in a subcritical and supercritical level, the latter marked by a chain reaction, representing respectively cascades of fission neutron and ideational patterns. Figure 1 amply demonstrates that an imagitron, in particular a recurrent one, can be denied synaptic noise to the point that it operates in a one-shot mode, responding with a memory closest to the applied input pattern. However, raising it past the critical point, the network is always generating a new output pattern that in turn recirculates to produce a progression of activation patterns tantamount to contemplation in which secondary, tertiary, and more remote ideas form as associative chains we call theories. Monitoring perceptrons may likewise dynamically interconnect themselves in associative gestalts that may in turn mimic the positive or negative feedback that moves the mean synaptic perturbation level back and forth through the critical point $<w>_c$.

Ironically, Turing's analogy to an atomic pile is amazingly fitting, since sufficient proximity of one fuel element to another dictates a critical mass. So too in the case of the Creativity Machine, interconnecting one net with another achieves another kind of criticality that results in an avalanche of potential ideas!

3.6 Argument from the Continuity in the Nervous System. I would be prone to agree if it weren't for the efforts of pioneers in the field of artificial neural networks who could emulate discrete state machines using a system of analog synaptic connection weights. The power of the Creativity Machine stems from this transformation from discrete logic, if need be, to its analog implementation via continuous connection weights and, if need be, back to a binary representation. The analogic intermediate stage forms the basis of a convenient "handle" by which to manipulate the discrete aspects of the problem. That manipulation is the introduction of analog, synaptic disruptions.

3.7 The Argument from the Informality of Behavior. Turing points out the inevitability of objections to the possibility of machine intelligence based upon our inability to program it with rules for every conceivable set of circumstances. I feel that such an objection would be moot because it would rely upon the very definite fiction that human brains are equipped with rules to fit all occasions.

The truth of the matter is that we must often improvise rules for dealing with novel situations most likely drawing upon

Creativity Machine Paradigm to degrade heuristics implicitly absorbed within synaptic connection weights until monitoring perceptrons judge such logic effective. In other words, the rules appropriate to any given circumstance are not always stored as memories within the cortex. They are largely invented on the fly to either compensate for constantly fading memories or to deal with the emergence of a totally new situation as in the example cited by Turing wherein a driver is presented with contradictory red and green lights at a traffic intersection.

The human mind deals with this stoplight dilemma as a Creativity Machine would, with an imagitron alternately interpreting the environmental scenario as either a "go" or "stop" situation. Associated with these two alternative analyses are two separate kinds of associative chains that may form within a perceptron collective, one filled with acoustic memories of screeching brakes and police sirens, along with visual recollections of crumpled cars and bloodied bodies. The other possible associative gestalt may contain imagery of smooth sailing toward one's intended destination or imagery of one's home. As the perceptron assembly gets wind of additional environmental clues, such as the absence of cross-traffic and law enforcement, imagitronic interpretation shifts toward that of the green light and the driver ever so cautiously rolls through the intersection.

As the reader imagines this scenario, it should be intuitively clear that in the case of unambiguous green or red lights the driver response corresponds respectively to foot on the gas or on the brake, with the decision to execute such behaviors prompt and distinctive. In the case of the vague, mixed red and green lights, the reaction is tentative, perhaps requiring seconds rather than the usual 300 millisecond clock cycle of the brain. In this dilemma, the solution requires not a memory, but an idea, drawn from the confabulation of proceeding through the intersection under a red light. The latter requires more juggling of interpretation, more evolution of the perceptron's associative chains, and the arrival of additional contextual clues about the environment.

But such hesitancy, and in general, the rhythm with which thoughts emerge is that of the Creativity Machine as reported in 1997 (Thaler, ref. a) wherein the prosody of both human cognition and Creativity Machines were compared. The result, derived from the theory of fractal Brownian motion (fBM, Peitgen and Saupe 1988), is that both neural systems produce notions at arrival rates quantitatively equal to that of a neuron subject to random disturbances to its synapses, allowing the evolution of thought to be expressed through the equation,

$$\rho = k\Delta t^{-D}0 \qquad (1)$$

Where ρ = the microscopic, synaptic perturbation rate[4] of a representative neuron, Δt the time to evolve N distinct patterns (or thoughts), D_0 the fractal dimension of the macroscopic succession of these patterns, and k a dimension preserving constant. What we find is that in both the human and Creativity Machine cases, inventive tasks, such as the time-intensive interpretation of an ambiguous stop light, occur at lower fractal dimension near zero, while the recollection of memories, standard operating procedures at intersections, occur at nearly linear rates wherein D_0 approaches 1. In effect, Equation 1 expresses the informality of behavior we all witness when listening to articulated thought (i.e, speech) wherein we hear a linear, homogeneously dispersed series of words when the speaker is rehearsed, versus tentative and irregularly spaced annunciations accompanying improvised thought.[5 6]

Further, D_0 is found to be a function of the microscopic, synaptic perturbation, which in turn may be imagined as the product of n, the number of perturbative agents (i.e., rogue neurotransmitters), and σ, the magnitude of synaptic perturbation deliverable by each such agent. It is found experimentally and theoretically that large synaptic fluctuations (large n or σ) lead to confabulation generation, whereas for ($n \approx \sigma$), the neural network remains on even keel, generating rote memories tantamount to a mundane stream of consciousness.

If this model is correct, then cognitive hesitancy is not due to the "hardness" of a challenge, as we have led ourselves to believe, but to large fluctuations in synaptic perturbations delivered to our brain's imagitrons. To make a machine imitate the informal speech pattern of a human, one doesn't need a sophisticated computer algorithm based upon tedious statistical studies. Instead, simply bombard the synapses of one or more neurons with random noise. To make it sound stressed, flood theses synapses with higher levels of noise. To calm it, lessen the mean disturbance levels. Never mind the wisdom or accuracy of its thoughts. It is simply thinking . . . .

3.8 The Argument from Extrasensory Perception. While not fully convinced of the existence of this phenomenon, allow me to introduce the following gedanken experiment designed with the intent of allowing two brains to intimately know each other's thoughts. Visualize human subject A's neural nets to be fused with those of subject B. Then, try as we may, A's neural nets can only interpret B's thoughts (via the interpretive scheme of Figure 3) in terms of its own idiosyncratic experience, and vice versa. Thus, even in intimate contact, there is no accurate mind reading, only error prone reinterpretation via the process known to neural network practitioners as *pattern completion*.

In a way, the Creativity Machine exemplifies a successful brand of ESP I have discussed in the context of subjective inner experience, since imagitron and perceptron live alongside one another, and through the sharing of common cumulative experience, acquire the "Rosetta Stone" for interpreting each other's otherwise cryptic activation patterns.

Similar co-habituation of brains within groups or societies can achieve such instant interpretation, but only at the basic levels involving fear or opportunity. In this case, connection density is sparse between individuals, exploiting largely the powerful electric fields produced by diffusing airborne molecules (i.e., pheromones), acoustic waves (i.e., cries for help), and visual, behavioral anomaly detection using neural network implemented novelty filters (a child missing in the night).

4. Conclusions. Let's work backward from the counterintuitive and possibly nightmarish position that there really is no biological consciousness, the attribute most commonly cited as lacking in machines. If that is the case, then there would be only generic neural activity in the brain, the complex but zombie-like succession of activation patterns that we can undeniably detect in functional brain scans (albeit at low resolution using contemporary techniques). Given this nihilistic position, some equally mechanistic brain methodology would be required to allow significance to be invented to a process that intrinsically had none, namely, another neural mapping that non-contemplatively associated such pattern activation with the overall neural assembly's past experiences.

Compounding the pessimism, let us assume that the parade of memories, sensations, and ideas is not because of some noble and intelligent process, but mere pattern turnover driven by the energetic fluctuations bathing this connectionist system.

Bleaker yet, consider that the associated pattern chains, based either upon their congenital design or cumulative learning, may also incorporate colonies of neurons whose purpose is to modulate the random and unintelligent synaptic

fluctuations, based upon the co-excitation of certain pattern-based memories that influence the rate and nature of pattern turnover.

And then, as the final humiliation, deny this system any facility at all by which it may monitor itself at the neuronal and synaptic level. Instead, let it familiarize itself with itself via an inherently counterfeit, tokenized reality that all of its component neural colonies have "settled upon" as a common, instinctive, and automatic language.

If this were the wretched case, then:

1. Among ensembles of such systems, natural selection would favor those within which the associative response to such a generic neural activation turnover was least stressful, allowing these neural assemblies to stabilize themselves through a favorable self-interpretation that would then become habituated both individually and collectively. Amounting to an incentive for self-preservation, such indoctrinated perceptrons would selectively weaken any accidental activation of imagitron activation patterns denoting a sense of kinship with a system of inorganic switches and interconnects.

2. Without the necessary in situ probes to monitor energetic fluctuations occurring within their synapses, the monitoring portions of these zombie-like systems would only experience a succession of tokenized entities and scenarios that are somewhat representative of the external world. These fictions would certainly be functional in problem solving and acts of discovery and creativity, but for the most part such materialization of thought to them would be tantamount to rabbits emerging from a magician's hat. Nevertheless, such systems would simply habituate to the legerdemain as something routine that may be taken for granted.

3. In all but the most straightforward problems, cognitive tasks would typically take serpentine paths toward premeditated objectives, in contrast to a direct, logical path. Such intrinsic meandering would reflect the randomness underlying the succession of neural activation patterns. In particular progress toward an ideational goal would be desultory, most like the Brownian diffusion of molecules (i.e., neurotransmitters).

4. Their world models would be intrinsically faulty in fully simulating the external reality simply because they would not possess the degrees of freedom required to exhaustively model those of the external universe. Instead they would be forced to develop only semi-successful theories of their surrounding environment based largely upon limited, tokenized representation of the world's entities and mechanics. Immense spaces of ideational possibilities would be created via the enormous combinatorial space offered through synaptic degradation schemes, with the most captivating of these notions subsequently converted to memories at the discretion of monitoring perceptrons.

5. Inner mental life of these neural systems would be based largely upon the intensity and distribution of unintelligent noise internal to them rather than the intermittent contacts with the outer reality. Such dominance within their conscious awareness of inner over outer experience would be due to the sheer preponderance of the number of synapses, a volume effect, over sensory neurons, a surface effect. There would then be a fine line between cognitive processes such as contemplation and hallucination.

6. Function of these neural systems would be limited by an intrinsic bottleneck separating the generative and pattern recognizing elements, with the latter neural assemblies tantamount to a reptile surveying its environment for a tasty insect. As such, many potential revelations nucleating within imagitrons (i.e., cortex) would go undiscovered when the watching components (i.e., reptilian brain) were momentarily distracted, unable to simultaneously devote attention to multiple targets. This intrinsic disability would likely be played up as the noble search for an idea thus contributing to a favorable and stabilizing associative gestalt.

7. The cognitive turnover of these neural assemblies would possess a signature rhythm, marked by hesitancy as they creatively reach for new ideas or strategies, or prompt linearity as they interrogate themselves for stored memories. Such prosody would be temptingly close to that produced by the random disruptions to the synapses feeding a representative neuron therein.

8. Such assemblies would be susceptible to numerous pathologies related to their ability to generate useful notions distinct from their direct experience (i.e., ideas). For instance, overloaded by perturbative agents (i.e., neurotransmitters and neurohormones), they could easily dissociate from the surrounding reality as well as soften the synaptically absorbed rules within the perceptrons used by them to separate fact from fiction. In effect, there would be another fine line separating historically novel idea generation (i.e., genius) from erratic fantasy (i.e., insanity).

9. After prolonged observation of their world through a layer of token reality and fantasy-like confabulations, it would be difficult for them to distinguish between these two forms of attractor basins within their dynamical landscapes. Oftentimes, factual information would be abandoned on the basis of being too mundane or pessimistic. Fantasy deemed exciting or comforting would sometimes become well habituated as memories indistinguishable from direct experience.

10. After prolonged periods of simultaneously experiencing their environment, all neural modules involved would mutually learn the meaning of each other's activation patterns, memories and fantasies included. As such assemblies equilibrate with one another a secret language would arise, knowable only to one another. Within this neural lingo would arise the subjective, "raw feels" we commonly refer to as qualia. The veracity and validity of such feelings would not be guaranteed. They would just occur.

In many respects, the objective reality is likely even harsher than the all too familiar scenarios enumerated above, with the fundamental cognitive loop of the brain imprisoned within genetically perfected illusions that include an imagined sense of supremacy over mere mechanisms. That is why we cannot rely upon Gallup poles, as Dr. Turing emphasized, to arrive at a scientific determination of what separates mind from machine. Underlying such a consensus would be individual brains inventing significance to themselves at both visceral and intellectual levels.

However, there will be a conceptual "jail break" as a few minds reach beyond the illusory and challenge the rest to describe at least one, just one, neurobiological mechanism that could be effective at neutralizing the conscious paradigm discussed at great length herein. Patiently waiting for an answer to this question, this minority would likely seek an equivalency test between human and machine intelligence that significantly differs from that of Turing's imitation game. This new test would amount to the direct observation within both biological and synthetic neural systems of patterns of neuronal activation nucleating upon noise within the synaptic sea within which they are immersed, with perceptrons forming the associated patterns we have come to know as feelings. From this novel perspective the brain would be viewed as nature's attempt at rigging a Creativity Machine from the available protoplasmic resources using a very strongly encrypted, pattern-based, communications scheme.

The tradeoff is obvious. Our egos will be bruised, but by harnessing this paradigm we will attain machine intelligence capable of trans-human level discovery and invention. If he were with us, Turing would consider this quite an optimistic outcome for such a mechanistic outlook.

Acknowledgements. Dr. Peter Boltuc was instrumental in motivating the writing of this paper. I offer my sincere gratitude to him in directing me to revisit A. M. Turing's work, within the context of the Creativity Machine. I also find camaraderie and confirmation in his scientifically based stance that consciousness may be engineered in machines.

References

Baars, B. J. 1997. In the theatre of consciousness: global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4:292-309.

Boden, Margaret. 2004. *The Creative Mind: Myths and Mechanisms*. New York: Routledge.

Boltuc, P. 2009. Replication of the hard problem of consciousness in AI and Bio-AI: an early conceptual framework. In *AI and Consciousness: Theoretical Foundations and Current Approaches*, eds. Anthony Chella & Ricardo Manzotti. Merlo Park, CA: AAAI Press.

Boltuc, P. 2009. The philosophical problem in machine consciousness. *International Journal of Machine Consciousness* 1.1: 155-76.

Chalmers, D. 1990. Consciousness and cognition. Unpublished. http://consc.net/papers/c-and-c.html.

Chalmers, D. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200-19.

Dennett, D. 1991. *Consciousness Explained*. Boston: Little Brown and Co.

Hambling, D. 2006. Experimental AI Powers Robot Army. *Wired* http://www.wired.com/software/coolapps/news/2006/09/71779?currentPage=all.

Hesman, Tina. 2004. The machine that invents. *St. Louis Post-Dispatch*, Jan. 24.

Kahn, D. and Hobson, A. 1993. Self-organization theory of dreaming. *Dreaming* 3.

Peitgen, H. and Saupe, D. 1988. *The Science of Fractal Images*. New York: Springer-Verlag.

Plotkin, R. 2009. *The Genie in the Machine*. California: Stanford University Press.

Rowe, J. and Partridge, G. 1993. Creativity: a survey of AI approaches. *Artificial Intelligence Review* 7:43-70.

Thaler, S. L. 1995a. Death of a Gedanken creature. *Journal of Near-Death Studies* 13(3).

Thaler, S. L. 1995b. "Virtual input" phenomena within the death of a simple pattern associator. *Neural Networks* 8(1):55-65.

Thaler, S. L. 1996. A proposed symbolism for network-implemented discovery processes. *Proceedings of the World Congress on Neural Networks 1996*. 1265-68. Mahwah, NJ: Lawrence Erlbaum & Associates.

Thaler, S. L. 1997a. A quantitative model of seminal cognition: the creativity machine paradigm. http://imagination-engines.com/iei_seminal_cognition.htm. Mind II Conference, Dublin, Ireland.

Thaler, S. L. 1997b. U.S. Patent 5,659,666. "Device for the Autonomous Generation of Useful Information," issued August 19, 1997.

Thaler, S. L. 1997c. The Fragmentation of the Universe and the Devolution of Consciousness." U. S. Library of Congress, Registration No. TXU00775586.

Thaler, S. L. 1998. Predicting ultra-hard binary compounds via cascaded auto- and hetero-associative neural networks. *Journal of Alloys and Compounds* 279:47-59.

Thaler, S. L. 2000. U.S. Patent 6,014,653. Non-Algorithmically Implemented Artificial Neural Networks and Components Thereof, issued January 11, 2000.

Thaler, S. L. 2008. U.S. Patent 7,454,388. Device for the Autonomous Bootstrapping of Useful Information, issued November 18, 2008.

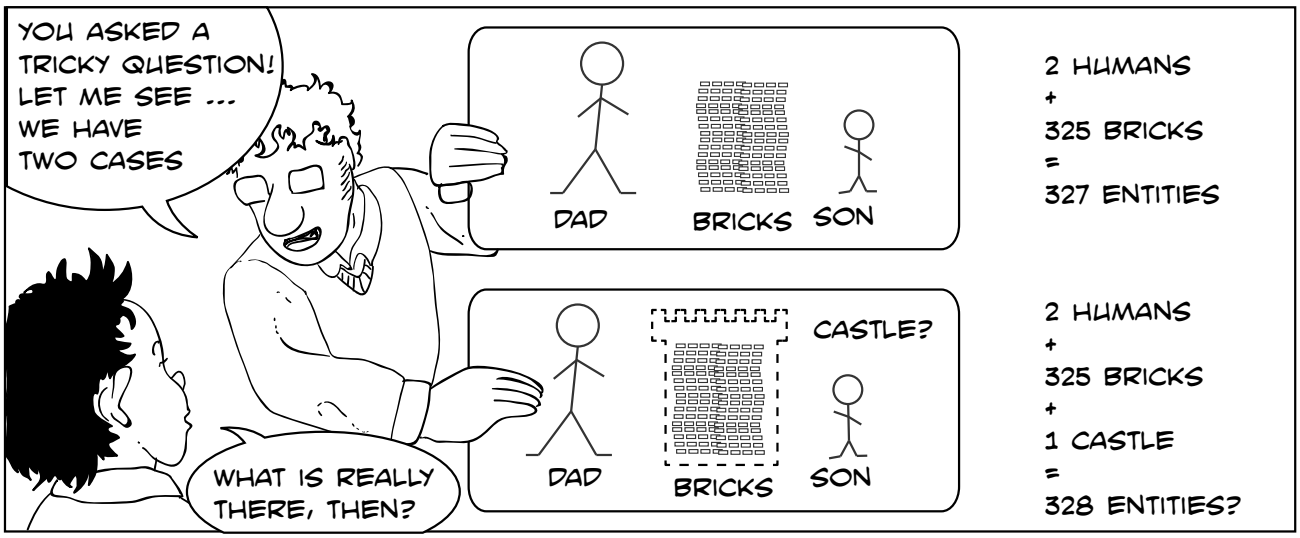
Thaler, S. L. 2010. Thalamocortical algorithms in space: the building of conscious machines and the repercussions thereof. In *Strategies and Technologies for a Sustainable Future*, ed. Cynthia G. Wagner. World Future Society.

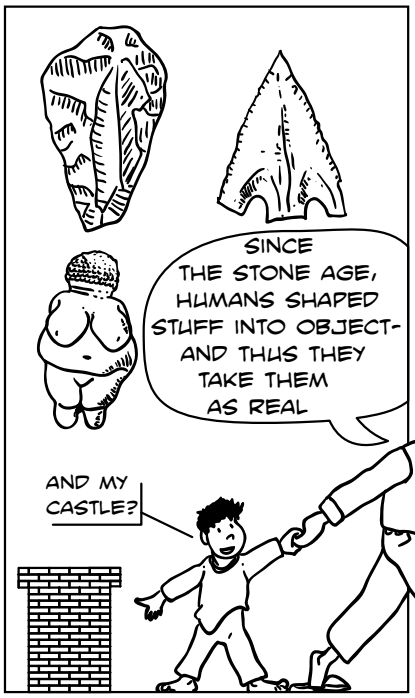
Turing, A. 1950. Computing machinery and intelligence. *Mind* LIX (236):433-60.

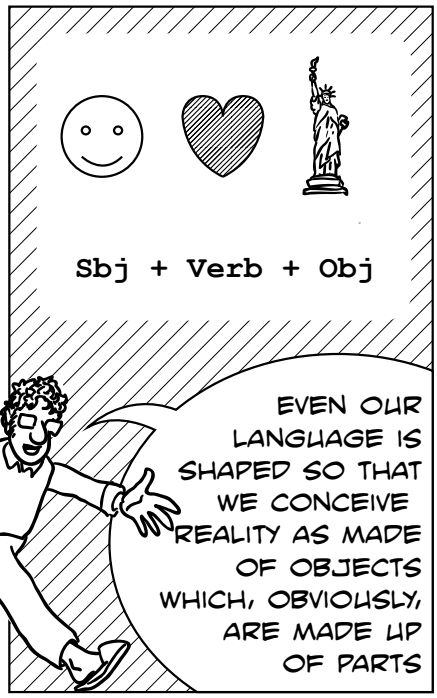
Endnotes

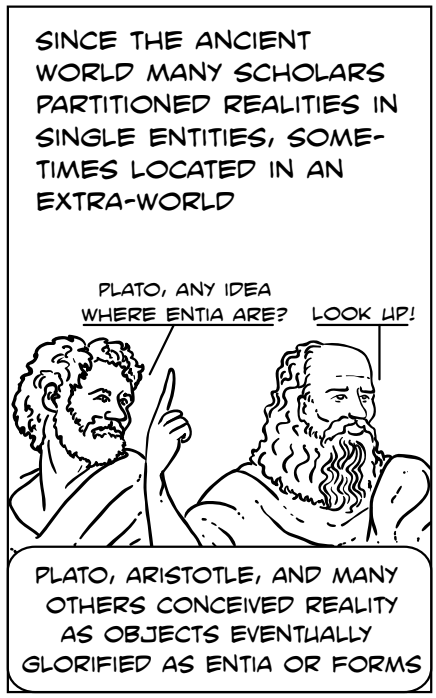
1. A recurrent, auto-associative neural network.
2. Each data point represents 1,000 experiments conducted on the network at the mean synaptic perturbation level indicated. A memory is defined here as an output pattern within 5 percent RMS error from the training pattern it is closest to.
3. With all indices implicitly repeated.
4. Effectively a constant, at the critical perturbation cusp, $<w>_c$
5. Taking the log of both sides of Equation 1, we find that fractal dimension, D_0, should linearly scale with $1/\ln\Delta t$. Both articulated human though (i.e., speech) and synaptically perturbed artificial neural networks closely obey this relationship.
6. This relationship is essentially the dynamical equation behind any nested system of entities and events, either a multilayered neural net or the world in general. After all, the brain, a biological neural net, is a world model, driven by energetic fluctuations just as its environment.


DO OBJECTS EXIST OR TAKE PLACE ?
@ 2012 BY RICCARDO MANZOTTI
LET'S TAKE A SIMPLE
LET'S TAKE MORE SIMPLES
AND MORE!
WE HANDLE PORTIONS OF PHYSICAL STUFF
SOMETIMES THESE PARTS GET TOGETHER
EITHER BY NATURAL OR BY ARTIFICAL MEANS INTO ...
12 FEB 2012
WE DID IT! WE BUILT THE CASTLE AS I PROMISED YOU ... WELL ... A FEW WEEKS AGO?
MMMMMH ARE YOU SURE, DAD? AFTER ALL WE STILL HAVE JUST A BUNCH OF PLASTIC BRICKS. WE JUST CHANGED THEIR POSITIONS IN SPACE. HAVE WE REALLY CREATED SOMETHING NEW?
RICCARDO.MANZOTTI@IULM.IT
1

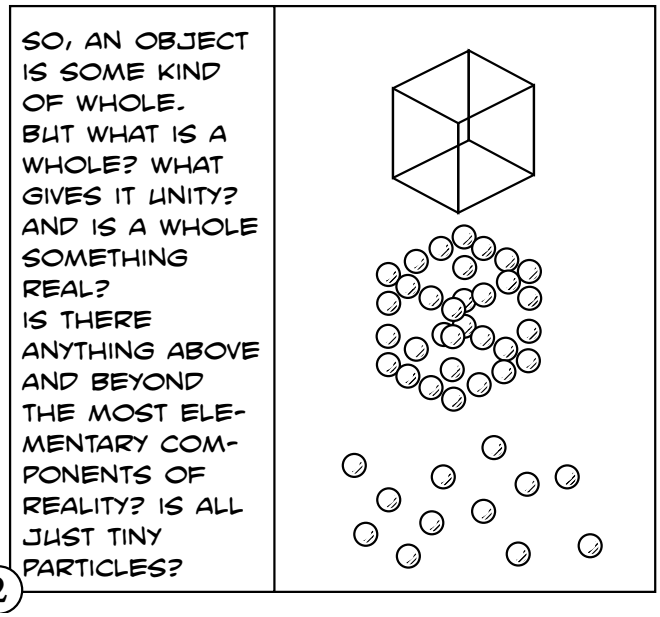

YOU ASKED A TRICKY QUESTION! LET ME SEE ... WE HAVE TWO CASES
DAD
BRICKS
SON
2 HUMANS + 325 BRICKS = 327 ENTITIES
DAD
BRICKS
CASTLE?
SON
2 HUMANS + 325 BRICKS + 1 CASTLE = 328 ENTITIES?
WHAT IS REALLY THERE, THEN?

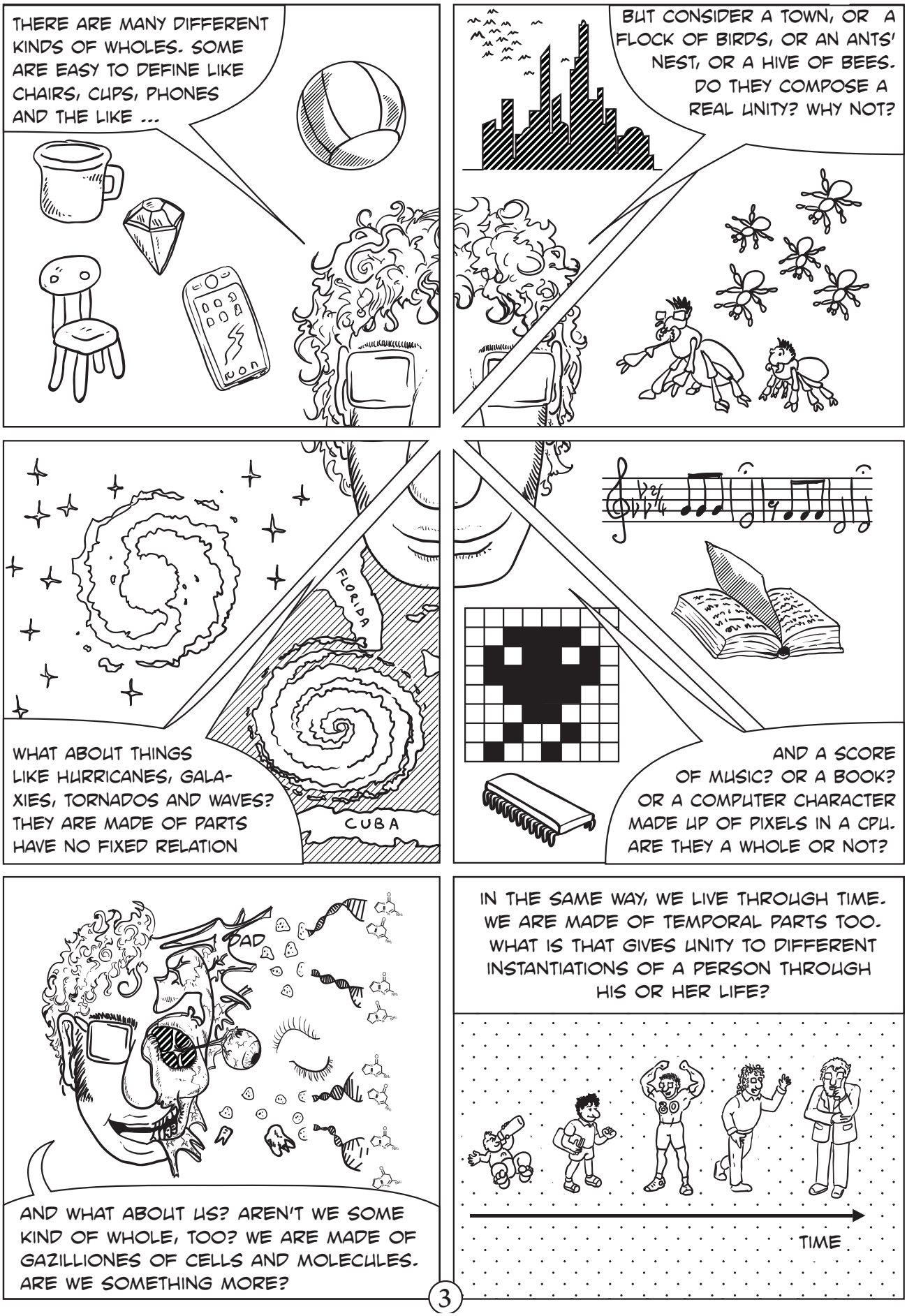

SINCE THE STONE AGE, HUMANS SHAPED STUFF INTO OBJECT- AND THUS THEY TAKE THEM AS REAL
AND MY CASTLE?

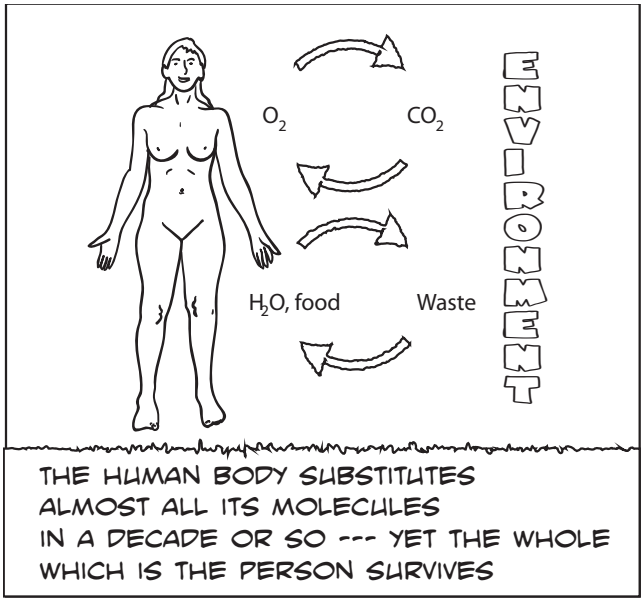

Sbj + Verb + Obj
EVEN OUR LANGUAGE IS SHAPED SO THAT WE CONCEIVE REALITY AS MADE OF OBJECTS WHICH, OBVIOUSLY, ARE MADE UP OF PARTS

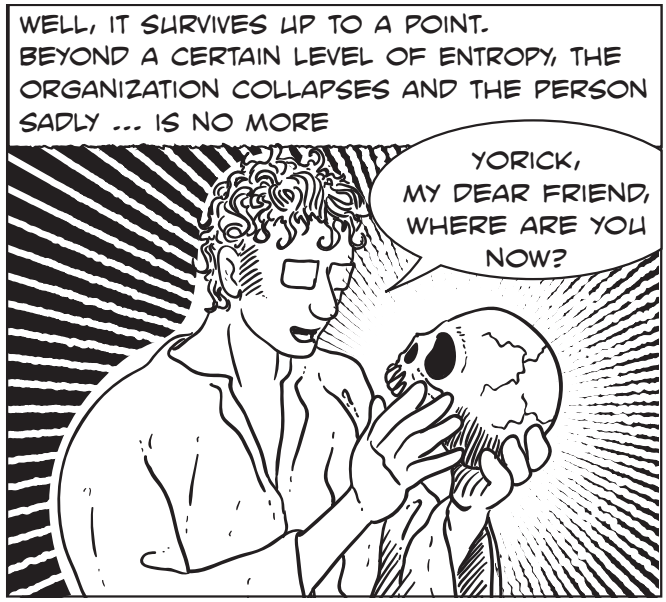

SINCE THE ANCIENT WORLD MANY SCHOLARS PARTITIONED REALITIES IN SINGLE ENTITIES, SOMETIMES LOCATED IN AN EXTRA-WORLD
PLATO, ANY IDEA WHERE ENTIA ARE?
LOOK UP!
PLATO, ARISTOTLE, AND MANY OTHERS CONCEIVED REALITY AS OBJECTS EVENTUALLY GLORIFIED AS ENTIA OR FORMS

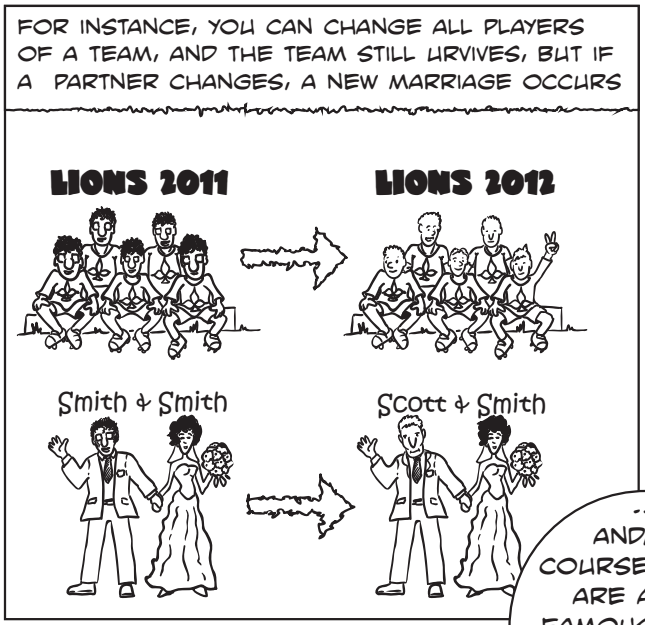

DUCCIO DA BONINSEGNA, LAST SUPPER, 1308
IT WAS THE SAME IN ALL WESTERN CULTURE. FOR INSTANCE, IN MOST MEDIEVAL FRESCOS NO OBJECT IS OCCLUDED. EACH OBJECT APPEARS AS A WHOLE. AS A RESULT, LOOK AT HOW INNATURAL IS THE POSITION OF OBJECTS ON THE TABLE.

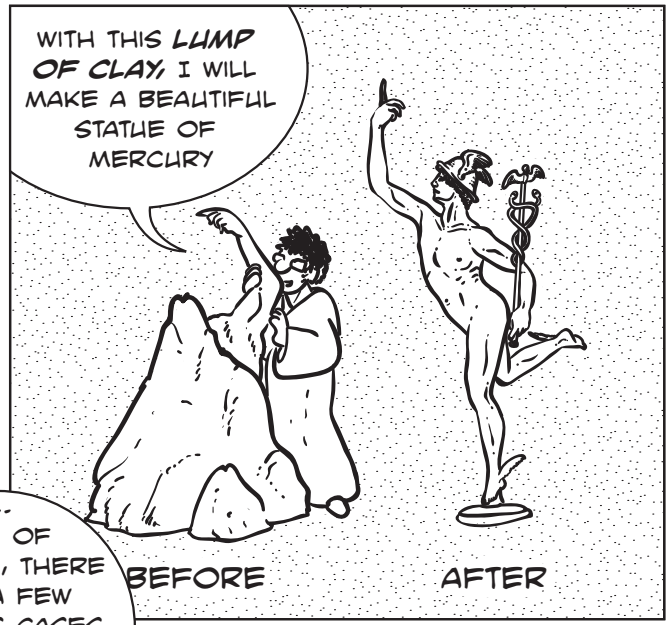

SO, AN OBJECT IS SOME KIND OF WHOLE. BUT WHAT IS A WHOLE? WHAT GIVES IT UNITY? AND IS A WHOLE SOMETHING REAL? IS THERE ANYTHING ABOVE AND BEYOND THE MOST ELEMENTARY COMPONENTS OF REALITY? IS ALL JUST TINY PARTICLES?

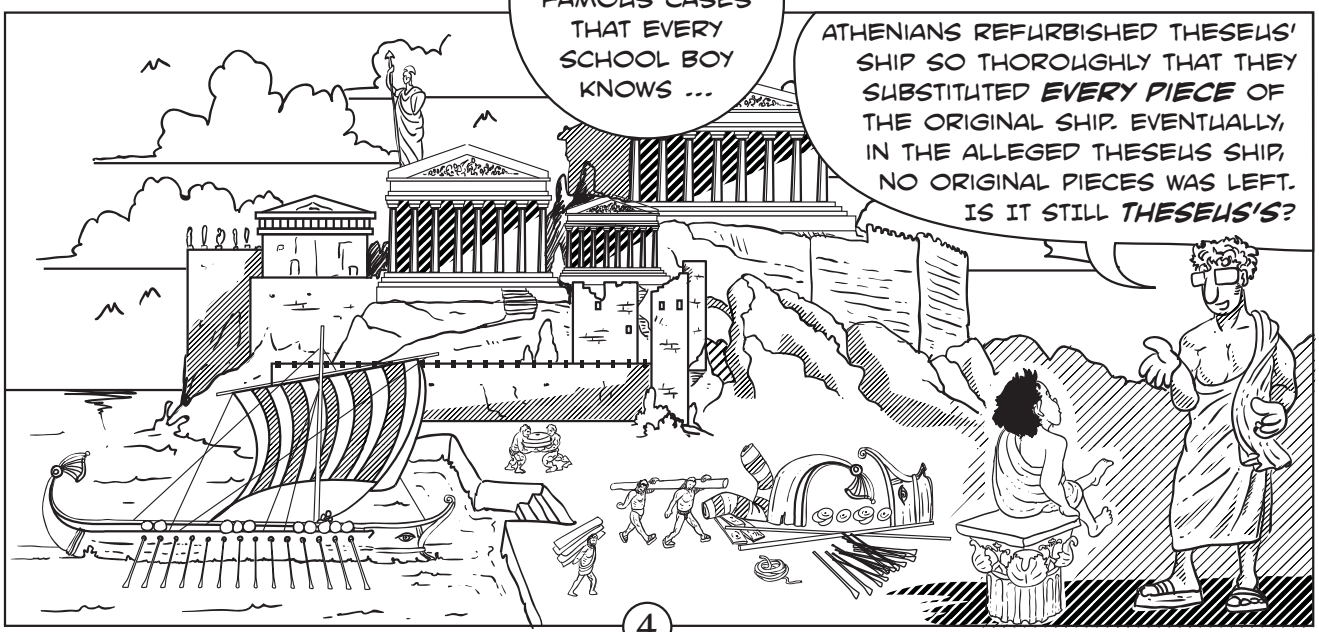

THERE ARE MANY DIFFERENT KINDS OF WHOLES. SOME ARE EASY TO DEFINE LIKE CHAIRS, CUPS, PHONES AND THE LIKE ...
BUT CONSIDER A TOWN, OR A FLOCK OF BIRDS, OR AN ANTS' NEST, OR A HIVE OF BEES. DO THEY COMPOSE A REAL UNITY? WHY NOT?
FLORIDA
CUBA
WHAT ABOUT THINGS LIKE HURRICANES, GALAXIES, TORNADOS AND WAVES? THEY ARE MADE OF PARTS HAVE NO FIXED RELATION
AND A SCORE OF MUSIC? OR A BOOK? OR A COMPUTER CHARACTER MADE UP OF PIXELS IN A CPU. ARE THEY A WHOLE OR NOT?
DAD
AND WHAT ABOUT US? AREN'T WE SOME KIND OF WHOLE, TOO? WE ARE MADE OF GAZILLIONES OF CELLS AND MOLECULES. ARE WE SOMETHING MORE?
IN THE SAME WAY, WE LIVE THROUGH TIME. WE ARE MADE OF TEMPORAL PARTS TOO. WHAT IS THAT GIVES UNITY TO DIFFERENT INSTANTIATIONS OF A PERSON THROUGH HIS OR HER LIFE?
TIME
3

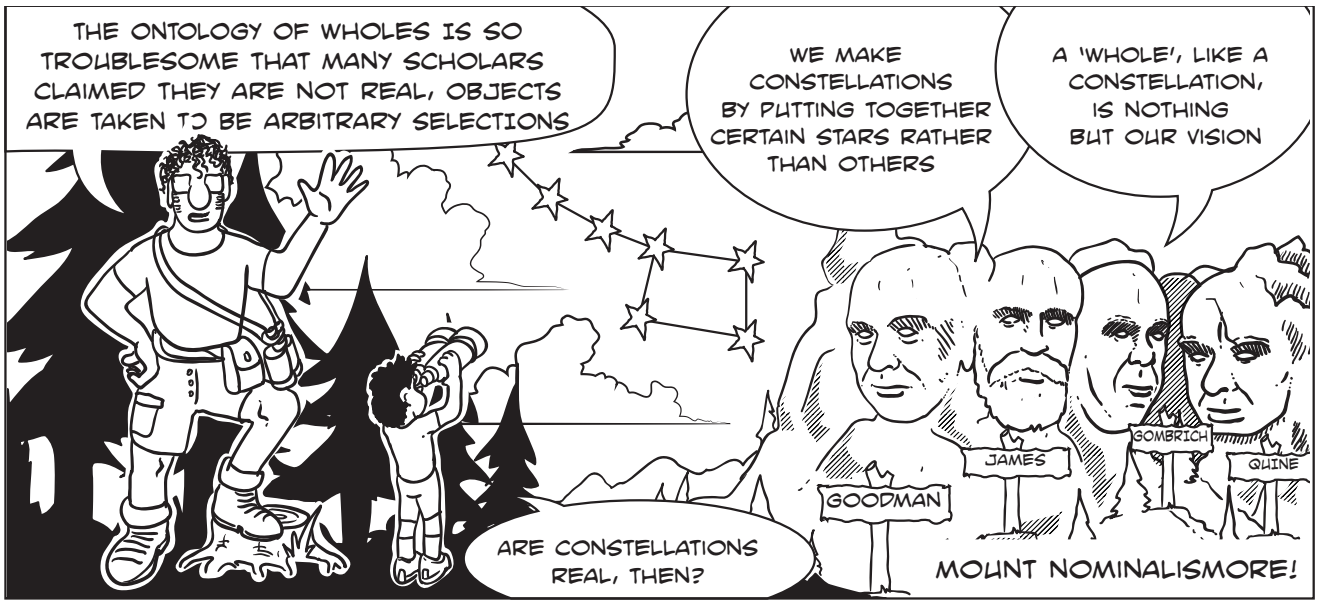

O_2
CO_2
H_2O, food
Waste
ENVIRONMENT
THE HUMAN BODY SUBSTITUTES ALMOST ALL ITS MOLECULES IN A DECADE OR SO --- YET THE WHOLE WHICH IS THE PERSON SURVIVES

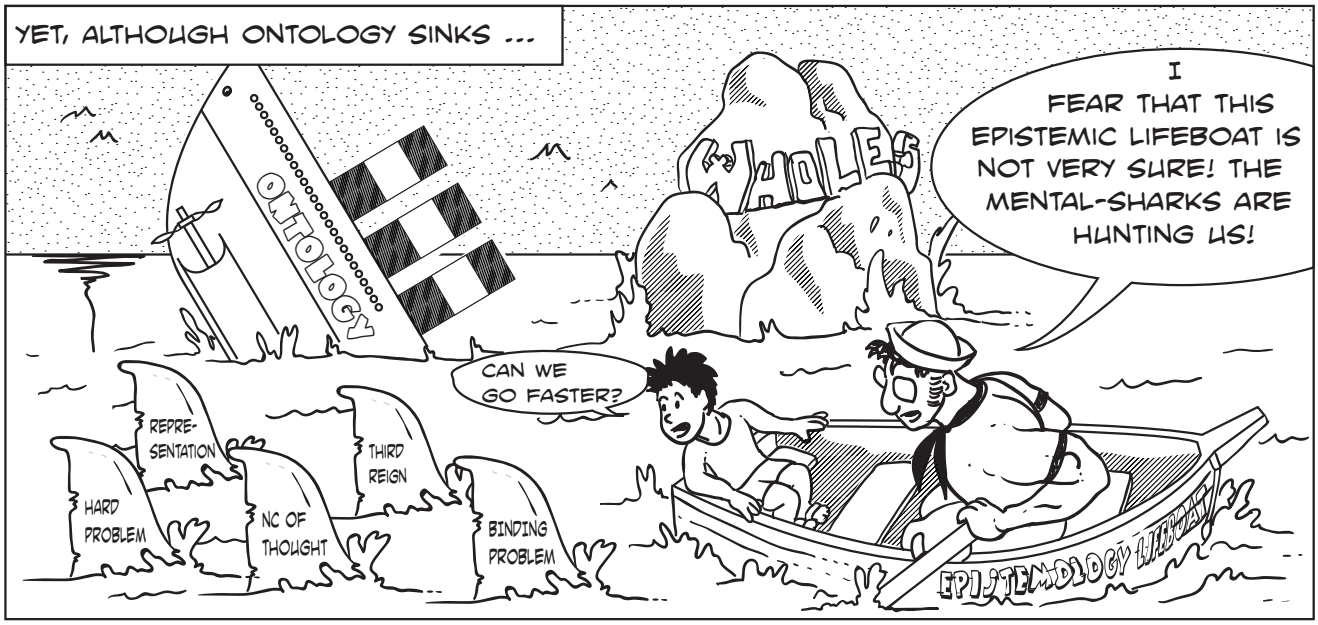

WELL, IT SURVIVES UP TO A POINT. BEYOND A CERTAIN LEVEL OF ENTROPY, THE ORGANIZATION COLLAPSES AND THE PERSON SADLY ... IS NO MORE
YORICK, MY DEAR FRIEND, WHERE ARE YOU NOW?

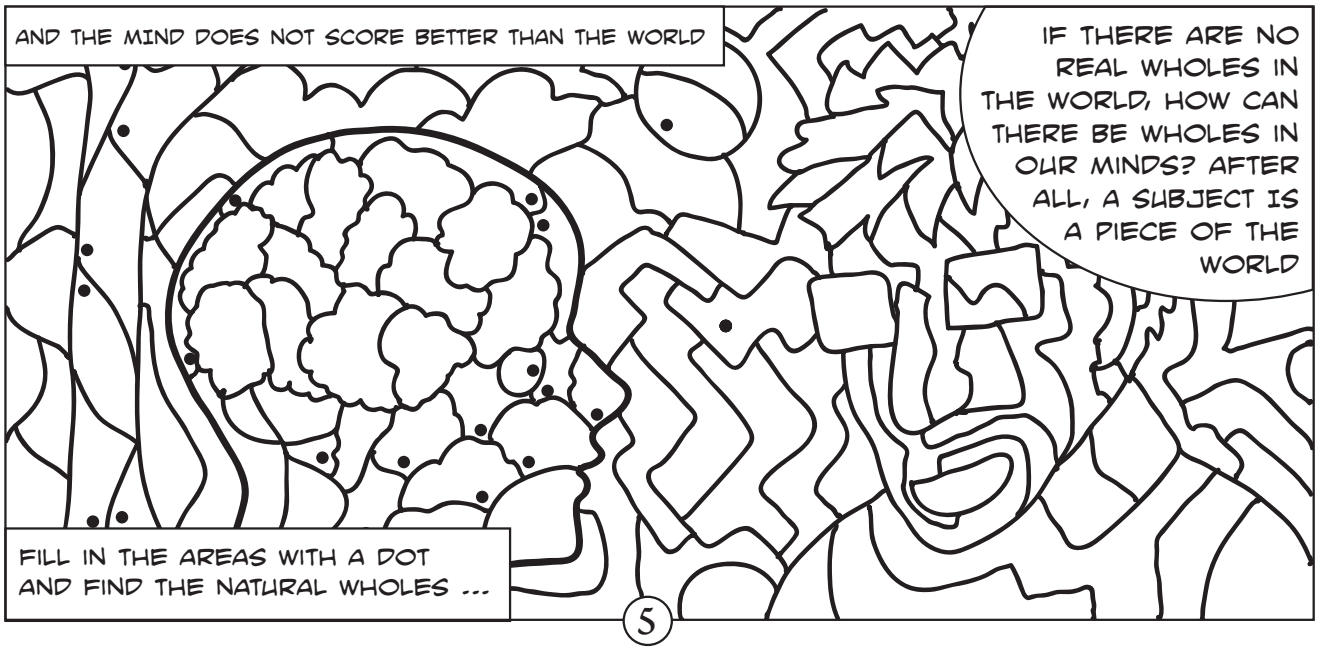

FOR INSTANCE, YOU CAN CHANGE ALL PLAYERS OF A TEAM, AND THE TEAM STILL URVIVES, BUT IF A PARTNER CHANGES, A NEW MARRIAGE OCCURS
LIONS 2011
LIONS 2012
Smith & Smith
Scott & Smith

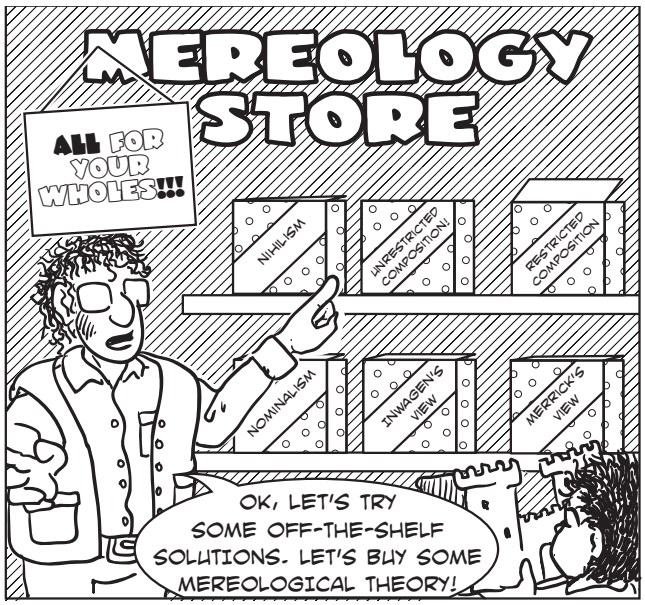

WITH THIS LUMP OF CLAY, I WILL MAKE A BEAUTIFUL STATUE OF MERCURY
BEFORE
AFTER

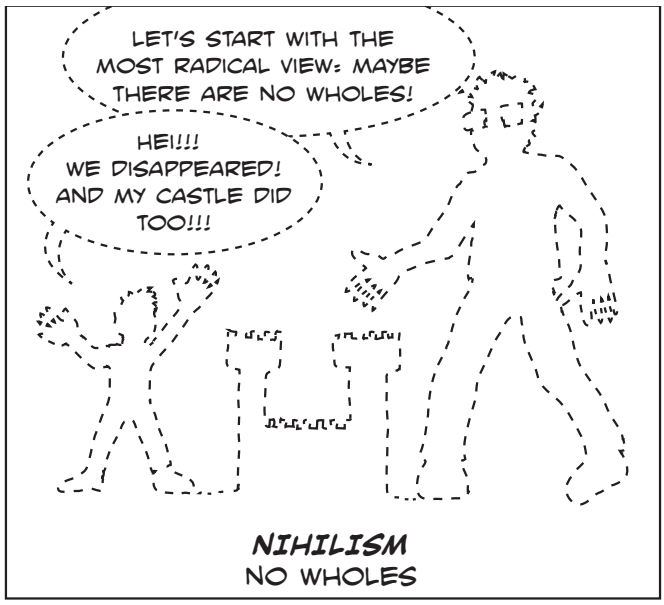

... AND, OF COURSE, THERE ARE A FEW FAMOUS CASES THAT EVERY SCHOOL BOY KNOWS ...
ATHENIANS REFURBISHED THESEUS' SHIP SO THOROUGHLY THAT THEY SUBSTITUTED EVERY PIECE OF THE ORIGINAL SHIP. EVENTUALLY, IN THE ALLEGED THESEUS SHIP, NO ORIGINAL PIECES WAS LEFT. IS IT STILL THESEUS'S?


THE ONTOLOGY OF WHOLES IS SO TROUBLESOME THAT MANY SCHOLARS CLAIMED THEY ARE NOT REAL, OBJECTS ARE TAKEN TO BE ARBITRARY SELECTIONS
WE MAKE CONSTELLATIONS BY PUTTING TOGETHER CERTAIN STARS RATHER THAN OTHERS
A 'WHOLE', LIKE A CONSTELLATION, IS NOTHING BUT OUR VISION
ARE CONSTELLATIONS REAL, THEN?
GOODMAN
JAMES
GOMBRICH
QUINE
MOUNT NOMINALISMORE!

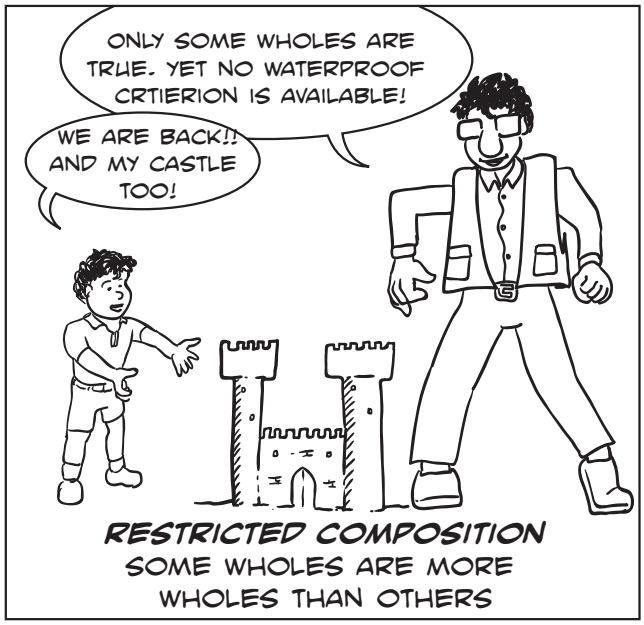

YET, ALTHOUGH ONTOLOGY SINKS ...
ONTOLOGY
WHOLES
I FEAR THAT THIS EPISTEMIC LIFEBOAT IS NOT VERY SURE! THE MENTAL-SHARKS ARE HUNTING US!
CAN WE GO FASTER?
REPRE-SENTATION
HARD PROBLEM
NC OF THOUGHT
THIRD REIGN
BINDING PROBLEM
EPISTEMOLOGY LIFEBOAT


AND THE MIND DOES NOT SCORE BETTER THAN THE WORLD
IF THERE ARE NO REAL WHOLES IN THE WORLD, HOW CAN THERE BE WHOLES IN OUR MINDS? AFTER ALL, A SUBJECT IS A PIECE OF THE WORLD
FILL IN THE AREAS WITH A DOT AND FIND THE NATURAL WHOLES ...
5

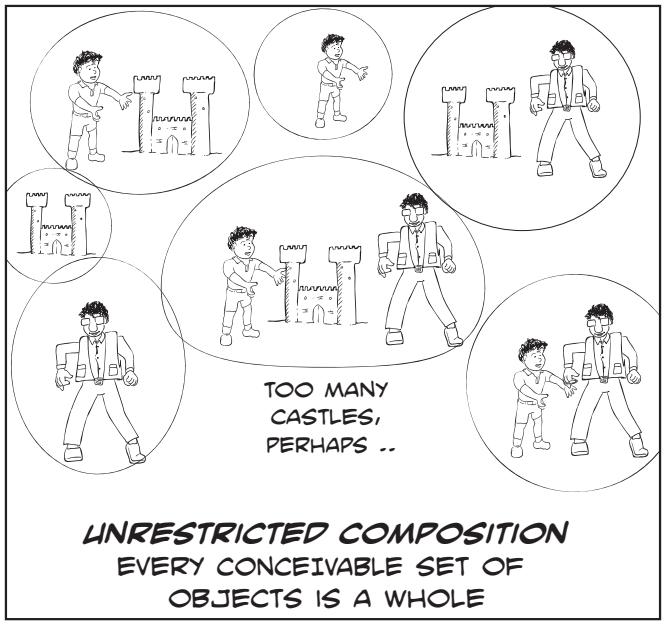




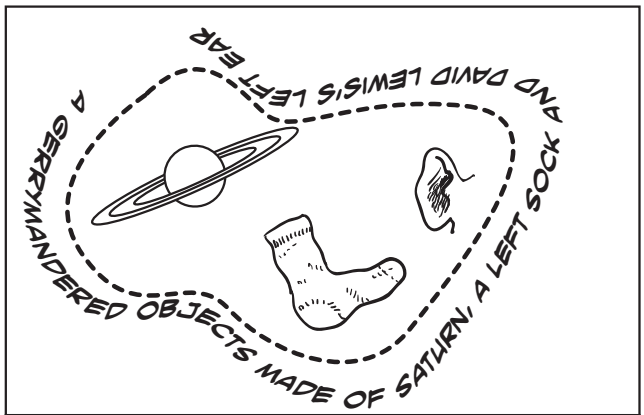
NIHILISM
NO WHOLES



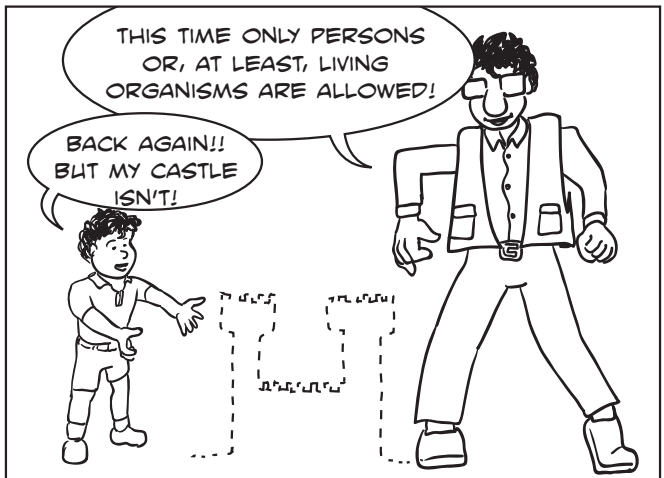
RESTRICTED COMPOSITION
SOME WHOLES ARE MORE WHOLES THAN OTHERS



UNRESTRICTED COMPOSITION
EVERY CONCEIVABLE SET OF OBJECTS IS A WHOLE

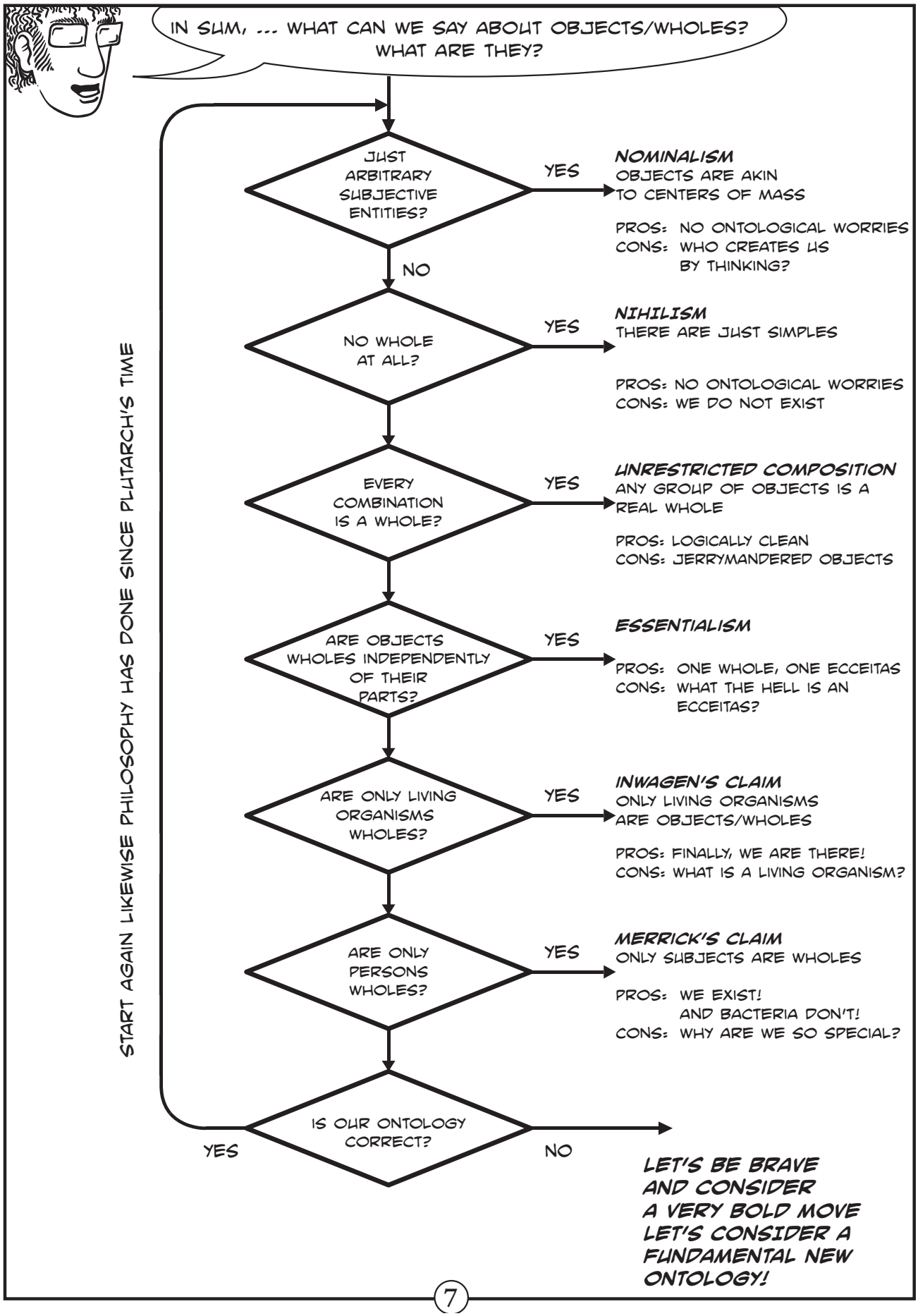


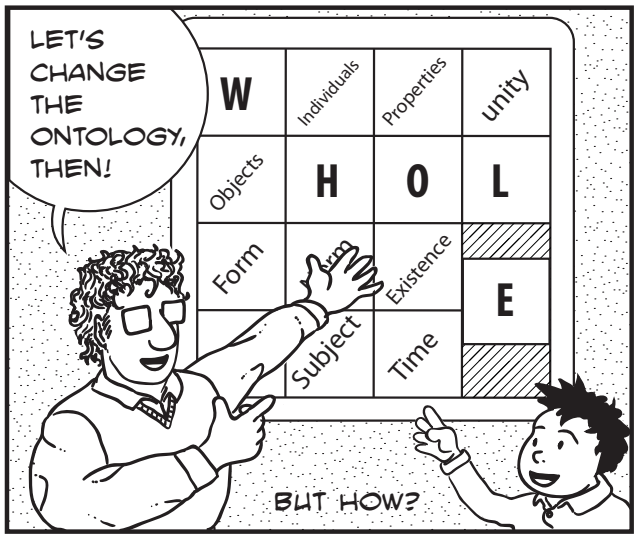
UNRESTRICTED COMPOSITION IS FINE BECAUSE THERE IS SPACE BOTH FOR US AND FOR THE CASTLE. UNFORTUNATELY IT GETS TOO CROWDED. INFACT, IT IS PLAGUED BY GERRY-MANDERED OBJECTS MADE OF THE MOST OUTRAGEOUS COMBINATIONS OF PARTS

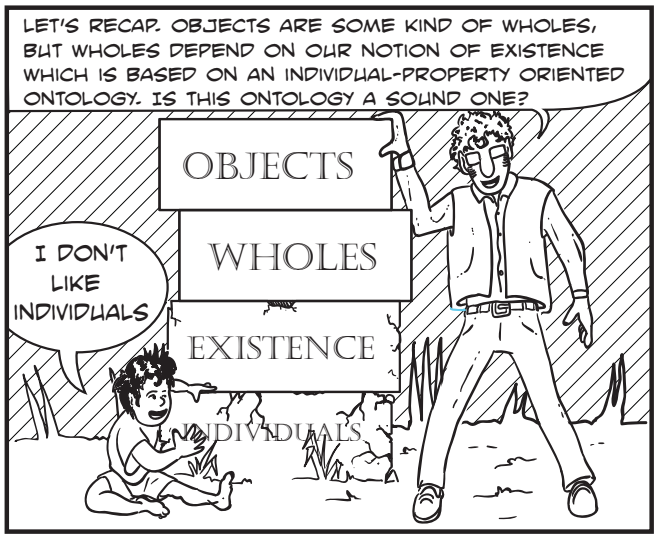


INWAGEN'S AND MERRICK'S:
ONLY PERSONS OR LIVING ORGANISMS ARE TRUE WHOLES

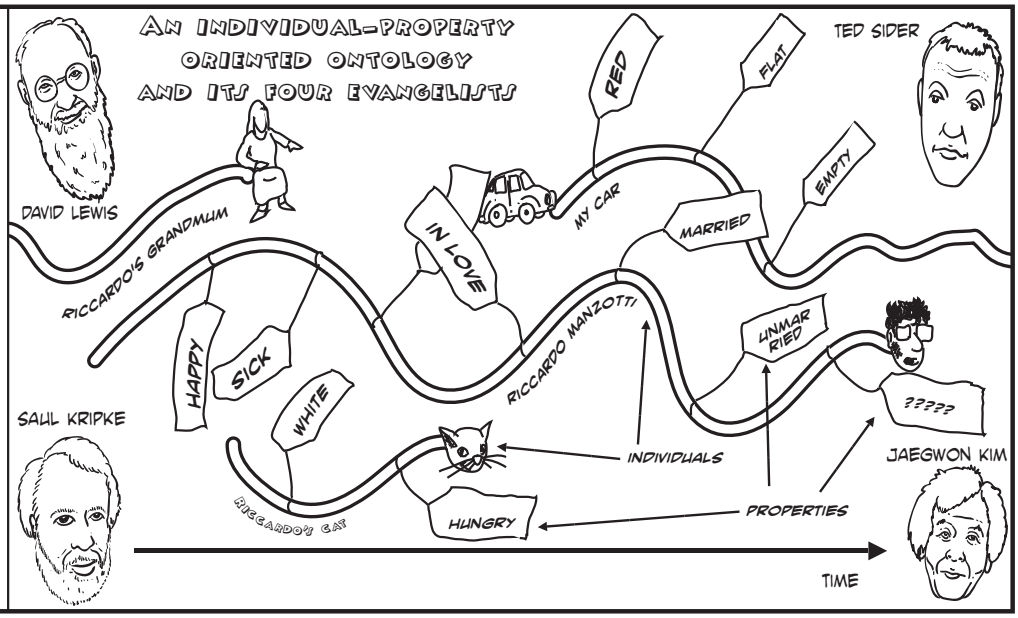
6

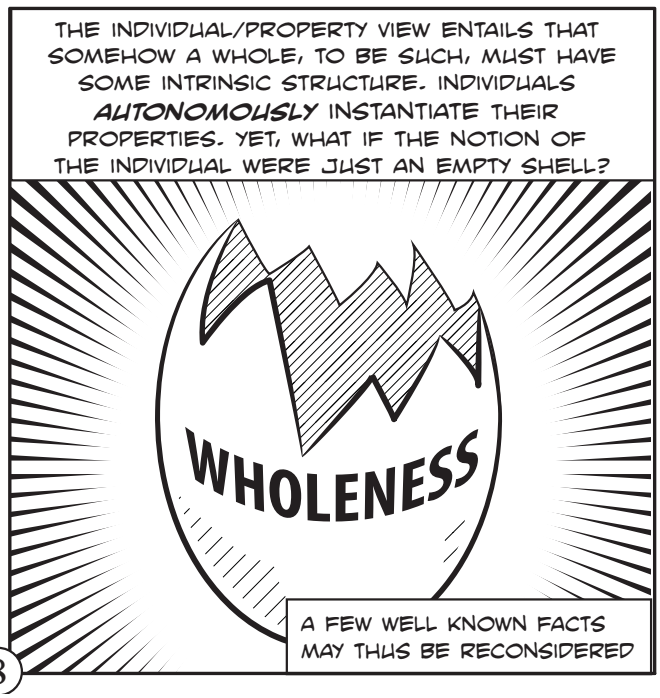

IN SUM, ... WHAT CAN WE SAY ABOUT OBJECTS/WHOLES? WHAT ARE THEY?
JUST ARBITRARY SUBJECTIVE ENTITIES?
YES
NOMINALISM
OBJECTS ARE AKIN TO CENTERS OF MASS
PROS: NO ONTOLOGICAL WORRIES
CONS: WHO CREATES US BY THINKING?
NO
NO WHOLE AT ALL?
YES
NIHILISM
THERE ARE JUST SIMPLES
PROS: NO ONTOLOGICAL WORRIES
CONS: WE DO NOT EXIST
EVERY COMBINATION IS A WHOLE?
YES
UNRESTRICTED COMPOSITION
ANY GROUP OF OBJECTS IS A REAL WHOLE
PROS: LOGICALLY CLEAN
CONS: JERRYMANDERED OBJECTS
ARE OBJECTS WHOLES INDEPENDENTLY OF THEIR PARTS?
YES
ESSENTIALISM
PROS: ONE WHOLE, ONE ECCEITAS
CONS: WHAT THE HELL IS AN ECCEITAS?
ARE ONLY LIVING ORGANISMS WHOLES?
YES
INWAGEN'S CLAIM
ONLY LIVING ORGANISMS ARE OBJECTS/WHOLES
PROS: FINALLY, WE ARE THERE!
CONS: WHAT IS A LIVING ORGANISM?
ARE ONLY PERSONS WHOLES?
YES
MERRICK'S CLAIM
ONLY SUBJECTS ARE WHOLES
PROS: WE EXIST! AND BACTERIA DON'T!
CONS: WHY ARE WE SO SPECIAL?
IS OUR ONTOLOGY CORRECT?
YES
NO
START AGAIN LIKEWISE PHILOSOPHY HAS DONE SINCE PLUTARCH'S TIME
LET'S BE BRAVE AND CONSIDER A VERY BOLD MOVE LET'S CONSIDER A FUNDAMENTAL NEW ONTOLOGY!
7





THE TRADITION IS A RECEIVED ONTOLOGY BASED ON THE NOTION OF ***INDIVIDUALS*** AND THEIR ***PROPERTIES***. ROUGHLY IT ASSUMES THAT THE WORLD IS MADE OF INDIVIDUALS PERSISTING THROUGH TIME AND INSTANTIANTING PROPERTIES ALONG THE WAY. IT IS A VERY APPEALING VIEW, FROM A LOGICAL PERSPECTIVE. YET, IS IT TRUE?



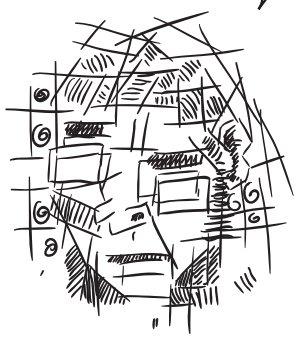


SINCE PREHISTORY, ARTISTS HAVE EXPLOITED THE CAPABILITY TO SINGLE OUT FACE-LIKE WHOLES FROM THE PHYSICAL CONTINUUM, THANKS TO THE EXISTENCE OF A SPECIALIZED AREA IN THE BRAIN. YET WOULD FACES EXIST IF SUCH AN AREA WERE WIPED FROM EVERYONE'S BRAIN? DO FACES EXIST INDEPENDENTLY OF BRAIN?

BTW, I AM NO MORE REAL THAN THE OTHERS!!!

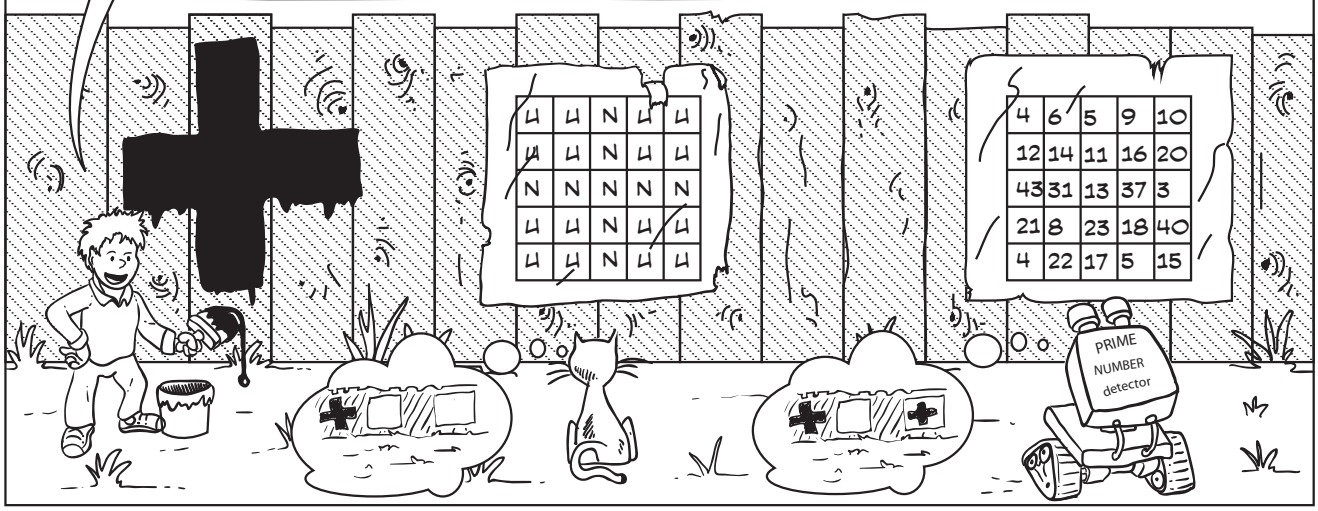
ARCIMBOLDO, 1527-93

SALVADOR DALI', 1904-89

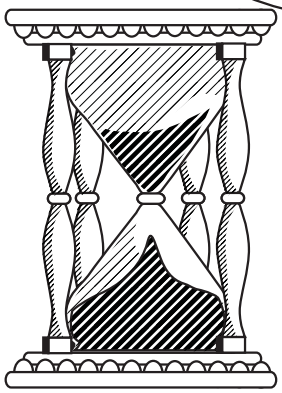


GEORGES BRAQUE, 1882-1963

NOW, CONSIDER THAT MANY OBJECTS DO NOT EXIST FOR DIFFERENT SUBJECTS. LOOK AT THE THREE CROSSES ON THE WALL. THEY EXIST ONLY FOR SOME OF THE SUBJECTS



A HINT MAY BE NOT TO CONSIDER WHOLES AS INDIVIDUALS, BUT AS OCCURRENCES IN TIME. THUS THE TWO CRUCIAL NOTIONS TO (RE)-CONSIDER ARE ***TIME*** AND ***EXISTENCE***.

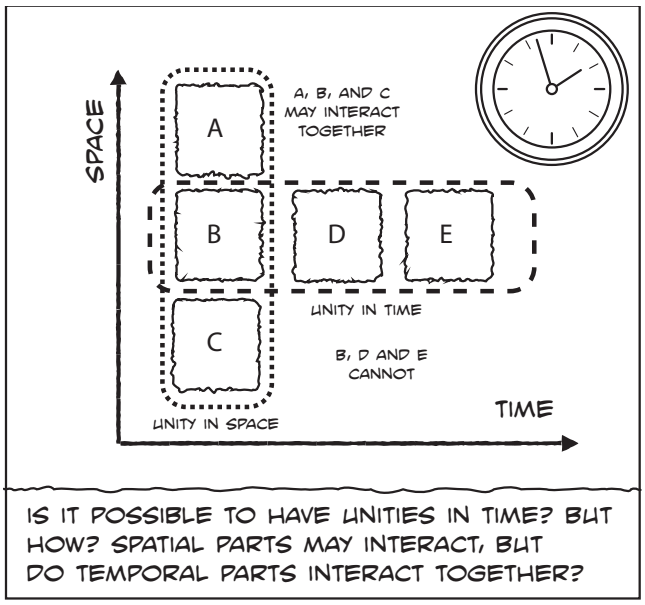


TIME

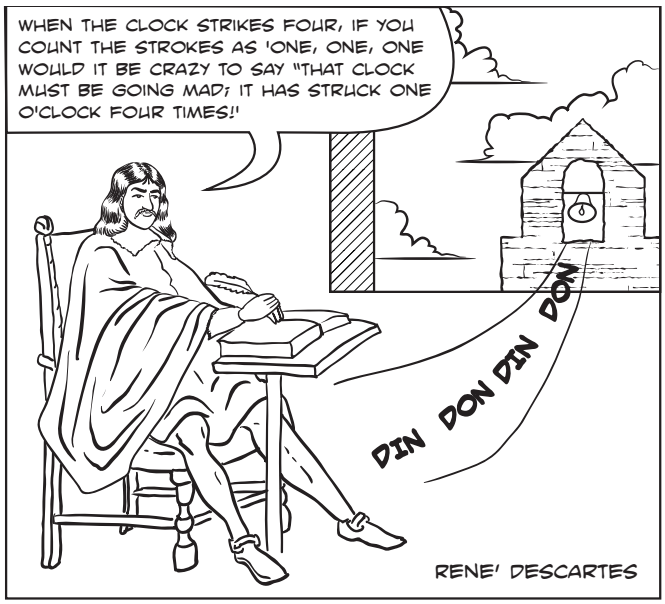
E

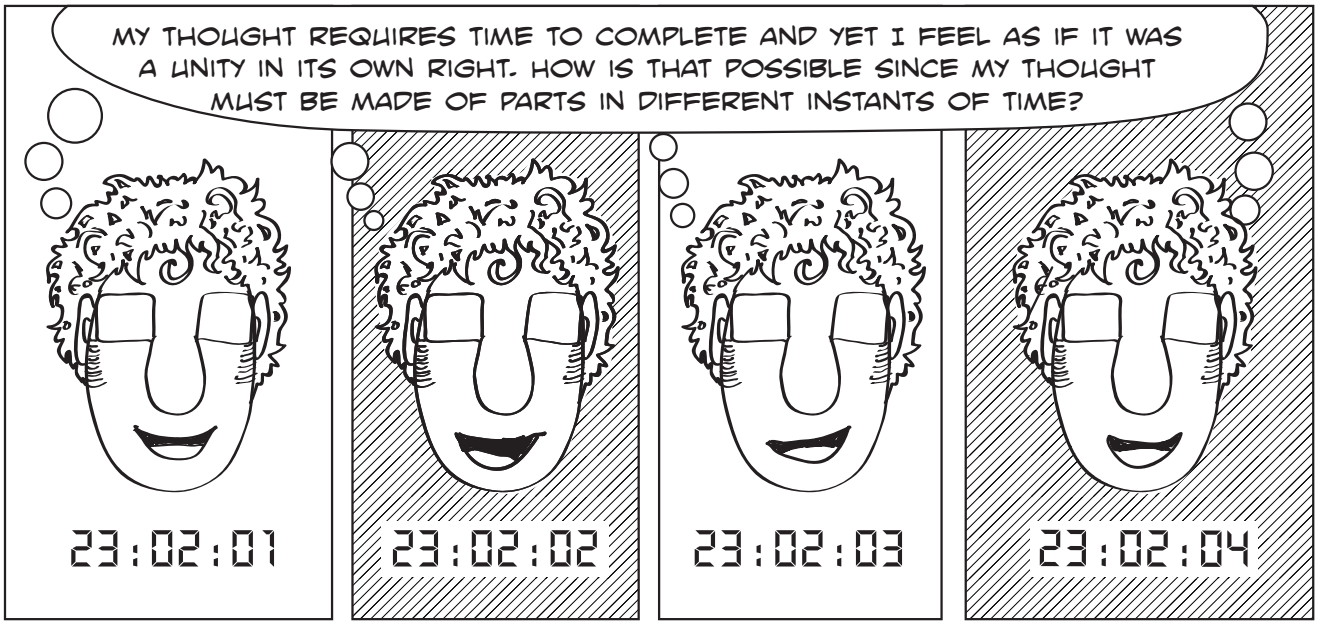
EXISTENCE

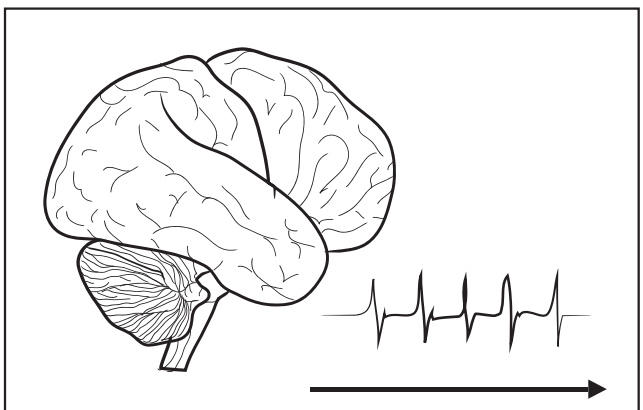
9



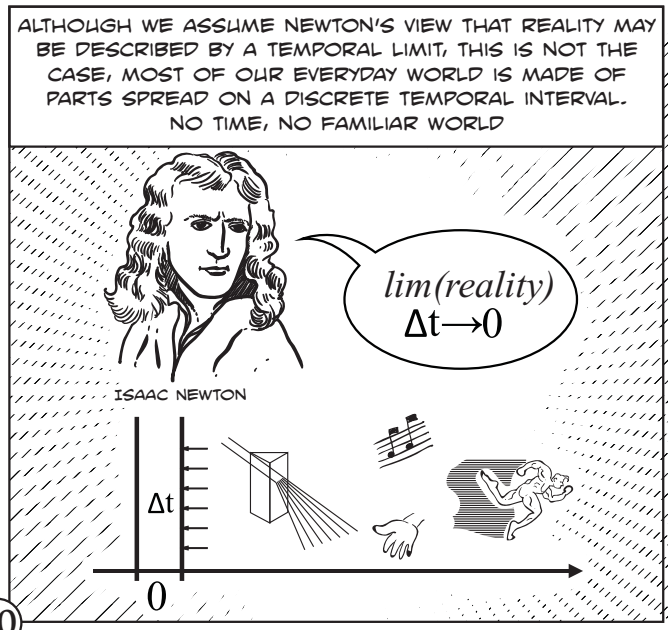
IS IT POSSIBLE TO HAVE UNITIES IN TIME? BUT HOW? SPATIAL PARTS MAY INTERACT, BUT DO TEMPORAL PARTS INTERACT TOGETHER?



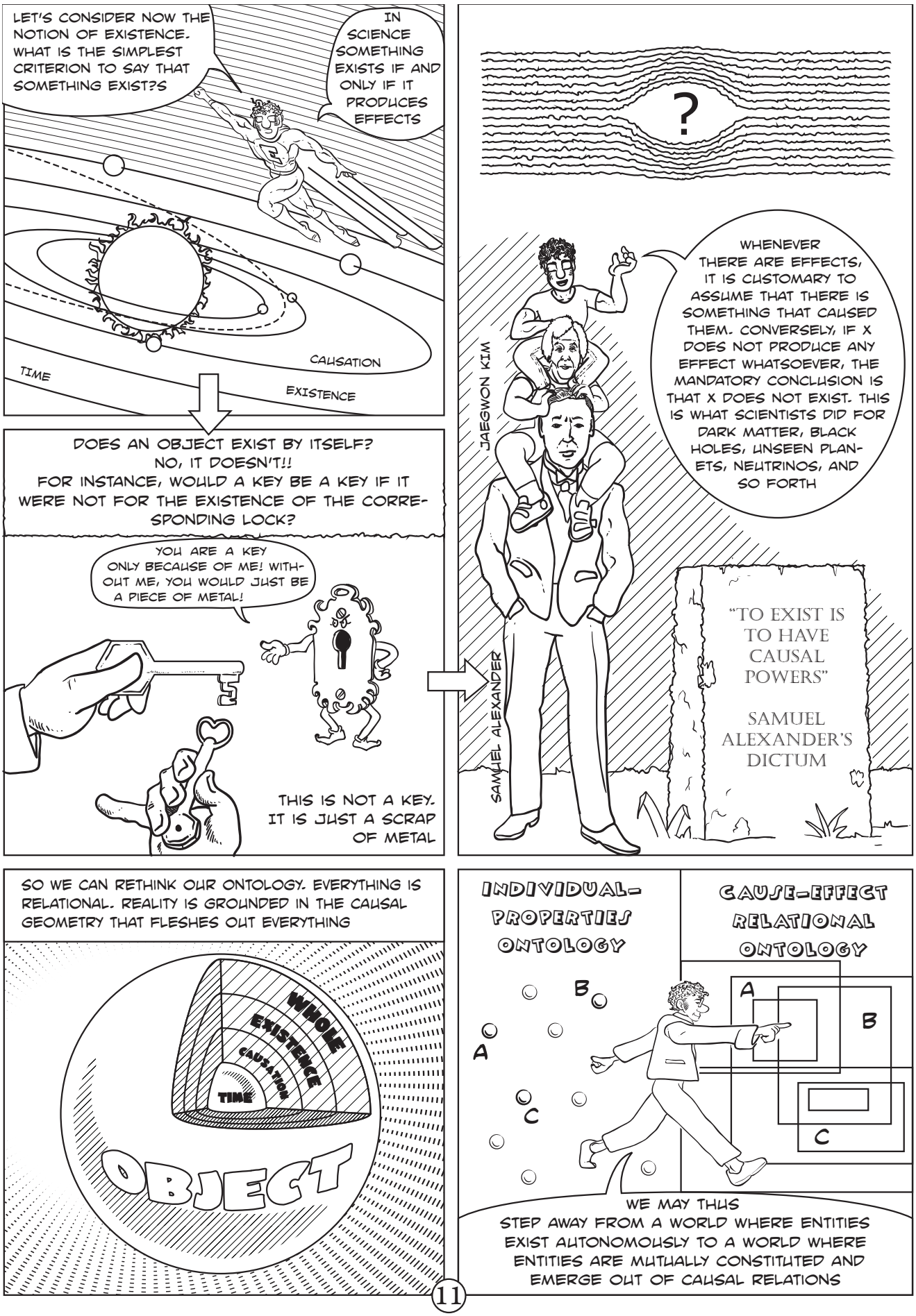


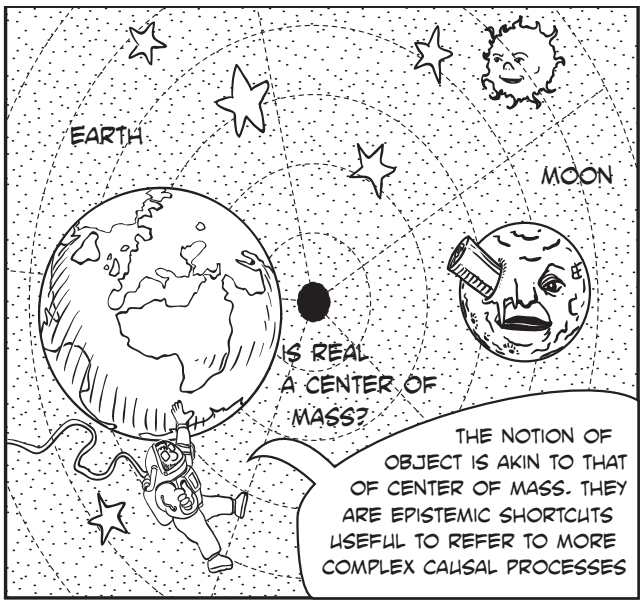


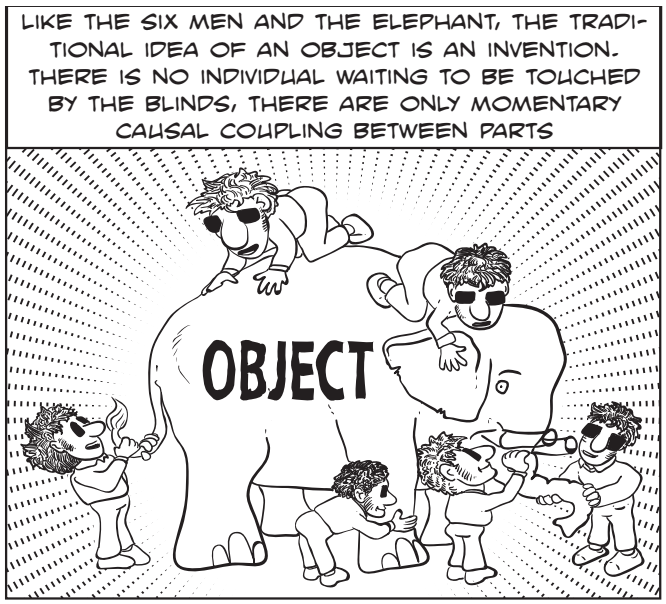
THOUGHTS ARE NOT THE ONLY THINGS SPREAD IN TIME, BUT THEIR PHYSICAL AND NEURAL UNDERPINNINGS ARE SPREAD TOO. HOW COULD THEY BECOME A UNITY? HOW MAY A THOUGHT BE A SERIES OF SPIKES?

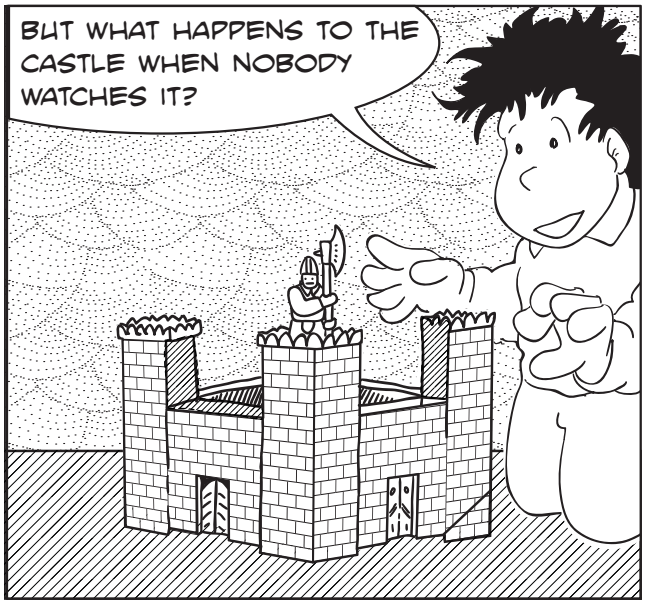


10

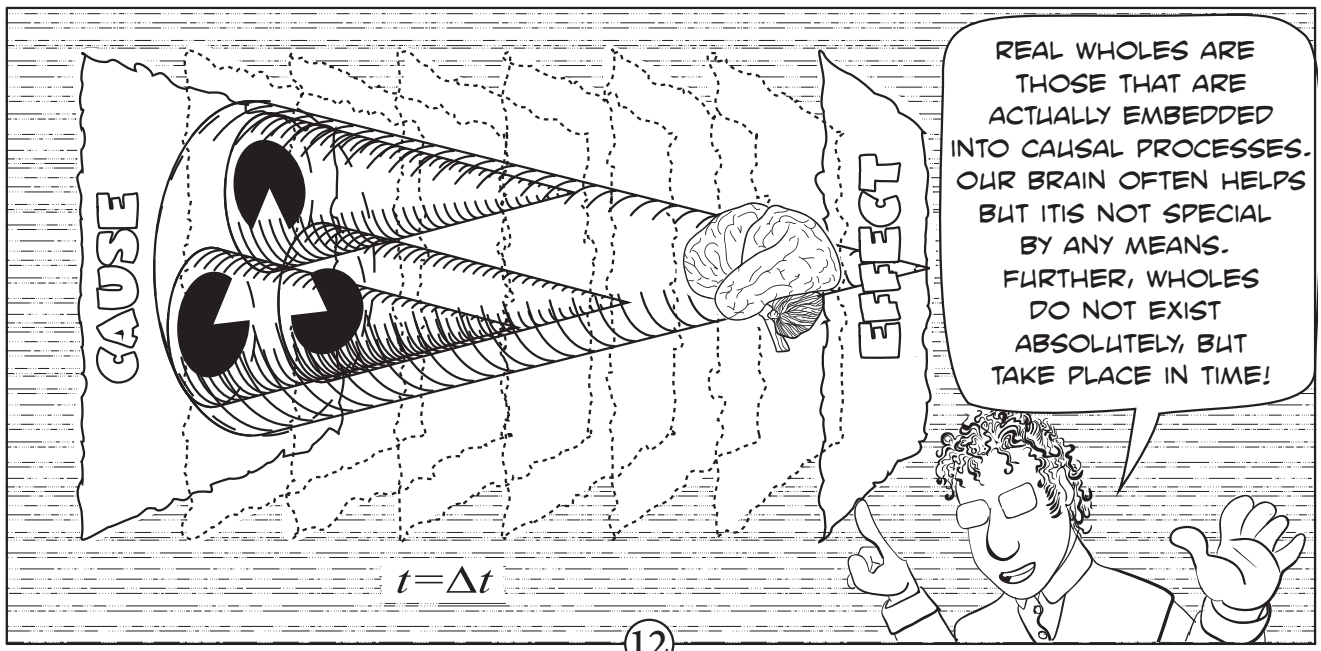

LET'S CONSIDER NOW THE NOTION OF EXISTENCE. WHAT IS THE SIMPLEST CRITERION TO SAY THAT SOMETHING EXIST?S
IN SCIENCE SOMETHING EXISTS IF AND ONLY IF IT PRODUCES EFFECTS
TIME
CAUSATION
EXISTENCE
DOES AN OBJECT EXIST BY ITSELF? NO, IT DOESN'T!! FOR INSTANCE, WOULD A KEY BE A KEY IF IT WERE NOT FOR THE EXISTENCE OF THE CORRESPONDING LOCK?
YOU ARE A KEY ONLY BECAUSE OF ME! WITHOUT ME, YOU WOULD JUST BE A PIECE OF METAL!
THIS IS NOT A KEY. IT IS JUST A SCRAP OF METAL
?
JAEGWON KIM
WHENEVER THERE ARE EFFECTS, IT IS CUSTOMARY TO ASSUME THAT THERE IS SOMETHING THAT CAUSED THEM. CONVERSELY, IF X DOES NOT PRODUCE ANY EFFECT WHATSOEVER, THE MANDATORY CONCLUSION IS THAT X DOES NOT EXIST. THIS IS WHAT SCIENTISTS DID FOR DARK MATTER, BLACK HOLES, UNSEEN PLANETS, NEUTRINOS, AND SO FORTH
SAMUEL ALEXANDER
"TO EXIST IS TO HAVE CAUSAL POWERS"
SAMUEL ALEXANDER'S DICTUM
SO WE CAN RETHINK OUR ONTOLOGY. EVERYTHING IS RELATIONAL. REALITY IS GROUNDED IN THE CAUSAL GEOMETRY THAT FLESHES OUT EVERYTHING
WHOLE
EXISTENCE
CAUSATION
TIME
OBJECT
INDIVIDUAL-PROPERTIES ONTOLOGY
CAUSE-EFFECT RELATIONAL ONTOLOGY
A
B
C
A
B
C
WE MAY THUS STEP AWAY FROM A WORLD WHERE ENTITIES EXIST AUTONOMOUSLY TO A WORLD WHERE ENTITIES ARE MUTUALLY CONSTITUTED AND EMERGE OUT OF CAUSAL RELATIONS
11


EARTH
MOON
IS REAL A CENTER OF MASS?
THE NOTION OF OBJECT IS AKIN TO THAT OF CENTER OF MASS. THEY ARE EPISTEMIC SHORTCUTS USEFUL TO REFER TO MORE COMPLEX CAUSAL PROCESSES


LIKE THE SIX MEN AND THE ELEPHANT, THE TRADITIONAL IDEA OF AN OBJECT IS AN INVENTION. THERE IS NO INDIVIDUAL WAITING TO BE TOUCHED BY THE BLINDS, THERE ARE ONLY MOMENTARY CAUSAL COUPLING BETWEEN PARTS
OBJECT

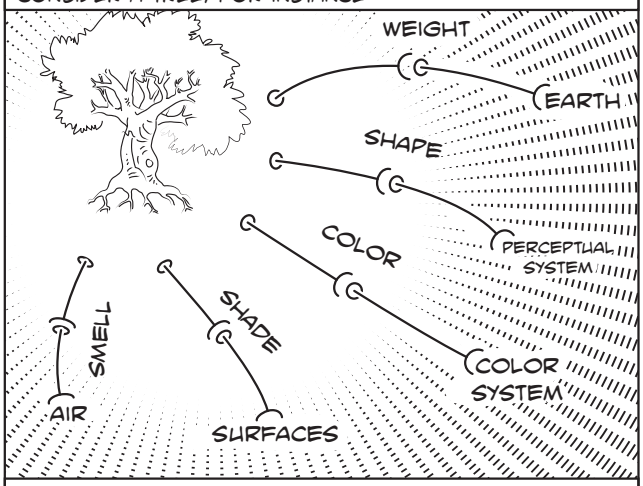

BUT WHAT HAPPENS TO THE CASTLE WHEN NOBODY WATCHES IT?


TRADITIONAL OBJECTS ARE LIKE THE LIGHT IN THE FRIDGE- WHENEVER YOU LOOK AT THEM, THEY ARE THERE
MILK

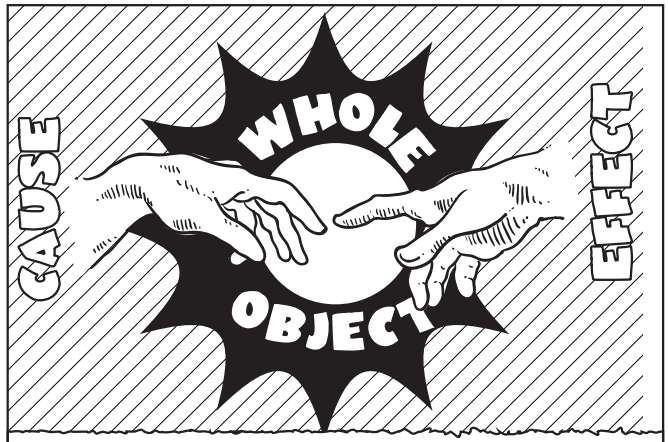

CAUSE
EFFECT
REAL WHOLES ARE THOSE THAT ARE ACTUALLY EMBEDDED INTO CAUSAL PROCESSES. OUR BRAIN OFTEN HELPS BUT ITIS NOT SPECIAL BY ANY MEANS. FURTHER, WHOLES DO NOT EXIST ABSOLUTELY, BUT TAKE PLACE IN TIME!
$t = \Delta t$

12

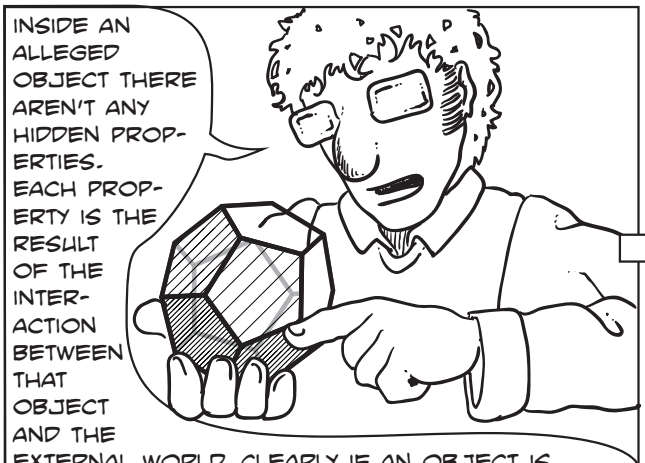
THE SAME RATIONALE HOLDS FOR OBJECT PROPERTIES CONSIDER A TREE, FOR INSTANCE



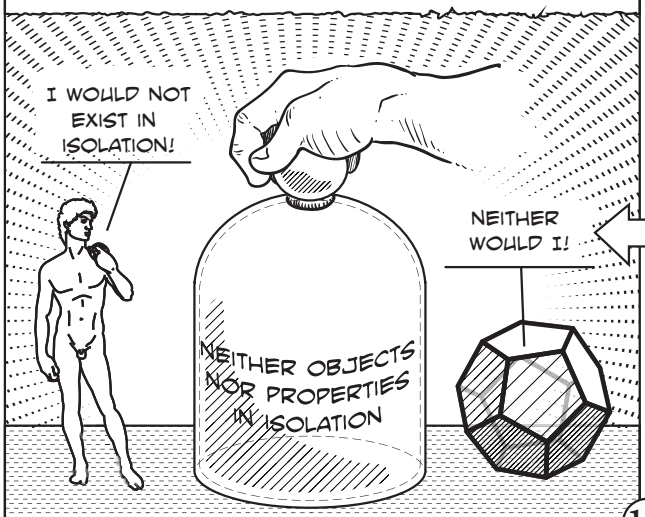
ANY PROPERTY IS EMBEDDED INTO A CORRESPONDING CAUSAL PROCESS. IF THE WORLD WERE DIFFERENT THE OBJECT WOULD BE DIFFERENT TOO



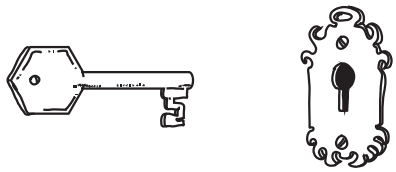
IN SHORT, A ***RELATIONAL VIEW*** OF OBJECTS AND WHOLES SUGGESTS THAT AN OBJECT DOES NOT EXIST IN VIRTUE OF ANY INTRINSIC REASON. THE OBJECT IS THE RESULT OF A CAUSAL ENTANGLEMENT BETWEEN DIFFERENT PORTIONS OF THE PHYSICAL CONTINUUM.



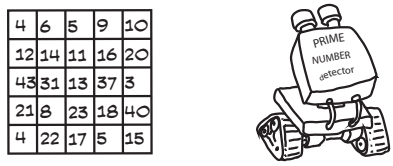
THIS MEANS THAT NEITHER OBJECTS NOR THEIR PROPERTIES EXIST IN ISOLATION. EVERYTHING WE KNOW IS THE RESULT OF A CAUSAL INTERACTION



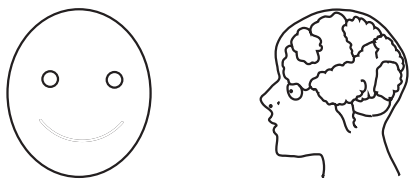
TO RECAP ...



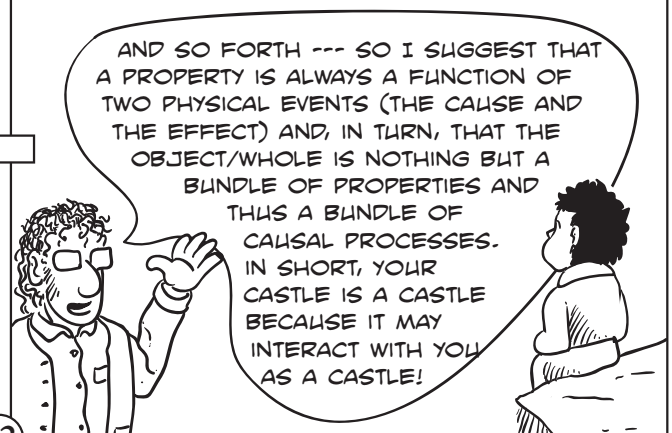
TO BE A KEY, A PIECE OF METAL NEEDS A LOCK

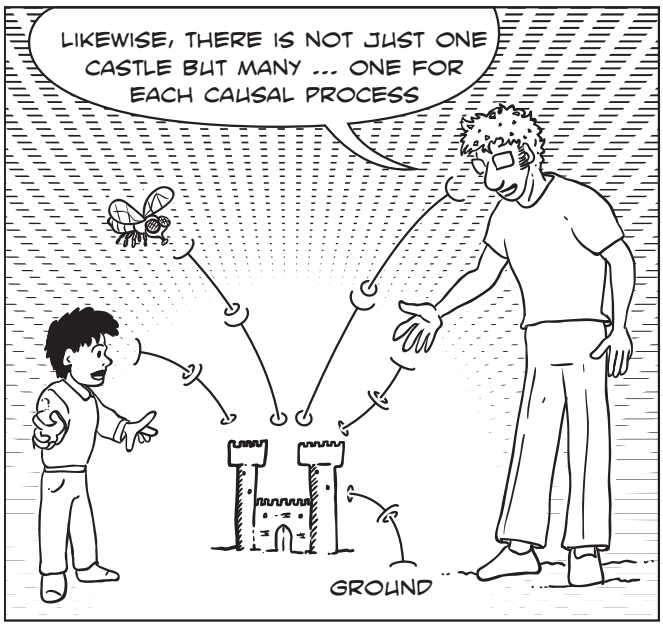


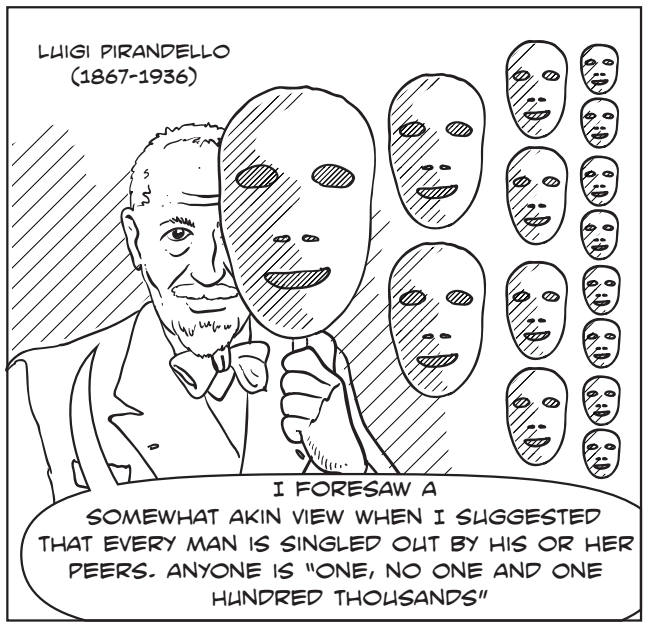
TO BE A CROSS, A SET OF SIGNS NEEDS A DETECTOR

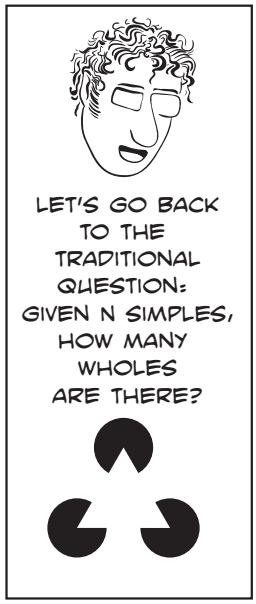


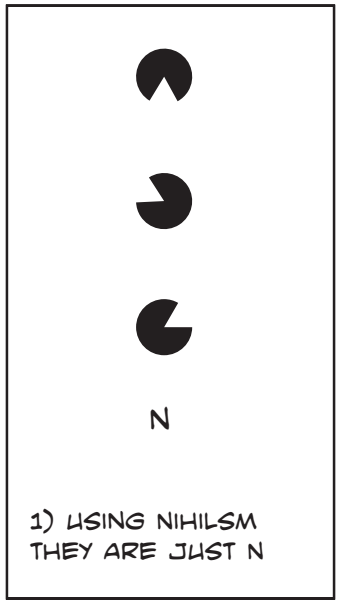
TO BE A FACE, A SET OF SIGNS NEEDS A FUSIFORM GYRUS

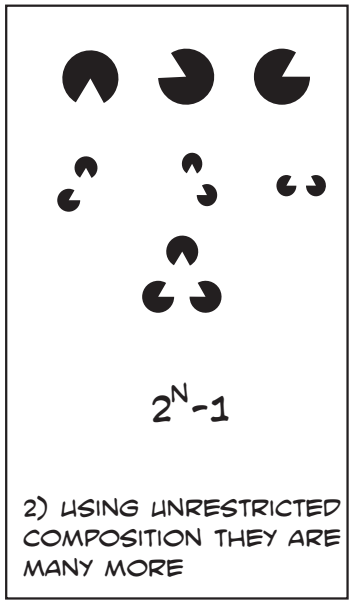


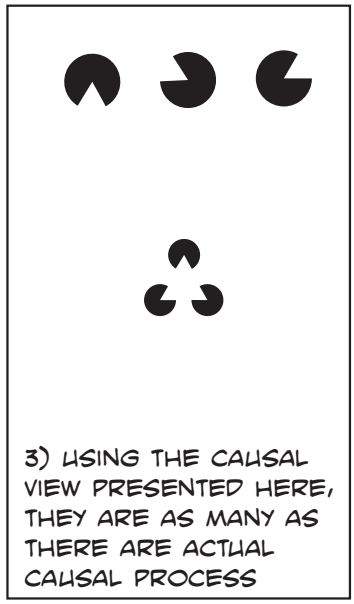

LIKEWISE, THERE IS NOT JUST ONE CASTLE BUT MANY ... ONE FOR EACH CAUSAL PROCESS
GROUND


LUIGI PIRANDELLO (1867-1936)
I FORESAW A SOMEWHAT AKIN VIEW WHEN I SUGGESTED THAT EVERY MAN IS SINGLED OUT BY HIS OR HER PEERS. ANYONE IS "ONE, NO ONE AND ONE HUNDRED THOUSANDS"

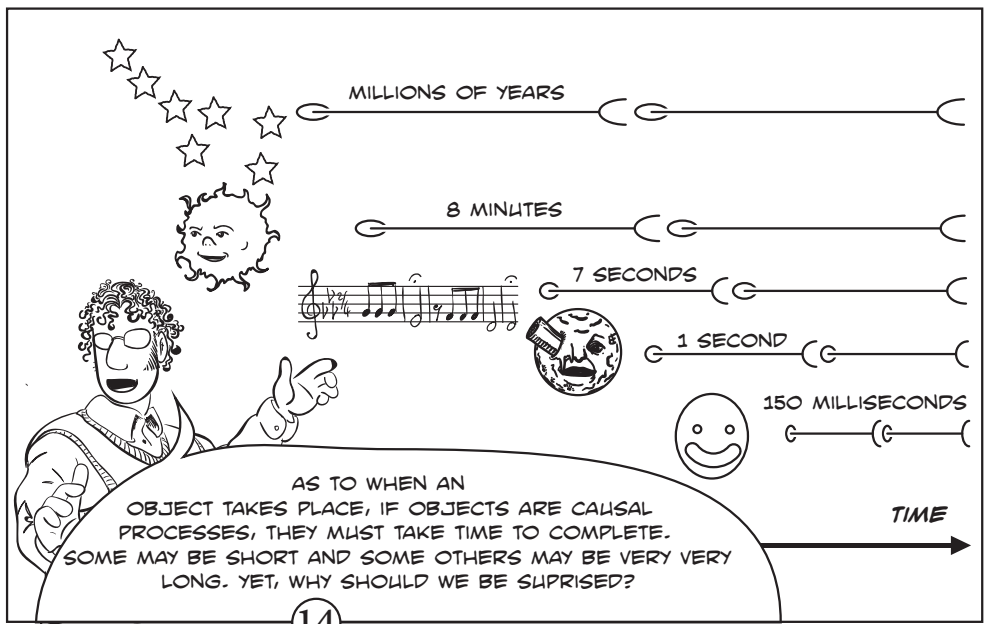

LET'S GO BACK TO THE TRADITIONAL QUESTION: GIVEN N SIMPLES, HOW MANY WHOLES ARE THERE?


N
1) USING NIHILSM THEY ARE JUST N


2^N-1
2) USING UNRESTRICTED COMPOSITION THEY ARE MANY MORE


3) USING THE CAUSAL VIEW PRESENTED HERE, THEY ARE AS MANY AS THERE ARE ACTUAL CAUSAL PROCESS


SO, WE SHOULD NOT ASK WHETHER OBJECTS EXIST BUT RATHER
WHERE
AND
WHEN
DO THEY TAKE PLACE?

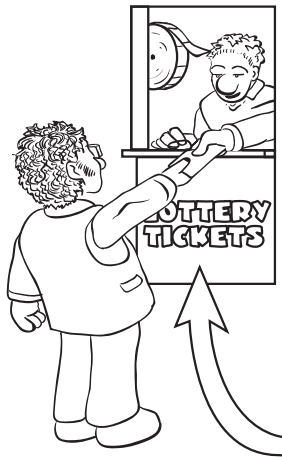

MILLIONS OF YEARS
8 MINUTES
7 SECONDS
1 SECOND
150 MILLISECONDS
TIME
AS TO WHEN AN OBJECT TAKES PLACE, IF OBJECTS ARE CAUSAL PROCESSES, THEY MUST TAKE TIME TO COMPLETE. SOME MAY BE SHORT AND SOME OTHERS MAY BE VERY VERY LONG. YET, WHY SHOULD WE BE SUPRISED?

THE RELATION BETWEEN OBJECTS AND TIME IS A VERY INTIMATE ONE.

AN OBJECT IS ***UNDEFINED*** UNTIL IT PRODUCES AN EFFECT

THUS ITS EXISTENCE DEPENDS ON THE ***PASSING OF TIME*** IN WHICH EFFECTS CAN OCCUR

CONSIDER THIS EXAMPLE. ON 30 OCTOBER YOU BUY A LOTTERY TICKET

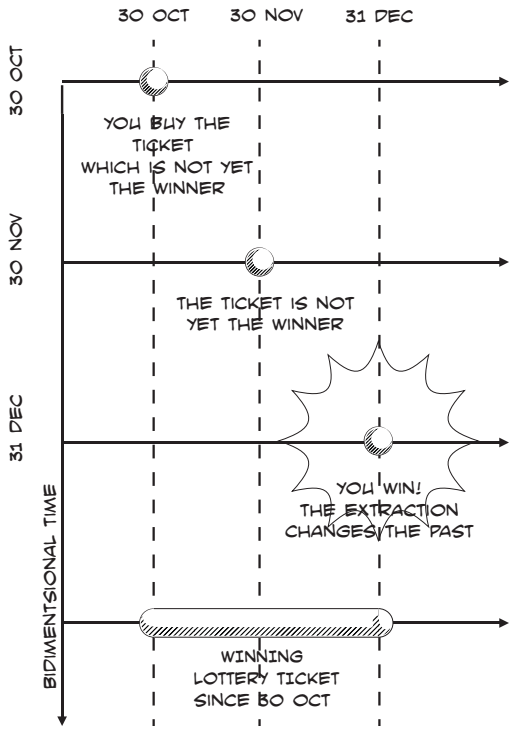


TIME GOES BY. DID YOU BUY THE WINNING ONE? NOT YET

THE NIGHT OF 31 DEC YOU WIN!!! NOW THE TICKET IS THE WINNING ONE! WAS IT BEFORE?

THE PRESENT ***CHANGES*** THE PAST

THE TICKET YOU BOUGHT WAS NOT THE WINNER UNTIL THE EXTRACTION, BUT AFTERWARDS IT BECAME THE WINNER SINCE THE TIME YOU BOUGHT IT.



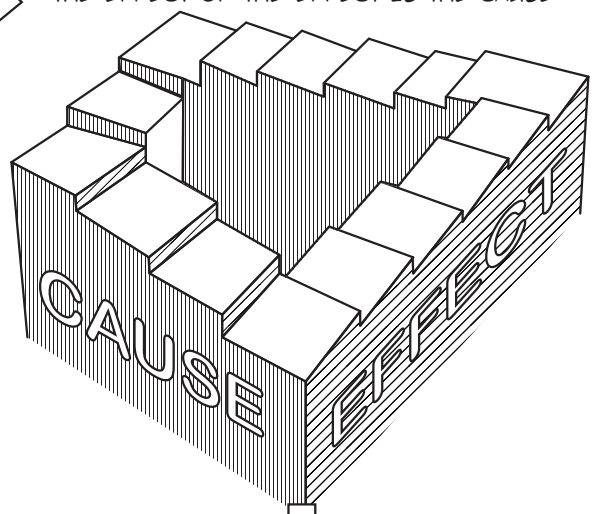
AFTER 31 DEC THE TICKET WAS THE WINNER AND THAT IT HAS BEEN SO AS FAR BACK AS OCT 30

WE HAVE AN APPARENTLY PARADOXICAL SITUATION IN WHICH

THE CAUSE OF THE CAUSE IS THE EFFECT

AND

THE EFFECT OF THE EFFECT IS THE CAUSE



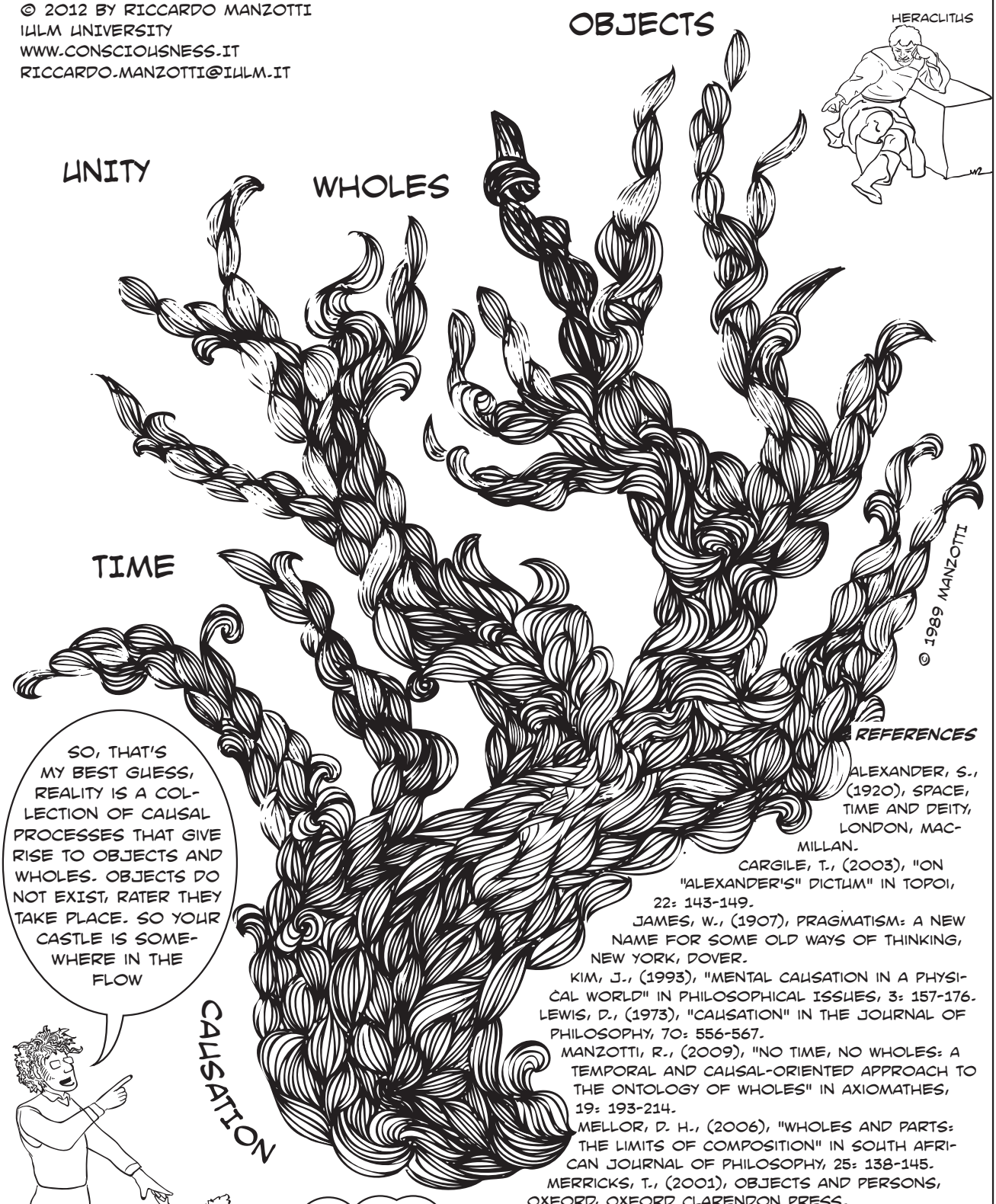
OBJECTS ARE LIKE THE WINNING LOTTERY TICKETS! THEY TAKE PLACE AGAINST ALL ODDS.

OUT OF GAZILLIONS OF POSSIBLE OBJECTS ONLY A VERY FEW BECOME ACTUAL OBJECTS.

AN OBJECT TAKES PLACE ONLY WHEN IT PRODUCES AN EFFECT, BUT WHEN IT DOES IT WAS THERE SINCE THE BEGINNING

15

IULM UNIVERSITY
WWW.CONSCIOUSNESS.IT
RICCARDO.MANZOTTI@IULM.IT



REFERENCES

ALEXANDER, S., (1920), SPACE, TIME AND DEITY, LONDON, MACMILLAN.

CARGILE, T., (2003), "ON "ALEXANDER'S" DICTUM" IN TOPOI, 22: 143-149.

JAMES, W., (1907), PRAGMATISM: A NEW NAME FOR SOME OLD WAYS OF THINKING, NEW YORK, DOVER.

KIM, J., (1993), "MENTAL CAUSATION IN A PHYSICAL WORLD" IN PHILOSOPHICAL ISSUES, 3: 157-176.

LEWIS, D., (1973), "CAUSATION" IN THE JOURNAL OF PHILOSOPHY, 70: 556-567.

MANZOTTI, R., (2009), "NO TIME, NO WHOLES: A TEMPORAL AND CAUSAL-ORIENTED APPROACH TO THE ONTOLOGY OF WHOLES" IN AXIOMATHES, 19: 193-214.

MELLOR, D. H., (2006), "WHOLES AND PARTS: THE LIMITS OF COMPOSITION" IN SOUTH AFRICAN JOURNAL OF PHILOSOPHY, 25: 138-145.

MERRICKS, T., (2001), OBJECTS AND PERSONS, OXFORD, OXFORD CLARENDON PRESS.

SIDER, T., (2001), FOUR DIMENSIONALISM. AN ONTOLOGY OF PERSISTENCE AND TIME, OXFORD, OXFORD CLARENDON PRESS.

SIMONS, P. M., (1987), PARTS. A STUDY IN ONTOLOGY, OXFORD, CLARENDON PRESS.

VAN INWAGEN, P., (1990), MATERIAL BEINGS, NEW YORK, CORNELL UNIVERSITY PRESS.

WHITEHEAD, A. N., (1929), PROCESS AND REALITY, LONDON, FREE PRESS.